Extracting economic narratives using Natural Language Processing: A temporal perspective

Dissertation

by

Kai-Robin Lange

born in Friedrichroda

in partial fulfillment of
the requirements for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

Submitted:	Dortmund, June 2025
Primary referee:	Prof. Dr. Carsten Jentsch
Secondary referee:	Prof. Dr. Jörg Rahnenführer

Day of oral examination:	September 9, 2025

Kai-Robin Lange

Extracting economic narratives using Natural Language Processing: A temporal perspective

Dissertation
in partial fulfillment of
the requirements for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

Submitted to the
Department of Statistics of the
TU Dortmund University

Dortmund, June 2025

Primary referee: Prof. Dr. Carsten Jentsch
Secondary referee: Prof. Dr. Jörg Rahnenführer

Day of oral examination: September 9, 2025

I sincerely thank Carsten Jentsch for supervising my thesis and providing me with a sense of direction when I needed it most. I greatly appreciate his confidence in my abilities as well as his guidance. I am also grateful to Jonas Rieger, who never hesitated to offer valuable advice whenever I had questions during my research.

Ich danke meinen Eltern für all die Unterstützung, die ihr mir in meinem Leben habt zuteil werden lassen, und alle Möglichkeiten, die ihr mir damit eröffnet habt.

Ich danke meiner Freundin dafür, dass du in der stressigsten Phase der Promotion meine Stütze warst.

Contents

Abstract

As our society has become increasingly digitized across the last decades, the need to analyze unstructured data sources has grown in many sectors. In particular, economic and political researchers have been interested in quantitatively uncovering information found in texts, as theoretical research in these fields points towards the great effects of narratives on economic decision making. While such economic narratives can follow a myriad of definitions, they can generally be described as sense-making stories that can influence the reader's economic or political decisions in the future. Developing quantitative methods that can extract and analyze narratives from texts is therefore a major step to better understand market behavior or the decisions of policy makers, among others.

In this cumulative dissertation, I outline the research I have conducted concerning economic and political narratives with a particular focus on using diachronic language modeling that enables me to analyze narratives not in a vacuum, but rather observe narrative shifts over time. To do this, I first properly define economic narratives and provide an overview of language models I used in my research. I then proceed to summarize the methodology and contributions to the current research proposed in my papers, starting with works that do not consider a temporal component when extracting narratives from texts. These works show the development of narrative extraction techniques over time, starting from an unsupervised text classification method and a large pipeline of models specifically designed to handle the task, and ending with the use of state-of-the-art Large Language Models to extract narratives utilizing their great language understanding capabilities.

After covering atemporal narrative extraction methods, I focus on my works that utilize diachronic language modeling. I present two diachronic change detection methods, designed to identify points in time at which we can suspect a narrative shift. The first method uses the frequency of mentions of an entity in the media over time to detect a change, enabling an analysis of narratives surrounding that entity. The second method detects changes in the topics of the topic model `LDA`, allowing for an analysis of the corpus at large rather than a specific entity. I then propose a method to combine temporal and atemporal methods, with atemporal methods extracting narratives (or narrative shifts) and the detected change points. I further present software publications that contain all diachronic methods proposed in my works. Lastly, I conclude by giving an outlook on future research.

Zusammenfassung

Mit der zunehmenden Digitalisierung unserer Gesellschaft in den letzten Jahrzehnten ist der Bedarf an der Analyse unstrukturierter Datenquellen in vielen Bereichen gestiegen. Insbesondere Forscherinnen und Forscher der Wirtschafts- und Politikwissenschaften sind an der quantitativen Extraktion von Informationen aus Texten interessiert, da die theoretische Forschung in diesen Bereichen Narrativen eine große Wirkung auf ökonomische Entscheidungsprozesse zuschreibt. Es gibt viele unterschiedliche Definitionen für ökonomische Narrative, jedoch lassen sie sich als sinnstiftende Geschichten, die die zukünftigen wirtschaftlichen oder politischen Entscheidungen des Lesenden beeinflussen können, zusammenfassen. Die Entwicklung quantitativer Methoden, mit denen Narrative aus Texten extrahiert und analysiert werden können, ist daher ein wichtiger Schritt, um unter anderem das Marktverhalten oder die Entscheidungen politischer und ökonomischer Entscheidungsträger besser zu verstehen.

In dieser kumulativen Dissertation skizziere ich die Forschung, die ich zu ökonomischen und politischen Narrativen durchgeführt habe, mit besonderem Schwerpunkt auf der diachronischen Sprachmodellierung, die es ermöglicht, Narrative nicht nur unabhängig voneinander zu analysieren, sondern auch narrative Veränderungen im Laufe der Zeit zu beobachten. Zu diesem Zweck lege ich zunächst einige Definitionen ökonomischer Narrative dar und gebe einen Überblick über die Sprachmodelle, die ich in meiner Forschung verwendet habe. Anschließend fasse ich die Methodik und die Beiträge meiner Papiere zur aktuellen Forschung zusammen, beginnend mit meiner Forschung, die keine temporale Komponente bei der Narrativextraktion betrachtet. Diese beschreibt die Entwicklung von Narrativextraktionsmethoden über die Zeit, da sie mit einer Methode zur unüberwachten Sentimentanalyse und einer großen Pipeline aus explizit für den Zweck der Narrativextraktion aneinandergeschalteten Sprachmodellen beginnt, und mit der Verwendung von Large Language Models, die den aktuellsten Stand der Technik widerspiegeln, endet.

Nach dem Kapitel über atemporale Methoden fokussiere ich mich auf meine Forschung zur diachronischen Sprachmodellierung. Ich stelle zwei diachronische Methoden zur Erkennung von Strukturbrüchen vor, mit denen Zeitpunkte identifiziert werden können, an denen ein Narrativwandel stattgefunden haben könnte. Die erste Methode erkennt Strukturbrüche in der Häufigkeit der Erwähnungen einer Entität in den Medien, was es ermöglicht, Narrative zu analysieren, die mit dieser Entität zusammenhängen. Die zweite Methode erkennt Strukturbrüche in den Topics des Topic Modells `LDA`, welche es erlauben, den gesamten Korpus statt nur einer einzelnen Entität zu analysieren. Anschließend stelle ich eine Methode vor, die atemporale mit temporalen Methoden kombiniert, in der atemporale Methoden Narrative (oder Narrativwandel) an den Zeitpunkten der gefundenen Strukturbrüche extrahieren. Weiterhin stelle ich Softwareveröffentlichungen vor, die alle diachronischen

Methoden dieser Dissertation und einen Datensatz enthalten. Schließlich gebe ich einen Ausblick auf zukünftige Forschung in den Themenbereichen dieser Dissertation.

Abbreviations

BERT	Bidirectional encoder representations from transformers
CEN	Collective economic narrative
COT	Chain-of-thought prompting
GloVe	Global Vectors for Word Representation
GPT	Generative Pre-Trained Transformer
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
Llama	Large Language Model Meta AI
NPF	Narrative Policy Framework
NLP	Natural language processing
PEFT	Parameter-Efficient Fine-Tuning
QLORA	Quantized Low-Rank Adaptation
Regex	Regular expressions
USA	United States of America

Notation

α	Prior for the document-topic distribution θ
B	Number of Bootstrap samples
cos	Cosinus similarity
$\mathcal{C}$	Number of classes for a text classification
$C_k^{(t)}$	Number of detected changes in topic k up until time t
d_k	Dimension of the key and query matrices
d_{model}	Number of embedding dimensions
d_v	Dimension of the value matrix in the transformer architecture
$\boldsymbol{d}_n^{(m)}$	Word embedding of the nth word in $D^{(m)}$
$\boldsymbol{d}^{(m)}$	Document embedding of $D^{(m)}$
$D^{(m)}$	mth Document in the corpus with $m \in \{1, \ldots, M\}$
$\mathcal{D}^{(m)}$	Static embedding matrix of $D^{(m)}$ in the transformers architecture
$\tilde{\mathcal{D}}^{(m)}$	Positional embedding matrix of $D^{(m)}$ in the transformers architecture
η	Prior for the word-topic distribution ψ
$\boldsymbol{f}_{k\vert t}$	Word frequency vector for topic k at time chunk t
h	Number of attention heads
$\mathbb{I}$	Indicator function
$\mathcal{K}$	Number of topics
$K_i^{(m)}$	Key matrix of the ith attention head $i \in \{1, \ldots, h\}$
$\lambda_i^{(t)}$	Poisson parameter for entity i at time t
$\mathcal{L}_c$	List of tokens of class $c \in \{1, \ldots, C\}$ in a dictionary analysis
M	Number of documents in the corpus
$\mathcal{M}$	Memory parameter of `RollingLDA`
$N^{(m)}$	Number of tokens in the mth document in the corpus
L	Maximum length of Doc2Vec's context window
$l_n^{(m)}$	Effective length of Doc2Vec's context window for the nth token in $D^{(m)}$
ln	Natural logarithm
p	Mixture parameter
$P^{(m)}$	Positional encoding matrix for $D^{(m)}$
$\pi_i^{(t)}$	Inflation parameter for the zero-inflated Poisson distribution
ψ_k	Word-topic distribution of the topic $k \in \{1, \ldots, \mathcal{K}\}$

q and $q_k^{(t)}$	Threshold for the entire corpus and for topic k at time chunk t
$Q_i^{(m)}$	Query matrix of the ith attention head with $i \in \{1, \ldots, h\}$
$\mathcal{T}$	Number of time chunks
$\boldsymbol{T}$	Set of all possible topics $\{T_1, \ldots, T_{\mathcal{K}}\}$
$\boldsymbol{T}^{(m)}$	Vector of topic assignments of the tokens in $D^{(m)}$
$\boldsymbol{T}_n^{(m)}$	Topic assignment of the nth token in $D^{(m)}$
θ_m	Document-topic distribution of $D^{(m)}$
$\mathcal{V}$	Size of the vocabulary
$V_i^{(m)}$	Value matrix of the ith attention head with $i \in \{1, \ldots, h\}$
$w_n^{(m)}$	nth token in $D^{(m)}$
$W_n^{(m)}$	Random variable for the nth word in $D^{(m)}$
W_i^K, W_i^Q, W_i^V	Key, Query, and value weight matrices of the ith attention head with $i \in \{1, \ldots, h\}$
W^O	Multi-head weighting matrix
$z_{\max}$	Maximum look-back distance
$z_k^{(t)}$	Effective look-back distance for topic k and time chunk t

1 Introduction

Along with the expanded use of personal hardware, the way in which major parts of our society receive and spread information about political or economic events has changed. In the age of smart devices with constant internet connection being in people's direct proximity, the speed in which information is spread and processed has increased through social media and online newspapers. The information circulating online does not only cover factual retellings of the aforementioned political or economic events, but also opinionated statement pieces, in which the authors may show varying levels of expertise in the subject. People often resort to assume causal connections between events that they deem to be the most logical or that resonate most with them on an emotional level, even if there is no hard proof of that connection available due to the difficulty on making causal connections with absolute certainty in our complex political and economic systems (Shiller, 2017). As everyone in our society, including important economic or political actors, might base their decisions on opinions rather than facts to a certain degree, it is important for researchers to also consider opinion-based decision making when attempting to analyze the actions of such actors. Economic researchers collectively gathered definitions of this phenomenon and ways to analyze it under the umbrella term *economic narratives*. While economic narratives can follow one of numerous definitions, depending on the research in question, they almost always share core aspects with small variations. They must contain a story of economic relevance and a form of personal opinion, which itself follows slightly different definitions and is sometimes called a "moral" (Shanahan et al., 2018), an "interpretation" (Shiller, 2017), or a "sense-making" component (Roos et al., 2024). Tracing economic narratives across online platforms might help us to understand how widely they are propagated in the population, how they circulate, and even how bad faith actors might use online platforms to spread potentially harmful narratives with malicious intent. To do all this however, we first need to find ways to extract narratives from media quantitatively to deal with the large amount of data that is generated each day online.

The following two examples illustrate both the nature and importance of economic narratives, as both center around narratives that affected the decisions of economic and political actors. On April 2nd 2025, the president of the United States of America, Donald J. Trump, announced tariffs on countries around the world. While the tariffs proposed in this announcement did have an unprecedented scale, they were not unexpected.

During his presidential election campaign in 2024, Donald Trump said “We’re going to bring the companies back. [...] And we’re going to protect those companies with strong tariffs” (Aratani, 2024). His reasoning, that foreign companies would stop exporting goods to the USA and instead start producing goods within the USA, has however been put into question by many, including qualified economists, who feared that the tariffs would lead to higher prices for both companies and consumers, create a climate of uncertainty, and cause companies to delay investments within the USA (Wachtel, 2025). With these two conflicting narratives about the effects of tariffs, it begs the question as to where they originate from, how they developed over time, how large the proportion of the population is that shares them, and why such notable actors in the economy believe in them.

The example provided above is just one of many cases in which the decision of an actor has been affected by believing in a certain narrative. Another example in which not one very relevant actor but many comparatively small actors acted based on a narrative is the GameStop short squeeze that occurred in 2021. Users of the `r/wallstreetbets` subreddit caused the GameStop stock to increase in value, leading more than $6 billion in losses for short-sellers (Laboe, 2021). The Reddit users were driven by a narrative that they, as small investors, could challenge major economic actors by leveraging their digital platform if they all purchased and held the stock (Van Kerckhoven et al., 2023). This example shows how rapidly narratives can spread to a large amount of people due to access to the internet.

While the aforementioned narratives are both well-known and analyzed examples, such economic narratives can be created by anyone and affect anyone in any economic decision. Analyzing them on a large scale is therefore necessary if one plans to capture a broad picture of existing narratives and evaluate their impact on economic actors. While narratives have been studied qualitatively by economists for decades (Roos et al., 2024), the large amount of data produced on the internet every day raises the need to quantitatively analyze economic narratives. To do this, researchers have started to extract information not just from tabular data that contains concrete factual measures, but also from unstructured data that contains opinions along with factual statements. In particular, the field of Natural Language Processing (NLP), so the quantitative analysis of text as data, has seen promising improvements in recent years. The great language understanding capabilities of modern pre-trained models based on the transformer architecture (Vaswani et al., 2023) allow users to perform state-of-the-art inference on corpora without the need to train the models from scratch themselves. The distinction between encoder-only models such as `BERT` (Devlin et al., 2019) and decoder-only models such as `GPT` (Brown et al., 2020) also enables a wide variety of analyses to be performed. Additionally, so called diachronic language models allow researchers to model the evolution of language over time (Kutuzov et al., 2018). However, despite all advancements in this area, extracting economic narratives from texts remains a complex task that requires precise definitions of what to look for rather than just looking for any kind of linguistic information.

In this dissertation, I will present the research that my co-authors and I conducted that contributes to the existing literature on the quantitative extraction of economic narratives from text corpora. We do so by not only designing methods to extract narratives from individual texts, but also by designing methods to find change points in text corpora, enabling us to analyze how narratives change over time and thus view narratives from a temporal perspective. In the rest of this chapter, I will provide an overview of related work that properly defines economic narratives and NLP methods that serve as the foundation for my contributions. In the two following chapters, I present the contributions of the research conducted by my co-workers and myself. Section 2 covers methods to analyze texts atemporally. In Section 3 I present our contributions intended to process corpora that show signs of narrative shifts over time.

1.1 Economic Narratives

One can imagine that trying to define or even quantify the effect of the spoken or written word on the world of economics, which in itself is very complex, is no straightforward feat. In fact, over time, many economic researchers have come up with a variety of possible definitions for economic narratives, from definitions that are supposed to be all-encompassing to ones that are designed for a specific use case. Roos et al. (2024) claim to have found 436 economic papers from 1967 to 2022 containing the word "narrative" in their title using the Web of Science search engine. By May 2025, this number has increased to 704 entries, indicating a strong increase in research activity, in which researchers try to both properly define economic narratives from a theoretical perspective and extract them from media using qualitative or quantitative approaches. The interest in this field of research certainly increased when Shiller (2017) popularized the concept of economic narratives to a wider audience, describing a narrative as "a simple story or easily expressed explanation of events that many people want to bring up in conversation or on news or social media because it can be used to stimulate the concerns or emotions of others, and/or because it appears to advance self-interest" (Shiller, 2017, p. 968). In his work from 2020, Shiller offers a more precise definition of economic narratives as "stories that offer interpretations of economic events, or morals, or hints of theories about the economy" (Shiller, 2020, p. 792). This definition contains core aspects that other definitions would also build upon. Narratives must contain a story, which most often is a description of unfolding events. In addition to this story, there must be an interpretation: a story, which in itself could just be a factual retelling of events, is enriched with a subjective (most often the author's) point of view. For instance, this definition would not consider the sentence "The prices for manufacturing are high and there are supply chain issues." to contain an economic narrative, as the two events presented in this sentence are not interpreted. However, the sentence "The prices for manufacturing are high because of supply chain issues." would be considered to contain an economic narrative, as the author interprets the existence of both events and assumes a causal connection between them.

Such a causal connection is arguably the most common form of interpretation that bridges the gap between stories and economic narratives. Consequently, many works focus on this causal aspect to refine the definition of economic narratives to a more concrete form that is easier to work with in practice. For instance, Eliaz et al. (2020) use causal connections as identifying elements of economic narratives and display them in a Directed Acyclic Graph:

$$\text{Supply chain issues} \rightarrow \text{Manufacturing prices are high}$$

While causal connections are common ways to identify economic narratives, other authors (Eliaz et al., 2024; Roos et al., 2024; Shenhav, 2006) argue that not every narrative has to make a causal connection but can also contain a value judgment, more akin to the definition of Shiller (2017). Alternative approaches like the Narrative Policy Framework (NPF, Jones et al., 2010; Schlaufer et al., 2022; Shanahan et al., 2018) focus fully on value judgments. The NPF considers economic narratives to consist of four distinct elements: A narrative must contain a setting, characters, a plot, and a moral of the story. A setting can be a location, but also a larger sequence of events the narrative takes place in. Certain characters, such as policy makers, companies, or governments, must either actively or passively participate in the narrative with their actions or with what happens to them. Similar to the concept of stories, the plot forms a connection between the settings and the characters by describing events surrounding the characters taking place in the setting. When there is a plot, the NPF additionally requires a moral to be attached to it to be considered an economic narrative. This can be a literal moral judgment that considers actions to be either morally good or bad, but it can also consist of a key takeaway that the author deems to be important to the plot.

This concept of a moral of the story is further expanded by one of the more abstract and complex definitions put forth by Roos et al. (2024) who define a collective economic narrative (CEN) as "a sense-making story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions" (Roos et al., 2024, p. 13). The aspect of a sense-making story is comparable to the need for causal connections mentioned above, requiring the story to be enriched with a sense-making component. Additionally, this definition requires a narrative to suggest actions, which extends the concept of a moral of the story put forth as a part of the NPF.

Only considering those stories that suggest actions can help to directly relate CENs to decisions made by economic actors that might have been influenced by such narratives. Despite the possible benefits, deciding if a story contains a call to action is a highly subjective task, especially in cases in which the author does make an explicit but only implicit call. Some of the other information mandated, such as the requirement that a narrative must be shared by members of a group, is arguably not even extractable from a single text in a vacuum at all. Unless being written by more than one person, a single text does not contain the information if any other person, and thus a group, shares any potential narrative. This condition can only be addressed in a postprocessing step, in which all potential narratives are aggregated adequately. While this definition is one of the most advanced and promising definitions, especially when having the goal of analyzing the effect of economic narratives on economic actions, it is also one of the most difficult to work with practically.

It is important to note that, while the methodological contributions in this dissertation have each been created with a specific narrative definition in mind, they can also be adapted to extract economic narratives that follow a different definition instead. Similarly, narrative research is not restricted to the field of economics, as the definitions can also be altered to extract narratives of other research areas, such as political research. In our research, my co-authors and I do not just analyze purely economic corpora, but also political corpora. While not all narratives within these corpora have an explicit economic focus, the actions made by political actors affect the economy and are thus economically relevant and satisfy the definition of Roos et al. (2024), showing that the research in these areas can overlap.

1.2 Natural Language Processing

To extract economic narratives from texts, my co-authors and I utilize Natural Language Processing. Natural Language Processing, or NLP for short, is an umbrella term for ways to transform raw and unstructured texts based on a human language to something that computers can work with, enabling the machines to understand, interpret and handle the text according to a task they were trained on. For the purposes of this dissertation, I divide NLP methods into four major categories: frequency-based methods, (probabilistic) topic models, static embedding models, and contextualized embedding models. In this Section, I explain the core concepts and methods of each category with regard to the themes of this dissertation.

To analyze language, most NLP methods require *tokens*, which they consider to be fundamental building blocks of language. Similar to humans, most methods correspond tokens to words for their analysis and thus work with words as individually interpretable parts of the language. However, many modern models, such as contextualized embedding models, use subword-units as tokens more commonly than words. To distinguish these cases, this dissertation uses the term *word* when specifically referring to word-level tokens and *token* when referring to subword-units or when discussing groups of methods that might utilize either.

In the following Section, I assume to have a corpus of M documents with $\boldsymbol{D}^{(m)} = \{w_1^{(m)}, \dots, w_{N^{(m)}}^{(m)}\}$ for $m = 1, \dots, M$. Each document contains tokens $w_n^{(m)} \in W$ that are part of the vocabulary of size $\mathcal{V}$.

Frequency-based methods extract information by counting token occurrences in texts. These methods can range from simple dictionary-based classification methods (Hutto et al., 2014) to methods that use token frequencies as a basis for statistical language models (Jentsch et al., 2020a,b).

A simple dictionary-based classification assigns one of C classes to each text using a dictionary as an external information basis. This dictionary contains lists of tokens (most commonly words) $\mathcal{L}_1, \dots, \mathcal{L}_C$ for all $\mathcal{C}$ classes. Then, the class with the highest (sometimes weighted) number of occurrences in a document m is chosen as that text's class c_m

$$c_m = \arg\max_{c \in 1, \dots, \mathcal{C}} \sum_{w_n^{(m)}=1}^{N^{(m)}} \mathbb{I}_{w_n^{(m)} \in \mathcal{L}_c}$$

with the indicator function $\mathbb{I}$. One of the most popular use cases of such dictionaries is a sentiment analysis. In a (document-based) sentiment analysis utilizing a sentiment lexicon like the WKWSCI lexicon (Khoo et al., 2018), the objective is to create a model to predict the underlying emotional polarity of a document. While some sentiment analyses can become more complex, most often only two or three classes are considered: positive documents, negative documents and sometimes neutral documents. In these simple cases, a sentiment dictionary (also called a sentiment lexicon) can help by counting negative and positive words in a document. The class with the most occurrences is then estimated to be the most common in the document and to thus represent its most prevalent polarity. If the number of positive and negative words is identical, the document is either considered neutral or one of the two classes is sampled.

Regular Expressions (Regex) are also commonly used methods that do not model language but are rather used to search and filter texts for specific patterns (Jurafsky et al., 2025, pp. 4 ff.). This is helpful to prepare a corpus for the actual analysis during the preprocessing step. For instance, when looking for dates in a text using the German date format `DD.MM.YYYY`, we could use the Regex "`\b\d{2}.\d{2}.\d{4}\b`". The operator `\d{i}` for `i`$= 2, 4$ denotes that we only look for occurrences in which `i` digits are the specified places between the dots. `\b` clarifies that this pattern specifying the date must be surrounded by word-boundaries, such as spaces or punctuation marks. As the example above demonstrates, Regex allow us to specify the exact pattern we want while also giving us the freedom to define which parts can be altered, for instance by not having to look for specific digits, but being able to use an encompassing operator `\d` instead.

Topic models enable us to create $\mathcal{K}$ interpretable clusters of tokens. These clusters represent latent topics in the corpus, that help the user to answer the question "What is the document about?" by observing which topics are prevalent in that document. Topics are represented using so called "top tokens" (also called "top words" when using words as tokens), which contain the tokens that are most commonly assigned to that topic, relative to the other topics. For instance, the top tokens of a topic on inflation could look like this: *inflation, price, interest, fed, rate*. Topic models can, for the most part, be divided into probabilistic topic models and neural topic models. Probabilistic topic models like the Latent Dirichlet Allocation (`LDA`, Blei et al., 2003) often use Bayesian inference to update their prior believes about the topic distribution of the corpus at hand. Neural topic models like `BERTopic` (Grootendorst, 2022) utilize neural networks to infer topics from a corpus. It is worth noting that, while neural topic models often utilize modern, state-of-the-art NLP architectures, they are not necessarily considered "superior" to probabilistic topic models that rely on classic statistical principles. As evaluating what makes a topic model "better" than another topic model is not trivial, there are conflicting studies showing that either architecture can perform "better" than the other given the right circumstances (Hoyle et al., 2022; Li et al., 2024).

`LDA`'s (Blei et al., 2003) topic assignment is based on a latent probability model that assumes each word to be sampled according to two distributions: The document-topic distribution θ_m and the word-topic distribution ϕ_k, both following a Dirichlet distribution with priors α and η, which define the mixture of topics. The model assumes that for document m, a topic vector

$$\boldsymbol{T}^{(m)} = \left(T_1^{(m)}, \ldots, T_{N^{(m)}}^{(m)}\right), \quad T_n^{(m)} \in \boldsymbol{T} = \{T_1, \ldots, T_{\mathcal{K}}\}.$$

is sampled based on the document-topic distribution θ_m for $m = 1, \ldots, M$. Then, for each element of that topic vector $n = 1, \ldots, N^{(m)}$, a word $w_n^{(m)}$ is sampled from the word-topic distribution $\psi_{T_n^{(m)}}$ as a realization of the random variable $W_n^{(m)}$. We can formalize this process as (Griffiths et al., 2004)

$$W_n^{(m)} \mid T_n^{(m)}, \phi_k \sim \mathrm{Discr}(\phi_k), \quad \phi_k \sim \mathrm{Dir}(\eta), \quad T_n^{(m)} \mid \theta_m \sim \mathrm{Discr}(\theta_m), \quad \theta_m \sim \mathrm{Dir}(\alpha).$$

To symbolize this process, we can also use a plate notation, as visualized in Figure 1.

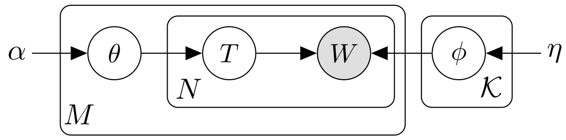


Figure 1: Plate notation of a Latent Dirichlet Allocation (Blei et al., 2003).

Static embedding models enhance the previously mentioned methods in their ability to assess the semantic similarity of tokens and documents. To demonstrate this, we can look at two distinct sentences that convey the same meaning using a different vocabulary:

"The German chancellor returned to the capital."

"Friedrich Merz came back to Berlin."

When comparing these sentences using, for instance, a frequency-based method, we would have no way of knowing that the two sentences talk about the same event, as the only word they share is "to". Static embedding models, following the `Word2Vec` model proposed by Mikolov et al.; Mikolov et al. (2013; 2013), attempt to address this issue.

Instead of perceiving every token in a vacuum, (static) embedding models create a latent vector space in which each token is given a fixed embedding vector that is trained to give semantically similar tokens a similar embedding. These models can be used for numerous tasks, including classification and clustering tasks, language generation as well as the identification of semantically similar tokens, while being computationally efficient with minimal data requirement, enabling researchers and users to train their own model for a specific use case. Popular models that specifically create word embeddings, but no embeddings for entire documents include `Word2Vec` (Mikolov et al., 2013a,b) and the Global Vectors for Word Representation model (`GloVe`, Pennington et al., 2014). Text embedding models extend these word embedding models by also training a specific embedding for each document in a corpus. That document embedding is seen as the vector representation of an entire document and can be used for document-level tasks, such as text classification. `fastText` (Bojanowski et al., 2016; Joulin et al., 2016) and `Doc2Vec` (Le et al., 2014) belong to the most commonly used text embedding models.

To create embeddings that capture semantic similarity, `Doc2Vec` (Le et al., 2014) uses a shallow neural network that is trained with a predictive language modeling task, training embeddings in the form of weights in the process. To perform the language modeling task, it processes the corpus of documents word by word. For each word, it considers a maximum context window size of $L \in \mathbb{N}$. In the frequently used gensim implementation (Řehůřek et al., 2010) for `Doc2Vec`, the effective size is then sampled from $\{1, \dots, L\}$ for every word and is denoted as $l_n^{(m)}$ for the document m at position n. Denoting the

n-th word embedding in document m as $\boldsymbol{d}_n^{(m)} \in \mathbb{R}^{d_{\text{model}}}$ and the document embedding as $\boldsymbol{d}^{(m)} \in \mathbb{R}^{d_{\text{model}}}$ for an embedding dimension $d_{\text{model}} \in \mathbb{N}$, the log-likelihood for the language modeling task

$$\sum_{m=1}^{M} \sum_{n=L}^{N^{(m)}-L} \ln \left(p \left(\boldsymbol{d}_n^{(m)} | \boldsymbol{d}_{n-l_n^{(m)}}^{(m)}, \ldots, \boldsymbol{d}_{n+l_n^{(m)}}^{(m)}, \boldsymbol{d}^{(m)} \right) \right)$$

is maximized using stochastic gradient descent. `Doc2Vec` calculates the probability $p(\cdot|\cdot)$ using a hierarchical softmax. While `Doc2Vec` is trained with a language modeling task, this is often not the task the model is used for, but instead simply a good way to teach a model semantics as semantically similar tokens are often used in close proximity to each other and will thus be trained to have similar embeddings. It is important to note that, similar to how probabilistic topic models require the number of topics $\mathcal{K}$ to be set by the user, `Doc2Vec` may also require hyperparameter-tuning by altering the maximum context window size L or the number of dimensions of the embedding vector space.

Contextualized embedding models are an extension of static embedding models that facilitate the embeddings of subword tokens to change depending on the context in which they are used in. To achieve this, they utilize large neural networks, often relying on the transformer architecture (Vaswani et al., 2023). Due to the size of the neural networks and the resulting computational cost, these models are often not trained by small research groups. Most often, existing, pre-trained language models are adapted and fine-tuned for specific use cases while preserving their language understanding capabilities. These models have outperformed static embedding models in almost all benchmarks conducted: Encoder-only models such as the Bidirectional Encoder Representations from Transformers model (`BERT`, Devlin et al., 2019) impress in classic inference tasks, such as text classification, semantic role labeling or named entity recognition. Decoder-only models such as the Generative Pre-trained Transformer (`GPT`, Brown et al., 2020) have seen massive mainstream appeal due to their high performance in language generation tasks.

In the transformer architecture (Vaswani et al., 2023), we map each token of a sequence of input tokens from document $D^{(m)}$ to a static embedding vector of length d_{model}, resulting in an input matrix $\mathcal{D}^{(m)} \in \mathbb{R}^{N^{(m)} \times d_{\text{model}}}$. The transformer then uses a positional encoding $P^{(m)} \in \mathbb{R}^{N^{(m)} \times d_{\text{model}}}$ to capture positional information. Often, a sinusoidal positional encoding

$$P_{n,2i}^{(m)} = \sin\left(\frac{n}{10000^{2i/d_{\text{model}}}}\right), \quad P_{n,2i+1}^{(m)} = \cos\left(\frac{n}{10000^{2i/d_{\text{model}}}}\right).$$

is used with the token positions $n \in \{1, \ldots, N^{(m)}\}$ and dimension $i \in \{1, \ldots, d_{\text{model}}\}$. Then, the positional encoding is added to the previously calculated static embedding $\tilde{\mathcal{D}}^{(m)} = \mathcal{D}^{(m)} + P^{(m)}$. Adding this fixed matrix to the input embeddings, that gives each position in the text a unique value, enables the model to more easily identify tokens based on their position in a text.

The positional embedding matrix $\tilde{\mathcal{D}}^{(m)}$ then serves as an input for the multi-head attention algorithm with $h \in \mathbb{N}$ attention heads. Given weight matrices $W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ for $i = 1, \ldots, h$, we calculate the Query, Key and Value matrices

$$Q_i^{(m)} = \tilde{\mathcal{D}}^{(m)} W_i^Q, \quad K_i^{(m)} = \tilde{\mathcal{D}}^{(m)} W_i^K, \quad V_i^{(m)} = \tilde{\mathcal{D}}^{(m)} W_i^V$$

with which we can calculate the attention matrix

$$\text{Att}(Q_i^{(m)}, K_i^{(m)}, V_i^{(m)}) = \text{softmax}\left(\frac{Q_i^{(m)} \left(K_i^{(m)}\right)^\top}{\sqrt{d_k}} \right) V_i^{(m)}.$$

Vaswani et al. (2023) claim that the scaling factor $\sqrt{d_k}$ is needed to counteract the growing dot-product with increasing dimensionality, leading to more stabilized gradients. W_i^Q, W_i^K and W_i^V contain learnable parameters that allow the model to learn which tokens attend to one another depending on the input matrix $\tilde{\mathcal{D}}^{(m)}$ and which of these relationships are important to pass on into later stages of the model.

The results of the h attention heads are then combined to generate one output matrix that best reflects which word attends to which other words.

$$\text{Multi-head}^{(m)} = \left(\text{Att}\left(Q_1^{(m)}, K_1^{(m)}, V_1^{(m)}\right), \ldots, \text{Att}\left(Q_h^{(m)}, K_h^{(m)}, V_h^{(m)}\right)\right) W^O$$

The matrix $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ is used to weight each attention head. These attention steps are the central part of the transformer architecture, as they allow the model to store linguistic information, that allows the model to understand contextualized information using matrix multiplications of the input $\tilde{\mathcal{D}}$ and the weight matrices W_i^Q, W_i^K and W_i^V. Using not just one, but multiple attention heads enables the model to at least theoretically store different kinds of contextualized information in different heads, which would make each head an expert at a certain type of syntactical relationship. The resulting matrix is sent through normalization and feed-forward layers into either the output or further encoder/decoder blocks, depending on the model.

The terms encoder and decoder refer to the two major building blocks of the transformer architecture that can be seen in Figure 2. The image shows an encoder-decoder architecture, in which the encoder feeds into the decoder. However, many popular models utilize either an encoder-only or decoder-only architecture. In the case of an encoder-only model, the encoder takes text as an input and returns embeddings for every token in the text. `BERT-base` (Devlin et al., 2019) is a popular encoder-only model that contains 12 encoder blocks, each of the form displayed on the left hand side in Figure 2, with $h = 12$ attention heads, the weight matrix dimensions $d_k = d_v = 64$ and an embedding size of $d_{\text{model}} = 768$. Encoder-only models perform very well on inference tasks that can use these embeddings as features. A decoder-only model like `GPT-4o` (OpenAI et al., 2024) also takes text as an input but responds in a text instead returning embeddings. Decoders are often used for open-ended questions that go beyond mere inference and can be very useful for zero-shot (Kojima et al., 2022; Wei et al., 2021) or few-shot learning (Song et al., 2023), in which a model is asked to complete a task not by fine-tuning it on thousands of examples beforehand, but by providing it with only a few or no examples.

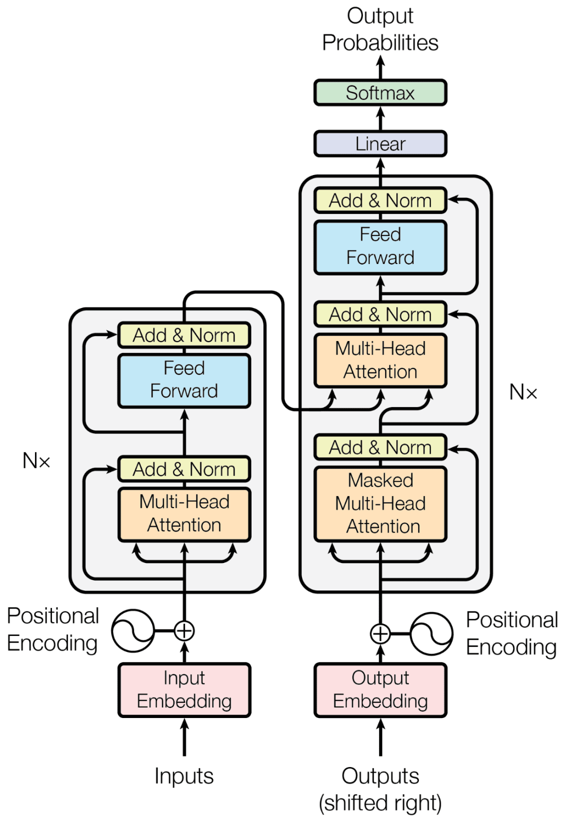


Figure 2: Visualization of the transformer architecture (Vaswani et al., 2023). The left hand side shows an encoder block while the right hand side shows a decoder block.

As generating language is a more complex task than understanding existing language, decoder-only models are often larger (in terms of the number of parameters) than encoder-only models. For example, one of the most commonly used encoder-only models `XLM-RoBERTa-large` (Conneau et al., 2020) has about 355 million parameters, while `Llama 4` (Dubey et al., 2024), one of the most prominent open source decoder-only models, comes with three size variations of about 109 billion, 400 billion or 2 trillion parameters. Due to the size of such decoder-only models, they are often also referred to as "Large Language Models" (LLMs). While this term is technically ambiguous and could be used for any language model with a high number of parameters, decoder-only models have often become synonymous with it. When mentioning LLMs in this dissertation, I therefore refer to decoder-only models that were trained on a task to fulfill the instruction of a user.

2 Atemporal narrative extraction

Many text corpora are temporally heterogeneous. Instead of portraying one homogeneous collection of texts, they most often contain texts that were written across a certain time span. During that time span, language can evolve, the topics the authors talk about can change, and, most importantly for my research, the narratives spread by the authors can change. Despite that, most classic NLP models such as topic models like `LDA` (Blei et al., 2003) or static embedding models like `Word2Vec` (Mikolov et al., 2013a,b) do not natively model temporal developments in the corpus. Even modern pre-trained contextualized embedding models like `BERT` (Devlin et al., 2019) only do so implicitly, by adapting embeddings to the context within a document and thus to the language of its time. This is most likely the case because temporal heterogeneity within a corpus is not important for most analyses, either because it does not matter for the task at hand (e.g. for a sentiment analysis of product reviews) or because the temporal heterogeneity is not believed to have much of an effect as the time frame covered is relatively short.

As more research has been conducted on creating language models that do not explicitly model the evolution of language over time, but rather look at documents in a vacuum, it is logical to try to tackle the complicated task of narrative extraction using such well-researched atemporal language models first. Instead of directly modeling temporal changes, one can then aggregate the results with regards to temporal developments later.

Many narrative extraction approaches rely on classic NLP tasks such as sentiment analysis (Tilly et al., 2021) or topic modeling (Macaulay et al., 2023). However, the narratives that can be extracted from such tasks alone do not meet the requirements of the economic narrative definitions discussed in Section 1.1. Santana et al. (2023) provide a detailed overview of narrative extraction techniques that have been developed in recent years. Most techniques discussed focus on identifying core elements of narratives, such as events and connecting them to form narratives. Such methodologies often incorporate common NLP tasks like semantic role labeling, part of speech tagging, event extraction, or named entity recognition and combine them all into one big pipeline. The `RELATIO` pipeline (Ash et al., 2024), which I cover in detail in Section 2.3, is one of the most prominent examples of this.

Alternatively, the language understanding capabilities of instruction LLMs can be used to analyze narratives more directly. Instead of creating one large pipeline with models that are originally trained not for narrative extraction but for small tasks that are then combined into a narrative extraction pipeline, LLMs can be prompted with the definition of narratives to extract them in one step. LLMs have shown in the past that they are able to solve complex tasks when provided with a detailed prompt. For instance, Ziems et al. (2024) show that LLMs can surpass human annotators in annotating certain political and economic documents. Furthermore, Gilardi et al. (2023) show that LLMs are able to extract narrative frames from tweets better than crowd workers. Other research also points towards their usefulness for narrative extraction or related tasks: Bornheim et al. (2024) show that LLMs can be fine-tuned to solve the related task of speaker attribution in German parliamentary speeches and Gueta et al. (2025) use LLMs for narrative extraction tasks, but utilize a loose narrative definition that does not satisfy the definitions mentioned in Section 1.1.

In this Section, I present works of my co-authors and myself that contribute to the existing literature of extracting economic narratives from corpora without explicitly accounting for temporal biases or stories that develop over time. These can both be seen as standalone extraction methods as well as a basis for later diachronic analyses.

2.1 Contributed publications

Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch (2022a). "Towards Extracting Collective Economic Narratives from Texts". In: *Ruhr Economic Papers* 963. DOI: `10.4419/96973127`.

Lange, Kai-Robin, Jonas Rieger, and Carsten Jentsch (2024). "Lex2Sent: A bagging approach to unsupervised sentiment analysis". In: *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*. Association for Computational Linguistics, pp. 281–291. URL: `https://aclanthology.org/2024.konvens-main.28/`.

Schmidt, Tobias, Kai-Robin Lange, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025). *Identifying economic narratives in large text corpora – An integrated approach using Large Language Models*. arXiv: `2506.15041 [econ.GN]`.

2.2 Sentiment analysis with `Lex2Sent`

While the sentiment of a document is not directly connected to the narrative definitions in Section 1.1, it can provide an interesting perspective on an existing narrative. As narrative extraction techniques focus on extracting the information the author of a document is trying to convey, a sentiment analysis can give insights into how the author conveys it. This tonality in which information is provided can be an interesting factor to consider in how narratives are spread. With research showing that social media algorithms focus on negative emotions to increase engagement (Abbas et al., 2021), it can be important to understand the sentiment of a document to infer how effective the narrative is conveyed to the reader.

In our paper on the unsupervised sentiment analysis method `Lex2Sent` (Lange et al., 2024), we combine the idea behind classic sentiment lexica with the static embedding model `Doc2Vec` (Le et al., 2014) for a binary sentiment classification task. We use the information stored in the sentiment lexica as a foundation for an unsupervised sentiment analysis that can be performed on any text or narrative that has been extracted without requiring labeled examples. For this, we consider both parts of a binary lexicon (so one negative and one positive word list) to be documents. They serve as a theoretical extreme example, containing only words that belong to their respective class. Thus, the document embedding of either lexicon part is considered to be representations of its class. When training a `Doc2Vec` model on a corpus, we can then use this characteristic of the resulting vector space to scale each document between the two dictionary parts. Formally, we calculate the cosine similarity between the embeddings of a document m and the word lists $\mathcal{L}_{pos}$ and $\mathcal{L}_{neg}$ of the classes *pos* and *neg* and calculate the difference between them:

$$\text{diff}^{(m)} = \cos\left(\boldsymbol{d}^{(m)}, \boldsymbol{d}^{(\mathcal{L}_{pos})}\right) - \cos\left(\boldsymbol{d}^{(m)}, \boldsymbol{d}^{(\mathcal{L}_{neg})}\right)$$

$\text{diff}^{(m)} > 0$ thus means that our `Doc2Vec` model places the document closer to the positive sentiment word list in its latent vector space. The larger $\text{diff}^{(m)}$ is, the more certain we are that the document m is written with a positive sentiment and vice versa.

To further improve the performance, we scale every document B times, each with a uniquely trained `Doc2Vec` model. Each `Doc2Vec` model is trained with different context window size, vector dimension size or number of training iterations. This eliminates the downside that an unsupervised `Doc2Vec` analysis would usually have of not being able to perform hyperparameter tuning and having to use potentially sub-optimal parameters as well as having to rely on one particular run of a non-deterministic model. Instead, we argue that many models trained with different hyperparameters will answer correctly more often on average than individually. Additionally, we train each `Doc2Vec` model on resampled documents, within which we draw words with replacement, similar to classic bootstrapping (Efron, 1979). We propose that this will evenly distribute words from the lexicon in each text, enabling `Doc2Vec`'s limited context window to better capture the information contained in these sentiment words and spread it to other words that commonly co-occur

Table 1: Average classification rates in percent of `Lex2Sent` with a WKWSCI-, Loughran McDonald- or Opinion Lexicon-base for the fixed threshold 0, compared to the rates of the traditional lexicon method on the same lexicon

	WKWSCI		Opinion Lexicon		Loughran McDonald	
	`Lex2Sent`	lexicon	`Lex2Sent`	lexicon	`Lex2Sent`	lexicon
iMDb	80.01	70.10	78.43	73.37	70.73	61.22
Amazon	76.83	65.15	77.68	69.28	69.27	61.32
Airline	72.42	63.29	71.96	68.33	72.06	53.18

with them. We then combine the diff-vectors of all B `Doc2Vec` models:

$$diff^{(m)}_{\text{mean}} := \frac{1}{B}\sum_{b=1}^{B} diff^{(m)}_{b}$$

The label for the document m is then defined as

$$\text{label}_d = \begin{cases} \text{positive}, & diff^{(m)}_{\text{mean}} - q > 0 \\ \text{negative}, & diff^{(m)}_{\text{mean}} - q < 0 \\ \text{at random}, & diff^{(m)}_{\text{mean}} - q = 0 \end{cases}$$

for some threshold $q \in \mathbb{R}$ (usually 0 in an unsupervised case). This combination of multiple `Doc2Vec` models leads to a bagging effect (Breiman, 1996), similar to Random Forests (Breiman, 2001), but using language models instead of decision trees as a foundation.

We evaluate `Lex2Sent` on three benchmark corpora with distinct characteristics: the iMdB corpus contains a lot of long documents (Maas et al., 2011), the Amazon review corpus contains just as many, but shorter documents (He et al., 2016) and the Airline tweet review corpus contains less than half as many and very short documents (*Twitter US Airline Sentiment* 2021). On these corpora, we use three different sentiment dictionaries as a foundations for our `Lex2Sent` model: the WKWSCI dictionary (Khoo et al., 2018), the Opinion dictionary (Hu et al., 2004) and the Loughran McDonald dictionary (Loughran et al., 2010). The results in Table 1 show that `Lex2Sent` improves the performance of all three dictionaries on all three corpora. Other experiments presented in the paper also suggested that `Lex2Sent` exceeds the performance of all baseline dictionary methods when the `Doc2Vec` models are given enough documents to train on.

We believe that `Lex2Sent` can be a scalable alternative to other unsupervised sentiment analysis methods, such as prompting LLMs. In the future, this can help us to uncover what type of emotion is being used to spread certain narratives in large corpora.

2.3 Narrative extraction using NLP pipelines

Commonly, economic narratives are extracted from documents using a pipeline that combines multiple classic NLP tasks, which are solved by static embedding and encoder-only models, to complete this more complex task. One popular example of this is the `RELATIO` pipeline proposed by Ash et al. (2024). The `RELATIO` pipeline extracts statements in the form of three individual building blocks called "agent," "verb," and "patient," which answer the question of "Who did what to whom?". The agent is an actor with agency within the statement that does something (specified by the verb) to the patient, who is being assigned a passive role. An example of this could be the following sentence: "The FED raised the interest rates." In this particular case, we consider "The FED" to be the agent and "the interest rates" to be the patient. They are connected using the the verb "raised." Figure 3 shows a visual representation of the individual steps of the `RELATIO` pipeline.

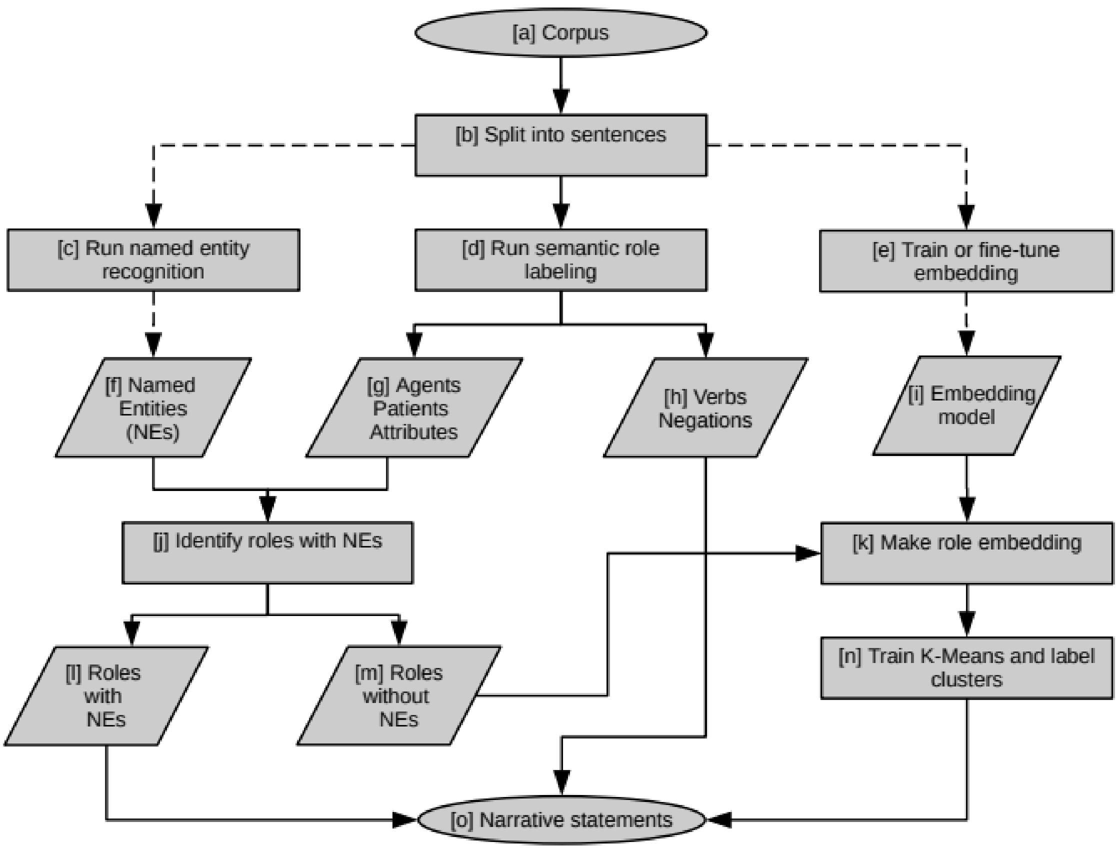


Figure 3: Flowchart of the `RELATIO` pipeline (Ash et al., 2024).

However, these statements do not align with the definitions of an (collective) economic narrative as presented in Section 1.1. While they can tell a story, they most often miss the core concept of the author's subjectivity. The statement "The FED raised the interest rates" mentioned above is a mere factual statement that contains no subjective interpretation, causal connection or call to action. Because of this, in Lange et al. (2022) we extend the `RELATIO` pipeline with additional steps to detect causal connections between statements (among other improvements) to generate output that more closely aligns with the definition of CENs provided in Section 1.1.

In addition to swapping some of the models that `RELATIO` uses for its respective pipeline steps for more modern, state-of-the-art models at the time like AllenNLP (Gardner et al., 2018) for semantic role labeling, we also add an additional preprocessing layer of coreference resolution to the pipeline. Coreference resolution denotes the task of tracing back and resolving references, such as pronouns that reference specific entities. For example, when given the sentence "The FED raised the interest rates, because it expects a high inflation period," the pronoun "it" would be resolved to its original reference "The FED". Thus, the sentence would be altered to "The FED raised the interest rates, because The FED expects a high inflation period." While this will not change much for a human reader when reading the entire sentence, it improves the resulting statements and enables us to aggregate them easier across a large corpus and across time. The second part of the sentence would then be interpreted as the `RELATIO` statement "The FED - expects - high inflation period" instead of "it - expects - high inflation period." This change enables us to understand what the statement is about out of context of what "it" refers to and aggregate it with statements from other documents. We perform coreference resolution using Spacy's Coreferee model (Honnibal et al., 2017).

We then extend the `RELATIO` pipeline with an additional layer after the statement extraction. For this, we save the location of each statement found by `RELATIO` in all documents. On a document basis, we then filter the statements produced by `RELATIO` with a greedy algorithm that leaves us with the maximum number of non-overlapping narratives. Without this step, `RELATIO` tends to produce narratives that can span an entire sentence, which can cause problems in its clustering step (step [n] in Figure 3), as `RELATIO` would try to cluster and aggregate an entire sentence into a single key word, which often fails to provide proper results. After creating our list of non-overlapping statements, we look for causal cue words that indicate a causal connection between the statements. Examples of such cues are words like "because," or "consequently." When we find such a causal cue between two `RELATIO` statements, we suspect them to be causally connected and extract them as narratives.

While our adaptation of the `RELATIO` pipeline improves upon a solid foundation, we still encounter major problems caused by the size of the pipeline. As the pipeline contains many individual tasks, only one of these tasks has to fail to create a mistake in the final output. Each individual model used is a state-of-the-art model in its respective task and performs well in the vast majority of cases, but the chance for an error is still relatively high due to the number of models used in the pipeline. Additionally, we encounter the problem that `RELATIO`'s statements do not cover all possible ways to describe individual events. For instance, the sentence "The interest rates are high" contains an agent and a verb but no patient according to the definition of Ash et al. (2024). The statement would thus not be coded as a part of our extended pipeline, even though it clearly is the description of an event or status that can be used as a building block of a story. We therefore decided to shift our focus to a different class of NLP models to solve the narrative extraction task.

2.4 Narrative extraction using LLM prompting

The advent of LLMs provided researchers with new ways to tackle particularly complex NLP tasks. This is also the case for narrative extraction tasks. In particular, an instruction LLM can theoretically solve the exact problems discussed in Section 2.3. Due to their great language understanding capabilities and their pre-training on generic instruction tasks, LLMs are able to solve complex tasks with just one prompt instead of having to rely on a lengthy pipeline. Additionally, as no semantic role labeling is applied, the output does not have to be restricted to agent-verb-patient statements, but can have any structure as long as the narrative definition applies.

To extract narratives from documents that stay true to the definition of collective economic narratives, in Schmidt et al. (2025) we create an entirely new codebook that is designed to cover as many edge cases as possible. The codebook is the results of many iterations of refinement, in which my colleagues and I served as three expert annotators that coded rules to best qualitatively extract narratives given a set of instructions and definitions. After a certain number of documents have been processed, we reconvened to revise the codebook to handle new edge cases and solve conflicts between the annotators. After the codebook was finalized, we annotated the narratives in 100 documents.

The codebook is designed to extract narratives that have a causal component of the form "event 1 - causes - event 2". The individual events can have a wide array of possible forms, such as occurrences, activities, conditions, future plans and policies. While our paper in Section 2.3 focused on explicit causal connections, the causal connections extracted here can be implicit, capturing many narratives that might have been missed otherwise. Among others, the codebook contains rules to deal with chained narratives and narrative forks. Narratives are chained, when the author implies a causal chain that leads from one event over another to a third event, such as "event 1 - causes - event 2" and "event 2 - causes - event 3". While some narrative extraction techniques would consider event 1 to be the cause for both other events, we consider it important to split causal chains into the smallest possible individual causal connections, as this allows us to aggregate the narratives better. Similarly, narrative forks describe a concept in which one event does in fact cause multiple other events or is caused by multiple events. In this case, it is important for the annotator to interpret whether the combination of events is important for the causal chain, in order to decide to either code each causal connection individually, or to combine them.

All of these steps mentioned above as well as some other aspects of the codebook are however not fully objective, hence we cannot expect the annotators nor the LLM to achieve the same results. It is therefore important to find a way to fairly evaluate the LLM's performance compared to the performance of the three annotators. After all annotators had labeled all 100 documents, we discussed each narrative found. If all three annotators agreed that the story in question is in fact a narrative according to the definition and the codebook, the narrative was added to a set of gold-standard annotations. The deviations from that gold-standard, that the narratives of the three individual annotators show, are used as a baseline for evaluating the output of the LLM. The deviations are divided into two categories: minor and major deviations. Minor deviations are mostly formal mistakes or subtle subjective differences from the gold-standard, where the core of the narrative has been annotated correctly. In contrast, major deviations contain cases in which a narrative from the gold-standard is missed, a narrative is extracted that does not exist in the gold-standard, or a narrative is coded entirely incorrectly. This differentiation helps us to more accurately evaluate the performance of our LLM instead of treating every deviation as equally bad, no matter the impact or gravity.

For our analysis, we use the model `GPT-4o` (OpenAI et al., 2024), the newest flagship model by OpenAI at the time of prompting. We decided to use this model, as initial experiments with open source models of the `Llama` model family (Dubey et al., 2024) didn't show satisfactory results. Because `GPT-4o` is not an open source model, the results of our paper might not be reproducible in the future if OpenAI decides to take the model offline. We do however believe that using this model is the best option to analyze the capabilities of current state-of-the-art LLMs and suspect that in the future, similar results will be possible with open source models, when their development has caught up to the current state of commercial models.

To find the optimal possible prompting strategy, we evaluate different prompts on 20 out of the 100 annotated documents. The remaining 80 documents are not used to tune our prompting, but are instead used only to generate narratives using the final prompt. This is done to fairly evaluate the out-of-sample performance of `GPT-4o`. To find the final prompt, we varied in what length we described the task and how many few-shot examples to give our LLM. As recent works suggest that LLMs suffer from a lost-in-the-middle effect (Liu et al., 2024), according to which information is lost in the middle of a prompt if the prompt is too long, we try to keep our prompt as concise as possible. To do that, we experimented with `GPT-4o` to see which parts of the codebook were common causes for errors and expended our explanations on these, while shortening our explanations for parts of the codebook that worked well. We then tested out different numbers of few-shot examples from our 20 annotated texts that we gave `GPT-4o` as additional input to learn from. We find that a high number of few-shots causes the aforementioned lost-in-the-middle effect, resulting in worse performance. Our zero-shot experiments, in which no examples were added to the prompt, were not promising either, as the model seems to lack critical codebook understanding without being given an example. When we gave the model only one or two examples, it tried to emulate the exact narrative structures of those examples and didn't show flexibility for other narratives. We thus decided in a trade-off between lack of flexibility with too few examples and an information loss due to a long prompt with too many examples, to choose seven few-shot examples, as this seemed to be the sweet-spot in our experiments.

Figure 4 visualizes the workflow that the LLM follows using our final prompt. We utilize a chain-of-thought prompting technique (Wei et al., 2022), in which the LLM is asked to not solve the entire extraction task in one step, but instead solve it in smaller steps, resembling the pipeline structure of previous narrative extraction approaches (see Section 2.3).

Table 2 shows the results of `GPT-4o` compared to our three annotators. We can see that the three annotators have a higher accuracy compared to the gold standard, ranging from 74% to 67% compared to `GPT-4o`'s 44%. This result shows that the task is very subjective, causing even annotators that have worked on this very topic together for years to still disagree in one out of three to four cases and that `GPT-4o` is not yet at the level of expert annotators, showing many major deviations from the gold standard. Many of these deviations are caused by a bias that the LLM exhibits: `GPT-4o` shows a lower standard deviation for the number of narratives found in a text, indicating that it has a prior expectation of how many narratives it is supposed to find. The three annotators instead analyzed each text without any expectations.

Table 2: Narrative counts per document and deviation rates for human experts and `GPT-4o` (Schmidt et al., 2025).

Measure	Expert 1	Expert 2	Expert 3	Model	Gold
# Narratives	2.22	2.36	2.29	2.32	2.91
# Narratives Standard deviation	1.95	2.12	2.04	1.15	2.35
Major-deviation rate	0.40	0.35	0.49	1.25	—
Unexpected major deviations	—	—	—	0.91	
Accuracy (vs. gold)	0.72	0.74	0.67	0.44	—
Jaccard similarity	0.59	0.60	0.59	0.40	—

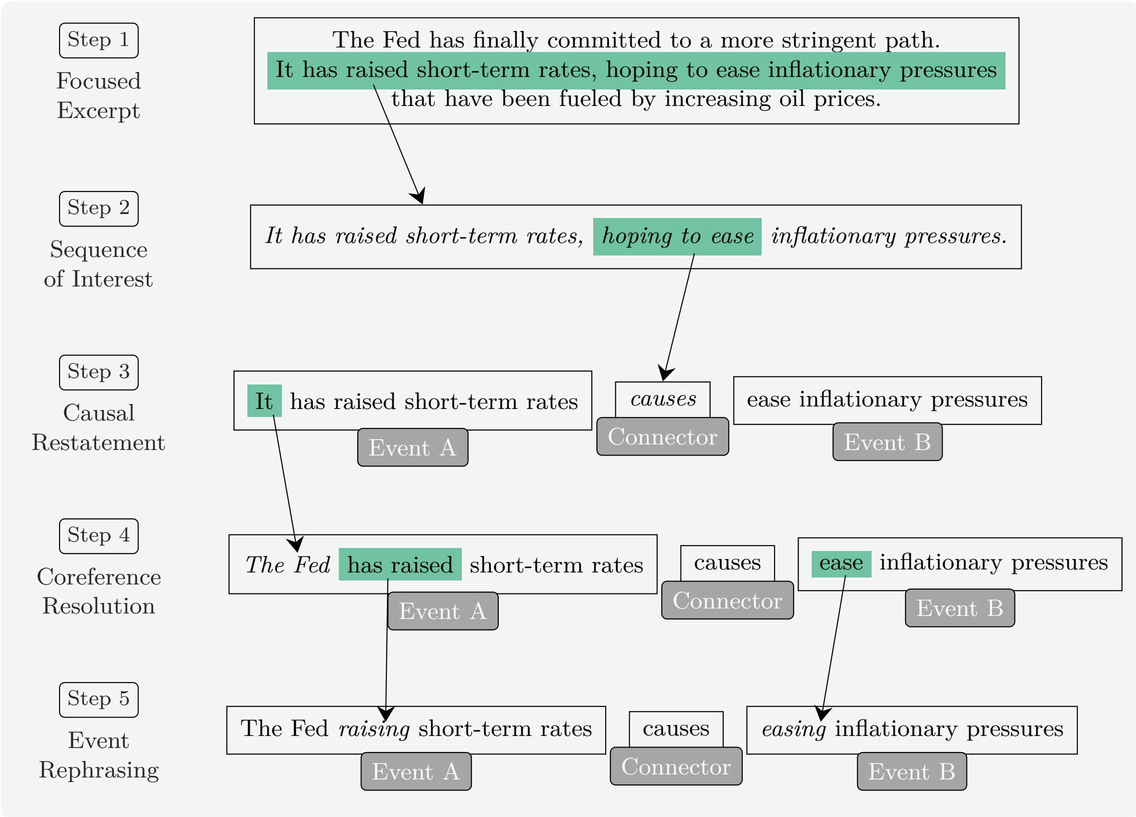


Figure 4: A diagram depicting the Chain-of-Thought transformations of our prompting strategy. Segments to be changed in the subsequent step are highlighted in green and arrows map the source segments to their corresponding results (Schmidt et al., 2025).

We conclude this part of our narrative extraction that currently, LLMs are not capable of reaching expert performance in the task of narrative extraction and that a human in the loop is therefore still needed. Despite this, the results of our LLM look very promising, particularly when looking at the individual texts and reading its narratives, and will likely further improve to approach the level of experts with future developments in NLP.

Lastly, we demonstrate how the results of narrative extraction tasks can be aggregated in postprocessing. We perform a scaling technique similar to `Lex2Sent` (see Section 2.2) to assign each narrative to a set of macroeconomic categories. To do this, we first use another LLM prompt to assign the events found to a particular semantic topic (e.g. inflation expectations). Afterwards, we use the encoder-only model `all-MiniLM-L6-v2`, that is specialized on creating embeddings for entire documents, to find the closest of the pre-defined macroeconomic categories for each semantic topic. We then translate the valence indicated in the event into a downward or upward arrow (e.g. rising inflation becomes $\uparrow$ inflation). This process is visualized in Figure 5.

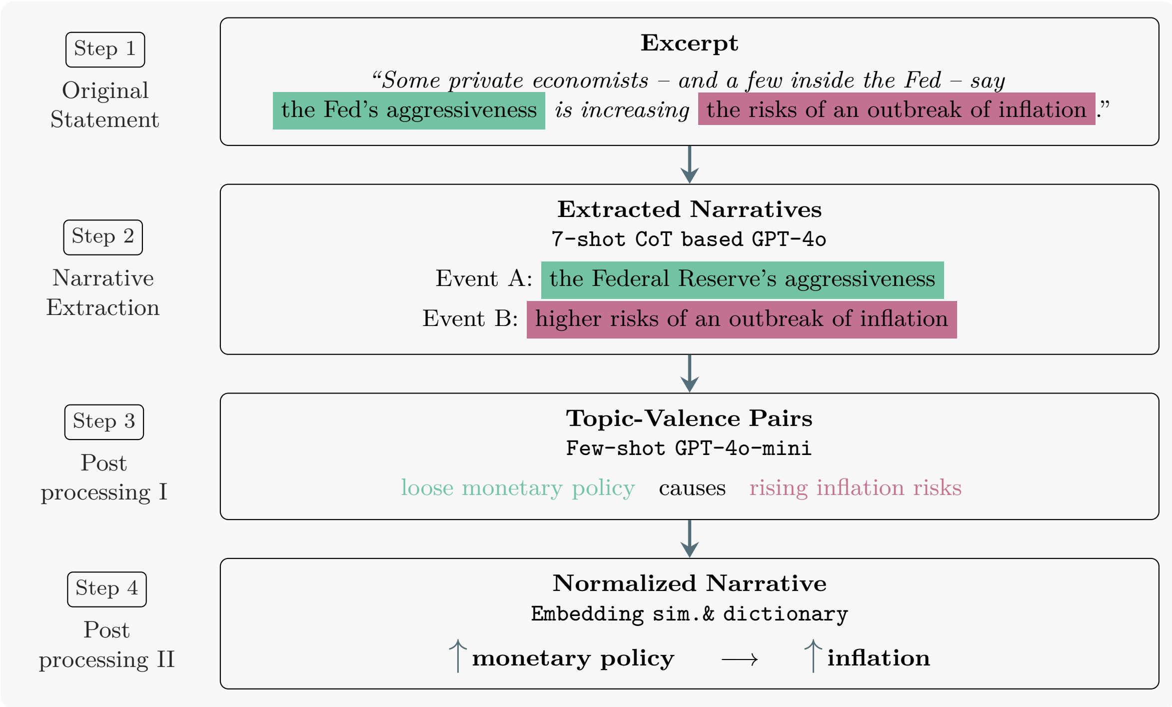


Figure 5: Illustration of our narrative abstraction pipeline using a real example from the corpus (Schmidt et al., 2025).

We find that, even in such a small data set, we can already observe the type of aggregated narratives that we would expect to find when analyzing a corpus of newspaper articles. For instance, the most common aggregated narrative found is "↑ government spending - causes - ↑ inflation".

We believe that both this narrative extraction as well as the aggregation technique need to be applied to a larger corpus in the future to fully understand their potential.

2.5 Outlook: Atemporal narrative extraction

Ultimately, I propose that LLMs are already capable at performing such a complex task at a decent level, but need to be used carefully to account for possible biases with a human in the loop to check on them. I suspect that, while the performance of LLMs can still improve, their performance will likely be limited due to the high subjectivity that is integral to the task itself. As the annotations of three expert annotators still show noticeable differences, even after refining the codebook together for years, it raises the question how much we should rely on gold-standards to extract narratives when following complex definitions like the one of Roos et al. (2024). Further analyses will need even more refined codebooks that cover even more edge cases to improve upon the result. But even with more refined codebooks, human subjectivity in interpreting documents and the message conveyed by the author will not fully fade.

Similarly, the automated extraction of narratives using instruction LLMs can theoretically be improved by improving the prompt and choosing "optimal" few-shot examples that perfectly weight different edge cases. The performance will likely also improve in the future due to the ongoing development of models like `GPT-41`, OpenAI's latest announced model as of writing this dissertation.

However, the largest improvement can most likely be performed by fine-tuning an instruction LLM to the specific task of narrative extraction. Preparing prompts and few-shot examples to let the model understand what the task is about is a unique ability of instruct LLMs, but users do not have to rely on this ability alone. Like other language models, they can be fine-tuned on a new task by altering the weights within their underlying neural network or by adding new trainable weights. The latter option is called Parameter-Efficient-Fine-Tuning (PEFT, Xu et al., 2023), in which the weights of the original model stay frozen while new trainable weights are added at certain spots within the model. This yields the advantage of limiting the environmental foot print and lowering the computational requirements of fine-tuning the model. For instance, the authors of the PEFT method Quantized Low-Rank Adaptation (`QLORA`, Dettmers et al., 2023) claim that that a 65B parameter `Llama` model only requires 48 GB of GPU RAM to fine-tune using their PEFT method. Utilizing such a PEFT method to fine-tune an open-source instruct LLM like `Llama 4` is a logical next step to improve the results of the narrative extraction task and to move away from the dependence on state of the art commercial LLMs.

With fine-tuning, it is also possible to tackle the obstacle of human subjectivity. As each annotator generates their own list of narratives, it is possible to train an LLM based on the narratives coded by each annotator separately. The results of the LLMs, which each represent one annotator, can then be aggregated, similar to how we combined multiple language models into one improved classifier in `Lex2Sent` (see Section 2.2).

3 Analyzing narratives using diachronic language modeling

While most narrative extraction methods and pipelines, such as `RELATIO` (Ash et al., 2024) can utilize post-processing to aggregate results to portray the shift of narratives over time, the language models used do not explicitly model the evolution of language over time. Diachronic language models instead incorporate temporal biases into the modeling process, enabling us to more effectively quantify the change of language and, as a result, narratives over time. Researchers have proposed adaptations of models in all four major model categories presented in Section 1.2.

Frequency-based methods often rely on statistical integer-based time series models. For instance, Jentsch et al. (2020a,b) propose the document scaling method called `Poisson Reduced Rank Model` that scales the documents on an politically ideological scale of one or more dimensions, representing an improvement over atemporal document scaling methods such as `Wordfish` (Slapin et al., 2008).

Diachronic topic models, most often referred to as dynamic topic models inspired by the `Dynamic Topic Model` proposed by Blei et al. (2006), model the change of topics over time. The model `RollingLDA` (Rieger et al., 2021) is particularly interesting for narrative extraction tasks on modern data sets, as it utilizes a rolling window approach. This approach allows the model to "remember" important events and patterns in the recent past, while also allowing it to "forget" older, less relevant events that would only add noise when trying to model new time chunks. It is thus well-suited to model periodic events such as quarterly trends. To do this, the model uses a memory parameter m, which denotes the size of the rolling window. In each time chunk, a new `LDA` is trained but receives the topic assignments from the previous $\mathcal{M}$ time chunks as an additional input to automatically align its new predictions with the topics created in previous time chunks. As the topic assignments in the initialization period thus play an important role, it uses `LDAPrototype` (Rieger et al., 2020), a method that trains multiple `LDA`s and chooses an adequate representative (called prototype) from among them to combat `LDA`'s non-deterministic training.

Diachronic static and contextualized embedding models enable users to detect semantic changes over time. Hamilton et al. (2016) propose laws of semantic change and use diachronic static embeddings to detect such changes. They train a `Word2Vec` model (Mikolov et al., 2013a,b) on each pre-defined time chunk and align the resulting vector spaces using appropriate rotation matrices. Assuming that the vector spaces are just linear combinations of each other when no semantic change occurs, the authors thus look for the words whose embedding changed the most, indicating that that particular word's meaning has changed over time.

As contextualized embedding methods usually utilize a pre-trained language model, their diachronic variants are not trained from scratch in each time chunk. Instead, researchers use the contextualized nature of the embeddings and create an aggregated measure of the similarity between the embeddings from one time chunk to the next (Kutuzov et al., 2020, 2022). This similarity can be used to infer if the context in which a word is used changes over time. A sudden drop in self-similarity over time thus points towards a sudden change in context, indicating a semantic change.

In this Section, I present works of my co-authors and myself that contribute towards the literature by proposing diachronic narrative extraction techniques. While all aforementioned diachronic language modeling techniques have unique advantages and use cases, we primarily focus on frequency-based methods as well as Dynamic topic models in our work.

3.1 Contributed publications

Benner, Niklas, Kai-Robin Lange, and Carsten Jentsch (2022). "Named entity narratives". In: *Ruhr Economic Papers* 962. DOI: `10.4419/96973126`.

Lange, Kai-Robin, Niklas Benner, Lars Grönberg, Aymane Hachcham, Imene Kolli, Jonas Rieger, and Carsten Jentsch (2025a). *ttta: Tools for Temporal Text Analysis*. arXiv: `2503.02625 [cs.CL]`.

Lange, Kai-Robin and Carsten Jentsch (2023). "SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments". In: *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*. Association for Computational Lingustics, pp. 19–28. URL: `https://aclanthology.org/2023.cpss-1.3/`.

Lange, Kai-Robin, Jonas Rieger, Niklas Benner, and Carsten Jentsch (2022b). "Zeitenwenden: Detecting changes in the German political discourse". In: *Proceedings of the 2nd Workshop on Computational Linguistics for the Political and Social Sciences*, pp. 47–53. URL: `https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf`.

Lange, Kai-Robin, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025b). “Narrative Shift detection: A hybrid approach of Dynamic Topic Models and Large Language Models”. In: *Proceedings of the Text2Story’25 Workshop*. Best Paper Award (runner-up). URL: https://www.di.ubi.pt/~jpaulo/Text2Story2025/paper6.pdf.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). “Dynamic change detection in topics based on rolling LDAs”. In: *Proceedings of the Text2Story’22 Workshop*. Vol. 3117. CEUR-WS, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/.

3.2 Temporal narrative analysis with a focus on named entities

In Benner et al. (2022), we assume that most economic narratives revolve around entities, such as politicians, companies, or governments. While this is implicitly the case for many definitions that require a story, which usually contains such entities, some definitions like the NPF (Jones et al., 2010) explicitly require “characters” that take an active or passive role in a narrative. Based on this assumption, we propose an algorithm that detects events surrounding entities in the media.

To do this, we analyze how often certain entities have been mentioned in newspaper articles and then look at the words that they most often co-occur with to be able to interpret the stories surrounding them. We first perform named entity recognition, a classic NLP task to detect entities within a text. While this task is traditionally performed by using fine-tuned versions of pre-trained models like `BERT` (Devlin et al., 2019) to detect the spans of entities, we instead opt for the more easily scalable option of cross checking each document for the mention of an entity that has its own Wikipedia article. As we are interested in narratives popular enough to not only appear in newspaper articles but to also have narratives surrounding them, we argue that each entity that popular will most likely have a Wikipedia article attached to it. Thus, our performance does likely not decrease decisively, when using this more scalable option.

After identifying the entities, we focus on specific entities for which we observe how frequently they appear in the corpus over time using daily time chunks. We assume the frequency in which entity i is mentioned at time chunk t to be a realization of a zero-inflated Poisson distributed random variable $X_i^{(t)}$ (Jazi et al., 2012)

$$P(X_i^{(t)} = x) = \begin{cases} \pi_i^{(t)} + (1 - \pi_i^{(t)}) e^{-\lambda_i^{(t)}}, & \text{if } x = 0 \\ (1 - \pi_i^{(t)}) \dfrac{e^{-\lambda_i^{(t)}} \left(\lambda_i^{(t)}\right)^x}{x!}, & \text{if } x \in \mathbb{N} \end{cases}$$

with a Poisson parameter $\lambda_i^{(t)} \in (0, \infty)$ and an inflation parameter $\pi_i^{(t)} \in (0, 1)$ that denotes the probability of the entity not appearing in any document in a given time chunk. These two parameters are modeled dynamically using a rolling window of 365 days with dynamic weighting. We detect an event when the observed frequency exceeds the 99.9% quantile of this theoretical distribution. This detection returns specifically those points in time in which an entity appears extraordinarily often in the news. Figure 6 shows the frequency in which the entity "Wladimir Putin" occurred in the news articles of the Süddeutsche Zeitung with vertical bars representing the observed number of news articles containing his name. The 99.9% quantile of the theoretical distribution, and thus our event threshold, is displayed using a dotted line. If the daily mentions of Wladimir Putin exceed the threshold, the vertical bar extends beyond the dotted line and is represented in red color.

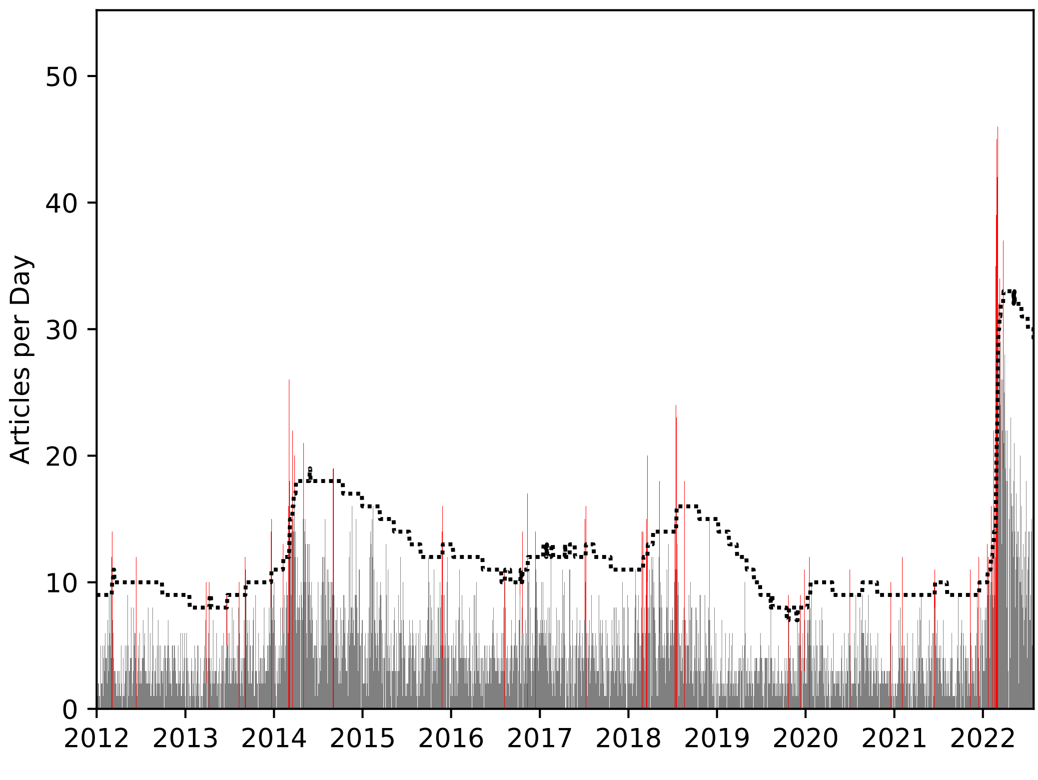


Figure 6: Daily number of newspaper articles from the Süddeutsche Zeitung containing the entity Wladimir Putin from 2012 to 2022 (Benner et al., 2022). The dotted line represents the 99.9% quantile of the theoretical distribution and the vertical bars represent the observed frequencies. If a vertical bar exceeds the dotted line, an event is detected and the bar is colored red.

To interpret why the events are detected, we utilize leave-one-out cosine similarities. We compare the frequencies of all words co-occurring with the named entity for which an event was detected in the month before and after the event. When the event had a long lasting impact, the word frequencies after the event should be more similar to the day at which the event occurred as to the word frequencies of the month before the event occurred. We compare these similarities while always removing one word from the vocabulary vector. In theory, if the word that is removed was important for the event and for the time following the event but not for the time before the event, the similarities should move closer to another. Utilizing this, we are able to get an idea of what happened during the event that had a long lasting impact on the news reporting.

Using such leave-one-out cosine similarities, we observe that, for instance, after the first event detection concerning Wladimir Putin in early 2022, the words “krieg” (engl. war) and “angriffskrieg” (engl. war of aggression) were used with increasing frequency despite Russian officials calling it a “military operation” and not a war. Such instances allow us to infer the framing and stance the authors portrayed about the war at the time.

However, this approach does not come without its downsides. Using this approach, we detect events for entities that we specifically look for. While this allows us to analyze that specific entity over time, we most likely will not find events and thus narratives that surround other entities that we do not specifically look for. We therefore would need to have a specific target for a narrative analysis, instead of being able to capture narrative shifts in the entire corpus. To address this issue, in the following sections, I present work that instead focuses on finding changes in the content of the articles and then refer it back to entities and narratives.

3.3 Topical Change Detection

Instead of focusing on named entities, diachronic narrative analyses can also be performed by focusing on topics. In Rieger et al. (2022), we utilize `RollingLDA` (Rieger et al., 2021) to detect changes in topics over time. This `Topical Changes` method monitors the word-topic distributions of `RollingLDA` in each time chunk and detects a change if the distribution changes more than expected.

Given a word frequency vector for topic $k \in \{1, \dots, \mathcal{K}\}$ in time chunk $t \in \{0, \dots, \mathcal{T}\}$

$$\boldsymbol{f}_{k|t} = \left(f_{k|t}^{(\bullet 1)}, \dots, f_{k|t}^{(\bullet V)}\right)^T \in \mathbb{N}_0^{\mathcal{V}}.$$

we monitor the cosine-distance between the observed word count vectors of a time chunk and previous time chunks. For this, we do not just consider the last time chunk, but up to $z_{\max}$ aggregated previous time chunks. $z_{\max}$ allows us to prevent our model from detecting known seasonal trends, such as quarterly earning calls that might be represented in the data. However, to assure that the word frequency vectors of the previous time chunks all follow the same distribution, we use only time chunks since the last detected change. Formally, for a calibrated threshold $q_k^{(t)}$ we define the set of changes in topic k up to time t as

$$C_k^{(t)} = \left\{ u \mid 0 < u \leq t \leq \mathcal{T} : \cos\left(\boldsymbol{f}_{k|u}, \boldsymbol{f}_{k|(u-z_k^{(u)}):(u-1)} \right) < q_k^t \right\} \cup 0,$$

with $\boldsymbol{f}_{k|(t-z_k^{(t)}):(t-1)} = \sum_{z=1}^{z_k^{(t)}} \boldsymbol{f}_{k|t-z}$ and choose $z_k^{(t)} = \min\left\{ z_{\max}, t - \max C_k^{(t-1)} \right\}$.

Utilizing the estimated word-topic distributions $\hat{\psi}_k^t$ of time chunk t and topic k as well as $\hat{\psi}_k^{(t-z_k^{(t)}):(t-1)}$ for the reference period of time chunk t and topic k, we then calculate the dynamic threshold $q_k^{(t)}$. To do this, we first create an aggregated distribution of word change that we would expect to see. For this, we introduce a mixture parameter $p \in (0, 1)$ that is chosen by the user and represents how much the user assumes the topics to change over time. As language and what people talk about always changes, we would almost always detect a change when just testing for any change in word frequencies. This mixture parameter allows us to instead assume that a certain amount of change will always happen and does not need to be detected. Instead, we detect fewer but more impactful changes. We use the mixture parameter to create a mix of the estimated word-topic distribution of our current time chunk for which we test and its reference time chunks:

$$\tilde{\psi}_k^{(t)} = (1-p)\, \hat{\psi}_k^{(t-z_k^{(t)}):(t-1)} + p\, \hat{\psi}_k^{(t)}.$$

We then simulate this distribution $\tilde{\psi}_k^{(t)}$ a total of B times to generate parametric Bootstrap samples $\tilde{\boldsymbol{f}}_{k|t}^{(b)}, b = 1, \ldots, B$ and set our threshold $q_k^{(t)}$ to the 1% quantile of the cosine similarity vectors between our samples and our reference distribution

$$q_k^{(t)} = \left(\cos\left(\tilde{\boldsymbol{f}}_{k|t}^{(1)}, \boldsymbol{f}_{k|(t-z_k^{(t)}):(t-1)} \right), \ldots, \cos\left(\tilde{\boldsymbol{f}}_{k|t}^{(B)}, \boldsymbol{f}_{k|(t-z_k^{(t)}):(t-1)} \right) \right)_{(0.01)}$$

where $x_{(p)}$ denotes the p-quantile of x. We then perform a Bootstrap percentile test using this threshold and the observed similarities following the formula of $C_k^{(t)}$.

We test our method on a corpus of CNN news articles dating from January 22nd 2020 to December 12th 2021 concerning the COVID-19 pandemic. We chose to model $\mathcal{K} = 12$ topics with a mixture parameter of $p = 0.85$. The results can be seen in Figure 7. The image shows the observed cosine similarities between the current time chunk and the reference period for all topics in all time chunks as line graphs. Each block represents one topic, containing the topic's top words at the bottom for the purpose of interpretation. The red lines represent the threshold $q_k^{(t)}$ corresponding to that topic and time and the blue lines display the observed cosine similarities. If the blue line showing the observed

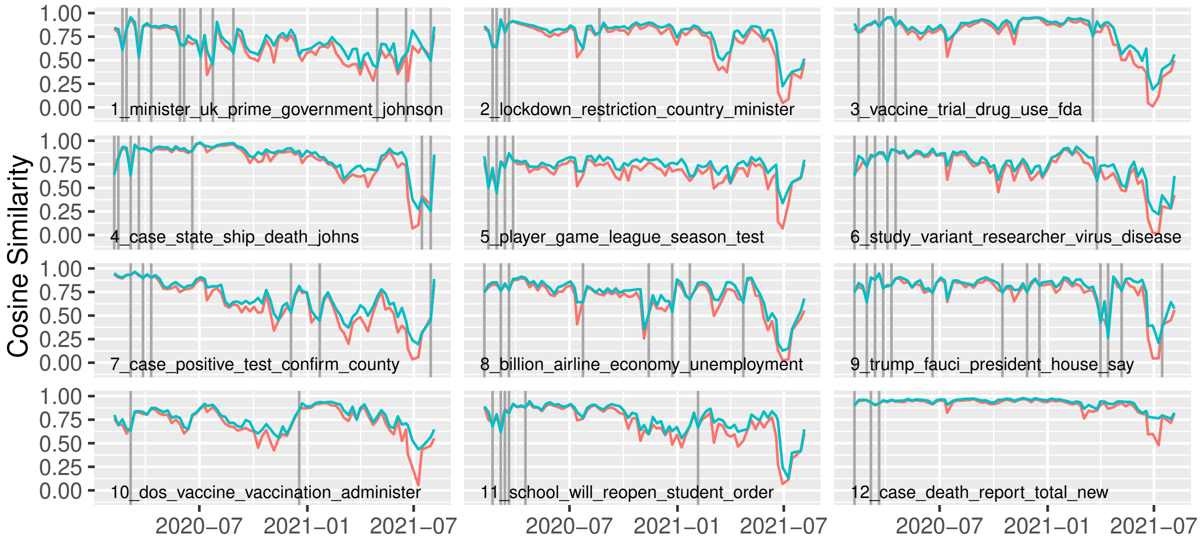


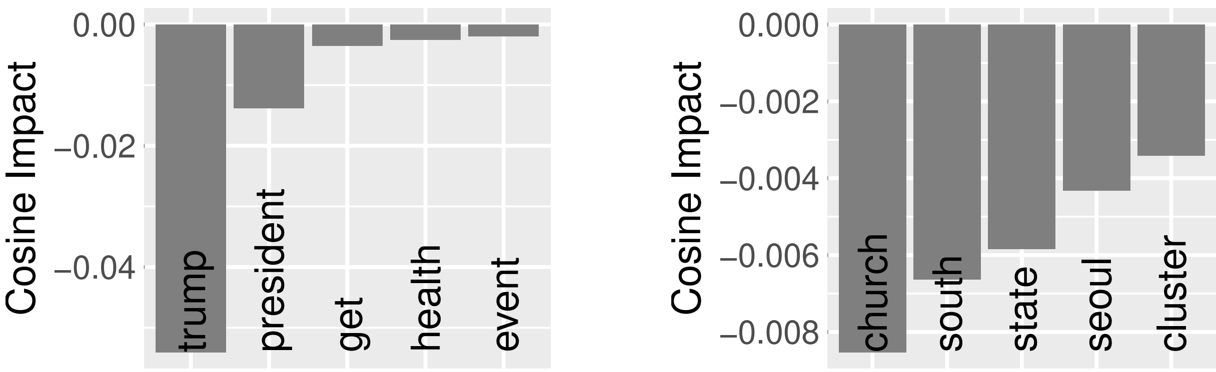
Figure 7: Result of applying the `Topical Changes` method on a corpus of COVID-19 articles from CNN showing the observed similarity (blue), thresholds $q_k^{(t)}$ (red) and detected changes C_k (vertical lines, gray) over the observation period for all topics $k \in \{1, \ldots, 12\}$ (Rieger et al., 2022).

similarities drops below the threshold in red, we detect a change in that topic and the graph shows a gray vertical line.

To interpret the changes, we use leave-one-out word impacts, similar to the technique presented in Section 3.2. We compare the similarities of a topic word count vector at a change point with the word count vector of the same topic within the reference period. We calculate this similarity by removing the same word from both vocabularies and see how much the similarity changes when that word is removed. When the observed similarity of word vectors is higher when the word is removed, the usage frequency of the word in question changed between the reference period and the change point. The higher this change in frequency is relatively to the word's base frequency, the larger of an impact removing the word from the similarity calculation will have. We thus use this leave-one-out word impacts to find those words whose frequency change in a topic have the highest impact on the detected change. Two such examples can be seen in Figure 8. Figure 8a on the left shows a change detected when Donald Trump, at that time in his first term as US President, returned to office after being diagnosed with COVID. Figure 8b on the right symbolizes the media coverage after a mass spreader event in a South Korean church.

3.3.1 Example: Detecting Topical Changes in the German Bundestag

To further test our model proposed in Rieger et al. (2022), in Lange et al. (2022) we perform a change detection on a corpus of German Bundestag speeches. As this corpus covers a



(a) Leave-one-out word impacts for topic 9 (2020-08), caused by Donald Trump's return to office after being diagnosed with COVID-19.

(b) Leave-one-out word impacts for topic 2 (2020-10), caused by the a COVID-19 outbreak caused by a mass spreader event in a South Korean church.

Figure 8: Leave-one-out impacts of two selected changes from applying the `Topical Changes` model on a corpus of COVID-19 articles (Rieger et al., 2022).

larger time span with arguably more and broader talking points than the COVID-19 corpus of the method's original paper, this is a good corpus to see if the method is generalizable.

For this purpose, we scrape the protocols of all Bundestag plenary sessions and cut off their start until the first speech as well as the appendix. We then split the plenary sessions into individual speeches using Regex, resulting in a total of 335,065 speeches.

As we expect a broader array of topics, we compare the results of a grid of $\mathcal{K} = 20, \ldots, 35$ topics as well as a mixture parameter of $p = 0.9, \ldots, 0.95$. We decide to use $\mathcal{K} = 30$ topics with a mixture parameter of $p = 0.94$. An overview of our topics and their respective detected change points is shown in Figure 9.

Our experiment provides us with mostly reasonable and well interpretable changes within the last 80 years of German politics. For instance, Figure 10 shows the leave-one-out word impacts of two selected changes. Given the words and dates in which the change has been detected, one can easily figure out that these changes were caused by the Russian invasion of Ukraine in 2022 as well as the financial crisis in 2008. However, in the name-based topic 9, many changes occurred after election years, indicating that the changes are detected due to a change in members of parliament rather than any change in the discourse.

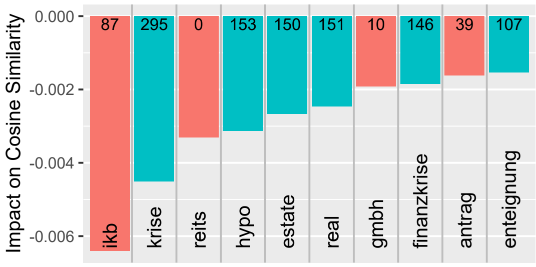
Ultimately, the experiment shows that the model can also be used on a corpus with different characteristics and still perform well, but has a unique sensibility to certain stop words, such as the names of the members of parliament in this example.

Figure 9: Result of applying the `Topical Changes` model on a corpus of Bundestag speeches. Observed similarity (blue), thresholds $q_k^{(t)}$ (red) and detected changes C_k (vertical lines, gray) over the observation period for all topics $k \in \{1, \ldots, 30\}$ (Lange et al., 2022b).

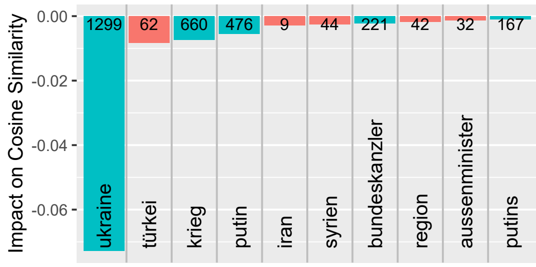
3.4 Extending Topical Changes to Narrative shifts

While the two aforementioned works indicated that the `Topical Changes` method is applicable to a wide array of corpora, the methodologies presented in Section 3.2 and Section 3.3, are still limited in their interpretation. The leave-one-out word impacts might tell us what happened during the time chunk of a time period but most of the time they do not provide us with a framing in combination with a story. Similar to `RELATIO` (see Section 2.3), the methods provide its user with fragments of a story, but make no attempt at adding a subjective element to the result. Thus, even if the user interprets the leave-one-out word impacts to construct a story out of the change, it most often still is no narrative. Instead, a human in the loop would be required to look into the documents themselves to extract narratives at the given change points.

In Lange et al. (2025), we attempt to add another layer to the `Topical Changes` model. By utilizing an instruction LLM that interprets our changes and categorizes them into



(a) Leave-one-out word impacts for topic 15 (2008-09), caused by the financial crisis.



(b) Leave-one-out word impacts for topic 21 (2021-22), caused by the Russian-Ukrainian war.

Figure 10: Leave-one-out impacts of two selected changes from applying the `Topical Changes` model on a corpus of Bundestag speeches (Lange et al., 2022b).

a narrative framework, we are able to move beyond mere topical changes to narrative shifts. This methodology can thus be interpreted as a combination of the strengths of the methodologies presented in Section 2.4 and Section 3.3. The `Topical Changes` method allows us to work with large corpora without having to let the instruction LLM work with every single document, making the method scalable to large corpora despite working with an LLM with billions of parameters. Instead, the scalable and computationally much less demanding dynamic topic model `RollingLDA` (Rieger et al., 2021) and the `Topical Changes` method are applied to each text. Then, the instruction LLM `Llama 3.1 8B` (Dubey et al., 2024) is only prompted to extract narratives for a select few documents that correspond to changes detected by the `Topical Changes` method.

Contrary to the procedure described in Section 2.4, we do not focus on individual documents in this work. Because the `Topical Changes` method returns changes in the word-topic distribution that manifested in many documents, we instead opt to provide `Llama` with multiple documents at once to allow it to get a bigger picture to analyze. We select these documents by utilizing the leave-one-out word impacts provided by the `Topical Changes` method: `Llama` is provided with documents of the time chunk in which a change was detected that contained the most words that are part of the 10 words that show the highest leave-one-out word impact. This way, we assure that the documents provided deal not only with the topic in which the change was detected, but also with exactly those words that are deemed to be responsible for the change detection. We also provide some additional information about the topic's top words during that and the previous time chunk as well the date at which the change occurred for context.

We then prompt `Llama` to determine if a shift in narrative has occurred from the last time chunk to the current one. If there is a shift in narrative, `Llama` is prompted to formalize it by utilizing the definition and narrative elements of the NPF (Jones et al., 2010; Schlaufer et al., 2022; Shanahan et al., 2018). To evaluate the model's performance, three expert annotators manually coded the detected changes. Just like `Llama`, they were tasked to decide if the change represents a change in narrative and, if so, formalize it according to the NPF.

The annotators coded 37 out of 68 detected changes to contain narrative shifts. `Llama` deemed 60 out of 68 detected changes to contain narrative shifts. This results in an accuracy of 57.35% and an f1-score of 0.7010. We deem the misclassifications to be most likely caused by `Llama`'s tendency to satisfy the user by fulfilling the task, similar to a confirmation bias, and interpreting too many changes as narrative shifts in the process. However, when a detected change was in fact a narrative shift according to the annotators, `Llama` correctly identified the narrative elements in 31 out of 37, so 83.78% of the cases.

3.5 Software publications

In addition to the previously covered methodological contributions, we published a new data set based on German parliamentary speeches as well as a Python library that aims to become a collection of diachronic NLP tools.

3.5.1 Curated data set: A German parliament speech corpus

While new textual data is generated every day, there is also an ongoing effort to digitize old documents. This also includes parliamentary speeches that are integral for political researchers. In the case of German state parliaments however, these speeches are mostly just published as pictures of the documents written with typewriters and also only on the website of the respective state parliament. If researchers attempt to perform any kind of quantitative research, this can be a major obstacle.

Along with the paper "SpeakGer: A meta-data enriched speech corpus of German state and federal parliament" (Lange et al., 2023), we published a collection of speeches from the German Bundestag as well as all 16 German state parliaments, covering the time frame from 1947 to 2022.

To collect this data, we scraped both the PDFs of all state parliament proceedings and meta-data about the members of each parliament at any given time from Wikipedia. The sources we used to scrape the protocols from can be seen in Table 3.

Table 3: Sources and links to all protocols that were analyzed. If the protocols of a parliament cannot be found in one place, we provide multiple sources for all possible legislative periods.

Parliament (English name)	Legislative period	Source
Baden-Württemberg (Baden-Wuerttemberg)	12-17 1-11	Landtag von Baden-Württemberg Württembergische Landesbibliothek
Bayern (Baveria)	1-18	Bayrischer Landtag
Berlin	12-19 6-11 1-5	Abgeordnetenhaus Berlin Zentral- und Landesbibliothek Berlin Zentral- und Landesbibliothek Berlin
Brandenburg	8-10 1-7	Landtag Brandenburg Parlamentsspiegel
Bremen	18-20 7-17	Bremische Bürgerschaft Parlamentsspiegel
Bundestag	1-20	Deutscher Bundestag
Hamburg	20-22 6-19	Hamburgerische Bürgschaft Parlamentsspiegel
Hessen	1-20	Hessischer Landtag
Mecklenburg-Vorpommern (Mecklenburg-Western Pomerania)	1-8	Landtag Mecklenburg-Vorpommern
Niedersachsen (Lower Saxony)	17-18 8-16	Landtag Niedersachsen Parlamentsspiegel
Nordrhein-Westfalen (North Rine Westfalia)	1-18	Landtag Nordrhein-Westfalen
Rheinland-Pfalz (Rhineland Palatinate)	1-18	Landtag Rheinland-Pfalz
Saarland	14-17 7-13	Landtag des Saarlandes Parlamentsspiegel
Sachsen (Saxony)	1-8	Sächsischer Landtag
Sachsen-Anhalt (Saxony-Anhalt)	6-8 1-5	Landtag von Sachsen-Anhalt Parlamentsspiegel
Schleswig-Holstein	1-20	Schleswig-Holsteiner Landtag
Thüringen (Thuringia)	4-7 1-4	Thüringer Landtag Parlamentsspiegel

Afterwards, we used the Optimal Character Recognition engine `tesseract` (Kay, 2007) to map characters on the PDF to the pixels they are written on. While `tesseract` makes mistakes when trying to recognize characters, particularly for old documents with bad picture quality, this is an important step to turn this collection of image data into textual data. We then cut off everything until the start of the parliamentary session as well as the appendix and split the speeches. These splits are based on heuristic criteria, which we inferred by the structure in which the documents were written.

For instance, for most protocols, a new speech is denoted by the name of the new speaker, followed in a colon and their speech in a new colon. We detect these split points by looking for the name of a member of parliament based on the data we scraped from Wikipedia using Regex, similar to our named entity recognition in Section 3.2. These heuristics can however fail, as such a form of the name of an member of parliament followed by a colon can also occur as a part of a speech when that member of parliament is addressed directly. Additionally, the name of a new speaker might not be recognized at all, when `tesseract` made a mistake and altered the name so that it cannot be matched with the Wikipedia data or if it misinterpreted the colon for a different character. Still, these heuristics are currently the best way to detect such speech splits.

Ultimately, the SpeakGer corpus contains 17,784,802 speeches from 17 parliaments. I hope that this corpus will help researchers to conduct deep analyses on German politics and political economics in the future.

3.5.2 Python library ttta

One of the biggest obstacles in working with temporal NLP methods is the lack of a Python library that serves as a continuously updated collection of such methods. The most commonly used NLP methods can be used with the gensim (e.g. `LDA` and `Word2Vec`, Rehurek et al., 2011) or the transformers library (e.g. `BERT`, `Llama`, Wolf et al., 2020). These libraries provide their user base, including many researchers, with code that is consistently updated to work with newer Python versions. This alone makes using the code for research a lot easier, since Python code is particularly prone to show dependency issues when upgrading to a new Python version. In addition to this, within these libraries, the methods follow a set syntax, enabling users to easily switch between models without the need to completely rewrite their code.

While some of the methods hosted as parts of these libraries can be considered temporal models, such as the `Dynamic Topic Model` (Blei et al., 2006) in gensim, these are not a major focus, nor do the libraries cover a broad array of different models to choose from.

Our Python library "ttta: Tools for Temporal Text Analysis" (Lange et al., 2025a) aims to establish a collection of temporal NLP methods and implement them with a common code interface, which is continuously updated to not only work with the newest Python version, but to also include more models over time.

In its current state, as of writing this dissertation, it contains the `Topical Changes` method (see Section 3.3) as well as the `Narrative Shift` method (see Section 3.4), a diachronic static embedding method, a diachronic contextualized embedding method, as well as two document scaling methods. The `Topical Changes` method is based on the topic models `LDAPrototype` (Rieger et al., 2020) and `RollingLDA` (Rieger et al., 2021), which have also been translated from their respective R packages (Rieger, 2020, 2021) with a new and fast C-implementation of the underlying `LDA` (Blei et al., 2003) algorithm. The diachronic static embedding method is based on the work of Hamilton et al. (2016), who use rotation matrices to align the latent vector spaces of `Word2Vec` models (Mikolov et al., 2013a,b) trained on different time chunks to make the embeddings comparable across time. The contextualized embedding model is based on the work of Hu et al. (2019), who use contextualized embeddings to relate the occurrences of polysemous words back to one of their word senses. This allows users to analyze the development of word sense usage over time. Lastly, the two document scaling methods cover the two iterations of the `Poisson Reduced Rank Model` (Jentsch et al., 2020a,b), which use an integer auto regressive model to model the frequency of word usage across time for different authors.

The ttta library is available on GitHub and on the Python Packaging Index PyPi.

3.6 Outlook: Combining scalable models and LLMs

Due to the complex nature of economic narratives, the works presented in this Section should mainly be seen as the foundation for future research in this area. Our paper on narrative shifts (Lange et al., 2025b) shows that it is possible to combine scalable diachronic NLP models with instruction LLMs to analyze the change of narratives across time. I deem this combination to be very promising for diachronic narrative extraction for two reasons. Firstly, we lack the possibility to train LLMs to specific time chunks on a given corpus to create diachronic versions of these models. This is the case not only because fine-tuning these models is very costly, but also because a corpus rarely comes with a instruction task the model could be trained on. Secondly, the inference of LLMs is currently not scalable to large corpora for small research groups. Even when it is possible to use them on a large corpus, it is questionable if the positive impact on research outweighs the negative environmental impact of the computational power needed (Strubell et al., 2019). Yet, I argue that they are needed as their language understanding capabilities yield a large advantage over every other current NLP model for extracting narratives. In the future, this combination of scalable diachronic models and instruction LLMs could yield to numerous new and promising approaches. The LLM part of the pipeline can be adapted with newly released or fine-tuned models as well as improved prompts for increasingly complex narrative definitions, like I describe in Section 2.5. On the side of the diachronic models, many variations are conceivable as well. For instance, one could use diachronic extensions of static or contextualized embedding models, our named entity event detection in Section 3.2 or document scaling methods like those implemented in the ttta library (Lange et al., 2025a) as a foundation rather than a dynamic topic model.

This way, different kinds of changes across time can be analyzed: A dynamic topic model might cover changes in what the authors talk about, but embedding-based methods would better capture the context in which individual tokens appear. They can for instance be used to track how the authors talk about certain entities, extending upon the idea of our paper on Named Entity Narratives (Benner et al., 2022), or about a specific, pre-defined topic, such as inflation.

Furthermore, the analysis of the temporal developments of economic narratives does arguably not go into enough detail about how and by whom narratives are spread. In addition to observing temporal changes, narrative extraction methods could also consider the effect of other meta-data, such as spatial data or the political alignment of the author. Observing and modeling this meta-data effectively with methods like the `Structural Topic Model` (Roberts et al., 2013) would help to trace back narratives to its origins and to evaluate how common they are in certain parts of the population.

Lastly, all works presented here use either journalistic articles or political speeches as corpora. While these corpora represent major narrative spreaders and are thus important for economic developments, they only contain documents stemming from professionals in their respective field. They do not cover social media, in which people without professional experience can spread narratives to large parts of the population, which arguably has a large effect to the narratives common within the population at large. While social media data is particularly costly to come by, analyzing it would be equally valuable for understanding how narratives are spread in this third major channel.

References

Abbas, Muhammad J., Lafi S. Khalil, Abdulah Haikal, Miriam E. Dash, Gauthier Dongmo, and Kelechi R. Okoroha (2021). "Eliciting Emotion and Action Increases Social Media Engagement: An Analysis of Influential Orthopaedic Surgeons". In: *Arthroscopy, Sports Medicine, and Rehabilitation* 3.5, e1301–e1308. ISSN: 2666-061X. DOI: 10.1016/j.asmr.2021.05.011.

Aratani, Lauren (Oct. 2024). *Trump vows to impose tariffs as experts warn of price hikes and angry allies*. The Guardian. URL: https://www.theguardian.com/us-news/2024/oct/15/trump-tariffs-price-hikes-warnings?CMP=share_btn_url (visited on 06/12/2025).

Ash, Elliott, Germain Gauthier, and Philine Widmer (2024). "Relatio: Text Semantics Capture Political and Economic Narratives". In: *Political Analysis* 32.1, pp. 115–132. DOI: 10.1017/pan.2023.8.

Benner, Niklas, Kai-Robin Lange, and Carsten Jentsch (2022). "Named entity narratives". In: *Ruhr Economic Papers* 962. DOI: 10.4419/96973126.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd ICML-Conference*. ACM, pp. 113–120. DOI: 10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). "Enriching Word Vectors with Subword Information". In: arXiv: 1607.04606 [cs.LG].

Bornheim, Tobias, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski (Mar. 2024). "Speaker Attribution in German Parliamentary Debates with QLoRA-adapted Large Language Models". In: *Journal for Language Technology and Computational Linguistics* 37.1, pp. 1–13. DOI: 10.21248/jlcl.37.2024.244.

Breiman, Leo (1996). "Bagging predictors". In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Brown, Tom B. et al. (July 2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. arXiv: 1911.02116 [cs.CL].

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 NAACL-Conference, Volume 1 (Long and Short Papers)*. ACL, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dubey, Abhimanyu et al. (July 2024). *The Llama 3 Herd of Models*. en. arXiv: 2407.21783v2.

Efron, Bradley (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. ISSN: 0090-5364. DOI: 10.1214/aos/1176344552.

Eliaz, Kfir, Simone Galperti, and Ran Spiegler (Nov. 2024). "False Narratives and Political Mobilization". In: *Journal of the European Economic Association*. ISSN: 1542-4774. DOI: 10.1093/jeea/jvae047.

Eliaz, Kfir and Ran Spiegler (Dec. 2020). "A Model of Competing Narratives". In: *American Economic Review* 110.12, pp. 3786–3816. DOI: 10.1257/aer.20191099.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer (July 2018). "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/W18-2501.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (July 2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences* 120.30. ISSN: 1091-6490. DOI: 10.1073/pnas.2305016120.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235. DOI: 10.1073/pnas.0307752101.

Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv: 2203.05794 [cs.CL].

Gueta, Almog, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart (Apr. 2025). "Can LLMs Learn Macroeconomic Narratives from Social Media?" In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 57–78. ISBN: 979-8-89176-195-7. URL: https://aclanthology.org/2025.findings-naacl.4/.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (Aug. 2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1489–1501. DOI: 10.18653/v1/P16-1141.

He, Ruining and Julian J. McAuley (2016). "Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering". In: *WWW '16: Proceedings of the*

25th International Conference on World Wide Web, pp. 507–517. ISBN: 9781450341431. DOI: 10.1145/2872427.2883037.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing".

Hoyle, Alexander Miserlis, Rupak Sarkar, Pranav Goel, and Philip Resnik (Dec. 2022). "Are Neural Topic Models Broken?" In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5321–5344. DOI: 10.18653/v1/2022.findings-emnlp.390.

Hu, Minqing and Bing Liu (2004). "Mining and summarizing customer reviews". In: *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ISBN: 1-58113-888-1. DOI: 10.1145/1014052.1014073.

Hu, Renfen, Shen Li, and Shichen Liang (July 2019). "Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3899–3908. DOI: 10.18653/v1/P19-1379.

Hutto, Clayton and Eric Gilbert (May 2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 1, pp. 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Jazi, Mansour Aghababaei, Geoff Jones, and Chin-Diew Lai (2012). "First-order integer valued AR processes with zero inflated poisson innovations". In: *Journal of Time Series Analysis* 33.6, pp. 954–963. DOI: 10.1111/j.1467-9892.2012.00809.x.

Jentsch, Carsten, Eun Ryung Lee, and Enno Mammen (July 2020a). "Poisson reduced-rank models with an application to political text data". In: *Biometrika* 108.2, pp. 455–468. ISSN: 0006-3444. DOI: 10.1093/biomet/asaa063.

Jentsch, Carsten, Eun Ryung Lee, and Enno Mammen (2020b). "Time-dependent Poisson reduced rank models for political text data analysis". In: *Computational Statistics & Data Analysis* 142, p. 106813. ISSN: 0167-9473. DOI: 10.1016/j.csda.2019.106813.

Jones, Michael D. and Mark K. McBeth (2010). "A Narrative Policy Framework: Clear Enough to Be Wrong?" In: *Policy Studies Journal* 38.2, pp. 329–353. DOI: 10.1111/j.1541-0072.2010.00364.x.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (2016). "Bag of Tricks for Efficient Text Classification". In: arXiv: 1607.01759 [cs.LG].

Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kay, Anthony (2007). *Tesseract: An open-source optical character recognition engine*. URL: https://www.linuxjournal.com/article/9676.

Khoo, Christopher SG and Sathik Basha Johnkhan (2018). "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons". In: *Journal of Information Science* 44.4, pp. 491–511. ISSN: 0165-5515. DOI: 10.1177/0165551517703514.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large language models are zero-shot reasoners". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3601883.

Kutuzov, Andrey and Mario Giulianelli (Dec. 2020). "UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova. Barcelona (online): International Committee for Computational Linguistics, pp. 126–134. DOI: 10.18653/v1/2020.semeval-1.14.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal (Aug. 2018). "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1384–1397. URL: https://aclanthology.org/C18-1117.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2022). "Contextualized embeddings for semantic change detection: Lessons learned". In: *Northern European Journal of Language Technology* 8. Ed. by Leon Derczynski. DOI: 10.3384/nejlt.2000-1533.2022.3478.

Laboe, Daniel (Jan. 2021). *WallStreetBets: The Big Short-Squeeze*. Zacks. URL: https://www.zacks.com/commentary/1252046/wallstreetbets-the-big-short-squeeze (visited on 06/12/2025).

Lange, Kai-Robin, Niklas Benner, Lars Grönberg, Aymane Hachcham, Imene Kolli, Jonas Rieger, and Carsten Jentsch (2025a). *ttta: Tools for Temporal Text Analysis*. arXiv: 2503.02625 [cs.CL].

Lange, Kai-Robin and Carsten Jentsch (2023). "SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments". In: *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*. Association for Computational Lingustics, pp. 19–28. URL: https://aclanthology.org/2023.cpss-1.3/.

Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch (2022a). "Towards Extracting Collective Economic Narratives from Texts". In: *Ruhr Economic Papers* 963. DOI: 10.4419/96973127.

Lange, Kai-Robin, Jonas Rieger, Niklas Benner, and Carsten Jentsch (2022b). "Zeitenwenden: Detecting changes in the German political discourse". In: *Proceedings of the 2nd Workshop on Computational Linguistics for the Political and Social Sciences*, pp. 47–53. URL: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf.

Lange, Kai-Robin, Jonas Rieger, and Carsten Jentsch (2024). “Lex2Sent: A bagging approach to unsupervised sentiment analysis”. In: *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*. Association for Computational Linguistics, pp. 281–291. URL: https://aclanthology.org/2024.konvens-main.28/.

Lange, Kai-Robin, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025b). “Narrative Shift detection: A hybrid approach of Dynamic Topic Models and Large Language Models”. In: *Proceedings of the Text2Story’25 Workshop*. Best Paper Award (runner-up). URL: https://www.di.ubi.pt/~jpaulo/Text2Story2025/paper6.pdf.

Le, Quoc and Tomas Mikolov (22–24 Jun 2014). “Distributed Representations of Sentences and Documents”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html.

Li, Zongxia, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber (Mar. 2024). “Improving the TENOR of Labeling: Re-evaluating Topic Models for Content Analysis”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian’s, Malta: Association for Computational Linguistics, pp. 840–859. URL: https://aclanthology.org/2024.eacl-long.51/.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). “Lost in the Middle: How Language Models Use Long Contexts”. In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173. DOI: 10.1162/tacl_a_00638.

Loughran, Tim and Bill McDonald (2010). “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks”. In: *SSRN Scholarly Paper* ID 1331573. DOI: 10.1111/j.1540-6261.2010.01625.x.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (June 2011). “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.

Macaulay, Alistair and Wenting Song (2023). “News media, inflation, and sentiment”. In: *AEA Papers and Proceedings*. Vol. 113. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 172–176. DOI: 10.1257/pandp.20231117.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). “Efficient Estimation of Word Representations in Vector Space”. In: *arXiv:1301.3781*. arXiv: 1301.3781 [cs.CL].

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’13, pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

OpenAI et al. (2024). *GPT-4o System Card*. arXiv: 2410.21276 [cs.CL].

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Rehurek, Radim and Petr Sojka (2011). “Gensim–python framework for vector space modelling”. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2.

Řehůřek, Radim and Petr Sojka (May 2010). “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. URL: https://is.muni.cz/publication/884893/en/.

Rieger, Jonas (2020). “ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations”. In: *Journal of Open Source Software* 5.51, p. 2181. DOI: 10.21105/joss.02181.

Rieger, Jonas (2021). *rollinglda: Construct Consistent Time Series from Textual Data*. R package version 0.1.0. DOI: 10.5281/zenodo.5266717.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2020). “Assessing the Uncertainty of the Text Generating Process Using Topic Models”. In: *ECML PKDD 2020 Workshops*. Vol. 1323. CCIS. Springer, pp. 385–396. DOI: 10.1007/978-3-030-65965-3_26.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2021). “RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data”. In: *Findings Proceedings of the 2021 EMNLP-Conference*. ACL, pp. 2337–2347. DOI: 10.18653/v1/2021.findings-emnlp.201.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). “Dynamic change detection in topics based on rolling LDAs”. In: *Proceedings of the Text2Story’22 Workshop*. Vol. 3117. CEUR-WS, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi (2013). “The Structural Topic Model and Applied Social Science”. In: *NIPS-Workshop on Topic Models: Computation, Application, and Evaluation*. URL: https://projects.iq.harvard.edu/files/wcfia/files/stmnips2013.pdf.

Roos, Michael and Matthias Reccius (2024). “Narratives in Economics”. In: *Journal of Economic Surveys* 38.2, pp. 303–341. DOI: 10.1111/joes.12576.

Santana, Brenda, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes (2023). “A survey on narrative extraction from textual data”. In: *Artificial Intelligence Review* 56.8, pp. 8393–8435. DOI: 10.1007/s10462-022-10338-7.

Schlaufer, Caroline, Johanna Kuenzler, Michael D Jones, and Elizabeth A Shanahan (2022). “The narrative policy framework: a traveler’s guide to policy stories”. In: *Politische Vierteljahresschrift* 63.2, pp. 249–273. DOI: 10.1007/s11615-022-00379-6.

Schmidt, Tobias, Kai-Robin Lange, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025). *Identifying economic narratives in large text corpora – An integrated approach using Large Language Models*. arXiv: 2506.15041 [econ.GN].

Shanahan, Elizabeth A, Michael D Jones, Mark K McBeth, and Claudio M Radaelli (2018). “The narrative policy framework”. In: *Theories of the policy process*. Routledge, pp. 173–213. DOI: 10.4324/9780429494284.

Shenhav, Shaul R. (2006). “Political Narratives and Political Reality”. In: *International Political Science Review* 27.3, pp. 245–262. ISSN: 0192-5121. DOI: 10.1177/0192512106064474.

Shiller, Robert J. (2017). “Narrative Economics”. In: *American Economic Review* 107.4, pp. 967–1004. DOI: 10.1257/aer.107.4.967.

Shiller, Robert J. (2020). “Popular economic narratives advancing the longest U.S. expansion 2009–2019”. In: *Journal of Policy Modeling* 42.4. SI: Growth and Trade in the United States and Europe, pp. 791–798. ISSN: 0161-8938. DOI: 10.1016/j.jpolmod.2020.03.005.

Slapin, Jonathan B and Sven-Oliver Proksch (2008). “A scaling model for estimating time-series party positions from texts”. In: *American Journal of Political Science* 52.3, pp. 705–722.

Song, Yisheng, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo (July 2023). “A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities”. In: *ACM Computing Surveys* 55.13s, pp. 1–40. ISSN: 1557-7341. DOI: 10.1145/3582688.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (July 2019). “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650. DOI: 10.18653/v1/P19-1355.

Tilly, Sonja and Giacomo Livan (July 2021). *Predicting market inflation expectations with news topics and sentiment*. en. arXiv: 2107.07155.

Twitter US Airline Sentiment (2021). URL: https://kaggle.com/crowdflower/twitter-airline-sentiment (visited on 08/21/2021).

Van Kerckhoven, Sven and Sean O’ Dubhghaill (2023). “Narrative Economics and YOLO Investors: r/WallStreetBets and the Gamestop Short Squeeze”. In: *Activist Retail Investors and the Future of Financial Markets Understanding YOLO Capitalism*. Routledge, pp. 99–115. ISBN: 9781032397252. DOI: 10.4324/9781003351085-9.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Wachtel, Ileana (Apr. 2025). *Economist explains impact of Trump’s tariff plans*. USCDornsife. URL: https://dornsife.usc.edu/news/stories/impact-of-tariff-plans-explained-by-economist/ (visited on 06/12/2025).

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (Sept. 3, 2021). *Finetuned Language Models Are Zero-Shot Learners*. Tech. rep. arXiv: 2109.01652 [cs.CL].

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). “Chain-of-thought prompting elicits reasoning in large language models”. In: *Proceedings of the 36th International Conference on Neural*

Information Processing Systems. NIPS '22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3602070.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xu, Lingling, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang (2023). *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. arXiv: 2312.12148 [cs.CL].

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang (Mar. 2024). "Can Large Language Models Transform Computational Social Science?" In: *Computational Linguistics* 50.1, pp. 237–291. DOI: 10.1162/coli_a_00502.

Contributed publications

The following publications are appended as originals or accepted manuscripts. Published articles have been reused with the permission of the copyright holder.

Benner, Niklas, Kai-Robin Lange, and Carsten Jentsch (2022). "Named entity narratives". In: *Ruhr Economic Papers* 962. DOI: 10.4419/96973126.

Lange, Kai-Robin, Niklas Benner, Lars Grönberg, Aymane Hachcham, Imene Kolli, Jonas Rieger, and Carsten Jentsch (2025a). *ttta: Tools for Temporal Text Analysis*. arXiv: 2503.02625 [cs.CL].

Lange, Kai-Robin and Carsten Jentsch (2023). "SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments". In: *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*. Association for Computational Lingustics, pp. 19–28. URL: https://aclanthology.org/2023.cpss-1.3/.

Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch (2022a). "Towards Extracting Collective Economic Narratives from Texts". In: *Ruhr Economic Papers* 963. DOI: 10.4419/96973127.

Lange, Kai-Robin, Jonas Rieger, Niklas Benner, and Carsten Jentsch (2022b). "Zeitenwenden: Detecting changes in the German political discourse". In: *Proceedings of the 2nd Workshop on Computational Linguistics for the Political and Social Sciences*, pp. 47–53. URL: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf.

Lange, Kai-Robin, Jonas Rieger, and Carsten Jentsch (2024). "Lex2Sent: A bagging approach to unsupervised sentiment analysis". In: *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*. Association for Computational Linguistics, pp. 281–291. URL: https://aclanthology.org/2024.konvens-main.28/.

Lange, Kai-Robin, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025b). "Narrative Shift detection: A hybrid approach of Dynamic Topic Models and Large Language Models". In: *Proceedings of the Text2Story'25 Workshop*. Best Paper Award (runner-up). URL: https://www.di.ubi.pt/~jpaulo/Text2Story2025/paper6.pdf.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). "Dynamic change detection in topics based on rolling LDAs". In: *Proceedings of the Text2Story'22 Workshop*. Vol. 3117. CEUR-WS, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/.

Schmidt, Tobias, Kai-Robin Lange, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025). *Identifying economic narratives in large text corpora – An integrated approach using Large Language Models.* arXiv: 2506.15041 [econ.GN].

Dynamic change detection in topics based on rolling LDAs

Jonas Rieger[1], Kai-Robin Lange[1], Jonathan Flossdorf[1] and Carsten Jentsch[1]

[1]Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany

Abstract
Topic modeling methods such as e.g. Latent Dirichlet Allocation (LDA) are popular techniques to analyze large text corpora. With huge amounts of textual data that are collected over time in various fields of applied research, it becomes also relevant to be able to automatically monitor the evolution of topics identified from some sort of dynamic topic modeling approach. For this purpose, we propose a dynamic change detection method that relies on a rolling version of the classical LDA that allows for coherently modeled topics over time that are able to adapt to changing vocabulary. The changes are detected by assessing the intensity of word change in the LDA's topics over time in comparison to the expected intensity of word change under stable conditions using resampling techniques. We apply our method to topics obtained by applying the RollingLDA to Covid-19 related news data from CNN and illustrate that the detected changes in these topics are well interpretable.

Keywords
change point, event, shift, narrative, story, evolution, monitoring, Latent Dirichlet Allocation

1. Introduction

While change detection is an active field in modern research, the application for text data poses even further obstacles due to its unstructured nature. And yet, an effective method for change detection would have many use cases. Particularly, when dealing with large text corpora collected over time, an online detection approach will be useful to analyze the evolution of narratives or to spot a shift in a discourse about certain topics. For this purpose, we propose an online change detection method for text data by analyzing the change of word distributions within topics of Latent Dirichlet Allocation (LDA) models. As we are dealing with time series of textual data, we make use of a rolling version of the classical LDA, called RollingLDA [1]. The method is designed to construct coherently interpretable topics modeled over time that are allowed to adapt to a changing vocabulary. The changes are detected by dynamically assessing the change intensity in word usage in the LDA's topics over time in comparison to the change intensity expected in stable periods using resampling techniques.

The main goal of change detection is to identify possible anomalies in a process. Typically,

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022
rieger@statistik.tu-dortmund.de (J. Rieger); kalange@statistik.tu-dortmund.de (K. Lange); flossdorf@statistik.tu-dortmund.de (J. Flossdorf); jentsch@statistik.tu-dortmund.de (C. Jentsch)
0000-0002-0007-4478 (J. Rieger); 0000-0003-1172-9414 (K. Lange); 0000-0003-2153-0281 (J. Flossdorf); 0000-0001-7824-1697 (C. Jentsch)

there are the two perspectives towards this issue: offline and online applications. Our approach is applicable for both tasks, but, for each time point, it relies exclusively on the text data that has already been observed. Hence, we focus on the usually more relevant task of online monitoring. In traditional schemes for change detection [2, 3], control charts are applied to visualize the monitoring procedure using a control statistic which is successively calculated for each time point. An alarm is triggered whenever the statistic lies outside of some control limits. In practice, there are a variety of different control charts including memory-free setups (e.g. Shewhart charts) and memory-based charts (e.g. EWMA, CUSUM), However, these traditional procedures can not be applied to textual data off the shelf because of the high dimensionality of large text corpora. In addition, an in-control state to reliably calculate the control limits is frequently not available due to the strong dynamics in text data, e.g. newspaper articles. To overcome these issues, we propose to use a control statistic based on a similarity metric that represents the resemblance of topic's word distributions over consecutive time points. Control limits are derived by a resampling procedure using word count vectors based on time-variant topics modeled by RollingLDA [1].

In a similar context, the usage of LDA was proposed for change point detection for topic distributions in texts [4], which is based on a modified version of the wild binary segmentation algorithm [5] designed for offline detection setups. There is also work considering Bayesian online monitoring [6] for textual data using a document-based model [7] and an approach based on similarity metrics, which aims to detect global events in topics in offline settings [8]. There is also work which analyzes the transitions of narratives between topics [9]. In contrast, the rolling window approach of RollingLDA constructs coherently interpretable topics modeled over time and allows the resulting dynamic change detection method to become applicable in online settings. Compared to the mentioned related methods, our method is designed to detect changes in word distributions of topics over time rather than global changes in topic distributions (of sets) of documents [e.g. 4, 7, 8] or sentiments in topics [e.g. 10] or (in contrast) changes in topic distributions of words [e.g. 11]. This results in a more refined monitoring procedure that allows for the detection of narrative shifts that are changing the word usage within a certain topic instead of measuring the frequency of a topic over time within the whole corpus. Building on this, we aim that our proposed method can provide groundwork for the extraction and temporal localization of narratives in texts.

2. Methodological framework

For the proposed change detection algorithm, we make use of the existing method of a rolling version of the classical LDA (RollingLDA) to construct coherent topics over time and measure similarities of topics for consecutive time points using the well-established cosine similarity.

2.1. Latent Dirichlet Allocation

The classical LDA [12] models distributions of K latent topics for each text. Let $W_n^{(m)}$ be a single word token at position $n = 1, \dots N^{(m)}$ in text $m = 1, \dots, M$ of a corpus of M texts. Then, a single text is given by

$$\boldsymbol{D}^{(m)} = \left(W_1^{(m)}, \ldots, W_{N^{(m)}}^{(m)}\right), \quad W_n^{(m)} \in \boldsymbol{W} = \{W_1, \ldots, W_V\}, V = |\boldsymbol{W}|$$

and the corresponding topic assignments for each text are given by

$$\boldsymbol{T}^{(m)} = \left(T_1^{(m)}, \ldots, T_{N^{(m)}}^{(m)}\right), \quad T_n^{(m)} \in \boldsymbol{T} = \{T_1, \ldots, T_K\}.$$

From this, let $n_k^{(mv)}, k = 1, \ldots, K, v = 1, \ldots, V$ denote the number of assignments of word v in text m to topic k. Then, we define the cumulative count of word v in topic k over all texts by $n_k^{(\bullet v)}$ and denote the total count of assignments to topic k by $n_k^{(\bullet\bullet)}$. Using these definitions, the underlying probability model [13] can be written as

$$W_n^{(m)} \mid T_n^{(m)}, \phi_k \sim \mathrm{Discr}(\phi_k), \quad \phi_k \sim \mathrm{Dir}(\eta), \quad T_n^{(m)} \mid \theta_m \sim \mathrm{Discr}(\theta_m), \quad \theta_m \sim \mathrm{Dir}(\alpha).$$

For a given parameter set $\{K, \alpha, \eta\}$, with the Dirichlet priors α and η defining the type of mixture of topics in every text and the type of mixture of words in every topic, LDA assigns one of the K topics to each token. A word distribution estimator per topic for $\phi_k = (\phi_{k,1}, \ldots, \phi_{k,V})^T \in (0,1)^V$ can be derived through the collapsed Gibbs sampler procedure [13] by

$$\hat{\phi}_{k,v} = \frac{n_k^{(\bullet v)} + \eta}{n_k^{(\bullet\bullet)} + V\eta}. \tag{1}$$

2.2. RollingLDA

RollingLDA [1] is a rolling version of classical LDA. New texts are modeled based on existing topics of the previous model. Thereby, not the whole knowledge of the entire past of the model is used, but only the information of the topics from more recent texts based on a user-chosen memory parameter. For each time point, based on the topic assignments within this memory period, the topics are initialized and modeled forward. This form of modeling preserves the topic structure of the model so that topics remain coherently interpretable over time. At the same time, constraining the knowledge of the model to the user-chosen memory period allows for changes in topics based on new vocabulary or word choices. There are other dynamic variants of the LDA approach [14, 15, 16, 17, 18] deliberately designed to model gradual changes, and therefore not as well suited to detect abrupt changes. We use the update algorithm RollingLDA to make our proposed change detection method applicable in an online manner. Thereby, a text is assigned to a time point on the basis of its publication date. The step size of the and is chosen on a weekly basis in the present case as this seems natural for journalistic texts.

2.3. Similarity

Our change detection algorithm builds on a similarity measure for word count vectors. Following up on the notation from Section 2.1 the word count vector for topic $k \in \{1, \ldots, K\}$ at one time point $t \in \{0, \ldots, T\}$ is given by

$$\boldsymbol{n}_{k|t} = \left(n_{k|t}^{(\bullet 1)}, \ldots, n_{k|t}^{(\bullet V)}\right)^T \in \mathbb{N}_0^V = \{0, 1, 2, \ldots\}^V.$$

Then, monitoring the similarity of topics over time for (consecutive) time points t_1 and t_2 is done using the cosine similarity

$$\cos\left(\boldsymbol{n}_{k|t_1}, \boldsymbol{n}_{k|t_2}\right) = \frac{\sum_v n_{k|t_1}^{(\bullet v)} n_{k|t_2}^{(\bullet v)}}{\sqrt{\sum_v \left(n_{k|t_1}^{(\bullet v)}\right)^2} \sqrt{\sum_v \left(n_{k|t_2}^{(\bullet v)}\right)^2}}. \tag{2}$$

The choice of cosine similarity is common in the context of change point detection for text data [e.g. 8, 19]. Compared to other similarity measures such as the Jaccard coefficient, Jensen-Shannon Divergence, χ^2-, Hellinger and Manhattan Distance, the cosine similarity fulfills some typical conditions required for monitoring a similarity measure [1].

3. Change detection

In combination with the existing method RollingLDA and cosine similarity, our contributed method for change detection relies on classical resampling approaches to identify changes within topics. We estimate the realized change in a topic based on the similarity between the current and previous count vectors of word assignments and compare the resulting similarity score to resampling-based similarity scores which are generated under stable conditions, such that no extraordinary changes occurred in the topic.

3.1. Set of changes

Suppose we consider K topics over T time points to be monitored. If the actual observed similarity of the word vector of some topic $k \in \{1, \dots, K\}$ at some time $t \in \{0, 1, \dots, T\}$ given by $\boldsymbol{n}_{k|t}$, compared to the frequency vector of the topic over a predefined reference time period $t - z_k^t, \dots, t-1$, given by

$$\boldsymbol{n}_{k|(t-z_k^t):(t-1)} = \sum_{z=1}^{z_k^t} \boldsymbol{n}_{k|t-z}, \tag{3}$$

is smaller than a threshold q_k^t which is calibrated based on similarities under stable conditions (see Section 3.2), then we identify a change within topic k at time t. The set of identified changes in topic k up to time point t can then be defined as

$$C_k^t = \left\{u \mid 0 < u \le t \le T \,:\, \cos\left(\boldsymbol{n}_{k|u}, \boldsymbol{n}_{k|(u-z_k^u):(u-1)}\right) < q_k^t\right\} \cup 0, \tag{4}$$

where the time point $t = 0$ is always included for technical reasons, to compute the current run length without a change $z_k^t = \min\left\{z_{\max}, t - \max C_k^{t-1}\right\}$. Thus, the reference period spans the last $z_{\max}$ time points if no change was detected during that time, and spans the time that has passed since the last change, otherwise. The parameter $z_{\max}$ is to be chosen by the user and is intended to smooth the similarities to prevent from detecting false positives.

3.2. Dynamic thresholds

For the calculation of the threshold q_k^t, the estimated word distribution of a topic k at some time point t, as well as over the corresponding reference period $t - z_k^t, \dots, t-1$ are needed. For this, let $\hat{\phi}_k^t$ and $\hat{\phi}_k^{(t-z_k^t):(t-1)}$ be defined by

$$\hat{\phi}_{k,v}^t = \frac{n_{k|t}^{(\bullet v)} + \eta}{n_{k|t}^{(\bullet\bullet)} + V\eta} \quad \text{and} \quad \hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} = \frac{n_{k|(t-z_k^t):(t-1)}^{(\bullet v)} + \eta}{n_{k|(t-z_k^t):(t-1)}^{(\bullet\bullet)} + V\eta} \tag{5}$$

analogously to Equation (1).

The application of the change point detection algorithm is designed for text data, more precisely for empirical word distributions of K topics modeled by LDA in a given text corpus. Since word choice - especially in journalistic texts - varies considerably over time, a situation in which there is no change in the word distribution within topics across consecutive time points does not reflect the expected situation. Rather, it is to be expected that topics change gradually on an ongoing basis. Accordingly, our method aims to identify not the numerous customary changes in the topics, but the unexpectedly large ones. To do so, we define an expected word distribution $\tilde{\phi}_k^{(t)}$ for time point t under stable conditions that include the customary changes as a convex combination of the two estimators of the word distribution of topic k, one for the reference time period $t - z_k^t, \dots, t-1$ and one for the current time point t. Using the mixture parameter $p \in [0,1]$, which can be tuned based on how substantial the detected changes should be, the intensity of the expected change is considered in the determination of this estimator by

$$\tilde{\phi}_k^{(t)} = (1-p)\,\hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} + p\,\hat{\phi}_{k,v}^{(t)}. \tag{6}$$

Our method uses the estimator $\tilde{\phi}_k^{(t)}$ to simulate R expected word count vectors $\tilde{\boldsymbol{n}}_{k|t}^r, r = 1, \dots, R$ based on a parametric bootstrap approach. In this process, each word is drawn according to its estimated probability of occurrence regarding $\tilde{\phi}_k^{(t)}$ and each sample r consists of $n_{k|t}^{(\bullet\bullet)}$ draws, the number of words assigned to topic k at time point t. Then, we calculate the cosine similarity

$$\cos\left(\tilde{\boldsymbol{n}}_{k|t}^r, \boldsymbol{n}_{k|(t-z_k^t):(t-1)}\right) \tag{7}$$

for each of the $r = 1, \dots, R$ bootstrap samples and set the threshold q_k^t equal to the 0.01 quantile of these simulated similarity values generated under stable conditions. Combinations of topics and time points for which the observed similarity is smaller than the corresponding quantile are classified as change points according to Equation (4).

4. Analysis

For conducting the real data analysis, the data set under study was created with Python, whereas the preprocessing, the modeling, all postprocessing steps and analyses are performed using R. The scripts for all analysis steps can be found in the associated GitHub repository github.com/JonasRieger/topicalchanges.

4.1. Data and study design

To assess the quality of our change point algorithm, we use the TLS-Covid19 data set [20]. It is generated using Covid-19 related liveblog articles of CNN, collected from January 22nd 2020 up until December 12th 2021. Each liveblog is interpreted to belong to a topic and comprises texts and key moments. The texts form a time line containing events, which are summarized by its key moments. The resulting corpus consists of 27,432 texts and 1,462 key moments. Although the data set contains multiple key moments per day on average, we do not consider all them a change point as our aim is to detect larger changes based on aggregated weekly texts. However, these key moments serve well as indicators, which enable us to check whether the detected changes are actually related to real events or if they are false positives.

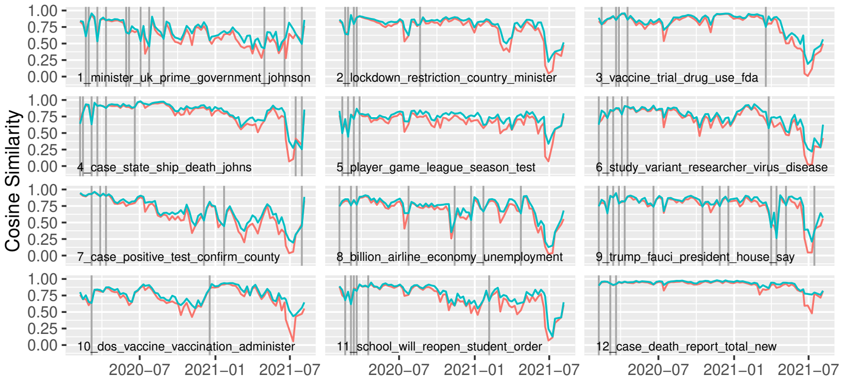
We use common NLP preprocessing steps for the texts, i.e. characters are formatted to lowercase, numbers and punctuation are removed. Moreover, a trusted stopword list is applied to remove words that do not help in classifying texts in topics, we use a lemmatization dictionary (github.com/michmech/lemmatization-lists) and neglect words with less than two characters.

We model the CNN data set using RollingLDA on a weekly basis, starting on Saturday of each week, and we consider the previous week as initialization for the model's topics. The first 10 days of modeling, Wednesday, January 22nd 2020 until Friday, January 31st 2020, serve as the initial chunk corresponding to $t = 0$. During this period, 605 texts were published. In the data set, there are weeks that do not contain any texts. In this case, the corresponding time point is omitted. Then, to model the texts of the following chunk, at least the last 10 texts are used, as well as all other texts published on the same date as the oldest of these 10 texts. As parameters, we assume $K = 12$ topics, define the reference period of the topics to the last $z_{\max} = 4$ weeks, and choose $p = 0.85$, since these values are accountable by plausibility and seem to yield reasonable results. For other parameter choices, i.e. $K = 8, \dots, 20$, $z_{\max} = 1, \dots, 20$, $p = 0.5, \dots 0.8, 0.81, \dots, 0.90$, results can be found in our associated repository.

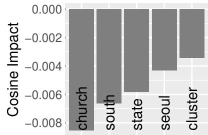
4.2. Findings

The results of our chosen model are displayed in Figure 1. Fig. 1a shows the detected changes by vertical gray lines, which are the weeks in which the observed similarity (blue curve) is lower than the expected one (red curve). Furthermore, for two changes we show which words are mainly causing the detection of the change. The score of a word in a topic at a given time point is calculated by the topic's similarity without considering this word and subtracting it from the actual realized similarity. These leave-one-out cosine impact scores for the words with the five most negative scores are shown in Fig. 1b and 1c. In general, most of the changes we detect occur within the first four months of 2020. This is because the wording was constantly changing, as the Covid-19 epidemic turned into a pandemic over the course of these months. New people and organizations were associated with Covid-19, which is why we detect a bunch of consecutive changes in every topic. As the pandemic reached out into further countries, the detected changes became less frequent for most topics. In the following we share our interpretation of some exemplary detected changes.

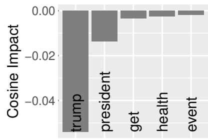
The third topic, containing information about vaccination and testing procedures, shows a change in the week starting on the 13th of March 2021. In this week, the AstraZeneca



(a) Similarities (blue), thresholds (red) and detected changes (vertical lines, gray).



(b) T2: 2020-08-21.



(c) T9: 2020-10-16.

Figure 1: Similarity values, thresholds, and detected changes over the observation period for all $K = 12$ topics, as well as the five most influential words for two selected change points in topics 2 and 9.

vaccination process in several EU-states was stopped due the risk of causing blood clots.[1] The sixth topic, a topic about medical studies and research, shows a change in the following week, in which AstraZeneca presented a study about the effectiveness of its vaccine.[2] Another interesting detection is the change in the vaccination-related topic 10 in December 2020, just as the vaccination process started in the US.[3]

Political changes are also detected in several topics, such as the start of Joe Biden's presidential era in late January 2021 in topic 11, the return of Donald Trump to office after his Covid-19 infection in October 2020[4] in topic 9 (cf. Fig. 1c) or the discussion about the origin of the virus after a WHO report in late March 2021 in topic 9.[5] A Covid-19 outbreak in the South Korean Sarang-jeil church in August 2020[6] is detected in topic 2 (cf. Fig. 1b).

While these topics detect changes across the entire time span, the twelfth topic, representing the report of the current number of Covid cases, does not detect a single change after March

[1] CNN online, 2021-03-15 3:03 p.m. ET, "Spain joins Germany, France and Italy in halting AstraZeneca Covid-19 vaccinations", https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-15-21/h_d938057f2ef588f74565bdbb01f12387, visited on 2022-01-20.

[2] CNN online, 2021-03-25 2:48 a.m. ET, "New AstraZeneca report says vaccine was 76% effective in preventing Covid-19 symptoms", https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-25-21/h_9f01e2e53b62873f1c742254d27fbf5f, visited on 2022-01-20.

[3] CNN online, 2020-12-14 10:08 p.m. ET, "The first doses of FDA-authorized Covid-19 vaccine were administered in the US. Here's what we know", https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-12-15-20/h_32be1a72dc05f874eda167c95c8f1bba, visited on 2022-01-20.

[4] CNN online, 2020-10-12 12:01 a.m. ET, "Trump says he tested 'totally negative' for Covid-19", https://edition.cnn.com/world/live-news/coronavirus-pandemic-10-12-20-intl/h_7570d53b184a5b1d6ec97ce67330e4c9, visited on 2022-01-20.

[5] CNN online, 2021-03-29 11:22 a.m. ET, "Upcoming WHO report will deem Covid-19 lab leak extremely unlikely, source says", https://www.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-29-21/h_1f239fee1b0584ca9a5b6085357ac907, visited on 2022-01-20.

[6] CNN online, 2020-08-20 12:55 a.m. ET, "South Korea's latest church-linked coronavirus outbreak is turning into a battle over religious freedom", https://edition.cnn.com/world/live-news/coronavirus-pandemic-08-20-20-intl/h_288a15acd1b29e732c4e10693641088a, visited on 2022-01-20.

2020. This is most likely because, after the pandemic had reached the US and Europe in early 2020, the number of cases was consistently reported and the interpretations and implication of those case numbers are detected as changes in other topics. Even in the last months of the data set, in which the number of texts decreased and the results thus show a lower similarity, the twelfth topic retained a rather high similarity of above 0.75.

5. Discussion

In this paper, we presented a novel change detection method for text data. To construct coherently interpretably topics, we used RollingLDA to model a time series on textual data and compared the model's word distribution vectors with those of texts resampled under stable conditions. We applied our model on the TLS-Covid19 data set consisting of Covid-19 related news articles from CNN between January 2020 and December 2021.

Our method detects several meaningful changes in the evolving news coverage during the pandemic, including e.g. the start of vaccinations and several controversies over the course of the vaccination campaign as well as political changes such as the start of Joe Biden's presidential era. Out of 78 detected changes, we were instantly able to judge 55 (71%) as plausible ones based on manual labeling using the leave-one-out cosine impacts (cf. Fig. 1b, 1c and repository). The share increases to 78% if we exclude the turbulent initial phase of the Covid-19 pandemic and only consider changes since April 2020. While we cannot tell how many changes were missed out that could be considered as important as the ones mentioned above, our model contains a mixture parameter to calibrate the detection for general change of topics within a usual news week. If more, but less substantial or less, but more substantial changes are to be detected, this parameter p can be tuned accordingly. In combination with the maximum length of the reference period $z_{\max}$, the set $\{p, z_{\max}\}$ forms the model's hyperparameters to be optimized.

Acknowledgments

The present study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number PR-2019-0019. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

[1] J. Rieger, C. Jentsch, J. Rahnenführer, RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data, in: Findings Proceedings of the 2021 EMNLP-Conference, ACL, 2021, pp. 2337–2347. doi:10.18653/v1/2021.findings-emnlp.201.

[2] D. C. Montgomery, Introduction to statistical quality control, John Wiley & Sons, 2020.

[3] J. S. Oakland, Statistical process control, Routledge, 2007.

[4] A. Bose, S. S. Mukherjee, Changepoint analysis of topic proportions in temporal text data, 2021. arXiv:2112.00827.

[5] P. Fryzlewicz, Wild binary segmentation for multiple change-point detection, The Annals of Statistics 42 (2014) 2243–2281. doi:10.1214/14-AOS1245.

[6] R. P. Adams, D. J. MacKay, Bayesian online changepoint detection, 2007. arXiv:0710.3742.

[7] T. Kim, J. Choi, Reading documents for bayesian online change point detection, in: Proceedings of the 2015 EMNLP-Conference, ACL, 2015, pp. 1610–1619. doi:10.18653/v1/D15-1184.

[8] N. Keane, C. Yee, L. Zhou, Using topic modeling and similarity thresholds to detect events, in: Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, ACL, 2015, pp. 34–42. doi:10.3115/v1/W15-0805.

[9] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: An exploration of temporal text mining, in: Proceedings of the 11th SIGKDD-Conference, ACM, 2005, pp. 198–207. doi:10.1145/1081870.1081895.

[10] Q. Liang, K. Wang, Monitoring of user-generated reviews via a sequential reverse joint sentiment-topic model, Quality and Reliability Engineering International 35 (2019) 1180–1199. doi:10.1002/qre.2452.

[11] L. Frermann, M. Lapata, A Bayesian model of diachronic meaning change, Transactions of the Association of Computational Linguistics 4 (2016) 31–45. doi:10.1162/tacl_a_00081.

[12] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

[13] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235. doi:10.1073/pnas.0307752101.

[14] X. Song, C.-Y. Lin, B. L. Tseng, M.-T. Sun, Modeling and predicting personal information dissemination behavior, in: Proceedings of the 11th SIGKDD-Conference, ACM, 2005, pp. 479–488. doi:10.1145/1081870.1081925.

[15] D. M. Blei, T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, Hierarchical topic models and the nested chinese restaurant process, in: Advances in Neural Information Processing Systems, volume 16, MIT Press, 2003, pp. 17–24. URL: https://proceedings.neurips.cc/paper/2003/hash/7b41bfa5085806dfa24b8c9de0ce567f-Abstract.html.

[16] X. Wang, A. McCallum, Topics over time: A non-markov continuous-time model of topical trends, in: Proceedings of the 12th SIGKDD-Conference, ACM, 2006, pp. 424–433. doi:10.1145/1150402.1150450.

[17] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd ICML-Conference, ACM, 2006, pp. 113–120. doi:10.1145/1143844.1143859.

[18] C. Wang, D. M. Blei, D. Heckerman, Continuous time dynamic topic models, in: Proceedings of the 24th UAI-Conference, AUAI Press, 2008, pp. 579–586. URL: https://dl.acm.org/doi/10.5555/3023476.3023545.

[19] Y. Wang, C. Goutte, Real-time change point detection using on-line topic models, in: Proceedings of the 27th ACL-Conference, ACL, 2018, pp. 2505–2515. URL: https://www.aclweb.org/anthology/C18-1212.

[20] A. Pasquali, R. Campos, A. Ribeiro, B. Santana, A. Jorge, A. Jatowt, TLS-Covid19: A new annotated corpus for timeline summarization, in: Advances in Information Retrieval, ECIR 2021, volume 12656 of *LNCS*, 2021, pp. 497–512. doi:10.1007/978-3-030-72113-8_33.

Benner, Niklas; Lange, Kai-Robin; Jentsch, Carsten

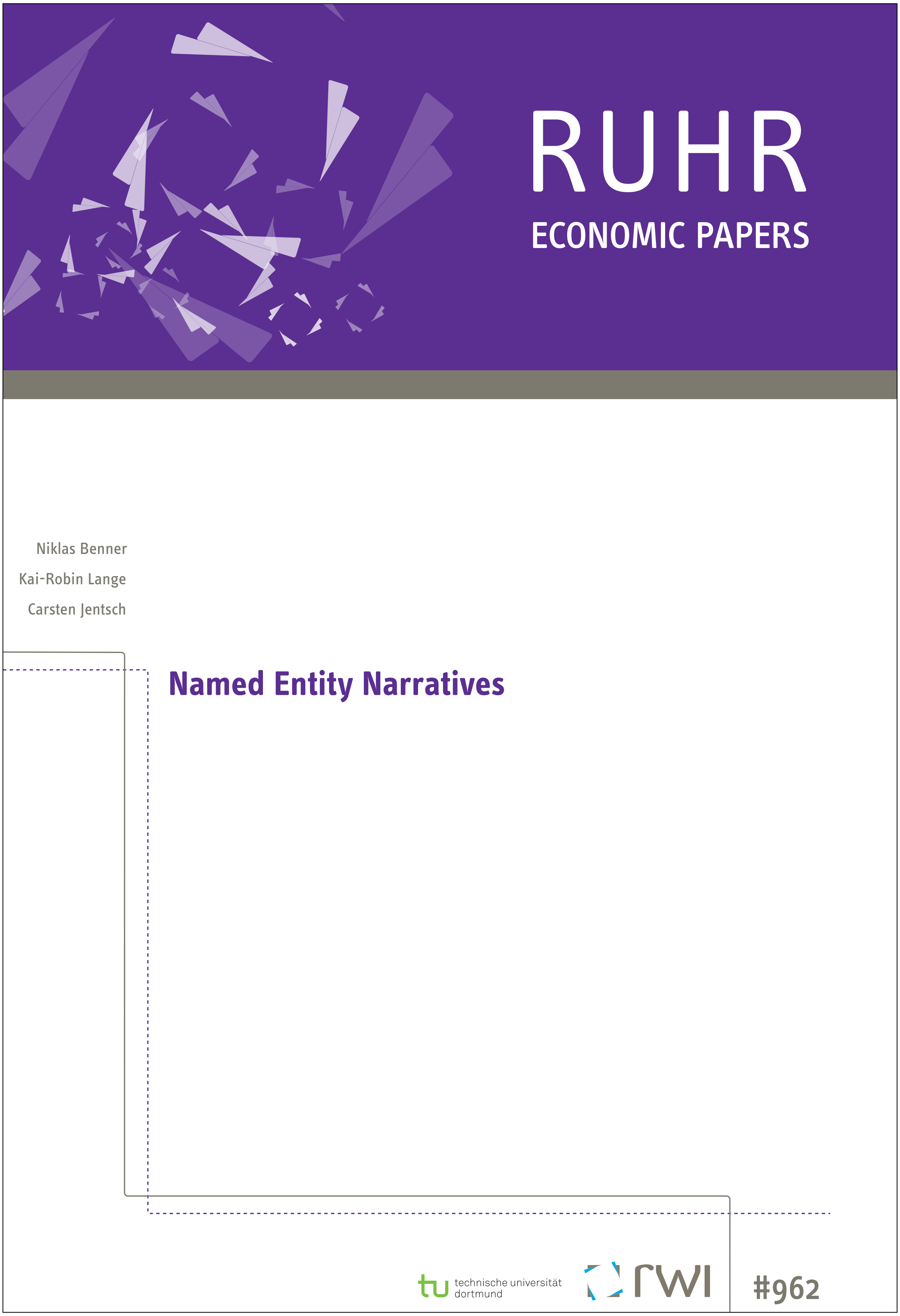
Working Paper

Named entity narratives

Ruhr Economic Papers, No. 962

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Benner, Niklas; Lange, Kai-Robin; Jentsch, Carsten (2022) : Named entity narratives, Ruhr Economic Papers, No. 962, ISBN 978-3-96973-126-0, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen,
https://doi.org/10.4419/96973126



RUHR
ECONOMIC PAPERS

Niklas Benner
Kai-Robin Lange
Carsten Jentsch

Named Entity Narratives

technische universität dortmund
rwi

#962

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Ruhr Economic Papers #962

Responsible Editor: Torsten Schmidt

ISSN 1864-4872 (online) – ISBN 978-3-96973-126-0

Ruhr Economic Papers #962

Niklas Benner, Kai-Robin Lange, and Carsten Jentsch

Named Entity Narratives

technische universität dortmund

rwi

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973126
ISSN 1864-4872 (online)
ISBN 978-3-96973-126-0

Niklas Benner, Kai-Robin Lange, and Carsten Jentsch[1]

Named Entity Narratives

Abstract

While the actions of economists and politicians can be influenced by facts, statistics or empirical predictions, narratives are becoming an increasingly important factor for the decision making in the field of economics and politics. Evaluating such narratives not at selective points in time but rather as a timeline can give us an insight on the effects of changing narratives on economic processes. We propose a model to detect two distinct types of temporal narratives by evaluating the relevance of entities in a timeline of newspaper articles. This methodology is based on the fundamental concept that all narratives are driven by and centered around certain entities. We provide a model to describe entity-based time dynamic media attention and detect both temporary (events) and permanent (structural break) changes of narratives by analyzing the number of appearances of an entity and the change in word frequency surrounding it. Our model detects several meaningful events and structural breaks, such as Mario Draghi's well known "Whatever it takes" speech in 2012 or the change of narrative surrounding Wladimir Putin due to start of the Russian-Ukrainian war in 2022. For instance, this enables us to detect the narrative shift contained in newspaper articles about the Russian Federation from being a German business partner and gas trader to being called a war mongering regime.

JEL-Codes: C43, C55, C89, E71

Keywords: Event detection; time series for count data; text mining; econometrics; narrative

August 2022

1 Niklas Benner, RWI and TU Dortmund; Kai-Robin Lange, TU Dortmund; Carsten Jentsch, TU Dortmund. *- This study is part of a project of the Dortmund Center for databased Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.* *– All correspondence to: Niklas Benner, RWI, Hohenzollernstr. 1/3, 45128 Essen, Germany, e-mail: niklas.benner@rwi-essen.de*

1 Introduction

With the Russian-Ukrainian war affecting both global politics and markets, a suitable interpretation of the actions of the involved parties is a crucial tool to predict future diplomatic and economic developments. *Will Wladimir Putin continue to supply the European Union with gas and oil? Will Wolodymyr Selenskyj and Wladimir Putin agree on a deal to deliver Ukraine's massive wheat production into the rest of the world to avoid famines?* Also, when looking back at the economy and world politics of the last few years, narratives about the current COVID-19 situation, with government restrictions decided by high-ranking politicians being an important factor, can have a large impact on the economy Harrs et al. (2021). Such questions and forecasts often evolve around singular entities, mainly being politicians, major companies or banks, and do so beyond the scope of the aforementioned war and pandemic. Whenever such influential entities take part in decision making concerning major incidents, narratives start to form. This process of narrative creation does not stop alongside the event, but rather evolves over time. A narrative's lifetime can often be separated into three stages: it starts out as the prediction of the actions of an entity before said action is taken. It then changes when the action has already been taken, now containing a possible explanation for the entity's action based on its past narratives and related incidents. This narrative often develops into a retrospective, describing the effects of the action in the following years.

We propose a model based on textual data that analyzes the relevance of entities and the words commonly surrounding them to detect narratives. The focus on entities enables us to suppress noise that would usually be part of narrative extraction as only texts incorporating a specific entity are relevant to said entities narrative and are therefore analyzed. While this pipeline is prone to our definition of an entity (i.e. if only persons or also organizations are included) and to the number of entities we decide to analyze, it enables us to detect events and thus narrative changes, even for less relevant entities. A person that barely comes up in the news may not be as popular, but there might still exist a narrative surrounding it. Our pipeline is able to detect when such an entity has a spike in relevance due to an event surrounding it, which enables us to check the development of narrative over its lifetime. While most narrative extraction pipelines focus on the most prevalent narratives, our pipeline is also able to extract narratives and their changes for less prevalent narratives that surround less important entities. This can for instance be used when investigating specific companies or economic branches.

To be able to extract and detect changes in narratives, the term has to be properly defined first. The concept of narratives started becoming relevant to political science in the 1980s (Fisher, 1984). In economics, the interest in narratives has greatly increased since the work of Shiller (2017, 2019). In this paper, we leverage our definition of the term "narrative" from Roos and Reccius (2021), who define a collective economic narrative as "*a sense-making story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions*". Such a collective economic narrative is thus a complex theoretical construct that is hard to detect. Due to the nature of the newspaper articles we use for our analysis, we assume that the narratives that we detect are shared by a group and proliferates in social interaction, as they would otherwise not be relevant enough to be published in an established German news outlet. Thus, we will focus on detecting "sense-making stories about economically relevant topics". We do so by analyzing the most important aspect of each stories: the entities that drive said story and are responsible for any actions within it. Even if said entity is a politician rather than an economist or CEO of a company, its actions still often have economical impact. Narratives about politicians can thus still be considered to revolve around economically relevant topics. Even celebrities, who might have narratives surrounding them as well, might have an economical impact on a smaller scale, due to for instance brand deals.

The rest of this paper is structured as follows. In Section 2, we describe our pipeline, which we evaluate in Section 3 on our data set. We conclude our paper and give an outlook to future research in Section 4.

2 Model

Narratives are usually born on the basis of events or structural breaks and evolve from their subsequent interpretation. We aim to detect both structural breaks and singular events by evaluating the appearances of named entities every single day over the course of 21 years of German newspaper articles.

2.1 Identification of Named Entities

Instead of trying to find named entities found on the basis of semantic properties, we aim to detect them more reliably by a supervised approach. We therefore use data from the German Wikipedia, as every entity that is relevant enough to have a narrative revolved around it will likely also be relevant enough to have a Wikipedia article. We use the Wikipedia dump from July 22nd 2022 (a local copy of every Wikipedia article of the time), including 5.2 million articles[1]. The articles are categorized, enabling us to filter for entities with specific properties, such as politicians, economists or celebrities as well as non-person entities such as companies or countries (see Table 1). This makes it possible to perform the analysis not only at the level of the individual entities, but also as an aggregation of generic terms or combinations of categories. Using non-economical or non-political articles is thus not problematic for our models, as we can simply filter the texts for entities of interest by their profession.

[1] `https://dumps.wikimedia.org/dewiki/20220720/`

In this analysis, we focus on persons rather than companies and countries. An extension to such non-person entities is a possible next step to improve our detection and identify a larger variety of narratives. While incorporating different categories of entities is not a difficult task, for this analysis, we filter the Wikipedia articles by the categories "man" or "woman", yielding approximately 857 thousand articles and thus persons that can be used as entities. These rather strict categories are suitable for a basic analysis of our model to detect events and structural breaks. Still, other categories enable a more complex analysis to link the narratives of, for instance, two politicians of the same party. Then we will be able to distinguish between a party-wide narrative that affects all party members, and a personal narrative, that affects only one politician. These additional categories are thus a possible improvement for future research.

A text is assigned to a named entity as soon as the name of the entity is mentioned in the text. For this, at least the first and last token of the name have to be mentioned in a row. I.e. for the entity *Angela Merkel* to be assigned to a text, both the words *Angela* and *Merkel* have to appear in it in direct succession. For entities with multiple given names, such as *Wladimir Wladimirowitsch Putin* to be assigned to a text however, only the term *Wladimir Putin* has to appear. As the second given name of entities is rarely mentioned, requiring it to be part of the texts would lead to many missed entities.

2.2 Identification of Events

We differentiate between two types of narratives we aim to detect, both initiated by an incident. We define an incident as a situation of public interest that, in our case, translates into articles being written about it. It can be political or economical, like the outbreak of the Russian-Ukrainian war in 2022 or the announcement of bankruptcy of Lehman Brothers in 2008. Ultimately it serves as a starting point for narratives to be born. The narratives of entities being involved subsequently change, either temporarily or permanently. This section describes our model to detect an event, while the following section will cover our model to detect a long lasting structural break. We define an event as a sudden change in distribution that is quickly undone, so that the distribution returns back to its original form.

If an event is associated with an entity and the entity is sufficiently relevant to the media, this should be reflected in an increase in the intensity of short term media coverage. The objective of this analysis is to identify those events that have a significant impact on the entity and as a result of which the view on the entity has expanded or changed.

The basis of the statistic modeling of medial attention is the daily number of articles in which the entity is mentioned. We assume our count data to be zero-inflated Poisson distributed (Jazi et al., 2012). Let X_i represent the zero-inflated poisson distributed random variable for day i for $i \in \{1, \dots, n\}$ with the poisson parameter $\lambda_i \in (0, \infty)$ and the additional inflation parameter $\pi_i \in (0,1)$. Then we define the probability density function of X_i as

$$P(X_i = k) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & \text{if } k = 0 \\ (1-\pi_i)\dfrac{e^{-\lambda_i}\lambda_i^k}{k!}, & \text{if } k \in \mathbb{N} \end{cases}$$

where π symbolizes an additional chance to yield a value of 0 in addition to the probability of the base line poisson distribution. In our context, this additional chance to yield zero articles per day reflects the fact that (less relevant) entities are not discussed daily in the media, resulting in too many days with zero articles containing their name to be poisson-distributed. We estimate our parameters $\pi_{i,t}$ and $\lambda_{i,t}$ individually for each entity. Different types of media attention cannot only differ between entities, but are also dynamic in time. Therefore, the distribution parameters for each entity are calculated dynamically based on the article counts of ρ days before the potential event using a rolling window. Since we assume that the general media attention of an entity and thus the distribution of article counts manifests itself over a longer period of time, and we consider the short-term outliers as events, we have chosen a window of $\rho = 365$ days. Since closer days are more important to estimate the expected media attention on a day, we weight the observations linearly ascending from with a weight 0 for the observation 365 days before to 1 for the observation of the last day.

Using linear weighting over our rolling window period, the distribution parameters $\pi_{i,t}$ and $\lambda_{i,t}$ can then be calculated as

$$\pi_{i,t} = \frac{s^2 - \bar{x}}{s^2 + \bar{x}^2 - \bar{x}}, \qquad \lambda_{i,t} = \frac{s^2 + \bar{x}^2 - \bar{x}}{\bar{x}}$$

with the weighted mean $\bar{x}$ and weighted variance s^2 of the number of articles (Jazi et al., 2012).

While in statistics, often a test level of 0.05 is used, we are using a monitoring approach. A level 0.05 test would however yield an type I error 5% of the time. As we test each entity daily and separately over the course of 21 years of media coverage, we need to test to a lower level of significance to avoid false detections and to only be notified, if an actual event occured and needs to manually checked. We therefore determine the 99.9%-quantile of the theoretical distribution and use it as our critical value, as this represents a test to the level 0.001. If the number of articles on this day is above the quantile limit, we assume an event for this day. If the limit is exceeded for several days in a row[2], these together form an event.

[2]If a single day within a chain is under the limit, the event chain will be continued anyway.

2.3 Identification of structural Breaks

The second type of detection we propose, is a detection of structural breaks. While events are sudden changes in distribution that only manifest over a short period of time, structural breaks can be identified as persistent changes in distribution that do not return to the original state.

Similar to the approach of Rieger et al. (2022), we assume that such structural breaks manifest in the context in which the entity is used, as a changing context indicates a new and relevant incident has taken place, involving the entity. Rieger et al. (2022) identify structural breaks in the entire corpus using a rolling window version (Rieger et al., 2021) of the topic model LDA (Blei et al., 2003). They demonstrate their change detection on a data set of the German Bundestag, showing the relevance of this method to detect changes in political discussions and therefore likely also in narratives (Lange et al., 2022b). This strategy is however hard to combine with individual entities, as structural breaks in the texts of a specific entity do not always align with structural breaks of a specific topic of the model. Thus, we quantify the change of word distribution and thus entity context by differences in word structure, using cosine similarity (Li and Han, 2013) of absolute word frequencies in different time chunks. A low cosine similarity would implicate that either the incident the entity is involved in or the reporting about the entity itself has changed. This might answer interesting questions as to, for instance, how the style of reporting about the entity Wladimir Putin has changed since the outbreak of the Russian-Ukrainian war. Using changing word frequencies enables us to interpret the longevity of the incidents to differentiate if the narrative changed permanently and thus applies to our definition of a structural break. Still, creating of combination topic modeling for specific entities might be a possibility to improve our model in future research. Cosine similarity is denoted by

$$S_C(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

for two vectors A and B, each containing the total count of each word in the entire vocabulary for the corresponding time periods of the data set Li and Han (2013). The cosinus similarity allows us to use high dimensional vectors without major problems that might cause for other similarity metrics. Also we do not need to differentiate between absolute and relative word frequencies, as the difference between those terms simply cancels out due to the multiplicative properties of cosine similarity.

To be able to assign the change in the word structure to an unique incident, we analyze three time intervals for each day that is analyzed. The first interval denotes the time in which the incident affecting the entities has not yet taken place and the word distribution is only affected by older narratives. Then the incident takes place and affects the entities narratives directly for β days. A structural break differs from an event due to the next γ days, in which the incident still affects the narrative. This time line is shown in Figure 1

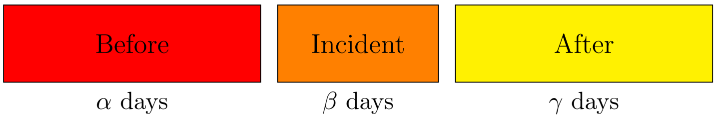


Figure 1: Life span of a narrative defined by a structural break. After the incident, the narrative still affects the word frequencies for γ days.

In theory, a persistent change in the word structure occurs during the days of the event, the word structure during the incident is significantly more similar to the word structure after the event than compared to the structure before the event. For a given point in time t, we use

$$\begin{aligned}
\Delta_t =& S_C\ (\text{event}, \quad \text{after}) - S_C(\text{event}, \quad \text{before}) && (2)\\
=& S_C\left(\sum_{i=t}^{t+\beta-1}\Gamma_i, \sum_{i=t+\beta}^{t+\beta+\gamma-1}\Gamma_i\right) - S_C\left(\sum_{i=t}^{t+\beta-1}\Gamma_i, \sum_{i=t-\alpha}^{t-1}\Gamma_i\right) && (3)
\end{aligned}$$

where Γ_i with $|\Gamma_i| = V$ is the vector of the total count for each word on day i as an indicator for how much the context of the text containing our entity have changed due to the incident.

By calculating the leave-one-out cosine similarity of the absolute word frequencies in texts with our entity in question, the words that have the greatest influence on the change of the associated words can be determined. Here, a distinction can be made between words that are less frequently associated with the entity after the incident and those that co-occur more frequently with the entity. Under the assumption that narratives are more likely to be born in a short amount of time and regress slowly by decreasing interest or merging with other narratives, especially those words whose frequency increases are of special interest. Consequently it is possible to examine how long these words are associated with the entity, i.e. how long the narrative remains relevant in this form.

3 Evaluation

The basis of our study is a combined data set with both print and online articles from "Süddeutsche", "Welt" and, "Handelsblatt". It comprises over 2.8 million articles in a period from 2001-2022 (see Figure 5). Handelsblatt is a daily newspaper specialized for economic topics while "Süddeutsche" and "Welt" offer articles to a wider range of topics. Thus, our data set contains a mix of political, economical and other texts, providing us with a large list of entities from different professions to analyze.

Following the approach described in Section 2.1, we found over 90,000 named entities in the articles. Due to the high computational cost of the search for these entities in the entire corpus, we searched for the complete set of 857 thousand entities in a random 1% sample of the entire corpus and registered the entities we found. If an entity does not appear in randomly selected 280,000 articles, it is likely not relevant enough for our analysis. Since a sufficient number of articles is necessary to model medial attention analysis, in the following analysis, we focus only on the 1000 out of 90,000 people who are mentioned most frequently over the entire corpus.

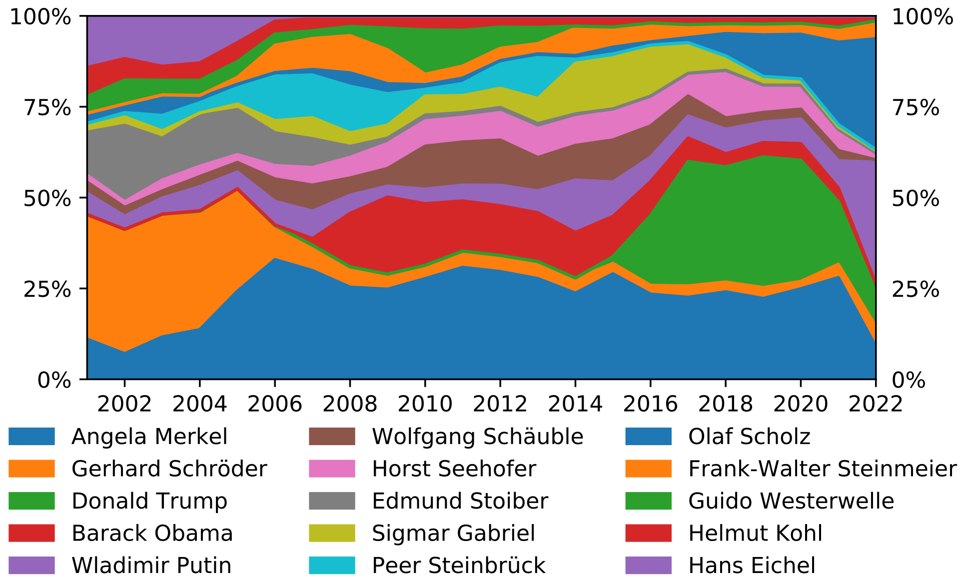


Figure 2: Relative occurrences of the 15 most frequent named entities over the entire data set.

Figure 2 shows the 15 most prominent individuals in the articles analyzed, all of them being politicians, including 12 Germans, 2 U.S. presidents and Russian President Wladimir Putin (see Table 2 for more categories.) The most frequently mentioned person is "Angela Merkel", with over 82 thousand articles. Her relative importance rose after she became the German Chancellor in 2005 and stayed relatively stable during her 16 years in office. Despite being in power for only one term, Donald Trump appears more often than Barack Obama, as he receives more media attention than any other person during the years of his presidency.

For a more detailed analysis of our entities, we created a network for community visualization with Yifan Hu's algorithm using Gephi (Bastian et al., 2009), displaying all connections between entities with a jaccard coefficient (Jaccard, 1912) of at least 0.02

(i.e. both entities appear in the same text at leas 2% of the time). The results can be found in our GitHub.

Figure 3 shows the number of articles per day in which Mario Draghi is mentioned. The dotted line displays the 99.9%- quantile of the theoretical Poisson distribution which was calculated on the basis described in Section 2.2. The days on which the quantile limit is exceeded are marked in red. A total of 57 such event days can be identified in the period between 2010 and 2022. If the entity would be mentioned following the Poisson distribution based on the last 365 days with no true events, the test would only detect an event in one out of one thousand days on average, which would result in a total of four to five detections over the entire time period. 57 detections compared to the expected four to five however suggest that there are a lot of true events that can be interpreted.

The first spike of media attention was triggered when Draghi took over the office of President of the ECB in November 2011. The most significant moment in his presidency was when he announced that ECB will "do whatever it takes to preserve the euro" on July 26, 2012. This is also evident in the article number, as the event identified in the following four days (see also Figure 6).

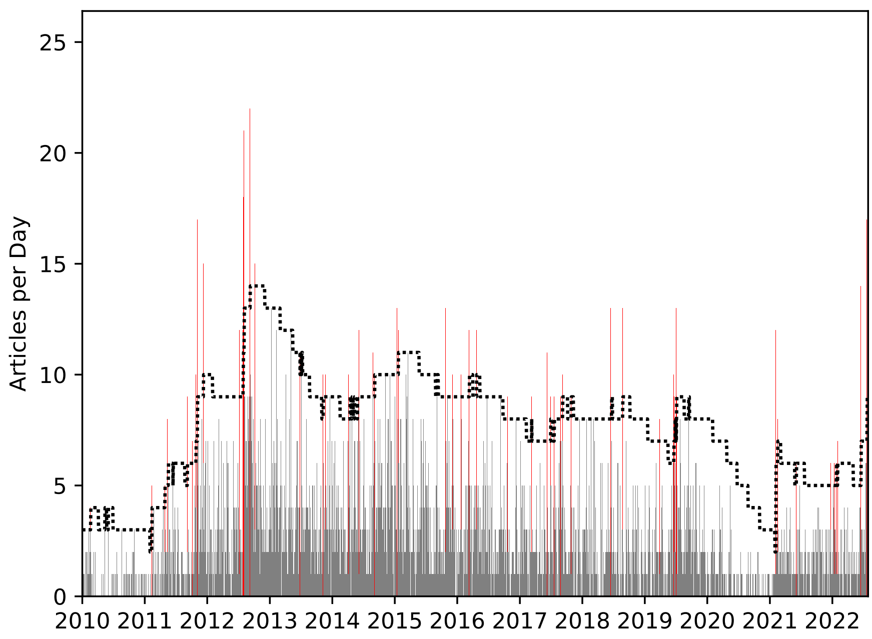


Figure 3: Number of articles mentioning "Mario Draghi" per day

Comparing the words associated with Draghi during the $\alpha = 30$ days before July 27th with those during the $\gamma = 30$ days after July 31st, we find by taking Δ_t that the words used during the event are much more similar to those used after the event than before ($\Delta_t = 89.64\% - 78.83\% = 10.81\%$). So this turning point in the history of the euro also represents one for the media's view of Mario Draghi.

If we use the leave-one-out cosine similarity to determine the words that have the greatest importance for the occurrence of the structural break we can recognize possible narratives: After the event, "bundesbank" and "weidmann" come into connection with Draghi much more frequently as Draghi got into a conflict with the German National Bank and its president, which was also magnified by the (German) media. The issue of possible illegal "staatsfinanzierung" (engl. state financing) would accompany Draghi for many years to come. Draghi as a person became more political ("politik") and was criticized by German media and politicians like "dobrindt" and "seehofer". On the other side, the uprise of "retten" (engl. rescue) could point to the narrative's birth of Draghi as the savior of the euro.

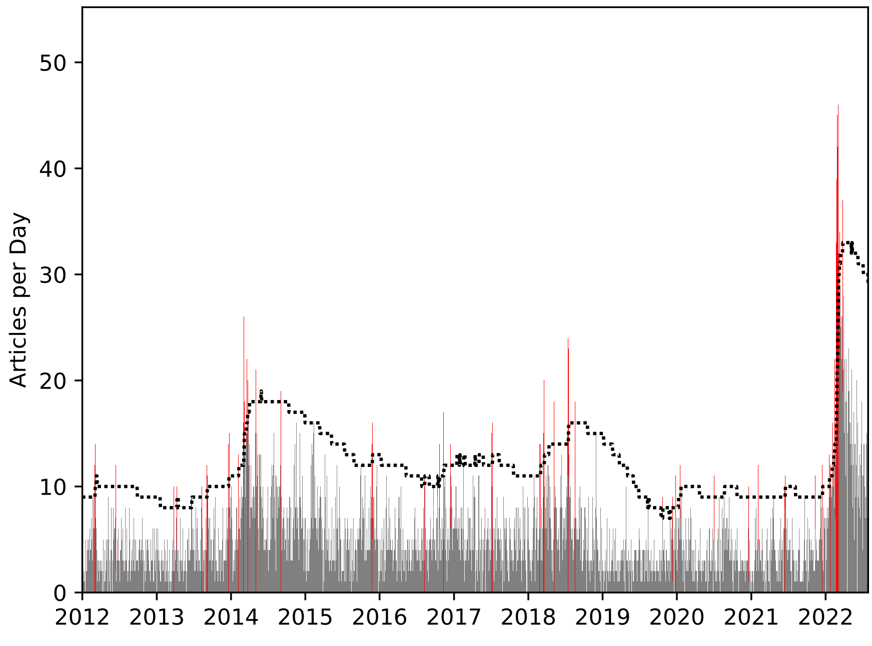


Figure 4: Number of articles mentioning "Wladimir Putin" per day

Analogous to Figure 3, Figure 4 shows the number of articles containing the entity

Wladimir Putin over the time period. While there are a couple of events detected throughout, two time periods contain particularly many events: late 2014 and early 2022. During these time periods, the entity rose in relevancy. Both dates mark two phases of Russia's invasion of the Ukraine – the annexation of Crimea in 2014 and the Russian-Ukrainian war of 2022. The latter is a long lasting incident, as the appearances of the entity continue to stay at its new level, which leads to a total of 19 detections (see Figure 7). Along with the change in word distribution surrounding this entity, we can assume that this narrative change was a structural break rather than a short-term event. Terms like "krieg" (engl. war), "angriffskrieg" (engl. war of aggression) are much more frequent than before the war started, even if Russia officially called it a "military operation" rather than a war. The words "menschen" (engl. the people), "flüchtlinge" (engl. refugees), "mariupol" and "butscha" started being mentioned alongside Wladimir Putin, as authors started to write about the victims of this war.

4 Conclusion

Identifying and evaluating narratives is becoming an increasingly important task in modern research. Depending on the definition and specifications, narratives can for instance be used to analyze expert opinions on economical actions, such as ways to resolve the high inflation in 2022, or the public opinion on parties and their politics. All these narratives revolve around entities, such as persons, parties, organizations or companies. We therefore seek to detect narratives using these entities.

We propose two models to detect narratives in a timeline by analyzing either the number of entity appearances or the change of word frequencies surrounding those entities. We use these different models to detect two distinct types of narrative. The narrative can have a very short life span, as it is linked to a temporary (perhaps seasonal) event that has no long lasting consequences. Or it can have a long lasting life span, starting out as an incident that changes the narrative of an entity permanently.

We evaluate our models on a data set of German news paper articles and detect several meaningful narratives surrounding major entities like the Russian president Wladimir Putin and Mario Draghi.

In future research we aim to improve our model by incorporating not only persons but also organizations and companies as entities to analyze. An additional pipeline to investigate the texts shortly after an event or structural break has been found, such as the pipeline proposed by Lange et al. (2022a), can be useful to extract the essence of changes that occured more easily. We also plan to compare our structural break detection using real word frequencies with word frequencies of topic models, which was proposed by Rieger et al. (2022) to detect structural breaks for possible narrative detection.

References

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Fisher, W. R. (1984). The Narrative Paradigm: in the Beginning. *Journal of Communication*, 34(1):74–87.

Harrs, S., Müller, L. M., and Rockenbach, B. (2021). How optimistic and pessimistic narratives about covid-19 impact economic behavior. ECONtribute Discussion Papers Series 091, University of Bonn and University of Cologne, Germany.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.

Jazi, M. A., Jones, G., and Lai, C.-D. (2012). First-order integer valued ar processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, 33(6):954–963.

Lange, K.-R., Reccius, M., Schmidt, T., Müller, H., Roos, M., and Jentsch, C. (2022a). Towards extracting economic narratives. *Ruhr Economic Papers #963*.

Lange, K.-R., Rieger, J., Benner, N., and Jentsch, C. (2022b). Zeitenwenden: Detecting changes in the german political discourse. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)*.

Li, B. and Han, L. (2013). Distance weighted cosine similarity measure for text classification. In Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., and Yao, X., editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, Lecture Notes in Computer Science, pages 611–618. Springer.

Rieger, J., Jentsch, C., and Rahnenführer, J. (2021). RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2337–2347, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rieger, J., Lange, K.-R., Flossdorf, J., and Jentsch, C. (2022). Dynamic change detection in topics based on rolling ldas. In *Proceedings of the Text2Story'22 Workshop*.

Roos, M. and Reccius, M. (2021). Narratives in economics. *Ruhr Economic Papers #922.*

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

Shiller, R. J. (2019). *Narrative economics. How stories go viral & drive major economic events.* Princeton, Oxford: Princeton University Press.

Appendix

Man	718,836	Actor	48,592	Swiss	27,491
German	294,593	Briton	41,517	Literature (20. century)	27,370
Woman	147,249	Austrian	37,252	Literature (German)	26,485
US-American	109,366	Abbreviation	34,690	Europe by place	26,014
Author	65,805	French	34,288	Actor	24,304

Table 1: Most frequent categories of German Wikipedia articles

Man	884	US-American	113	Journalist (DE)	50
German	666	Author	74	Manager	49
Politician (21. c.)	166	National football player (DE)	59	SPD member	49
Politician (20. c.)	142	Actor	52	Lawyer	37
Woman	116	GDR citizen	52	French	35

Table 2: Distribution of categories for the 1000 most frequent persons

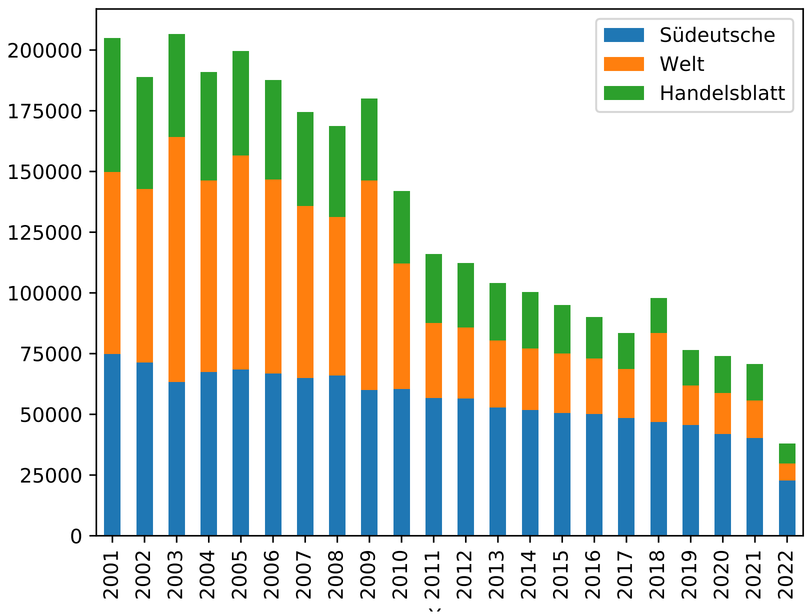


Figure 5: Number of articles published by "Süddeutsche", "Welt" and "Handelsblatt".

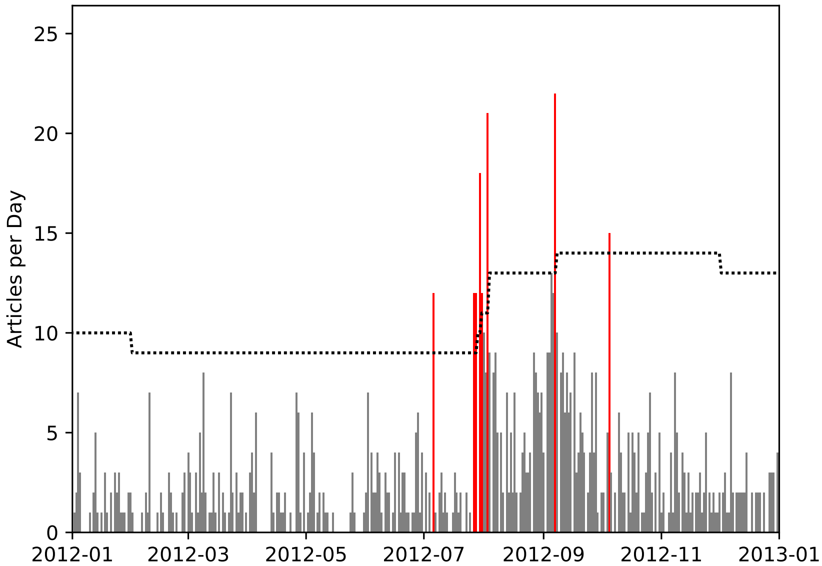


Figure 6: Number of articles mentioning "Mario Draghi" per day in 2012

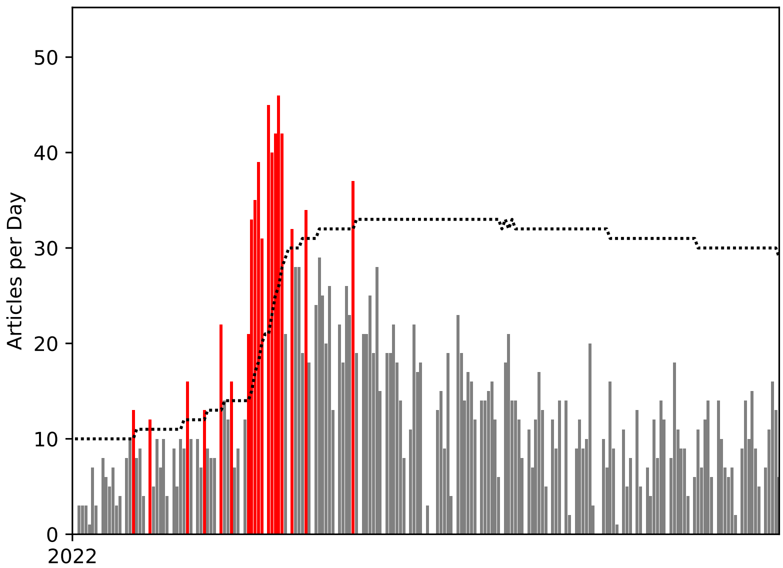


Figure 7: Number of articles mentioning "Wladimir Putin" per day in 2022

Lange, Kai-Robin et al.

Working Paper

Towards extracting collective economic narratives from texts

Ruhr Economic Papers, No. 963

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Lange, Kai-Robin et al. (2022) : Towards extracting collective economic narratives from texts, Ruhr Economic Papers, No. 963, ISBN 978-3-96973-127-7, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen,
https://doi.org/10.4419/96973127

RUHR

ECONOMIC PAPERS

Kai-Robin Lange
Matthias Reccius
Tobias Schmidt
Henrik Müller
Michael Roos
Carsten Jentsch

Towards Extracting Collective Economic Narratives from Texts

technische universität dortmund

RUB

#963

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Editorial Office

Ruhr Economic Papers #963

Responsible Editor: Ludger Linnemann

ISSN 1864-4872 (online) – ISBN 978-3-96973-127-7

Ruhr Economic Papers #963

Kai-Robin Lange, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch

Towards Extracting Collective Economic Narratives from Texts

technische universität dortmund

RUB

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973127
ISSN 1864-4872 (online)
ISBN 978-3-96973-127-7

Kai-Robin Lange, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch[1]

Towards Extracting Collective Economic Narratives from Texts

Abstract

Identifying narratives in texts is a challenging task, as not only narrative elements such as the actors and events have to be identified but their semantic relation has to be explained as well. Despite this complexity, an effective technique to extract narratives from texts can have a great impact on how we view political and economical developments. By analyzing narratives, one can get a better understanding of how such narratives spread across the media landscape and change our world views as a result. In this paper, we take a closer look into a recently proposed definition of a collective economic narrative that is characterized by containing a cause-effect relation which is used to explain a situation for a given world view. For the extraction of such collective economic narratives, we propose a novel pipeline that improves the RELATIO-method for statement detection. By filtering the corpus for causal articles and connecting statements by detecting causality between them, our augmented RELATIO approach adapts well to identify more complex narratives following our definition. Our approach also improves the consistency of the RELATIO-method by augmenting it with additional pre- and post-processing steps that enhance the statement detection by the means of Coreference Resolution and automatically filters out unwanted noise in the form of uninterpretable statements. We illustrate the performance of this new pipeline in detecting collective economic narratives by analyzing a Financial Times data set that we filtered for economic and inflation-related terms as well as causal indicators.

JEL-Codes: C18, C55, C87, E70

Keywords: Econometrics; narrative; text mining; coreference resolution; named entity recognition; causal linking

August 2022

1 Kai-Robin Lange, TU Dortmund; Matthias Reccius,RUB; Tobias Schmidt, TU Dortmund; Henrik Müller, TU Dortmund; Michael Roos, RUB; Carsten Jentsch, TU Dortmund. *- This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.* *– All correspondence to: Kai-Robin Lange, TU Dortmund University, 44221 Dortmund, Germany, e-mail: kalange@statistik.tu-dortmund.de*

1 Introduction

What has caused the rising inflation in early 2022? What exactly has lead to Russia's invasion of Ukraine in February 2022? Are the measures of the of the US government combating Covid 19 appropriate? Questions like these are so complex that individual reasoning will not reflect the full set of information that is relevant to them. Instead, opinion formation will be guided by narratives that abstract from the multitude of facts available. Narratives are cognitive instruments (Bruner, 1991) through which humans organize their experiences, explain the past and predict the future (Bénabou et al., 2018). In complex and uncertain situations, it is convincing narratives rather than facts and figures that enable agents to take action (Tuckett and Nikolic, 2017). As a result, the study of economic narratives is becoming an important component of empirical macroeconomics.

Fortunately from an empirical perspective, narratives are not only psychological. They also manifest as linguistic structures and are often used in textual and other media (Meer, 2022). However, despite the recent advances in quantitative language modeling, no empirical model yet exists that is able to automatically identify narratives in a corpus of textual data. One reason why such one-step analyses using word embeddings (Mikolov et al., 2013) or topic models (Blei et al., 2003) are not up the task is that narratives are quite particular constructs. Even though state-of-the-art language models can deliver impressive results in related and seemingly similar tasks like for instance text summarization, a narrative is more specific than a summary. Thus, any attempt to identify narratives inevitably necessitates a prior definition of the term.

In this paper, we leverage the definition of collective economic narratives (CENs) put forth by Roos and Reccius (2021) to extract inflation narratives from newspaper articles. According to this definition, a CEN is a sense-making story about the economy that is shared within a social group and suggests action. Based on this definition, we identify text passages that establish a causal link between economic events and actors and thus allow the reader to make sense of a concrete economic issue. In a general sense, the aim of this study is to start closing the gap between the theoretical understanding of narratives as provided by the cognitive psychology literature and the definition in Roos and Reccius (2021) and their empirical identification.

Often research on narrative extraction focuses on determining narratives through named entity recognition and event detection (Benner et al., 2022). Ultimately, these named entities and events can be collected and presented by displaying their connection (e.g. the verbs that connect a named entity with an event). One such example is the RELATIO-method by Ash et al. (2021), which clusters named entities with similar meanings. Words such as "health insurance", "health human service" and "health care system" can be collectively represented by the word "healthcare". As a result of this dimensionality reduction, RELATIO is less prone to tagging synonyms of the same named entity as different entities. Each extracted text passage corresponds to the structure "AGENT VERB PATIENT" (for instance "worker lose jobs"), so that all statements contained in the corpus can later be displayed as a network with the agents and patients as nodes and the verbs as edges (see Figure 2). The edges can than be weighted based on how often that exact narrative occurs. RELATIO works well for identifying economic events, phrases and statements which is why we use a modified version of it as part of our pipeline.

However, the structure that is recognized by RELATIO is missing the sense-making property that is central to the concept of CENs. RELATIO will commonly detect statements like "worker lose jobs" or "god bless america", that are either common phrases or simply facts. Such statements are not necessarily narratives in the sense of being driven by the narrator's world view or interacting with the recipient's belief system. Thus, we consider RELATIO to identify narrative building blocks that have to be combined in some way to yield full economic narratives. Combined statements like "workers lose jobs due to Covid 19" or "god bless america because america is best country in the world" would be considered complete CENs. Therefore, our pipeline is aimed at disregarding the noise created by incomplete narratives through the identification of causal indicators such as "because" that connect RELATIO-style narrative building blocks.

Being able to analyze the narratives attributed to a specific economic event or actor over time could offer insights on how they are viewed, how this perception changes and what actions may be taken towards them. Existing narrative extraction schemes often purely try to filter out events individually and display their factual connection as Directed Acyclic Graphs or temporal flow charts. In this paper however, we differentiate between narratives and factual description of events. That is, the sentence "XYZ died of Covid 19." is not considered a narrative, as it displays an event but does not contain an opinion or a sense-making element. On the other hand, the statement "XYZ died of Covid 19 because he was unvaccinated" is a narrative, as it is a counterfactual statement that is based on an assumption, opinion or personal world view rather than pure facts.

We propose a pipeline to identify such narratives by analyzing texts for causal structures. This pipeline identifies causal linkages between smaller statements, which can be seen as events, to provide the user with a collection of causally connected events to draw narratives from. For this purpose, first, we explain in detail, how our definition of a narrative differs from the literature in Section 2. We describe the data we use in Section 3, present each natural language processing task that is later used in our pipeline individually in Section 4 and combine them to propose our pipeline in Section 5. We present our results in Section 6 and provide an outlook to future research in Section 7. In Section 8, we conclude the paper.

2 Definition of an economic narrative

Different definitions of narratives coexist in several fields of research. Often times, the definitions used in applied work are not theoretically motivated and appear to be tailored to fit a specific research agenda. In some studies about narratives, the term is not defined at all. However, not subscribing to any specific definition of narratives allows for any type of verbal or textual information to be called and analyzed as a narrative. This is unfortunate because narratives possess specific characteristics that constitute their persuasiveness and differentiate them from other forms of text.

We leverage the definition put forth by Roos and Reccius (2021) according to which a collective economic narrative (CEN) is "*a sense-making story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions*" (Roos and Reccius, 2021, p. 13). The most fundamental part of this definition is the *sense-making story*. A story can be viewed as depicting a temporal sequence of (possibly independent) events without claiming any specific connection between them. For example, "Inflation started soaring and then my colleagues all quit" is a story (Forster, 1927). When those events are logically connected with each other, the story becomes a plot: "Inflation started soaring and then my colleagues all quit because our boss would not raise their salaries accordingly" is a plot. In this case, the causal connection provides the sense-making glue between the events. It allows the recipient of the narrative to deduce that it was the rise in inflation that ultimately caused the colleagues to quit. Similar to Andre et al. (2022), we exploit this sense-making function of causal reasoning as our primary way of identifying CENs. It is important to note, however, that an explicit causal connection is not necessarily required for a CEN. Sense-making connections can also be implied. An example is the sentence "Inflation shot up and the central bank raised interest rates.", which is a sense-making story to recipients with the proper background knowledge and economic education. In this case, the sense-making aspect of the narrative interacts with the belief system shared by most economically literate recipients.

Other elements of the definition by Roos and Reccius (2021), such as the emergence-property of narratives, are also quite demanding from an empirical standpoint because they cannot necessarily be detected or analyzed in a single text. This emergence-property for instance may entail the existence of proto-narratives – different versions and drafts of a narrative – that coexist, circulate and combine before a concensus is reached and a full CEN emerges in a group. This process cannot be captured by looking at single texts at a time. Instead, for this purpose, dynamic methods have to be used that can capture and represent narrative developments over time.

A distinction has to be made between personal narratives that people use to make sense of their own lives and collective narratives that serve a function for groups. The psychological literature tends to focus on the function of narratives in the decision-making of individuals. When faced with a decision, narratives thin out the vast space of possible actions that could be taken by favoring the option that best fit a simple and plausible narrative. While decision-making is obviously important from a macroeconomic standpoint, a narrative must transcend the individual and guide collective action in order to become relevant for the macroeconomy. In contrast to personal narratives, CENs can spur *collective* action. For example, "I need to renegotiate my wage because the soaring inflation lowers the purchasing power of my income" is a personal narrative insofar as it is employed by an individual to (i) analyze the current state of affairs, (ii) make predictions about the future and (iii) decide on a course of action. Only once it is shared and believed in within a larger group – of example a trade union – it emerges as a CEN and may affect macroeconomic dynamics such as wage-setting to a significant degree.

3 Data

The aim of our analysis is to identify CENs that are shared by a group of people and have the potential to become economically relevant. Like other scholars before us (Müller et al., 2022; Ter Ellen et al., 2021; Larsen and Thorsrud, 2019), we choose media data as our unit of study in this regard. We do so mainly for two reasons: i) journalistic texts are a pretty accurate proxy for debates that move society and ii) news articles are of high relevance for many market participants, especially for central banks. That is, narratives found there can not only potentially be interpreted as CENs, but also provide an excellent starting point for further research.

3.1 Narratives and the media

Journalists play a crucial role in the dynamics of narrative formation (Shiller, 2017). In order to meet the interest of their readership, journalists follow Twitter debates, evaluate letters to the editor, analyse the traffic on their online portals, track the reporting of their competitors - and thus know which topics move the public and which do not. If an already existing narrative gets big enough (for example, about villains and victims of inflation), it is likely that it finds its way into editorial conferences and thus (in some way) into reporting.

At the same time, this reporting leads to topics, frames[1] and narratives becoming entrenched in public debate.

We know from agenda-setting theory (McCombs and Shaw, 1972) that the media have a decisive influence on what we discuss in a society. When it comes to evaluate the general economic situation, media are even the most important source for many citizens (Lischka, 2015; Blinder and Krueger, 2004). So, while public debates regularly find their way into editorial conferences, it is primarily the media themselves that determine the popularity of a topic and/or a certain narrative. This dynamic leads us to see media texts as a natural and very promising textual basis for finding sufficiently popular narratives on inflation.

[1]In communication science frames are often described as the selection of "some aspects of a perceived reality [to] make them more salient in a communicating text [...]" (Entman, 1993, p.52). Just like narratives, frames are judgemental and have the potential to establish a social view that influences decision-making processes. Compared to narratives, however, frames tend to be static in the sense that frames focus on specific topics, whereas narratives cover a longer period of time. Following Müller et al. (2018) we propose the interpretation that "a frame is to a narrative what a still is to a movie" (p.559). More information on media frames and their link to public opinion in Scheufele (1999); De Vreese (2005); Scheufele and Tewksbury (2007); Matthes (2014).

3.2 Central banks and the media

Since central banks have recognised forward guidance as a central element of their monetary policy, media reports have increasingly become the focus of their attention. Forward guidance describes central banks' efforts to manage inflation expectations – and ultimately actual inflation – through targeted communication. It is "intended to correct faulty expectations, and thereby reduce misallocations of resources" (Blinder et al., 2008, p.22). Used wisely, it has the potential to move financial markets (Blinder et al., 2008), improve the predictive power of monetary policy (Haldane and McMahon, 2018), guide inflation expectations (Armantier et al., 2016; Binder and Rodrigue, 2018; Eusepi and Preston, 2010), and influence consumer behaviour (Armantier et al., 2015).

The media play a crucial role in this process (Berger et al., 2011). As a central bank wants consumers and decision-makers to interpret the economic situation correctly, it needs journalists to understand their analyses, and communicate them to the public in an understandable and captivating way. A central bank will hardly succeed in setting a certain narrative - for example, that rising commodity prices are responsible for inflationary dynamics - if most of the business press presents a different interpretation. Therefore, central banks make use of several communication channels (Monthly Bulletin, speeches, interviews, monthly press conferences etc.) to directly influence media coverage (Conrad and Lamla, 2007; Haldane and McMahon, 2018). However, despite all their efforts, they do not always succeed in getting their own narrative featured in the press: In a Speech at the 148th Baden-Baden Entrepreneurs' Talk, ECB Director Isabelle Schnabel expressed her displeasure that "many supposed experts and the media are again rousing people's fears without explaining the reasons behind the price movements" (Schnabel, 2021). The quote impressively illustrates how important the business press is for the ECB's forward guidance[2]. Above all, it shows how crucial it is for a central bank to recognise any (unwanted) narratives in the press. It is not surprising that central bank researchers are increasingly involved in research on narratives, especially narratives in the press (Ter Ellen et al., 2021; Nyman et al., 2021; Kalamara et al., 2020).

[2] The influence of media coverage on inflation expectations is well documented. See e.g. Larsen et al. (2021); Coibion et al. (2018b,a); Lamla and Lein (2014).

3.3 The media as data

Both, the high potential of media data to contain CENs and its relevance for the economy leads us to use the business press as our unit of investigation.

The medium that fits our research interest most is the Financial Times (FT). The weekday Financial Times newspaper is one of the world's leading business, politics and world-affairs newspapers. If there are any economic narratives circulating in a society, we are likely to find them in the Financial Times - either because they originated there or because they are reproduced in its reporting.

Our analysis corpus consists of all FT articles published in the English-language edition between 1/1/2010 and 18/07/2019 that depict economic uncertainty to some extent. The search term that narrows the corpus to economic uncertainty is based on the research of Baker et al. (2016). The keyword is of the the form econom* OR uncertain* and covers all articles that contain both patterns at least once.

In order to only capture narratives about the inflation, we further reduce the text base to those paragraphs in which the pattern *inflation* occurs. We define a paragraph as the sentence in which the pattern *inflation* appears as well as the six sentences before and the six sentences after it. As a further restriction we filter for paragraphs that also contain at least one causal indicator (see Table 1). Our text base comprises a total of 18,375 unique paragraphs. In this collection, we search for cause-and-effect statements that match our definition of an economic narrative.

4 Natural language processing sub-tasks

In the following sections we explain the process of some common tasks of natural language processing that are all incorporated into our pipeline. While Semantic Role Labeling, Named Entity Recognition and K-Means clustering are already used by the RELATIO-method Ash et al. (2021) which we cover in Section 4.5, in our pipeline, we perform an additional layer of Named Entity Recognition paired with Coreference Resolution to improve the method's consistency and make the results interpretable out of their textual context. After defining these components of our pipeline in this section, we will explain how we combine them with additional post-processing and filtering to create a narrative extraction technique in Section 5.

4.1 Semantic Role Labeling

Semantic Role Labeling is the process of assigning tokens to semantic roles that describe their effect within the sentence in relation to a specific verb (Jurafsky and Martin, 2020, pp.405 ff.). More specifically, in the RELATIO-method (Ash et al., 2021), which use as a part of our pipeline, Semantic Role Labeling is used to determine *who* did *what* to *whom*. In this case, *what* represents the verb and *who* as well as *whom* represent the active and receiving part of this action, which we call *agent* and *patient*.

The Semantic Role Labeling in this paper will be based on the state-of-the-art approach by AllenNLP (Gardner et al., 2018), which is also able to handle verb negations, allowing us to incorporate these into our statements and narratives.

4.2 Named Entity Recognition

After splitting each sentence into its semantic roles, we only want use the roles relevant to us, as these statements will represent the most important information of any event: *who did what to whom?*. We consider other roles to be unwanted noise, as anything beyond the description of the event would add minor details or linguistic mannerisms of the author, which might differ from text to text, to the statement. The statements thus only become needlessly large and distinct from another, which would greatly increase the efforts needed to identify common statements shared by multiple authors. Condensing short statements, many of which are overlapping, into a set of logical narratives is already a complex task. Hence, doing so to distinct and longer statements would be even harder. In the latter case, little mistakes from probabilistic models in the pipeline would yield an even larger amount of false statements. Hence, we will only use smaller and compact statements for our pipeline and leave narrative extraction using longer statements for future research. This compact form does however still contain unwanted noise, as many of the words applied to each semantic role may be representing the same underlying latent entity, but have a slightly different wording. For instance the terms "short term rate" or "short term interest rate" may both represent the same entity and can generalized by denoting them and similar entities simply as "interest rate".

Named Entity Recognition is the task of checking whether a semantic building block represents a real world entity, may it be a person, a specific building or an organization (Jurafsky and Martin, 2020, pp.164 ff.). In our context, it checks if an agent or patient itself is an entity or if it refers to a latent entity that is not specifically named. The Named Entity Recognition in this paper is performed by Spacy's Named Entity Recognition (Honnibal and Montani, 2017).

4.3 Coreference Resolution

Ultimately, instead of looking at each text in detail, we want to exclusively compare narratives from anywhere in the corpus to another without needing to read their respective contexts. This yields the problem that some sentence building blocks like pronouns cannot simply be interpreted out of context. For instance, a human reader may figure out that the "it" in the sentence "The fed tries to combat it by raising interest rates" represents the word "inflation". Our pipeline also needs to detect and resolve such coreferences, or else the resulting collection of narratives would still need a lot of manual work to interpret.

This automatic replacement of pronouns is called Coreference Resolution (Jurafsky and Martin, 2020, pp.445 ff). In this paper, we use Spacy's Coreferee Coreference Resolution model (Honnibal and Montani, 2017). The nouns within each sentence are detected and linked to nouns and pronouns in other sentences, based on which entity they represent. These linked nouns and pronouns are called a Coreference chain. The chain is then resolved by choosing one of the chain elements to replace all other chain elements in the original texts with. We will describe how we choose our element to resolve the chains with in Section 5.2.

4.4 K-Means clustering

As a large number of distinct entities is not easily interpretable, reducing them to a reasonably small number of latent entities is a crucial dimension reduction step. We cluster the resulting patients and agents that cannot be linked to entities by Named Entity Recognition. This way, one latent entity represents a larger number of agents and patients. In the context of natural language processing, clustering is performed in combination with word embeddings (Mikolov et al., 2013), which transform a word into a vector based on its semantic meaning. The word embeddings, each representing one word, are clustered using K-Means clustering. The word with an embedding closest to the resulting cluster mean, is chosen to represent said cluster. Thus, each non-recognized agent or patient is assigned to a latent entity that is determined by one of the cluster means.

The K-Means algorithm, given vectors $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$, assigns each vector to one cluster, based on the distance to the cluster's mean. Thus the training objective is to minimize the distance of all our vectors to their respective closest cluster means

$$\underset{S=\{S_1,S_2,\ldots,S_k\}}{\arg\min} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||$$

where $\mu_i \in \mathbb{R}^d$ denotes the mean of cluster S_i. The only parameter of this algorithm is k, the number of clusters to use (Lloyd, 1982). To make our clusters distinct enough, we chose $k = 300$ clusters for this paper.

4.5 RELATIO

RELATIO is a method developed by Ash et al. (2021) to identify narratives in large collections of text. However, the statements identified by RELATIO are narratives in a much broader sense than the CENs that we consider in this paper. To avoid confusion between the two definitions of a narrative, we will call the results of the RELATIO-method "statements" instead of narratives. Ash et al. (2021) extracts statements from a text, which are formed by three narrative building blocks called "agent", "verb" and "patient" as explained in Section 4.1. Although these building blocks do not suffice our definition of a narrative, they do provide important groundwork for us to be able to extract more complex narratives that conform more closely to the concept of a CEN.

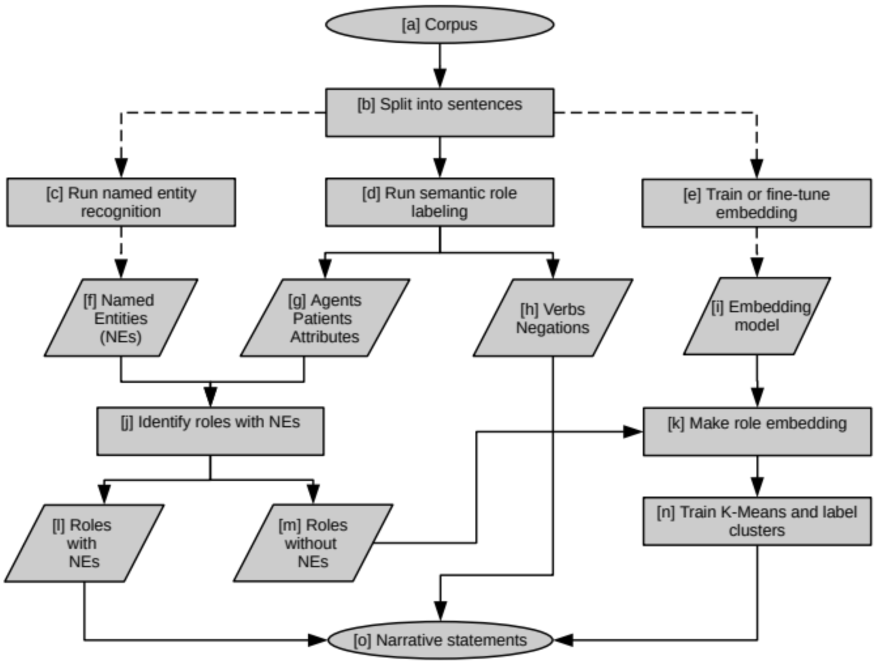


Figure 1: Flowchart of the RELATIO-method (Ash et al., 2021)

The RELATIO-method is displayed as a flowchart in Figure 1. It first splits each document into parts to evaluate each sentence separately. Then, Spacy's (Honnibal and Montani, 2017) Named Entity Recognition is used to detect named entities within each sentence. Parallel to this, Semantic Role Labeling is used to distinguish agents, patients and verbs. If the agents and patients align with named entities that were found, they are directly used as an agent of patient respectively. The remaining agents and patients that cannot be identified as named entities are likely a very high dimensional set of unique terms. That is, a lot of agents or patients will represent the same latent entity, but will have slightly different wordings. To reduce the dimension of these terms, they are clustered to a smaller set of latent entities using K-Means clustering. Thus, all agents and patients are assigned an entity. The final statements can be formed by combining agents and patients as well as the verbs that were identified during Semantic Role Labeling (verb negations included, if any).

5 Causal RELATIOns

In this section, we present our proposed pipeline by describing, how we combine and enhance the methods defined in Section 4 and proceed beyond the scope of the RELATIO-method. We detail our filtering process to only look at relevant articles, describe how we pre- and postprocess our data to be able to detect more complex statements and discuss how we intend to look for causal connections.

5.1 Detecting causal articles

As defined in Section 2, we are focusing on narratives displaying (explicit) causal connection between events. Filtering the corpus we seek to analyze for causal articles or paragraphs is thus a preprocessing step that will reduce unwanted noise. Like most tasks in natural language processing, this can be done using either a lexicon-based or learning-based approach.

We did train a BERT-model (Devlin et al., 2019) to detect causal articles (which we describe in further detail in Section 7), however, in this paper, we decided to filter our data using a lexicon-based approach. As we look for causally connected events, causality not only needs to be detected within articles in general, but, ultimately, when looking at two particular statements. While a learning-based approach that is specifically trained for this task will likely do this sufficiently well (see Section 7), causal indicators such as "because" can be used to link causal statements both more easily and more transparently. Crucially, lexicon-based identification reduces the chance for misinterpretation due to a failure of the learning-based model. Therefore we decided to use a lexicon approach to filter for causal articles, as implicit causality detected by a learning-based approach would not improve our pipeline, unless the model used to connect statements later on is also learning-based. The filtering process is described in Section 3.

5.2 Preprocessing and detecton of statements

After the texts to analyze have been filtered from the corpus, we use further preprocessing to improve our narrative extraction. For this we utilize a combination of Named Entity Recognition and Coreference Resolution, using both Spacy's Coreference Resolution and Named Entity Recognition on our original texts (Honnibal and Montani, 2017). We then look at each Coreference chain separately and use the Named Entity as an indicator on how to resolve the Coreference chain. If an entity detected by our Named Entity Recognition algorithm or at least 33% of its tokens (most tokens are simply words) are part of an element of the chain, the chain is resolved by inserting this entity into all of its references. The reasoning for the 33% threshold is that it enables us to identify three-part names, even if only one of its parts (e.g. only the last name of "Donald J. Trump") comes up in a chain element. If no entity is overlapping with the chain, the chain resolves using its first element, as this is likely the correct entity and later elements consist of pronouns referencing it. The texts with resolved chains are then saved as preprocessed texts.

After these preprocessing steps, the RELATIO-method is applied to the resulting texts. In addition to the usual pipeline, we return information about which token within a sentence represents the beginning and end of a statement. This is helpful in identifying smaller rather than larger statements. RELATIO often creates very long statements that tend to add additional noise and complicate interpretation. For instance, the sentence "economists say, the current inflation is caused by supply shortages and central banks must raise interest rates" will contain three entangled statements in the style of agent; verb; patient: *inflation; is caused; supply shortages*, *central banks; raise; interest rate* and *economists; say; the current inflation is caused by supply shortages and central banks must raise interest rates*. The first two statements are interesting for CEN extraction, but the third statement only repeats the entire sentence. This is because Semantic Role Labeling assigns the entire relative sentence as a patient to the verb "say". While this is factually correct, the resulting statement cannot be easily grouped alongside other statements due to its length and will only add unwanted noise to our resulting data. Instead of using all three statements, we only want to use the first two. We choose the maximum number of non-overlapping statements per sentence to represent the sentence – this task is similar to common scheduling-tasks in computer science. Thus, to filter out unwanted long statements, we use a greedy approach within sentences by always adding the current statements which stops first and does not overlap with any statement that was identified before (and thus starts at a later token than the previous statement had ended on) to our collection of filtered statements. This collection is our final statement list, which we use to display our results.

5.3 Causal direction

A meaningful narrative extraction pipeline should incorporate a way to detect a causal connection so that it can connect statements in a logical manner. This can be done using lexicon approaches or by analyzing the linguistic structure of the documents. While explicit causality like in the sentence "The inflation is high because of the supply shortage." are easily identifiable by lexicon approaches due to the word "because", implicit causality is not trivially detectable. However, lexicon-based methods that simply check a document for certain causal or temporal keywords are not flexible enough to detect causality in the statements "The Russo-Ukrainian war caused the gas prices rise." or "Donald Trump was a bad President. He divided the American people.", even though the causality is easily identifiable for a human.

While learning-based approaches, such as transformer-based models would be suitable for such a task, they also need to be trained correctly and carefully. By trying to detect implicit causality, such models can yield a high false-positive error rate and thus connect a lot of statements that are in reality not causally connected. This error rate can also not be easily resolved by persons checking the results manually, as such implicit causality might be hard to detect by the statements alone and the person thus would need to check the surrounding context of the statements. A learning-based model still remains a possible improvement in future research, but needs to be trained well to minimize the occurrences of false-positive errors. A lexicon approach on the other hand does not require additional training and the false-positive error is greatly reduced, as a explicit causal indicator directly implies the existence of causality, while implicit causality always relies on the context of the statements. These causal indicators can be found in Table 1 in the appendix. We look for these causal indicators in combination with two adjacent statements to combine them into a larger statement. We show examples of this in Section 6.2.

6 Evaluation

In this chapter, we evaluate the performance of our narrative extraction pipeline and compare it to the RELATIO-method in two steps. In Section 6.1 we show the graphical representation of our results using pre- and posprocessing compared to the original RELATIO-method. Then we show exemplary "complete" narratives that can be found using our lexicon approach for causality-filtering in Section 6.2.

6.1 Results

We display our resulting statements using the graph structure of the original RELATIO in Figure 2 and Figure 3. While Figure 2 shows the original detected actors and patients, Figure 3 shows the results after clustering the non-recognized entities to reduce dimensionality. In addition to this, Figure 7 shows the results when using the original RELATIO-method without any pre- or postprocessing. We can see that the most prominent statements of Figure 7 are similar to the ones of our adjusted pipeline in Figure 3, which was expected as the goal of Coreference Resolution was not to change the important statements of the corpus. Instead, it increased the number of already prominent statements, as common entities like *fed*, *interest rate* or *inflation* appear even more often. The pre- and postprocessing are also working as intended in context of their goal described in Section 5.2. Long statements that are too long to be generalized like `commitee; continiue; anticipate economic condition include low rate resource utilization...` (see Figure 7 at the bottom) are filtered out in Figure 2 and Figure 3.

While we see some statements of interest for our research in Figure 2, such as "fed eliminate inflation", "inflation adjust interest rate", "economy gain strength" or "economy need help", we can see there are a lot of distinct statements due to slight differences in wording. Entities like "interest rate", "benchmark interest rate", "bank benchmark interest rate", "fed benchmark rate", "short term rate", "short term interest rate" all refer to either the same or a very similar entity that can be generalized as "interest rate". Due to this, we use RELATIO's clustering algorithm to reduce dimensionality and work with less, more interpretable entities.

Figure 3 shows that a smaller amount of distinct statements improves our search for inflation narratives, as the clustering to latent entities reduces noise and also increases the number of inflation statements, as it is recognized as a latent entity. The clustering process behind the dimension reduction profits from the preprocessing as well as our filter for shorter statements. On the one hand, important entities like "inflation" appear more often, so that less agents and patients have to be clustered and are instead directly assigned to an entity. On the other hand, unreasonably long terms like the patients that can be found at the bottom of Figure 7 cannot simply be summarized into a single word. Removing them from model by filtering for shorter statements thus also increases the quality of the clustering process.

With these new statements we can filter out additional statements of interest like "oil price push inflation", "economist expect inflation", "government raise inflation", "inflation adjust unemployment".

Figure 2: Top 75 statements (39 distinct agents and patients) detected by the Causal RELATIOns pipeline without dimension reduction.

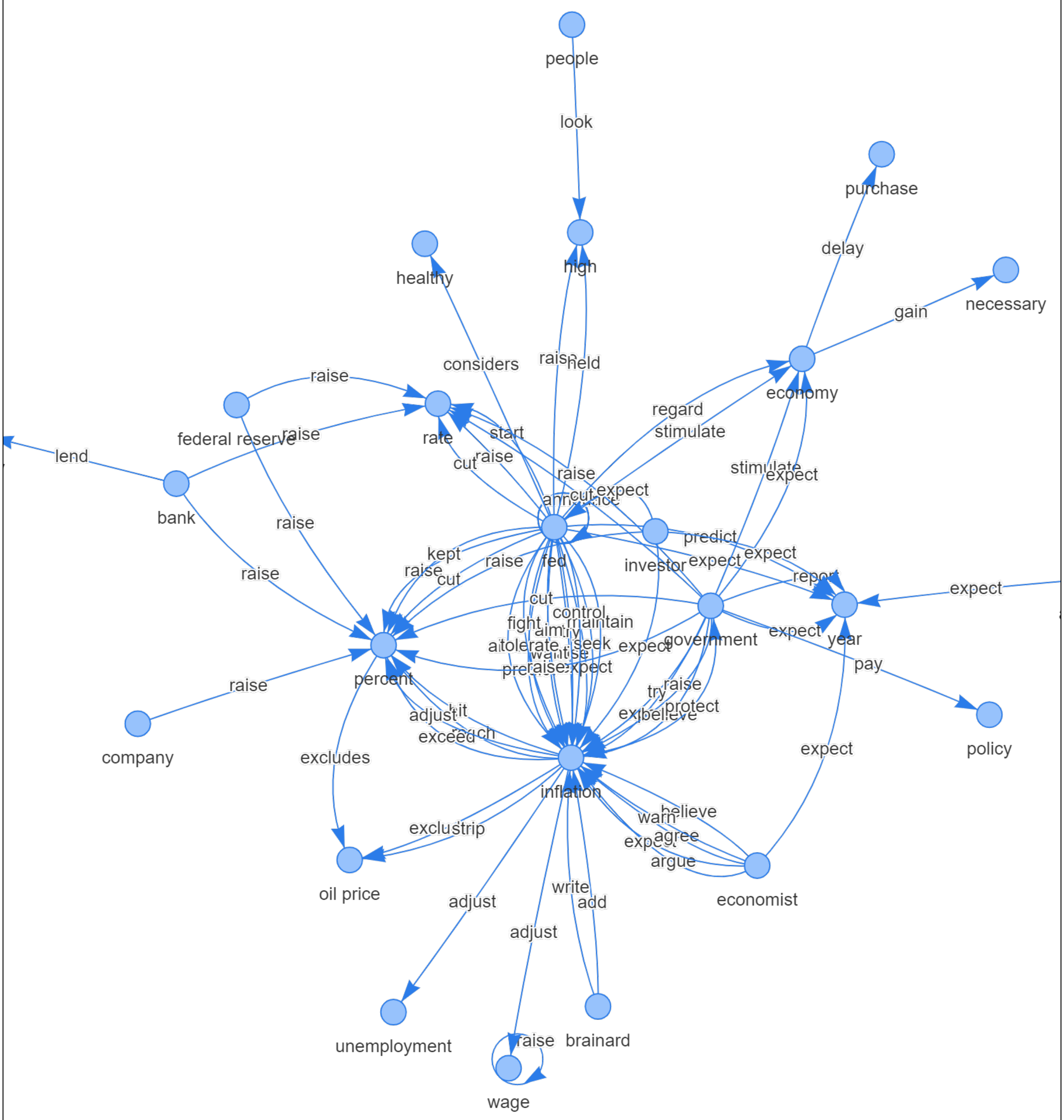


Figure 3: Top 75 statements (24 distinct agents and patients) detected by the Causal RELATIOns pipeline including dimension reduction.

We can also see statements that seem to stem from conflicting narratives such, that is "fed fight inflation" and "fed tolerate inflation". These two statements also show, why it is important to extract the entire narrative, rather than parts of the statements. Without further information we cannot tell, if these two statements are conflicting or whether they display the same narrative but show different perspectives, such as "The fed tolerates inflation in the livestock sector while it fights inflation in the energy sector". To identify these narratives, we need to analyze causal connections to relate the incomplete statements. We aim to display such narrative chains as graphs similar to the statements above. For now, we will display some exemplary CENs extracted from the text data in the following section.

6.2 CENs as chains of causally linked statements

The ultimate goal of the narrative extraction pipeline is to identify CENs in a fully automated way. Naturally, the requirement for a causal, sense-making link between narrative building blocks reduces the set of narratives that are extracted by the pipeline compared to RELATIO. This narrative structure also makes for more comprehensive and linguistically more complex narratives, as it increases the total number of featured tokens. Figure 6 visualizes two examples of a CEN as two causally linked statements that both consist of an AGENT, a PATIENT and VERB, where the latter specifies the relationship between the two former elements.

The CEN displayed in Figure 4 identifies a causal relation between the price-setting of fast-food restaurants and government subsidies for certain crops. Both halves of the CEN are identified separately and then connected through the causal indicator "because". In this case, coreference resolution was not utilized to attain the results because all entities and proper names in the narrative are stated and no pronouns are used.

The CEN in Figure 5 has indeed only been identified through coreference resolution. In the original text, the statement reads "*Yet funds are buying bonds because they (or their advisers) deem them low-risk*. The coreference resolution algorithm has correctly replaced the pronoun "*they*" with the named entity "funds" and the pronoun "them" with the term "bonds" before the narrative was identified. Without the prior use of coreference resolution, the second half of the CEN would not have been detected by RELATIO.

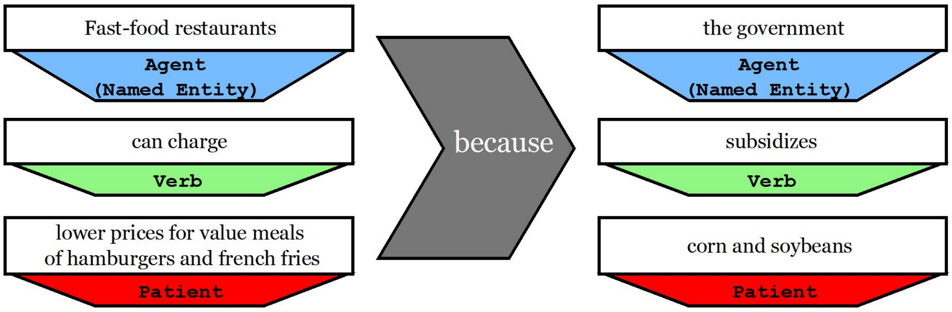


Figure 4

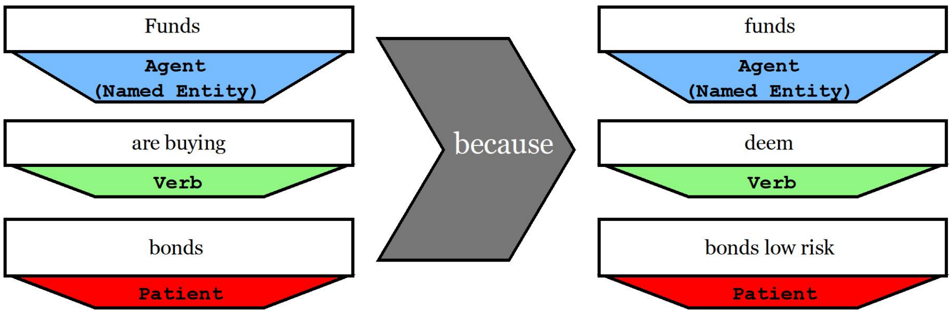


Figure 5

Figure 6: Examples of extracted CENs. Both sides of the plot show extracted statements that our pipeline links by the means of a causal indicator (here: "because").

It is important to note that, at this point, the CENs identified with our method have to strictly conform to the structure displayed in Figure 6. This empirical strategy minimizes the identification of false positives, as was the primary objective of this research. However, this approach also increases the false negative rate with regards to the sense-making criterion of CENs. For example, the statement "*The last 12 months have brought growing reports of mass layoffs, unpaid wages and factory closings because of outstanding government debts.*" is missed by the algorithm because it´s second half is a CEN, but does not conform to the AGENT, VERB, PATIENT structure. It would be very easy, however, to restate the narrative in a way that would be recognizable to the CEN detection algorithm: "*The last 12 months have brought growing reports of mass layoffs, unpaid wages and* *factory closings because* ***the government has accumulates a lot of debt***.". The two statements are semantically identical, but structurally different.

7 Outlook

This paper is designed to show our current work in extracting narratives from texts. As this work is not yet complete, we want to designate a section to provide an outlook to future research. As the extraction of narratives is a very complex process, we do have a lot of ideas of how to improve our pipeline, as there are many ways it can be adjusted step by step.

While we are able to detect explicit causality within chains by looking for causal indicators, displaying a large amount of narratives is a complicated task, as the causal connections open up a large variety of different narratives to be found. A graphic thus cannot simply show all detected narratives, but has to display them in a coherent way. In addition to this, we want our causality detection to not be based on lexicon approaches but rather learning-based approaches such as BERT, which could work like a variation of C-BERT (Khetan et al., 2020) that detects the flow of causality between two statements. As explained in detail in Section 2, the reason for this is that lexicon approaches can necessarily only identify explicit causal cues. Because sense-making stories can exist even when events are connected temporally or in other ways, a learning-based approach is required to improve the precision of the identification algorithm. However, the C-BERT model cannot be used directly, as it only detects the direction of causality within a statement and not between two statements. Creating such a model is thus an outlook to future research.

As soon as this causal detection between statements is not based on a lexicon any more, we will be able to filter our texts more freely. While researching for this topic, we fine-tuned a BERT model (Devlin et al., 2019) to detect (even implicit) causality within paragraphs using a self-created training data set based on the Yahoo Questions-and-Answers data set (Yahoo! Webscope, 2008). As far as we know, there is no data set available that displays causal versus non-causal texts in a journalistic setting, or at least not in a scale that is sufficient for training a BERT-model, which is why we use this Yahoo data set instead. This training data set uses the answers to questions asked on Yahoo Answers as its documents. We split the data into causal and non-causal texts by filtering by the type of question. An answer is considered causal if the question contains the word "why" and is considered non-causal if it does not contain the word "why" as well as a set of previously defined causal indicators. While a data set based on Yahoo Answers is not optimal for analyzing journalistic texts, it yields a starting point to work with. With the help of manual work, a journalistic data set can be created in future research, for instance by using a data set creation approach similar to the one of Pavllo et al. (2018).

Lastly, the clustering of entities used by RELATIO can be improved by using an embedding model that is fine tuned for economic texts rather than the general-use model proposed by Ash et al. (2021). In a similar matter, to analyze the current economic developments, we aim to use an updated version of the Financial Times data set, as our version only includes articles up until 2019. We could also improve the other aspects of the RELATIO-method or even our preprocessing in the form of Coreference Resolution or Named Entity Recognition. We are however currently content with their results and want to focus on the tasks that we are still missing to display a narrative in the form we want to. Choosing which Coreference Resolution, Semantic Role Labeling or Named Entity Recognition model we use can give our model the final touch, but we do not expect these changes to massively impact our results.

We also have ideas on how to proceed further, beyond the scope of this pipeline. While we are currently working on detecting and displaying narratives, we are also asking the question on how to monitor a narrative. For this, we are contemplating the idea of using a combination of this narrative extraction model in combination with change detection methods, which we already worked on (Rieger et al., 2022), to link change points in time to narrative shifts. We are also working on alternative perspectives to narrative detection, such as the detection of entity-related events constituting (new) narratives in a time line by focusing on entities rather than causal connections (Benner et al., 2022). As motivated in Section 2, a method based on change detection may also be able to take account of the dynamic aspects of narrative emergence. Such dynamic aspects are fundamental to CENs, but they elude cross-sectional methods by construction.

8 Conclusion

We present a pipeline that aims to extract economic narratives from a corpus of journalistic texts. We identify the definitions of narratives in the literature to be too broad for our understanding, as most of the definitions subsume common phrases or simple factual descriptions under the umbrella term "narrative". Instead, our definition puts the focus on statements that establish a sense-making connection between events and named entities that is based on a personal assessment, a world view or an opinion. Our pipeline uses the RELATIO-method as a groundwork and adapts it to detect complete CENs rather than smaller statements or narrative building blocks. We add Coreference Resolution in combination with Named Entity Recognition to our pipeline to improve the pipelines consistency and to make it interpretable out of context. By filtering the text for causal indicators, we remove unwanted noise and are able to detect complete narratives more easily by detecting causal links between pairs of RELATIO statements.

We evaluate the performance of our pipeline on a data set generated from articles of the Financial Times that is filtered for the word "inflation" and causal indicators. By equipping the RELATIO-method with our pre- and postprocessing we are able to remove unwanted and large statements and to include new statements that were found using Coreference Resolution. Consequently, our pipeline is able to detect meaningful economic narratives, for which we provide two examples.

The automatic detection of complete narratives that display a sense-making, causal connection between events still requires further research. This is mainly because of the inherent complexity of language and the often central role of subtext in establishing meaning in texts. However, we show that our adjustments improve the detection and consistency of the RELATIO-method for our purpose and provide some examples of narratives that we were able to manually extract using the causality detection algorithm.

References

Andre, P., Haaland, I., Roth, C., and Wohlfart, J. (2022). Narratives about the macroeconomy. CEBI Working Paper Series, (18/21).

Armantier, O., Bruine de Bruin, W., Topa, G., Van Der Klaauw, W., and Zafar, B. (2015). Inflation expectations and behavior: Do survey respondents act on their beliefs? International Economic Review, 56(2):505–536.

Armantier, O., Nelson, S., Topa, G., Van der Klaauw, W., and Zafar, B. (2016). The price is right: Updating inflation expectations in a randomized price information experiment. Review of Economics and Statistics, 98(3):503–523.

Ash, E., Gauthier, G., and Widmer, P. (2021). Text semantics capture political and economic narratives. arXiv preprint.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4):1593–1636.

Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, imperatives, and moral reasoning. NBER Working Paper, (24798).

Benner, N., Lange, K.-R., and Jentsch, C. (2022). Named entity narratives. Ruhr Economic Papers, 962.

Berger, H., Ehrmann, M., and Fratzscher, M. (2011). Monetary policy in the media. Journal of Money, Credit and Banking, 43(4):689–709.

Binder, C. and Rodrigue, A. (2018). Household informedness and long-run inflation expectations: Experimental evidence. Southern Economic Journal, 85(2):580–598.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. Journal of economic literature, 46(4):910–45.

Blinder, A. S. and Krueger, A. B. (2004). What does the public know about economic policy, and how does it know it? Princeton University's Industrial Relations Section Working Papers, 875.

Bruner, J. (1991). The narrative construction of reality. Critical Inquiry, 18(1):1–21.

Coibion, O., Gorodnichenko, Y., and Kamdar, R. (2018a). The formation of expectations, inflation, and the phillips curve. Journal of Economic Literature, 56(4):1447–91.

Coibion, O., Gorodnichenko, Y., and Kumar, S. (2018b). How do firms form their expectations? new survey evidence. American Economic Review, 108(9):2671–2713.

Conrad, C. and Lamla, M. J. (2007). An den lippen der ezb–der kof monetary policy communicator. KOF Analysen, 2007(4):33–45.

De Vreese, C. H. (2005). News framing: Theory and typology. Information design journal+ document design, 13(1):51–62.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. McQuail's reader in mass communication theory, 390:397.

Eusepi, S. and Preston, B. (2010). Central bank communication and expectations stabilization. American Economic Journal: Macroeconomics, 2(3):235–71.

Forster, E. (1927). Aspects of the Novel. Penguin Books Ltd.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Haldane, A. and McMahon, M. (2018). Central bank communications and the general public. In AEA papers and proceedings, volume 108, pages 578–83.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Jurafsky, D. and Martin, J. H. (2020). Speech and Language Processing. Third Edition.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. Journal of Applied Econometrics.

Khetan, V., Ramnani, R. R., Anand, M., Sengupta, S., and Fano, A. E. (2020). Causal-bert : Language models for causality detection between events expressed in text. arXiv.

Lamla, M. J. and Lein, S. M. (2014). The role of media for consumers' inflation expectation formation. Journal of Economic Behavior & Organization, 106:62–77.

Larsen, V. and Thorsrud, L. A. (2019). Business cycle narratives. CESifo Working Paper.

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. Journal of Monetary Economics, 117:507–520.

Lischka, J. A. (2015). What follows what? relations between economic indicators, economic expectations of the public, and news on the general economy and unemployment in germany, 2002-2011. Journalism & Mass Communication Quarterly, 92(2):374–398.

Lloyd, S. P. (1982). Least squares quantization in pcm. IEEE Transactions on Information Theory, 28:129–137.

Matthes, J. (2014). Framing. Nomos Verlagsgesellschaft mbH & Co. KG.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. Public opinion quarterly, 36(2):176–187.

Meer, D. (2022). Überlegungen zum begriff des narrativs – kommunikation von nachhaltigkeit am beispiel des european green deals. Technical report.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Müller, H., Schmidt, T., Rieger, J., Hufnagel, L. M., and Hornig, N. (2022). A german inflation narrative. how the media frame price dynamics: Results from a rollinglda analysis. DoCMA Working Paper.

Müller, H., von Nordheim, G., Boczek, K., Koppers, L., and Rahnenführer, J. (2018). Der wert der worte–wie digitale methoden helfen, kommunikations- und wirtschaftswissenschaft zu verknüpfen. Publizistik, 63(4):557–582.

Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: exploiting big data for systemic risk assessment. Journal of Economic Dynamics and Control, 127:104119.

Pavllo, D., Piccardi, T., and West, R. (2018). Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

Rieger, J., Lange, K.-R., Flossdorf, J., and Jentsch, C. (2022). Dynamic change detection in topics based on rolling ldas. In Proceedings of the Text2Story'22 Workshop, pages 5–13.

Roos, M. and Reccius, M. (2021). Narratives in economics. Ruhr Economic Papers, 922.

Scheufele, D. A. (1999). Framing as a theory of media effects. Journal of communication, 49(1):103–122.

Scheufele, D. A. and Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. Journal of communication, 57(1):9–20.

Schnabel, I. (2021). New narratives on monetary policy–the spectre of inflation. speech at the 148th.

Shiller, R. J. (2017). Narrative economics. American economic review, 107(4):967–1004.

Ter Ellen, S., Larsen, V. H., and Thorsrud, L. A. (2021). Narrative monetary policy surprises and the media. Journal of Money, Credit and Banking.

Tuckett, D. and Nikolic, M. (2017). The role of conviction and narrative in decision-making under radical uncertainty. Theory & Psychology, 27(4):501–523.

Yahoo! Webscope (2008). Yahoo! answers comprehensive questions and answers.

Appendix

In Table 1 we present the causal indicators that we used to filter for causal texts and determine the direction of a causal effect between two statements. This causal direction can be forward (event A caused event B), backward (event B caused event A) or simply associative (events A and B are associated, but not necessarily the cause of one another).

causal indicator	because	thus	therefore	after
direction	forward	forward	forward	backward
causal indicator	if	hence	consequently	resulting
direction	backward	forward	forward	forward
causal indicator	reason	due	then	since
direction	associative	backward	forward	backward

Table 1: List of causal indicators and the most likely direction of their causal effect. The direction might be forward, backward or associative.

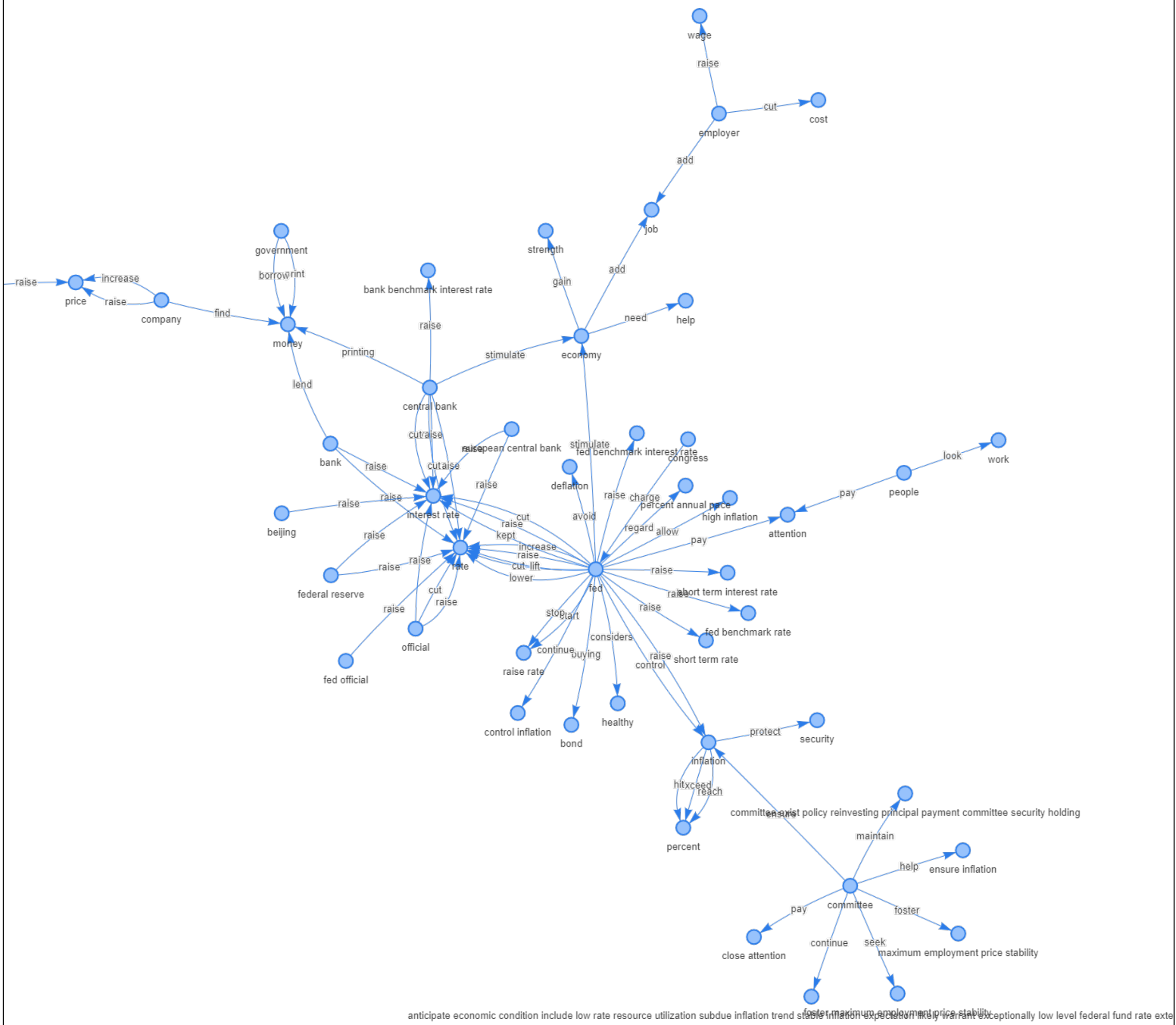


Figure 7: Top 75 statements (48 distinct agents and patients) of the original RELATIO on filtered texts, not including dimension reduction.

Zeitenwenden: Detecting changes in the German political discourse

Kai-Robin Lange and **Jonas Rieger** and **Niklas Benner** and **Carsten Jentsch**
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
{kalange, rieger, benner, jentsch} @statistik.tu-dortmund.de

Abstract

From a monarchy to a democracy, to a dictatorship and back to a democracy – the German political landscape has been constantly changing ever since the first German national state was formed in 1871. After World War II, the Federal Republic of Germany was formed in 1949. Since then every plenary session of the German Bundestag was logged and even has been digitized over the course of the last few years. We analyze these texts using a time series variant of the topic model LDA to investigate which events had a lasting effect on the political discourse and how the political topics changed over time. This allows us to detect changes in word frequency (and thus key discussion points) in political discourse.

1 Introduction

"Wir erleben eine Zeitenwende" – "We are witnessing a turn of eras". This quote by Germany's chancellor Scholz (Deutscher Bundestag, 2022) was the result of a turning point of political discourse in Germany after the outbreak of the Russian-Ukrainian war of 2022. This was an obvious turning point in German politics, as not only the discourse within the parliament, but also the decision making of the German government changed. For instance, 100 billion euros are planned to be spend as additional military expenses (Deutscher Bundestag, 2022). Throughout German history, there have been many changes and turning points in political discourse, but not all of them are as clearly reported or as obvious as this one. Many of them may not have been as clearly remembered because they were gradual rather than rapid changes, or because they did not have an immediate impact on real-world politics.

In this paper, we investigate all 72 years of plenary sessions of the German Bundestag to grasp the development of political discourse in Germany. We analyze these sessions by interpreting them as a time series of textual data for which we use a change detection method proposed by Rieger et al. (2022). For this, we use a rolling version of the topic model latent Dirichlet allocation (LDA), which is designed to construct topics that are coherent over time and allows for changing vocabulary. Within the resulting topics, we detect changes by analyzing the actual word usage in topics compared to theoretically expected word usage if no change occurs, which is determined using resampling. With this method, we are also able to differentiate between short-term changes and persistent ones. While "Zeitenwende" is a broad term, we refer to it as a persistent change in the way of how a political topic is discussed. This change in discussion can stem from differing speech patterns or from changing contents of said topic. The former is rarely detected as a change, as a changing speech pattern usually develops slowly. If it does change suddenly, this is, with minor exceptions, due to a change in formality, which is of no interest for our analysis as it does not affect the content of the discussions. The latter however is interesting for our analysis, as it symbolizes that the topic has changed for good due to economic, cultural or diplomatic events or developments. While this change may not affect the entirety of the political spectrum, we consider it to be a Zeitenwende, if it persistently changes one of the 30 most common and important topics of German political discussions, which we analyze using our topic model.

Data sets as large as the German parliament discussions are too large to be read and interpreted all manually. While qualitative expert analysis is needed to analyze German politics over the last decades, a quantitative text analysis can help in this regard by providing experts with ideas for what to look at and by verifying their results with an empirical basis. Our analysis aims to do just that, as the model used allows for changes of different magnitudes to be detected by simply adjusting a single parameter. Experts can use these findings to back up their qualitative results by pointing out the

This article is accepted for publication in *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)* held in conjunction with the 2022 KONVENS.

importance and long lasting effect of a change for the political discussions at the time or gather ideas for changes in uncommon topics. By tuning the so-called ”mixture”-parameter of this model, even rather niche changes can be found to interpret and to compare them based on the tuning needed to detect them (a high tuning parameter indicates a major change, a smaller one indicates a niche change).

Walter et al. (2021) use a similar data set to analyze ideological shifts throughout German history. This data set, DeuParl (Kirschner et al., 2021), does contain data from plenary sessions since 1867 to 2020. The data before 1949 are however less structured and contain clusters of incoherent text due to being automatically created by scanning old documents, which is why we do not use them in this analysis. While cleaning the data set is a task on its own, political analyses such as this one could greatly benefit from a "clean" version of this Reichstag data set as it enables to analyze German politics for an even larger period of time.

In a complementary approach, Jentsch et al. (2020, 2021) propose a (time-varying) Poisson reduced rank model for party manifestos to extract information on the evolution of party positions and of political debates over time.

2 Change detection via rolling modeling

We make use of a rolling version of the classical LDA (Blei et al., 2003) estimated via Gibbs sampling (Griffiths and Steyvers, 2004). This method is referred to as RollingLDA (Rieger et al., 2021) and allows new data to be added without manipulating the LDA assignments of the previous model. For this, a more reliable version of the classical LDA is used up to a date `init`. Then, according to a user-specified periodicity, minibatches of documents (`chunks`) are modeled using the data available up to that point. Moreover, the model's knowledge of previous documents is constrained in that it only uses the LDA assignments from a given time period (`memory`) to initialize the modeling of the new texts. Newly occurring vocabulary is added to the model vocabulary and subsequently considered as soon as it occurs more than five times in a minibatch. This flexibility enables the model to adapt for mutations of topics in the form of gradual or abrupt changes in word frequencies.

The minibatches are numbered in ascending order starting with the initialization batch: $t = 0,...,T$. Then, using the change detection algorithm by Rieger et al. (2022), we get our set of detected changes over time by

$$C_k = \left\{ t \,|\, \cos\left(n_{k|t}, n_{k|(t-z_k^t):(t-1)}\right) < q_k^t \right\},$$

where $0 < t \leq T$ refers to a specific minibatch and $k \in \{1, \ldots K\}$ to one topic. As proposed by Rieger et al. (2022), $q_k^t \in [0,1]$ denotes the 0.01 quantile of the set of cosine similarities when $n_{k|t}$ is replaced by $\tilde{n}_{k|t}^r, r = 1,...,500$, where $\tilde{n}_{k|t}^r$ denotes a resampled frequency vector under expected change and $n_{k|t}$ the observed vocabulary frequencies for each topic; analogously $n_{k|(t-z_k^t):(t-1)}$ refers to the sum of the count vectors from time points $t - z_k^t$ until $t - 1$. The algorithm has two parameters: the maximum length of the reference period to compare to, $z_k^{\max}$, and the intensity of the expected change under normal conditions p. Using the mixture-parameter $p \in [0,1]$, which can be tuned based on how substantial the detected changes should be, the intensity of the expected change is considered in the determination of this estimator by

$$\tilde{\phi}_k^{(t)} = (1-p)\hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} + p\hat{\phi}_{k,v}^{(t)}.$$

Depending on the choice of this parameter, we are able to gradually alter the magnitude of change needed to be detected by the model. While a large p only displays the most impactful changes, which are likely widely known, a smaller value for p allows for experts on this topic to identify more niche changes.

3 Evaluation

3.1 Data set

To analyze the German political landscape, we use the protocols of plenary sessions of the German Bundestag. These were collected over the course of 72 years, starting from the first plenary session of the Federal Republic of Germany on the 7th of September 1949 until the the 3rd of June 2022 in the 20th legislative period. Each protocol can be downloaded from the website of the German Bundestag (Deutscher Bundestag, 2016) and is provided in an XML-format, which contains, among other things, the date and entire plenary discussion in a text format. As one plenary session might contain multiple topics and points of discussion, we split these texts into smaller texts. Because there are a total of 4345 sessions we aim to split these texts automatically instead of manually and do so by splitting them into individual speeches using regular expressions. We

This article is accepted for publication in *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)* held in conjunction with the 2022 KONVENS.

also deleted the attachments and registers, as well as heckling and comments. This is an ongoing work but already provides better results than splitting the texts any arbitrary number of tokens or using the original plenary sessions as single documents. In total, the 4345 plenary sessions are split into 335 065 documents. The distribution of documents by legislative period is displayed in the appendix in Table 1. The chunks of RollingLDA are adjusted to match the legislative periods, where each period is split into eight chunks (approximately two chunks per year).

3.2 Study design

For this study, we examined the different topic numbers $K = 20, \ldots, 35$ each with $\alpha = \eta = 1/K$. For the RollingLDA we used the first legislative period as the initialization of the model. Starting from this, we modeled semi-annual minibatches, each using the last two years as memory. We applied the change detection algorithm with $p = 0.90, 0.91, \ldots, 0.95$ and $z_{\max} = 4$, i.e., for the detection of changes, a maximum of the previous 4 minibatches ($\sim$2 years $\approx$ `memory`) are taken as the reference period. If a change is detected for topic k at time t, z_k^{t+1} is set to 1, else to $\min\{z_k^t + 1, z_{\max}\}$.

3.3 Results

Upon inspection of the results for the different parameters, we choose to present the findings for $K = 30$ and $p = 0.94$ in detail, yielding an interpretable number of detected changes while providing logical topics which can be analyzed separately from another and consistently over time. The following results serve as a proof of concept, as for a more fine-grained analysis in the future, a lower value of p can be used. This way, the model will detect changes with a lesser impact on the topic, which will enable experts on German politics to identify changes that had an impact on German politics but may not be as well-known as the results we present here. All detected changes, corresponding top words, our interpretations and the results for other parameters can be accessed via the associated GitHub repository (K-RLange/Zeitenwenden).

The changes are displayed in Figure 1. The blue and red curves represent the observed similarities and the simulated quantile similarities, respectively. Each time the blue is below the red line, a change is detected as a gray vertical line.

The topics can be separated into political topics, which contain information about the current political discussions, and formality-topics which contain the names and titles of the parliament members as well as key words for common procedures, such as the voting process when deciding about a bill. While changes can be detected in either type of topics, changes in formality-topics will most likely not contain any information about the current political situation or discussion but rather about common political procedures or who is a current member of the parliament. Topic 9 for instance is a topic that is almost completely consisting of the names of parliament members. All 7 changes are detected at the start of a new legislative period, which is reasonable as new politicians join the parliament, but is not interesting for the sake of our analysis. Similarly, topics 4, 7 and 22 yield multiple changes that can be explained by a change in procedure or a different style of logging. Thus, we focus on the remaining 26 topics when looking for Zeitenwenden that were rooted in the topics of political discussions. We are able to link 22 of our 25 detected changes in relevant topics to interprable events. The remaining changes are caused by events that we were not able to interpret in retrospect.

Our model is able to detect some obvious events which affected political discourse such as the Russian-Ukrainian war (2022, topic 21, Deutscher Bundestag, 2022), the Covid 19 pandemic (2020-21, two changes in topic 17, Organization, 2020), the European financial crisis in 2008 (topic 15, Gode, 2021), the introduction of the Euro as Germany's currency (2002, topic 28, Directorate-General for Communication, 2022), the Kosovo-war (1999, topic 21, Beaumont and Wintour, 1999), the German Reunification (1989-91, topics 1, 12, 15, Schmemann, 1989), the founding of the Bundeswehr (1955, topic 16, Bundeswehr, 2022) and the Saar-referendum (1955, topic 3, Jaeckels, 2020). Events like these had a long lasting impact on German society and politics and could be called "Zeitenwenden". Interpreting the context of change is particularly easy, as the RollingLDA-model provides us with information about the overall topic of the change consistently over time. The Kosovo and Russian-Ukrainian war are for instance both detected in topic 21, which can be interpreted as the "war"-topic. To identify the exact reason for the change, we analyze the impact of each word using leave-one-out word impacts. Such word impact graphs are displayed in Figure 2 and Figure 3 for both mentioned wars. While most changes are caused by words being used more frequently due to a new event (blue

Figure 1: Observed similarity (blue), thresholds q_k^t (red) and detected changes C_k (vertical lines, gray) over the observation period for all topics $k \in \{1,...,30\}$.

bars), some are also caused by words that are used significantly less (red bars). The financial crisis of 2008 is a case in which the change is caused both by a change of focus within an event, as ”ikb” was mentioned far less frequency, while words such as ”krise” (crisis) started to emerge (see Figure 5).

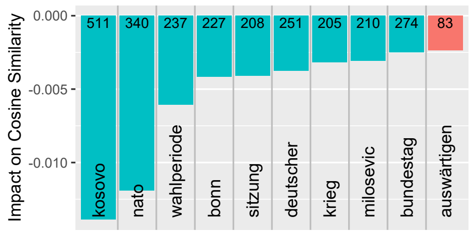


Figure 2: Leave-one-out word impacts for topic 21 (1998-99), caused by the Kosovo war.

While these were major changes which had a lasting impact on Germany, there are several smaller changes that are detected as well, such as the Bonn-Copenhagen declarations in 1955 recognizing the danish minorities in Schleswig-Holstein (topic 21, Federal Foreign Office, 2015), the removal of the statute of limitations on murder (1979, topic 19, Schmid, 2017) and the tax reform of 1998 (topic 26, Tagesschau, 2010).

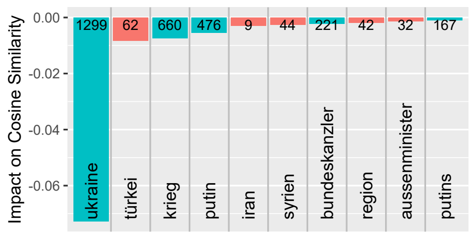


Figure 3: Leave-one-out word impacts for topic 21 (2021-22), caused by the Russian-Ukrainian war.

This article is accepted for publication in *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)* held in conjunction with the 2022 KONVENS.

Our results also enable us to reflect on the relationship between both German states BRD and DDR as well as BRD and western powers over the course of 40 years by interpreting the corresponding changes of major discussion points. West-Germany began major partnerships with western countries, such as the EGKS, a German-French cooperation that was founded according to the ”Schuman-Plan” making the coal and steel-industries of both countries a European rather than a national matter (topic 3, Zandonella, 2021). The Bonn-Paris conventions were also a result of both a closer connection towards western powers such as France and the troubled relationship between West-Germany and the eastern block (including East-Germany), as West-Germany became a member of NATO in 1955 (topic 1, Küsters, 2015) and introduced a mandatory conscription in 1956 (topic 27, Bundeszentrale für politische Bildung, 2016). All of this lead to the second Berlin-crisis in 1958, in which the Soviet Union demanded West-Berlin to become a free city rather than a part of West-Germany (topic 25, Barker, 1963). Still, the NATO was not left unquestioned though, as the piece demonstrations in Bonn in 1981 against the NATO Double-Track Decision were a major discussion point in the Bundestag (topic 15, Der Spiegel, 1981). In 1990 West- and East-Germany unified. This is detected in several topics (1, 12, 15), as it was an long process which had a lasting impact in almost every political sectors, such as financial politics, inner politics, outer politics and many more. In 1991, a distinction was made between the “Neue Bundesländer” and “Alte Bundesländer”, denoting the parts of former East- and West-Germany after the unification. This was important as the parts of former East-Germany needed additional financial help to stabilize and reach the economical level of the western parts (topic 12). Ultimately, the usage of the word “DDR” decreases heavily in 1991 after both states had dissolved (topic 1, see Figure 4).

4 Summary

To identify turning points in German political discourse, we analyzed plenary sessions of the German Bundestag from 1949 to 2021 using a change detection algorithm. This algorithm is based on a rolling version of the topic model LDA to create topics that are comparable across time. The changes detected reflect a significant change in the word distribution of the topics.

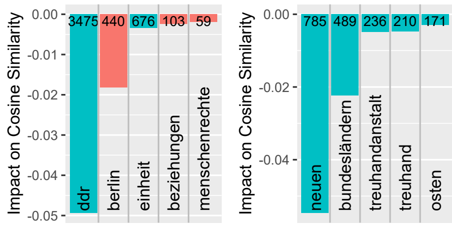


Figure 4: Leave-one-out word impacts for topics 1 (1989-90) and 12 (1990-91) concerning East-Germany.

Our algorithm detects several meaningful changes over the course of the last 68 years of plenary discussions, such as key moments of the relationship between West- and East-Germany as well as world political events like the Russian-Ukrainian war, the Covid 19 pandemic and the financial crisis of 2008.

While these changes are identifiable as “true changes”, we do not know how many changes we missed, as major political discussions in the 21st century such as the refugee crisis in 2014 are not detected. This might be caused by a mixture-parameter that was chosen too restrictively or by the inability of the algorithm used to detect changes in topic distribution (see Figure 6), as it is based on word distribution. Thus, topics that are suddenly a lot more relevant are not detected if the vocabulary used did not change. Identifying both would improve this analysis. Along with adjusting the mixture-parameter, this may enable a detailed analysis of Germany politics for experts on this topic. This can be further amplified by cleaning and using plenary sessions from 1867 to 1945, of East-Germany and of German state parliaments in addition to the Bundestag data set that we used here, as this would enable us to cover a broader spectrum of Germany’s political discourse and history.

Acknowledgments

The present study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Elisabeth Barker. 1963. The berlin crisis 1958–1962. *International Affairs*, 39(1):59–73.

Peter Beaumont and Patrick Wintour. 1999. Kosovo: the untold story. *The Guardian*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bundeswehr. 2022. The establishment of the Bundeswehr.

Bundeszentrale für politische Bildung. 2016. 1956: Das wehrpflichtgesetz tritt in kraft.

Der Spiegel. 1981. Schmeiß die Atomwaffen in die Gracht.

Deutscher Bundestag. 2016. Deutscher Bundestag - Open Data.

Deutscher Bundestag. 2022. Deutscher Bundestag - Bundeskanzler Olaf Scholz: Wir erleben eine Zeitenwende.

Directorate-General for Communication. 2022. History and purpose of the euro.

Federal Foreign Office. 2015. Joint german-danish declaration on the 60th anniversary of the bonn-copenhagen declarations.

Solveig Gode. 2021. Bilanz der finanzkrise in deutschland: Die rolle von georg funke, dem "gesicht der finanzkrise", und der hypo real estate bank.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Ulrich Jaeckels. 2020. Abstimmung 1955! - Ja oder Nein?

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2020. Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142:106813.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2021. Poisson reduced-rank models with an application to political text data. *Biometrika*, 108(2):455–468.

Celina Kirschner, Tobias Walter, Steffen Eger, Goran Glavas, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Deuparl.

Hanns Jürgen Küsters. 2015. Inkrafttreten der Pariser Verträge.

World Health Organization. 2020. Who director-general's opening remarks at the media briefing on covid-19.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2021. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, pages 2337–2347. ACL.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. 2022. Dynamic change detection in topics based on rolling LDAs. In *Proceedings of the Text2Story'22 Workshop*, volume 3117 of *CEUR-WS*, pages 5–13.

Serge Schmemann. 1989. Clamor in the east: Germans' special times; reunification next? *The New York Times*.

Sandra Schmid. 2017. Deutscher Bundestag - Historische Debatten (4): Verjährung von NS-Verbrechen.

Tagesschau. 2010. Jahresrückblick 1999: Sparpaket und Steuerreform.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš;, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–60. IEEE.

Bruno Zandonella. 2021. Schuman-plan. In *pocket europa. EU-Begriffe und Länderdaten*. Bundeszentrale für politische Bildung.

A Additional Material

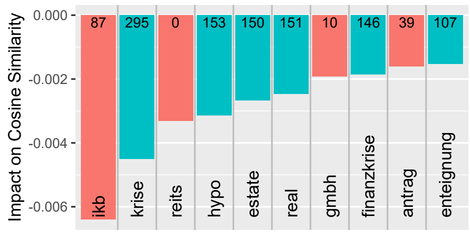


Figure 5: Leave-one-out word impacts for topic 15 (2008-09), caused by the financial crisis.

This article is accepted for publication in *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)* held in conjunction with the 2022 KONVENS.

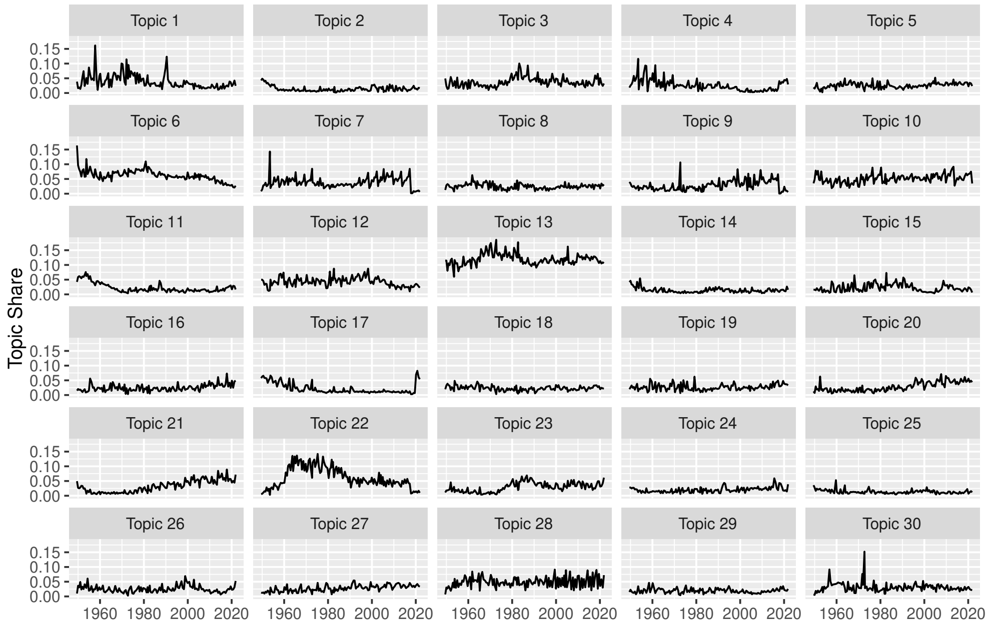


Figure 6: Topic shares per chunk: Relative number of assignments to the specific topic for a given time period.

Table 1: Approximate number of documents in relevant topics (see Section 3.3) and detected changes for each legislative period.

Period	Start date	Documents	Changes
1	1949-09-07	16868	NA
2	1953-10-06	7963	7
3	1957-10-15	7365	3
4	1961-10-17	10900	0
5	1965-10-19	17265	0
6	1969-10-20	16540	0
7	1972-12-13	19917	0
8	1976-12-14	18199	1
9	1980-11-04	10954	2
10	1983-03-29	20812	1
11	1987-02-18	20087	1
12	1990-12-20	19780	3
13	1994-11-10	20999	0
14	1998-10-26	18582	2
15	2002-10-17	13252	1
16	2005-10-18	19867	1
17	2009-10-27	25909	0
18	2013-10-22	20752	0
19	2017-10-24	25163	2
20	2021-10-26	3891	1

This article is accepted for publication in *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)* held in conjunction with the 2022 KONVENS.

SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments

Kai-Robin Lange and **Carsten Jentsch**
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
{kalange, jentsch} @statistik.tu-dortmund.de

Abstract

The application of natural language processing on political texts as well as speeches has become increasingly relevant in political sciences due to the ability to analyze large text corpora which cannot be read by a single person. But such text corpora often lack critical meta information, detailing for instance the party, age or constituency of the speaker, that can be used to provide an analysis tailored to more fine-grained research questions. To enable researchers to answer such questions with quantitative approaches such as natural language processing, we provide the SpeakGer data set, consisting of German parliament debates from all 16 federal states of Germany as well as the German Bundestag from 1947-2023, split into a total of 10,806,105 speeches. This data set includes rich meta data in form of information on both reactions from the audience towards the speech as well as information about the speaker's party, their age, their constituency and their party's political alignment, which enables a deeper analysis. We further provide three exploratory analyses, detailing topic shares of different parties throughout time, a descriptive analysis of the development of the age of an average speaker as well as a sentiment analysis of speeches of different parties with regards to the COVID-19 pandemic.

1 Introduction

In February of 2022, Germany's chancellor Scholz held a speech in the German Bundestag regarding the outbreak of the Russian-Ukrainian war of 2022. It was one of the most prolific speeches in a German parliament in the latest years due to its impact on Germany's foreign and defense policy, as it can be seen as the starting point for an increase in military spending and distancing towards the Russian government. But such decisions and speeches portrait the political stance of the speaker but not necessarily of the entire government or the speaker's party. We propose a data set with parliamentary debates from 16 German federal state parliaments as well as the German Bundestag over the time span of 76 years which is split into individual speeches with meta data to identify the current speaker. This meta data enables the analysis of topics, opinions and speech patterns of different politicians by party, political alignment, age, or constituency. We additionally identified comments from the audience, interrupting the speeches, to enable the analysis of crowd reactions to specific topics or speech patterns. We also labeled the speeches of session chairs: analyses can thus reduce the text corpus to only politically relevant speeches. As our data contains speeches from all 16 federal state parliaments, it can also be used to compare speeches across states to verify regional differences. We will publish the data set upon publication of this paper.

Further, we conduct an exploratory data analysis on the given corpus, using the "party" meta data to analyze party topic shares as well as the sentiment of the 7 Bundestag parties in COVID-19 related speeches. We then use the "age" indicator to analyze the developement of the average speaker age across time.

2 Related Work

In recent years, the interest in researching German political speeches by the means of Natural Language Processing has greatly increased. For instance, Lange et al. (2022a) identify important political change points in the German political discourse using RollingLDA (Rieger et al., 2021, 2022), a time-varying version of the topic model LDA (Blei et al., 2003), on a similar political data set of speeches of the German Bundestag. Another common research topic is the comparison of party positions (Ceron et al., 2022), estimation of political alignment or ideological clarity of German and European political parties by using document scaling techniques. Some follow a classical bag-of-

Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, pages 19–28
September 22, 2023. ©2023 Association for Computational Linguistics

word approach (Jentsch et al., 2020, 2021; Slapin and Proksch, 2008; Proksch and Slapin, 2010; Lo et al., 2016), while others use topic models such as Top2Vec (Angelov, 2020) to scale the available speeches or party manifestos (Diaf and Fritsche, 2022). Such analyses have also been extended to the predecessors of the Federal Republic of Germany, as Walter et al. (2021) analyze political biases throughout the years using Reichstag as well as Bundestag data by using diachronic embeddings. Recent developments have also demonstrated the importance of claims and frames for the analysis of party positions, as Blokker et al. (2022) exemplified using a data set of party manifestos. This aforementioned research does however often focus on the federal political level but disregards politics on the state level and below. And even at state level such analyses can often only differentiate their findings by party by using party manifestos over parliamentary speeches, as the available data sets used do not provide the necessary meta data. Goet (2019) also argues that such meta information is important to, for instance, measure political polarity in a supervised manner. The SpeakGer data set is meant to enable such fine-grained political research by meta-data enrichment.

In recent years, several similar data sets have been released which however lack some properties that are needed for quantitative text analysis of German parliaments. For instance, Open Parliament TV provide an interface for qualitative researchers for speeches in the German Bundestag from 2013 to 2023, split into individual speeches. This data set does however lack the speeches from the federal state parliaments and all Bundestag speeches prior to 2013. The ParlSpeech data set (Rauh and Schwalbach, 2020) provides split speeches of the German Bundestag from 1991 to 2018, but does not include speeches prior to this or from the 16 state parliaments. Still, Rauh and Schwalbach (2020) include information, to which agenda item the current speech refers to, which our corpus does not as of the publication of this paper, due to the different agenda and document structures across the 17 parliaments and the differences in stenographic reporting across 76 years. Abrami et al. (2022) provide a similar data set which also includes parliamentary documents of the German Bundestag and the German federal state parliaments. This data set is also only provided in already pre-processed and part-of-speech-annotated form, while we publish unprocessed data to enable all researchers to apply pre-processing of their liking. We also split our data set by speeches and equip it with meta data about all speakers to enable a more fine-grained political analysis which includes meta-data such as the constituency, the party and the year of birth of all speakers while also allowing users to filter out speeches e.g. by session chairs and comments from the audience. Additionally, our data set contains data of the first 10 legislative periods of the federal state parliament of Berlin and the first 8 legislative periods of the federal state parliament of Baden-Württemberg.

3 Data collection

We primarily recieved our data from the websites of the respective parliaments. However, some parliaments do not publish the documents of all legislative periods on their website, even if they are available. Thus, we collected additional documents from the Parlamentsspiegel-website and looked for additional digitized documents in corresponding local museums. Still, not every legislative period of every German federal state parliament is digitized, as Bremen, Hamburg and Niedersachsen are missing digitized versions of the first legislative periods. However, representatives of all three federal state parliaments assured us that the remaining protocols are planned to be digitized as a part of a retro-digitization project. We therefore aim to update our data set as soon as the missing protocols are available to us. The source of each protocol gathered is detailed in Table 1. To enable a time-dependent analysis, we collected the exact dates for each plenary session of all 17 parliaments and integrated these dates into our meta data. We directly received this information from the respective parliament officials we contacted.

3.1 Text extraction and spelling correction

Out of the available 240 legislative periods, the protocols of a total of 106 periods are either available as text files or pdf-files from which text can be extracted. Some of the remaining documents are scanned pdf files in which each page of the protocol is only displayed as a picture with no possibility for direct text extraction. To extract the text from these documents, we use Google's tesseract (Kay, 2007), a model for Optimal Character Recognition (OCR), with the German language option (and a Fraktur-option for the first legislative period of the state

Table 1: Sources and links to all protocols that were analyzed. If the protocols of a parliament cannot be found in one place, we provide multiple sources for all possible legislative periods.

Parliament (English name)	Legislative period	Source
Baden-Württemberg (Baden-Wuerttemberg)	12-17 1-11	Landtag von Baden-Württemberg Württembergische Landesbibliothek
Bayern (Baveria)	1-18	Bayrischer Landtag
Berlin	12-19 6-11 1-5	Abgeordnetenhaus Berlin Zentral- und Landesbibliothek Berlin Zentral- und Landesbibliothek Berlin
Brandenburg	8-10 1-7	Landtag Brandenburg Parlamentsspiegel
Bremen	18-20 7-17	Bremische Bürgerschaft Parlamentsspiegel
Bundestag	1-20	Deutscher Bundestag
Hamburg	20-22 6-19	Hamburgerische Bürgschaft Parlamentsspiegel
Hessen	1-20	Hessischer Landtag
Mecklenburg-Vorpommern (Mecklenburg-Western Pomerania)	1-8	Landtag Mecklenburg-Vorpommern
Niedersachsen (Lower Saxony)	17-18 8-16	Landtag Niedersachsen Parlamentsspiegel
Nordrhein-Westfalen (North Rine Westfalia)	1-18	Landtag Nordrhein-Westfalen
Rheinland-Pfalz (Rhineland Palatinate)	1-18	Landtag Rheinland-Pfalz
Saarland	14-17 7-13	Landtag des Saarlandes Parlamentsspiegel
Sachsen (Saxony)	1-8	Sächsischer Landtag
Sachsen-Anhalt (Saxony-Anhalt)	6-8 1-5	Landtag von Sachsen-Anhalt Parlamentsspiegel
Schleswig-Holstein	1-20	Schleswig-Holsteiner Landtag
Thüringen (Thuringia)	4-7 1-4	Thüringer Landtag Parlamentsspiegel

Bayern). We improve tesseract's performance by binarizing each page to a pure black-white format using Otsu's threshold (Otsu, 1979) and by correcting a possible skew of each page using OpenCV (Bradski, 2000). We found tesseract to best capture the text of two-column documents in a sample of our data we used as an experiment.

Such an OCR model is however not able to detect a text perfectly, but will, especially for older and less clean fonts, yield "spelling"-errors. That is, despite not literally spelling the word, single letters of a word can be misinterpreted as a different letter, having a similar effect to a misspelled word. The term "Bravo!" is for instance often misclassified as "Bravol" by tesseract. We contemplated using a prediction-based spelling correction, e.g. a masked word prediction based on BERT (Devlin et al., 2019), but due to frequent mistakes in particularly old documents, this context-based prediction yielded sub-optimal results. To correct the errors that are caused by such OCR models, we therefore aim to instead use a lexicon-based approach by using Symspell's (Garbe, 2012) German language dictionary which we additionally provided with the last names of all members of parliament (mps) of all 16 federal state parliaments to stop the spelling correction from affecting our speech-splitting. We detect every word in every OCR scanned document that is not part of this dictionary and determine, whether there is a word in the dictionary that is sufficiently similar to the misspelled word with regards to their Levenshtein-distance (Levenshtein et al., 1966), that is the number of character transformations needed to turn one misspelled token into a correctly spelled token. This distance is chosen dynamically, depending on the word's length. For instance, a word with 7 characters is allowed to have a larger levenshtein-distance to it's "correct spelling" than a word with just two characters. We publish both the spell-checked versions as well as the original processed documents.

3.2 Speech splitting

To identify speeches, we first gathered crucial information about possible speakers by scraping meta data about the first name, last name, year of birth, party, constituency and Wikipedia-links of each speaker, if available. For this, we used the Wikipedia-pages of each federal state parliament, detailing all participating members of parliament during each legislative period. To simplify the interpretation of smaller and regional parties, we also include the political alignment of the parties according to their Wikipedia-pages (e.g. left-populist, social democratic, liberal or conservative). The regular expressions used to identify the start and end of a plenary session as well as splitting the speeches can be found in our GitHub-repository. We will also use said GitHub-repository to detail link and update on the publication of the data set. In the following paragraphs, we describe how they are designed as well as their purpose.

To split the speeches, we first determine, where the plenary session starts and when it ends to cut off the table of contents and a possible appendix to the pdf-file. To account for possible OCR mistakes, we use Regular Expressions to identify either a comment such as "(Beginn: ... Uhr)" marking the start of a session, or, if this cannot be detected, the first appearance of common speech patterns, such as a greeting like "Meine sehr verehrten Damen und Herren". We also incorporate common OCR errors for those phrases in our Regular Expressions, such as misinterpreting an "B" as an "ß". To find the end of the session, we look for either a comment marking the end of the session similar to "(Ende: ... Uhr)" or we end the session when we detect common speech patterns, which are used to close a session like "die Sitzung ist damit geschlossen" or "Ich schließe damit die Sitzung". If none such indicators are found, which usually only happens in old documents with bad quality scans, we heuristically cut the last/first 1000 lines of our document to remove the table of contents and appendix.

After detecting in which part of the document the speeches take place, we split the remaining text into pieces with the use of Regular Expressions and our meta data. All documents have common styles which can be used to identify comments and the start of a speech.

Speeches can be identified by a string search for each line by looking for the last name of said mp, followed by a colon. There are some variations of this, such as including the word "Abgeordneter" or a title before stating the name ("Abgeordneter Dr. Mustermann:"), or the party of the mp ("Mustermann (SPD):"), but the last name of the mp as well as the colon are always present across all analyzed parliaments. Thus, we detect a change in speakers by scanning the lines for the last names of all possible mps in this legislative period paired with a colon. For this we use the names from the mps of the parliament and legislative period that are analyzed, which were scraped from Wikipedia. If we detect the word "Präsident" or get another indication that the speech is held by the chair of the session, we mark it accordingly, as it will likely only cover the organization of the plenary session and rarely contains political statements or arguments.

As a comment, we define additional information provided by the stenographer about the organization of the session (such as information on pauses when the parliament votes on a bill) as well as interjections from the audience during a speech. Such comments can be identified, as they are surrounded by either square or round brackets. Some contain an interjection from a specific member of the parliament, which is detected if the last name of an mp is used in the comment, or about reactions of certain parties, which are detected if said party names are used in the comment. Otherwise, the meta data regarding the speaker is set to "unknown" for such comments. We consider a speech that is interrupted by such a comment to be two separate speeches, before and after the comment, held by the same speaker. This is done to enable the analysis of interactions between comments and speeches such that the effect of a comment on the speech or vice versa can be analyzed.

4 Descriptive Analysis

In total, the SpeakGer data set contains 17,784,802 texts across the 16 German federal state parliaments as well as the German Bundestag, which include a total of 5,510,951 comments, 1,467,746 speeches of session chairs and 10,806,105 speeches of other mps. The total number of documents (in thousands), split into comments, speeches of the session chair and other speeches, separated by parliament are displayed in Table 2.

4.1 Topic shares per party

To determine topic shares per party over time, we use RollingLDA (Rieger et al., 2021), a rolling window approach to topic modeling that creates coherently interpretable topics modeled over time that are allowed to adapt to a changing vocabulary. We thus receive a topic model each year from 1950 to 2022. The years 1947 to 1950 are used to fit the initial model while later years update the model that came beforehand. For this, we consider $K = 30$ topics to give the topic model the opportunity to separate a wide range of political aspects in different topics but still enabling a clear analysis in the scope of this paper. We additionally set the parameters $\alpha = \gamma = \frac{1}{K}$ and the memory-parameter to 4, thus enabling the model to "remember" the previous 4 years to create topics in the current year. We fit our model on the data of all federal state parliaments simultaneously but only use speeches that were not classified as comments to prevent topics simply representing crowd reactions like applause.

The topic shares for each topic over time, separated by party are displayed in Figure 1. For this figure, we used the ggplot-package (Wickham, 2016) for the R programming language (R Core Team, 2022). For better visibility, we limited the plots to topic shares up to 15%, which only has minor implications for most topic. Only the topic share of the Baverian party CSU is off the charts for most of topic 10 and 11, as these cover topics extensively covered in the Baverian parliament. In the figure, the topics' top words over the entire time period are used to title the respective topics graphs. These overall top words most often are not the top words at all times, but still decently represent said topic as a whole. Speeches that are part of documents with particular bad scan quality often contain a lot of misspelled words, which leads to topics that are characterized by commonly misspelled words – this can be seen by observing the top words "dar", "dan", "ale" (which are likely misspelled versions of the words "das" and "alle") of the topics 8 and 15. This filtering aspect of the topic model allows us to focus on the other, relevant topics without the need to account for misspelled words - also due to the properties of RollingLDA, these topics "rotate out" as soon as the OCR errors disappear.

Due to the fact that we perform a topic analysis on documents from all 16 federal state parliaments, several German states have a specific topic designated to them, which can be interpreted as the talk about local affairs. Despite talking about different places, some of these topics overlap, possibly due to similar actions that need to be taken – the city states Berlin, Bremen and Hamburg have a joint topic dealing with city state affairs (Topic 18). We can further inspect the respective topics of these states to gather information about the most important discussion in said parliament at the time. To further analyze the contents of each parliament though, a detailed topic analysis can be performed on only those documents that belong to said parliament. Apart from these topics, which specifically define misspelled words or German states, topics 9, 14, 16, 19, 22, 24, 25, 26, 27 and 28 also cover more general topics that are of interest in every federal state, for instance education, climate change and state-finances. The rest of the topics cover parliament-specific vocabulary like "drucksachen" or "gesetzesentwurf" in topics 11 and 2 respectively.

The topics of political interest confirm several political assumptions to parties that can be made by observing the parties in the Bundestag and considering their party manifestos on a federal level. For instance, we can observe the green party Die Grünen, having the highest topic shares of all parties in topics 16 and 24 covering climate change and agriculture respectively. The party CSU that is only present in the federal state Bavaria, which contains a lot of rural areas, also talks a lot about agriculture while talking the least of all parties about renewable energies and climate change. Conversely, the liberal party FDP, whose party manifestos focus on new technology, have a high topic share in the topic about climate change and renewable energy, while barely talking about agriculture.

For the right-wing party AfD, we observe a high topic share in the topics 19 and 28. Starting from 2020, topic 19 covers the COVID-19 pandemic during which the AfD was very vocal about opposing the lockdowns and other restrictions of the government to prevent the spread of the virus. Topic 28 covers the refugee crisis in Germany starting in 2014, which has been one of the AfD's biggest topics since it was founded in 2014. In 2022, topic 28 transformed about a topic about the Russian-

Table 2: Total number of speeches in thousands in each parliament, divided by party of speaker and whether the speech is a comment or given by the chair of the session. As the party Die Linke is the successor of the parties SED and PDS, we look at the speeches of said parties combined.

Parliament	Chair	Comment	AfD	CDU	CSU	FDP	Grünen	SPD	Linke
Bundestag	378	1168	26	644	144	282	167	605	116
Baden-Württemberg	18	555	14	319	0	106	97	238	0
Bayern	110	312	2	0	263	14	30	115	0
Berlin	38	263	13	144	0	48	49	172	46
Brandenburg	39	75	9	34	0	4	9	83	37
Bremen	37	254	0	91	0	29	33	162	9
Hamburg	69	337	4	120	0	32	7	168	14
Hessen	110	390	7	247	0	74	95	220	24
Mecklenburg-Vorpommern	55	334	22	120	0	15	11	119	113
Niedersachsen	109	226	0	181	0	57	68	152	7
Nordrhein-Westfalen	133	503	8	201	0	73	72	271	5
Rheinland-Pfalz	56	253	7	110	0	29	24	132	0
Saarland	34	122	1	78	0	9	9	76	6
Sachsen	67	129	11	97	0	11	18	35	50
Sachsen-Anhalt	49	106	14	67	0	10	17	32	35
Schleswig-Holstein	87	327	0	157	0	68	29	170	2
Thüringen	62	152	9	69	0	11	8	28	41

Ukrainian war with major parts of the major German parties AfD and Die Linke supporting Russia in the conflict. This is also reflected in our topic models, as both these parties have the highest share of all parties in this topics.

Interestingly, the two biggest parties of Germany, the SPD and CDU barely dominate the shares in any topic. This is likely because these two parties are considered the most centrist parties, that cover a broad range of political topics without extensively focusing on a specific topic.

Overall, the behavior of the major 7 German parties on the state level reflects their behavior on the federal level in the Bundestag. This analysis however only demonstrates this while looking at all federal states combined, to investigate whether this applies only “on average” or in all parliaments, said parliaments need to be evaluated individually.

4.2 Sentiment Analysis

As a further descriptive analysis of our data set, we perform a party-based sentiment analysis across each parliament to see if any party’s speeches are particularly positive or negative in speeches regarding the COVID-19 pandemic. As there is no training data set available, we perform an unsupervised sentiment analysis. For this we use Lex2Sent (Lange et al., 2022b), an unsupervised sentiment analysis tool that uses Doc2Vec (Le and Mikolov, 2014) to enhance a lexicon-based sentiment analysis. This approach allows us to specify, how a positive or negative sentiment can be determined for political speeches compared to regular web documents as it is based on a sentiment lexicon specifically catered for this task. Lex2Sent further improves the classical lexicon-approach by measuring the distance of a document to both the positive and negative half of a lexicon using Doc2Vec, which is trained on resampled documents of the original corpus. This resampling leads to a bagging-effect which boosts the performance of this analysis. To enable a political analysis using Lex2Sent, we use the sentiment dictionary for German political language as a lexicon-base for Lex2Sent (Rauh, 2018).

In Figure 2, we display the average sentiment polarity, calculated by Lex2Sent, for each party in 2020 to 2022. The larger the sentiment polarity, the more positive a speech is estimated, with negative values indicating rather negative speeches. We can see that the average sentiment across all parties is rather negative, which is not surprising given the topic at hand. Terms such as “Pandemie” are generally considered to be negative and the speeches thus generally have a negative undertone. What is more interesting is the comparison of the parties’ sentiment. For instance, speeches of the right-wing

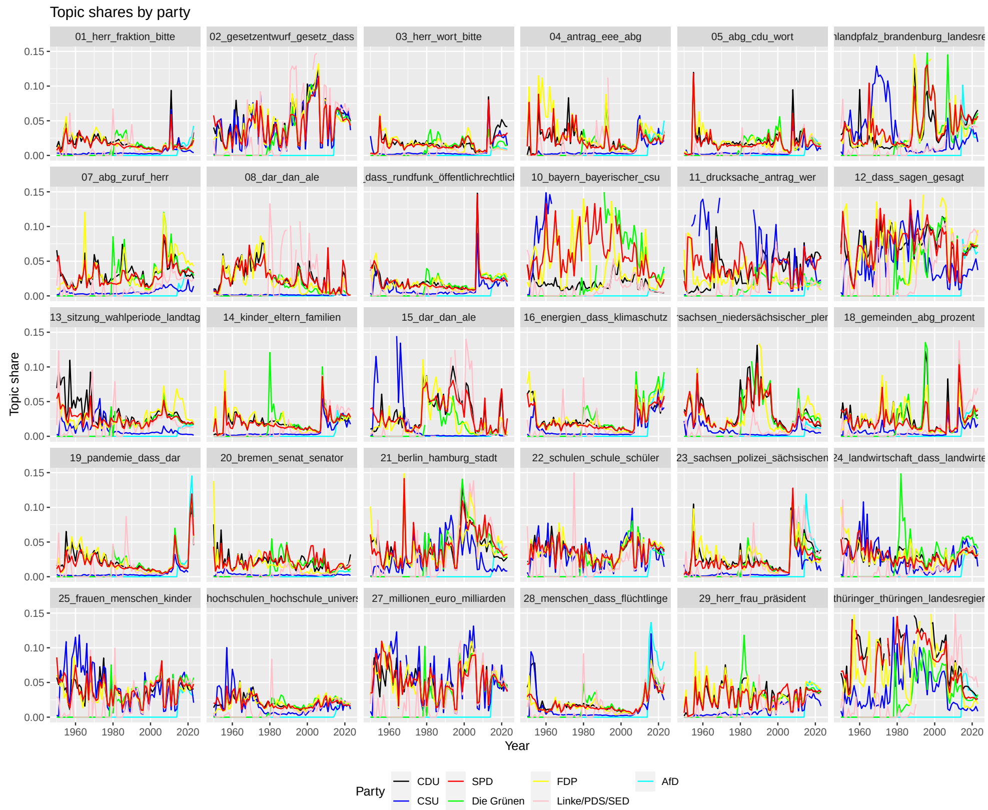


Figure 1: Topic shares of the current 7 Bundestag-parties from 1950 to 2022 in the 16 German federal state parliaments. As the party Die Linke is the successor of the parties SED and PDS, we look at the speeches of said parties combined.

party AfD, which heavily protested the COVID-19 lockdowns and restrictions, are considered to be the most negative in ten of the twelve observed quarters by the model.

In the last two quarters of 2022, the left-wing party Die Linke shows a more negative average sentiment compared to all other parties including the AfD. This is despite them generally delivering positive speeches until this point. One reason for this might be change of party doctrine following Russia's invasion of Ukraine. As mentioned before, major parts of Die Linke are considered to be Russian-favored. The debates resulting from the war outbreak might have thus caused the party to become more confrontational with other parties as a whole. This explanation should however be taken with caution, as the number of speeches concerning COVID-19 has greatly decreased in the last two quarters of 2022 and the observed negative sentiment could this be result of this low sample size.

The Bavarian party CSU also shifted its sentiment over time. As seen in Figure 2, the CSU starts off, having the most positive average sentiment in their speeches concerning COVID-19. During this time, the CSU were party of the government in both the Bundestag and the Bavarian state parliament. During this time the party, and especially their party leader Markus Söder, advocated in favor of hard lockdowns and restrictions. The CSU was thus very in-line with actions taken by the government to handle the pandemic. We see a shift in sentiment starting during the election campaign in the third quarter of 2021, worsening after the elections in 2021. This might be result of the CSU itself not being part of the German federal government anymore and thus not being so compliant with the actions of the government any more.

The same cannot be said for the CSU's sister party CDU however, as the conservative party's sentiment remains rather average across time. The same goes for Die Grüne, the FDP and the SPD.

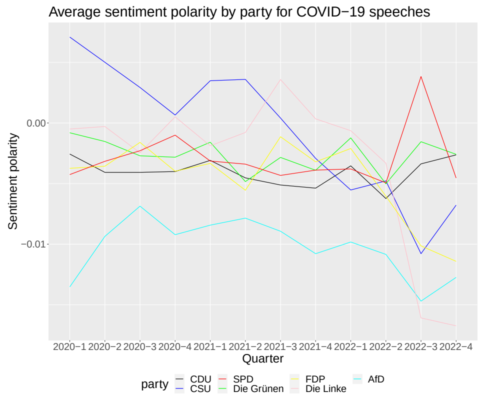


Figure 2: Average sentiment polarites of COVID-19 related speeches of the 7 Bundestag parties in all 17 state and federal parliaments of Germany from 2020 to 2022. The scores were calculated using Lex2Sent, where a negative value indicates a negative speech.

4.3 The age of speakers

In this subsection, we focus on the age of the speakers across Germany. The average age of all registered speakers in the SpeakGer data set from 1947 to 2022 is displayed in Figure 3. We can see that the average age of speakers started to decrease from 54.82 in 1963 to 48.17 in 1973. While the average age remained similar until 1991, the average speaker age started increasing after the German Reunification in 1990. Ultimately, the average speaker age continued to increase, reaching its maximum of 55.38 in 2022. This is partly due to the increasing age of CDU-speakers. While speakers of Germany's largest conservative party averaged at 54.37 years of age in 2018, this increased to an average of 60.04 years in 2021.

5 Summary

We propose the SpeakGer corpus, a comprehensive text data set detailing the long history of German parliamentary debates across 16 federal state parliaments as well as the German Bundestag, split into statements of the session chair, comments and interjections as well as speeches of members of the parliament. Each individual speech is equipped with rich meta data, such as the date of the speech, the party of the speaker and the political alignment of said party, the speaker's age and the speaker's constituency. In total, the SpeakGer data set contains 10,806,105 speeches. This enables researchers to perform fine-grained political analyses of the data set, in which different parties, age-groups and states can be compared. As an exemplary usage of the data set, we performed unsupervised sentiment analysis as well as time-dependent topic modeling to our data and demonstrate how even simple analyses can provide interesting results with the help of meta data. Our results indicate that regional alterations of Bundestag parties often follow the lead of the federal party, despite regional differences, as the sentiment and topics align with the behavior of the parties on a federal level. For instance, the left-wing party Die Linke appears to follow a more confrontational approach to speeches in federal state parliaments after the outbreak of the Russian-Ukrainian war, even in seemingly unrelated topics such as COVID-19 and despite being part of regional state governments themselves. This is however only a preliminary result of our exploratory analysis and should be inspected further.

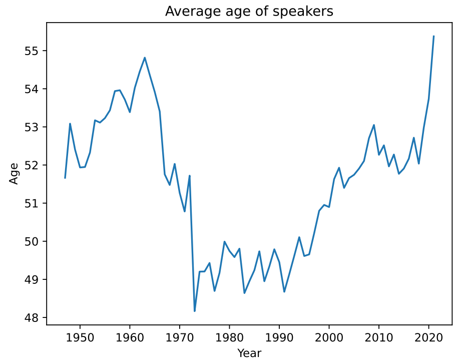


Figure 3: Average age of speakers in German parliaments from 1947 to 2022.

In future research, we aim to, among other possible research ideas, further use the SpeakGer data set proposed to inspect, validate and broaden our preliminary results on the differences between regional and federal versions of the same party. As we only focused on the "party" information in our exploratory research in this paper, in future research, we intend to use the remaining meta data, such as the age of the speaker or the speaker's constituency to perform analyses that take spatial and regional aspects into account.

Ethical considerations

We provide this data set with best intentions to enable researchers to gain a new perspective on German politics. We only use publicly available information to equip our corpus with meta data. We however cannot be certain that the data will not be misused to push political agendas by for instance framing a specific party. We do believe that the benefits of such a publicly available data set outweigh the possible negative aspects, as such malicious framing is commonly done without using a data set of federal state parliament speeches.

Limitations

As a result of sub-optimal document-scans in earlier legislative periods in almost all federal state parliaments, not all speeches and speakers could be correctly identified. In addition to this, old scans of the state Nordrhein-Westfalen contain not just one plenary session but multiple, which also had to be manually split. This session splitting might be sub-optimal due to the poor quality scans. While we contacted all federal state parliaments about the specific dates for all plenary sessions and most states were able to provide a complete list of all correct dates, the states Berlin, Niedersachsen and Schleswig-Holstein could only provide us with an incomplete list. Thanks to publicly available information on Wikipedia, we were able to estimate the dates for the missing plenary sessions of these states, which are however subject to some noise. Lastly, as a result of the meta-based splitting of speeches, we are not able to detect speeches of guests of the parliament, such as Wolodomyr Selensky speaking in the German Bundestag on March 17th 2023, as these guests' names are not part of our meta data containing only information about the mps of the parliament. We aim to improve on these aspects of the data set as soon as better OCR methods and the results of the retro-digitization project of the German federal state parliaments are released.

Acknowledgments

This paper is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Nico Blokker, Tanise Ceron, André Blessing, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, Gabriella Lapesa, and Sebastian Padó. 2022. Why Justifications of Claims Matter for Understanding Party Positions.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Tanise Ceron, Nico Blokker, and Sebastian Padó. 2022. Optimizing text representations to capture (dis)similarity between political parties.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Sami Diaf and Ulrich Fritsche. 2022. TopicShoal: Scaling Partisanship Using Semantic Search. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 167–174, Potsdam, Germany. KONVENS 2022 Organizers.

Wolf Garbe. 2012. SymSpell.

Niels D. Goet. 2019. Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, 27(4):518–539.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2020. Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142:106813.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2021. Poisson reduced-rank models with an application to political text data. *Biometrika*, 108(2):455–468.

Anthony Kay. 2007. Tesseract: An open-source optical character recognition engine.

Kai-Robin Lange, Jonas Rieger, Niklas Benner, and Carsten Jentsch. 2022a. Zeitenwenden: Detecting changes in the German political discourse. pages 47–53.

Kai-Robin Lange, Jonas Rieger, and Carsten Jentsch. 2022b. Lex2Sent: A bagging approach to unsupervised sentiment analysis.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

James Lo, Sven-Oliver Proksch, and Jonathan B. Slapin. 2016. Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos. *British Journal of Political Science*, 46(3):591–610.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Sven-Oliver Proksch and Jonathan B. Slapin. 2010. Position taking in European Parliament speeches. *British Journal of Political Science*, 40(3):587–611.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Christian Rauh. 2018. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Christian Rauh and Jan Schwalbach. 2020. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2021. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, pages 2337–2347. ACL.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. 2022. Dynamic change detection in topics based on rolling LDAs. In *Proceedings of the Text2Story'22 Workshop*, volume 3117 of *CEUR-WS*, pages 5–13.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Lex2Sent: A bagging approach to unsupervised sentiment analysis

Kai-Robin Lange and Jonas Rieger and Carsten Jentsch
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
{kalange, rieger, jentsch} @statistik.tu-dortmund.de

Abstract

Unsupervised text classification, with its most common form being sentiment analysis, used to be performed by counting words in a text that were stored in a lexicon, which assigns each word to one class or as a neutral word. In recent years, these lexicon-based methods fell out of favor and were replaced by computationally demanding fine-tuning techniques for encoder-only models such as BERT and zero-shot classification using decoder-only models such as GPT-4. In this paper, we propose an alternative approach: Lex2Sent, which provides improvement over classic lexicon methods but does not require any GPU or external hardware. To classify texts, we train embedding models to determine the distances between document embeddings and the embeddings of the parts of a suitable lexicon. We employ resampling, which results in a bagging effect, boosting the performance of the classification. We show that our model outperforms lexica and provides a basis for a high performing few-shot fine-tuning approach in the task of binary sentiment analysis.

1 Introduction

Most commonly, text classification is performed in a supervised manner by using a previously labeled data set to train a learning-based model to predict the sentiment of unlabeled documents. When a labeled data set is not available, an unsupervised labeling approach is useful to provide valuable initial information for an active learning approach or to label the texts right away, when a near-perfect classification is not strictly necessary. However, such unsupervised models often require financial backing or a high performing GPU to use on a large data set.

In this paper, we propose Lex2Sent, a model mainly designed for sentiment analysis, that can however be used for any binary text classification problem, where external resources in the form of lexica are available. We will thus define the model for any arbitrary binary classification. Lex2Sent uses text embedding models to estimate the similarity between a document and both halves of a given binary lexicon. These distances are calculated for multiple resampled corpora and are aggregated to achieve a bagging-effect. As Doc2Vec models are usually trained on the CPU, the method demonstrated here can be fully realized in low hardware resource environments that do not have access to a GPU or the financial means to let commercial models such as GPT label thousands of documents. As the Lex2Sent's architecture is not dependent on the language of choice, it can also be used in other languages than English, including low resource languages for which no powerful language models are available. To demonstrate that the results are generalizable, we compare them to the ones of traditional lexicon methods on three data sets with distinct characteristics. To assess the performance to the modern unsupervised classification state of the art, we compare Lex2Sent's results to GPT-3.5 on one data set. We also extend this active learning approach by fine-tuning a RoBERTa model on a sufficient subset of the labels predicted by Lex2Sent. This can be seen as an initial starting point for active learning approach.

The paper is structured as follows. In Section 2, we discuss previous approaches to text classification and research on resampling techniques for texts. Section 3 introduces our classification model by describing the Doc2Vec model, the unsupervised labeling approach and the resampling procedure used. The data sets and lexica used are specified in Section 4. In Section 5, the classification rates of Lex2Sent are compared to lexicon methods and the performance of Chat-GPT. We also show that we can use the results of Lex2Sent for an initial fine-tuning of a pre-trained language model in few-shot setting. In Section 6, we conclude and give an outlook to further research.

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 281–291

2 Related Work

When little to no labeled data is available, usually text classification is performed in one out of three ways. That is, by using either traditional lexicon methods, decoder-only models like GPT or parameter efficient fine-tuning methods to fine-tune pre-trained language models.

Traditionally, researchers used lexica/dictionaries that were meant to substitute the missing supervised label information by external information. For sentiment classification, such lexica contain both a list of positive and negative words, which could simply be counted within a text. Commonly used lexica include VADER (Hutto and Gilbert, 2014), Afinn (Nielsen, 2011), Loughran-McDonald (Loughran and McDonald, 2010), the WKWSCI lexicon (Khoo and Johnkhan, 2018) and the Opinion lexicon (Hu and Liu, 2004). Even within a specific task such as binary sentiment classification, these lexica are often designed for a specific use case. For instance, the Loughran-McDonald lexicon is designed for economic text data, while VADER is designed for social media data. Lange and Jentsch (2023) perform a sentiment analysis of German political speeches and use Lex2Sent with a lexicon base specifically designed for German political text data (Rauh, 2018).
This method is very resource-savvy, but yield worse performance than the other two methods. Nowadays, lexica are usually only used in low hardware resource environments or by researchers of social science disciplines, because they are white-box algorithms that are easy to interpret.

Alternatively, GPT-4 (Brown et al., 2020) or any other large language model (e.g. Llama 2 (Touvron et al., 2023), Mixtral (Jiang et al., 2024) or Jamba (Lieber et al., 2024)) can classify any document in a zero-shot manner due to their language understanding capabilities. Using GPT-4 or GPT-3.5 for large corpora requires financial backing not everyone has access to though and similarly, open source large language models need a GPU with large vram.

Lastly, a pre-trained Transformer model like BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), that was additionally fine-tuned on the task at hand, might help when a GPU is available, that cannot handle a large language model. This however yields the downside of using classification rules that are not based on the texts the model is meant to be used on. Instead, the model might carry a bias from a different subject over to the classification: the sentiment of a text might be based on completely different clues based on whether the text is a political speech or a social media post. This can be avoided by fine-tuning the model oneself, which, in turn, needs labeled data. To reduce the amount of data needed, active learning (Tharwat and Schenck, 2023) is increasingly being used in combination with few-shot learning techniques. Parameter-efficient fine-tuning (PEFT, Mangrulkar et al., 2022) uses adapter methods such as Low Rank Adaption (LoRA, Hu et al., 2021) to fine-tune language models with fewer training parameters than usual, and is thus suited to fine-tune on few-shot examples to achieve adequate results. Pattern exploiting training (PET, Schick and Schütze, 2021) uses the language understanding capabilities of language models to its advantage by "explaining" the task to the model. As Rieger et al. (2024) show, such methods can even be effectively combined into one.

Contrary to these approaches, we propose a fully unsupervised approach that can be used in low hardware resource environments, in which no access to a GPU is available and where there is no financial backing to let commercial models like GPT label thousands of documents. We do this by employing CPU-based embedding algorithms that leverage external information using lexica and are further improved by resampling, resulting in a bagging effect. Improving embedding-based text classification with the help of lexica has been explored by Shin et al. (2017), Mathew et al. (2020a) and Mathew et al. (2020b), but neither analyze a combination of embeddings and lexica for unsupervised analysis.

Xie et al. (2020) use resampling to improve the performance of supervised sentiment models by resampling words with certain probabilities based on their tf-idf-score or by translating the original document into another language and then translating it back to the original language. Similar augmentations can be performed with the nlpaug-package (Ma, 2019). This allows the user to, for instance, use embedding models, be it static models like Word2Vec (Mikolov et al., 2013) or contextual masked language models like BERT (Devlin et al., 2019). These types of data augmentation and resampling are most often used as additional training data for the embedding models and supervised methods. In this paper, instead of resizing the training set, we create multiple different training sets, on which one embedding model is trained each. Aggregating the information from these models into one combined classifier creates a bagging effect, improving the classification rate. Furthermore, we investigate the advantage of using such augmentation and resampling techniques in an unsupervised setting.

The procedures used by Xie et al. (2020) and Ma (2019) do however change the existing vocabulary. They either change the vocabulary by back-translation or resampling the document dependently from other documents due to the tf-idf-scoring or even introduce completely new words that are not part of the corpus at all by changing words based on similiar words in a given embedding space. This might be counter-productive for an unsupervised analysis, as the texts are not used as a training data set, but are supposed to be evaluated themselves. Changing the vocabulary might introduce a bias and hinder the classification performance, as the external information provided to classify the texts is given by the lexica, which are essentially word lists and thus more likely to work accurately to work with unchanged vocabulary. The resampling procedures in this paper are instead based on those employed by Rieger et al. (2020), who used resampling procedures to analyze the statistical uncertainty of the topic modeling method Latent Dirichlet Allocation. We chose those procedures, as they augment or resample the texts independently from another and do not add new words to the vocabulary.

3 Lex2Sent

In this section we propose Lex2Sent, a bagging model for unsupervised sentiment analysis. Lex2Sent is published as a Python package. The code can be found on GitHub[1].

3.1 Lexica

To perform unsupervised text classification, lexica can be used to interpret the words in a text without the need for previously labeled documents of a similar corpus, as they provide external information. This information is provided in the form of key words, which a lexicon assigns to a certain class.

For our analysis, we use binary lexica, that are used to separate words between two disjoint classes A and B. Such a lexicon assigns a value from an interval $[-s, s]$ with $s \in \mathbb{R}^+$ to all words, while assigning the value 0 to all neutral words. It assigns positive values to all words it deems to belong to class A and negative values to all words, it deems to belong to class B. To enable some words to have a larger weight during the classification process, lexica might give words different values. For instance, the word "fantastic" might receive a higher score than the word "good" when using a sentiment analysis lexicon, as it conveys an even stronger positive emotion. We modify such a binary lexicon to consist of two halves, one for each of the two classes. These halves are defined as lists of words in a way that each word that belongs to either class A or class B occurs exactly once in its respective half. Only neutral words are not assigned to a half. This enables the use of lexicon-based text embeddings.

As we use static embeddings, a key word's embedding is not changed, even if it is negated in a document. To incorporate the concept of negations into Lex2Sent, we merge negations with the following word during preprocessing. The term "not bad" is thus changed to "negbad". "negbad" is then added to the opposite lexicon half of the word "bad", so that Lex2Sent can interpret it correctly.

3.2 Lexicon-based text embeddings

Instead of looking only at key words, text embeddings can be used to analyze semantic similarities to other words. This enables us to identify the class of a text using words that are not part of the lexicon.

Text embedding methods create an embedding for each document, which represents the document as a real vector of some fixed dimension q. They are created using the word embeddings of all words in the current document and can be interpreted as an "average" word embedding. We thus interpret the text embedding of a lexicon half as an average embedding of a word of its respective class. Calculating the distance between the embedding of a document in the corpus and the embedding of a lexicon half is used as a measure of how similar a given document is to a theoretical document that is the perfect representation of that class.

As an alternative to the approach mentioned above, we also looked at the average distance of a document's text embedding to all word embeddings of the sentiment words that appear in the document itself, to analyze only its difference to the parts of the lexicon that are part of the document. However, this yielded a classification rate that is comparable to the traditional lexicon classification itself and does not provide substantial improvement over it.

[1] https://github.com/K-RLange/Lex2Sent

It will thus not be further reported in this paper.

The distance is calculated using the cosine distance

$$\mathrm{cosDist}(a,b) = 1 - \frac{\sum_{i=1}^{q} a_i b_i}{\sqrt{\sum_{i=1}^{q} a_i^2} \cdot \sqrt{\sum_{i=1}^{q} b_i^2}}$$

for two vectors $a = (a_1, \ldots, a_q)^T \in \mathbb{R}^q$ and $b = (b_1, \ldots, b_q)^T \in \mathbb{R}^q$ (Li and Han, 2013).

For our purposes of classifying documents into two classes, let A_d be the cosine distance of a text embedding of a document to the text embedding of the positive half of a sentiment lexicon and B_d be the cosine distance to the negative half. Then, the larger (smaller) the value

$$\mathit{diff}_d = B_d - A_d$$

is for a document d, the more confident the lexicon-based text embedding method is, that this document d in fact belongs to class A (B).

This method can be performed using any text embedding model in combination with any lexicon that enables a binary classification task. In this analysis, we choose Doc2Vec (Le and Mikolov, 2014b) as the baseline text embedding model and analyze texts for their sentiment.

3.3 Doc2Vec

Doc2Vec (Le and Mikolov, 2014a) is based on the word embedding model Word2Vec (Mikolov et al., 2013), which assigns similar vectors to semantically similar words by minimizing the distance of a word to the words in its context.

Since word embeddings are not sufficient to classify entire documents, the model is extended to text embeddings. A Doc2Vec model, using the Distributed Memory Model approach, uses a CBOW architecture (Mikolov et al., 2013) in which a document itself is considered a context element of each word in the document. The distance of the document vector to each word vector is minimized in each iteration, resulting in a vector that can be interpreted as a mean of each of its words. According to Le and Mikolov (2014a), these text embeddings outperform the arithmetic mean of word embeddings for classification tasks. In this paper, we use the Doc2Vec implementation of the gensim package in Python (Řehůřek and Sojka, 2010).

Formally, we consider D documents and denote by N_d the number of words in document $d \in \{1, \ldots, D\}$. Further, for $i \in \{1, \ldots, N_d\}$, let $w_{i,d}$ be the i-th word in document d and w_{doc} denote the document under consideration. To give larger weight to words that follow up on another than words that are far away from another, the window size is varied during training. For a Doc2Vec model, we denote by K the maximum size of the context window. For every word the effective size is then sampled from $\{1, \ldots, K\}$ and is denoted as $k_{n,d}$. With these windows, the log-likelihood

$$\sum_{n=K}^{N_d - K} \ln\left(p(w_{n,d} | w_{n-k_{n,d},d}, \ldots, w_{n+k_{n,d},d}, w_{\text{doc}})\right)$$

is maximized for the documents $d = 1, \ldots, D$ using stochastic gradient descent. $p(\cdot|\cdot)$ is calculated by the resulting probabilities from a hierarchical softmax (Mikolov et al., 2013).

We also investigated, if the Lex2Sent method would work when using a pre-trained language model, in this case RoBERTa-large (Liu et al., 2019), as the embedding-backend. For this, we used the CLS-vectors of the lexicon halves and the documents to create lexicon-based text embeddings (similar to Mathew et al. (2020b)). These results underperformed compared to Doc2Vec though, as they showed a bias for one of the two classes.

3.4 Text resampling

Word and text embedding models analyze the original text structure to create similar word embeddings for semantically similar words. We assume that lexicon-based text embeddings need an "optimal text structure" to identify the class of the text in the most efficient manner. Suppose a text contains a key word that is a strong indicatior for the classification task at hand and contained within the lexicon used. The location of such key words can be biased by the type of text. For instance, when analyzing reviews for their sentiment, most key words are located in the last third of the text, as this part draws the conclusion to the review. By resampling the text, we relocate the key words evenly within texts. In theory, this enables vocabulary that occurs more often in texts of a specific sentiment that is not part of any sentiment lexicon, such as topic-specific vocabulary, to be used for labeling texts more efficiently while training Doc2Vec.

We leverage resampling procedures proposed by Rieger et al. (2020), who used them to analyze the uncertainty of the Latent Dirichlet Allocation. Instead of analyzing our methods uncertainty, we use these procedures to create optimal text structures and create a bagging effect. For this, we interpret the original text as a bag of words in which words are drawn independently with replacement like observations when creating a bootstrap sample (Efron, 1979) or independently without replacement, resulting in a permutated text. We call these procedures BW (Bootstrap for Words) and BWP (Bootstrap for Word Permutation), respectively. We analyzed additional procedures, such as resampling sentences as a whole or resampling words only within sentences and variations of those, but these generally yielded lower classification rates than the procedure described above.

3.5 Bagging

In this subsection, we describe a technique to aggregate multiple text embeddings for the purpose of unsupervised sentiment analysis. In combination with resampled texts, this can be seen as a bagging method for unsupervised text classification (Breiman, 1996). Every text structure has an effect on the classification of lexicon-based text embeddings, as differing syntax and vocabulary change the resulting embeddings. However, identifying whether the texts already have an "optimal" structure is a difficult task, as this is an abstract concept that is not trivial to formalize. Instead of relying on the original texts' structure, resampling enables the possibility to create an arbitrary number of artificial texts. If we aggregate these text embedding models, they do not have to label a document correctly for one text structure (that is the original text), but instead only have to label a document correctly on average based on multiple differently structured texts. This aggregation also balances out the randomness of generating samples and the negative effect of missing out on a crucial word within documents in one resampling sample, as it will probably appear in other samples.

The aggregation is performed by calculating an average *diff*-vector using B resampling iterations. Let diff_d^b be the d-th element of the *diff*-vector for the b-th lexicon-based text embedding model with $d = 1, \ldots, D$. Then

$$\mathit{diff}_d^{\text{mean}} := \frac{1}{B} \sum_{b=1}^{B} \mathit{diff}_d^b$$

defines the d-th element of the averaged *diff*-vector.

3.6 Algorithm and Implementation

In training, the algorithm iterates over a grid, calculating models for different training epochs, context window sizes and embedding dimensions. For our application, we use a $3{\times}3{\times}4$-grid, which turns out to be sufficiently beneficial in application while remaining computationally feasible. The parameters are chosen from an equidistant set over reasonable parameter choices (see Algorithm 1 for the parameter choices). The grid can be adjusted according to the practitioner's problem at hand. For instance, a smaller grid is faster to train, but a larger grid will lead to more robust results. In each iteration, the parameter combination for the Doc2Vec model is chosen from the grid and the corpus is resampled. The resampled documents are sorted ascendingly by their respective absolute lexicon score. Then we train a Doc2Vec model and calculate the *diff* vector for all iterations. The classification task is performed by using the component-wise arithmetic mean of all the 36 *diff*-vectors. The algorithm is described as pseudocode in Algorithm 1.

Given a classifier $x = (x_1, \ldots, x_D) \in \mathbb{R}^D$, the document with the index $d \in \{1, \ldots, D\}$ is labeled

$$\text{label}_d = \begin{cases} \text{class A}, & x_d - t < 0 \\ \text{class B}, & x_d - t > 0 \ , \quad t \in \mathbb{R} \\ \text{at random}, & x_d - t = 0 \end{cases}$$

for some threshold $t \in \mathbb{R}$. This might be $t = 0$ or the empirical quantile $t = x_{(p)}$, where p is the estimated proportion of texts of class B based on a-priori knowledge. In the analyses of this paper, we assume to have no a-priori knowledge of the distribution of class labels, so we use $t = 0$.

4 Data sets and lexica

In this section, the two sentiment lexica and three data sets used to evaluate Lex2Sent are described.

4.1 Data sets

The three data sets considered in this paper are chosen to cover texts with distinct features. The iMDb data set consists of a large corpus with long documents and a strong sentiment compared to the other two data sets. The Airline dataset is more than four times smaller and the documents themselves are also shorter. The Amazon data set represents an intermediate case between these two data sets.

The texts are tokenized and stop words as well as punctuation marks and numbers are removed. Lemmatization is performed to generalize words with the same word stem, if the original word from the text does not already appear in the lexicon. The

Algorithm 1 Lex2Sent

```
procedure LEX2SENT(TEXTS, THRESHOLD, LEXICON, RESAMPLING)
    classifier ← [0] ∗ length(texts)
    for (epoch, window, dim) in Grid = ({5, 10, 15}, {5, 10, 15}, {50, 100, 150, 200}) do
        resampled_texts ← resampling(texts)
        sorted_resampled_texts ← sort(resampled_texts, lexicon)
        model ← Doc2Vec(sorted_resampled_texts, epoch, window, dim)
        emb ← lexicon_based_text_embeddings(model, resampled_texts)
        for i in 1:length(emb) do
            classifier[i] + = emb[i]
    return label_by_threshold(non_resampled_texts, classifier, threshold)
```

mentioned methods and stop word list are part of the Python package *nltk* (Bird et al., 2009).

iMDb data set The iMDb data set consists of 50,000 user reviews of movies from the website `iMDb.com`, provided by Stanford University (Maas et al., 2011). These are split into 25,000 training and test documents, each containing 12,500 positive and negative reviews. After preprocessing, each document in the data set is 120.17 words long on average.

Amazon Review data set The Amazon data set is formed from the part of the Amazon Review Data which deals with industrial and scientific products (He and McAuley, 2016). All reviews contain a rating between one and five stars. Reviews with four or five stars are classified as positive and reviews with one or two stars are classified as negative. We removed reviews with a rating of three stars from the data set because the underlying sentiment is neither predominantly negative nor positive. In addition, we filtered out reviews consisting of less than 500 characters. Out of the remaining documents, 52,000 documents are split into 26,000 training and 26,000 test documents, which are formed from 13,000 positive and 13,000 negative documents each. The average length of all documents in the training corpus is 85.51 words after preprocessing.

Airline data set The third data set consists of 11,541 tweets regarding US airlines and was downloaded from Kaggle (Crowdflower, 2015). The tweets are categorized into positive or negative tweets – 3099 neutral tweets are deleted to be able to use the data set for a two-label-case. We split this data set in half into a training and test set. The training set ultimately contains 5570 documents. On average, each document of the training set contains 10.60 words after preprocessing. In comparison to the other two data sets, where the labels are evenly split, in the Airline data set only 1386 and thus 24.02% of the documents are labeled positive.

4.2 Lexica

To demonstrate that the performance is not dependent on the lexicon chosen as a base, we show the performance for three lexica: The Opinion Lexicon (Hu and Liu, 2004) is used to represent as a review-specific sentiment lexicon, while the WKWSCI lexicon (Khoo and Johnkhan, 2018) is chosen as multiple-purpose lexicon. The Loughran-McDonald (Loughran and McDonald, 2010) lexicon was designed for economic texts and not for reviews, hence it represents the case in which a lexicon is used in a sub optimal domain. To make sure that Lex2Sent not only outperforms these two lexica, we also observed the classification rate when using VADER (Hutto and Gilbert, 2014) or Afinn (Nielsen, 2011) lexicon in the traditional way and compare these results to the one of Lex2Sent in Section 5.3.

We added four amplifiers and ten negations to improve the classification. If an amplifier occurs before a key word, its value is doubled and if a negation occurs, it is multiplied by -0.5. For traditional lexicon methods, the classifier is created by summing up the values of all words within a text.

5 Evaluation

The classification rates of Lex2Sent in this section are determined by evaluating 50 executions to observe the method's randomness and to get a metric for the average performance.

Table 1 displays the average classification rates of a WKWSCI-based Lex2Sent and the classification rate of the best performing sentiment lexicon for each data set, split by the classification-

Table 1: Average classification rate in percent of a WKWSCI-based Lex2Sent in comparison to the best lexicon method (in brackets), split into whether the fixed or proportion threshold is used

	WKWSCI-based Lex2Sent		Lexicon with the highest classification rate	
threshold	by proportion	0	by proportion	0
iMDb	80.93	80.01	76.82 (TextBlob)	73.32 (Opinion Lexicon)
Amazon	77.08	76.83	71.91 (VADER)	69.28 (Opinion Lexicon)
Airline	79.11	72.42	82.05 (VADER)	68.33 (Opinion Lexicon)

threshold used. The WKWSCI-lexicon is chosen as a basis for Lex2Sent as it is a multiple-purpose lexicon. Lex2Sent outperforms every of the 6 observed lexica on all three data sets when using the threshold 0, as it would usually be done in an fully unsupervised setting without a-priori knowledge. It also outperforms the lexica in two out of three cases in which the exact proportion of positive to negative documents is assumed to be known. Here it is only outperformed by VADER on the Airline data set, which is likely because this data set consists of short documents which do not give the Doc2Vec models much context to train on per document.

While Lex2Sent outperforms these lexica, it does not outperform Chat-GPT. Laskar et al. (2023) report that GPT-3.5 (`text-davinci-003`) reaches an 91.9% classification rate on the iMDb data set. While it is not known, if GPT-3.5 has seen this data set and its labels during training and it thus might have an unfair advantage by knowing the correct results (Li and Flanigan, 2024), due to its generally high performance on unsupervised classification tasks, we can assume that it will outperform Lex2Sent, at least on most data sets. Lex2Sent does yield the advantage of not requiring financial backing to analyze large data sets though. Only a CPU is needed.

5.1 Different resampling procedures

In this section, we investigate the effect of different resampling procedures on the performance of Lex2Sent. We examine the results of a WKWSCI-based Lex2Sent using either one of the resampling procedures defined in Section 3.4 or no resampling at all for the iMDb data set. Additionally we investigate the classification rate when using texts sorted by their absolute lexicon value (key words grouped at the end of a text). This serves as an ablation analysis to distinguish the effects of resampled, natural and sub optimal text structures (sorted texts).

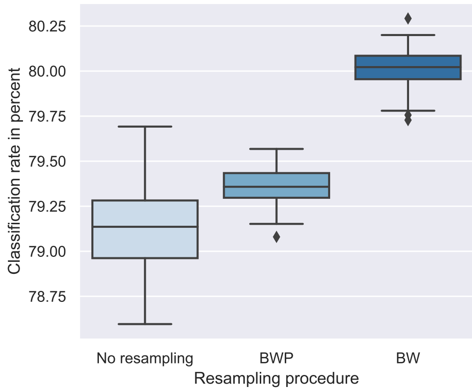


Figure 1: Results of the WKWSCI-based Lex2Sent on the iMDb data set for different resampling procedures

In comparison to the classification rates displayed as boxplots in Figure 1, this subotimal text structure results in a strongly decreased classification rate of 71.00%, which is in line with our interpretation of Section 3.4. The bagging-effect is visible for both procedures, as using either results in higher classification rates for the iMDb data set, with BW yielding the best performance. The method's stability is also increased, as the classification rates are more consistent, which can be seen by comparing the size of the respective box plots. Similar results (not reported) also occur for the other two data sets. For the rest of this paper, all further results are thus reported for Lex2Sent using BW resampling.

5.2 Evaluation on smaller corpora

As Lex2Sent requires training to accurately represent words with embeddings, it is important to determine how large a corpus needs to be for it to provide sufficient results. To analyze this, we evaluate Lex2Sent for subsamples of each data set. These include 10%, 25% or 50% of the original documents. The results of 50 repetitions are displayed in Table 2. The classification rates decrease for smaller corpora except for the Airline data set, in which it

Table 2: Average classification rates in percent of a WKWSCI-based Lex2Sent on subsets of the original data sets for the fixed threshold 0

subsample size	100%	50%	25%	10%
iMDb	80.01	79.73	79.43	78.88
Amazon	76.83	75.71	73.79	68.86
Airline	72.42	72.73	69.74	46.21

Table 3: Average classification rates in percent of Lex2Sent with a WKWSCI-, Loughran McDonald- or Opinion Lexicon-base for the fixed threshold 0, compared to the rates of the traditional lexicon method on the same lexicon

	WKWSCI		Opinion Lexicon		Loughran McDonald	
	Lex2Sent	lexicon	Lex2Sent	lexicon	Lex2Sent	lexicon
iMDb	80.01	70.10	78.43	73.37	70.73	61.22
Amazon	76.83	65.15	77.68	69.28	69.27	61.32
Airline	72.42	63.29	71.96	68.33	72.06	53.18

is slightly higher when examining only 50% of the data set. On the iMDb data set, Lex2Sent outperforms all lexica, even when using just 10% of all documents. On the Airline and Amazon data sets, the classification rate of Lex2Sent decreases to a larger extend for smaller subcorpora. This is likely caused by the short documents in these data set and indicates that it is meaningful to use Lex2Sent on smaller data sets if the documents themselves are long enough to train accurate embeddings.

5.3 Different lexicon-bases for Lex2Sent

So far, we focused on the WKWSCI-based Lex2Sent. In this section, we evaluate, how sensitive Lex2Sent is regarding its lexicon-base and if it improves the classification rate of other lexica as well. For this we compare it to Lex2Sent models based on the Opinion lexicon as well as the Loughran-McDonald lexicon. The average classification rates are displayed in Table 3. Lex2Sent improves the rates of all three lexica on all data sets. While WKWSCI is a general-purpose lexicon, the Opinion Lexicon is designed to analyze customer reviews. This specialization also affects Lex2Sent, as the Opinion Lexicon-based Lex2Sent outperforms every lexicon on every data set as well as the WKWSCI-based Lex2Sent on the Amazon data set, which consists of product reviews. Similarly, we see that Lex2Sent can improve the performance of a lexicon designed for a different domain, as it increases the classficiation rate for the Loughran-McDonald lexicon by at least 7.95 percentage points on all data sets. We recommend to use a general-purpose lexicon like WKWSCI or a lexicon with is domain-adapted to the data set under consideration as a lexicon base for Lex2Sent.

5.4 Lex2Sent as an initial fit

While Lex2Sent is designed for a low hardware resource environment without a GPU, it can still be benefitial to use it in combination with larger, pre-trained models like RoBERTa. To demonstrate this, we use Lex2Sent's beneficial property of displaying a degree of certainty in its results based on how high or low the value of $\mathit{diff}_d^{\text{mean}}$ is for $d = 1, \dots, D$. To create data set for our RoBERTa model to fine-tune on, we therefore only use 10% of the data set: the 5% documents that have the highest and 5% that have the lowest values of our training data set. We fine-tune this version of RoBERTa in 30 epochs using LoRA (Hu et al., 2021) with $r = 8$ and thus 1,838,082 trainable parameters.

To evaluate this approach, we use the iMDb data set, as it contains both a training data set for Lex2Sent to train and RoBERTa to fine-tune on and a test data set for out-of-sample observations that can be classified by RoBERTa. We repeated this procedure five times. On average, our fine-tuned model classified 85.47% of all test documents correctly. While this does not match GPT's classification rate, it does yield the advantage of being cost-efficient. This indicates that Lex2Sent can make for a good initial fit for an active learning approach. Starting from this classification rate, a human-in-the-loop style annotation might take place to improve the classification further.

6 Conclusion

Text classification is commonly performed in a supervised manner using a hand-labeled data set. Unsupervised classification can help when there is no such annotated data set available. This paper proposes the Lex2Sent model, which steers an intermediate course between learning-based and deterministic approaches to create an unsupervised classification, which can be created in a low hardware resource environment without access to a GPU. A binary lexicon is used as a replacement for the missing information that is usually represented by the annotations. The performance of this method is increased by aggregating the results from resampled data sets, which can be seen as a bagging effect.

Lex2Sent yields higher classification rates than all six analyzed sentiment lexica on all three data sets under study, no matter the lexicon-base. Our findings indicate that this might be caused by classifying documents in a more balanced way compared to traditional lexicon methods. Despite being a learning-based approach, the Lex2Sent method shows higher classification rates than traditional lexica on smaller data sets.

Ethical Considerations

While our model requires calculating multiple Doc2Vec models for a single analysis, we modified our model specifications and the number of executions to keep the computational budget manageable in the context of climate change (Strubell et al., 2019). Hence, we perform 50 executions in all of our experiments to ensure that the results are not affected by outliers, but the computational budget remains within reasonable boundaries. Our choice of using the fixed grid with 36 parameter combinations is also caused by this goal. Using this grid, each model finished training in less than two hours.

Limitations

While Lex2Sent improves the classification rate of lexica, it is not capable of reaching the classification rates of models like GPT, but should be seen as a much less resource intensive alternative for the specific task of binary text classification.

Lex2Sent's architecture is independent of the type of binary classification task at hand, so it should work similarly well for other classification tasks given suitable lexica. This is however a theoretical assumption, as we have tested Lex2Sent's capabilities for sentiment analysis specifically.

Lex2Sent has been designed for a two-label-case. To use it in a ordinally scaled multi-label-case, we would need to create multiple thresholds that determines the predicted class, instead of just one. This yields new challenges, as we can not heuristically choose the threshold as 0 like in a binary classification task.

While Lex2Sent's architecture does not depend on the language of the documents or the lexica, it should theoretically perform just as well in low resource languages without needing large training data sets like sophisticated language models. We have not tested this hypothesis though.

Acknowledgments

This paper is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was supported by the Mercator Research Center Ruhr (MERCUR) with project number Pe-2019-0044. In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165*.

Crowdflower. 2015. Twitter US airline sentiment.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW '16: Proceedings of the 25th International Conference on World Wide Web*, pages 507–517.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv:2401.04088*.

Christopher SG Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Kai-Robin Lange and Carsten Jentsch. 2023. SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 19–28.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

Quoc Le and Tomas Mikolov. 2014a. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196.

Quoc V. Le and Tomas Mikolov. 2014b. Distributed representations of sentences and documents. *arXiv:1405.4053*. Version: 2.

Baoli Li and Liping Han. 2013. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, Lecture Notes in Computer Science, pages 611–618.

Changmao Li and Jeffrey Flanigan. 2024. Task Contamination: Language Models May Not Be Few-Shot Anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. 2024. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv:2403.19887*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Tim Loughran and Bill McDonald. 2010. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *SSRN Scholarly Paper*, (ID 1331573).

Edward Ma. 2019. Nlp augmentation. `https://github.com/makcedward/nlpaug`.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020a. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of The Web Conference 2020*, WWW '20, page 1548–1558.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020b. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of The Web Conference 2020*, pages 1548–1558.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98.

Christian Rauh. 2018. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2020. Assessing the uncertainty of the text generating process using topic models. In *ECML PKDD 2020 Workshops*, pages 385–396. Springer International Publishing.

Jonas Rieger, Kostiantyn Yanchenko, Mattes Ruckdeschel, Gerret von Nordheim, Katharina Kleinen-von Königslöw, and Gregor Wiedemann. 2024. Few-shot learning for automated content analysis: Efficient coding of arguments and claims in the debate on arms deliveries to Ukraine. *SCM Studies in Communication and Media*, 13(1):72–100.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. *arXiv:2001.07676*.

Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–158.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Alaa Tharwat and Wolfram Schenck. 2023. A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. *Mathematics*, 11(4):820.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. *arXiv:1904.12848*. Version: 6.

The Journal of Open Source Software

ttta: Tools for Temporal Text Analysis

Kai-Robin Lange[1], **Niklas Benner**[2], **Lars Grönberg**[1], **Aymane Hachcham**[3], **Imene Kolli**[3], **Jonas Rieger**[1], and **Carsten Jentsch**[1]

1 TU Dortmund University **2** RWI - Leibniz Institute for Economic Research **3** University of Zurich

DOI: DOIunavailable

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Reviewers:

- @Pending Reviewers

Submitted: N/A
Published: N/A

Statement of need

Text data is inherently temporal. The meaning of words and phrases changes over time, and the context in which they are used is constantly evolving. This is not just true for social media data, where the language used is rapidly influenced by current events, memes and trends, but also for journalistic, economic or political text data. Most NLP techniques however consider the corpus at hand to be homogenous in regard to time. This is a simplification that can lead to biased results, as the meaning of words and phrases can change over time. For instance, running a classic Latent Dirichlet Allocation (Blei et al., 2003) on a corpus that spans several years is not enough to capture changes in the topics over time, but only portraits an "average" topic distribution over the whole time span.

Researchers have developed a number of tools for analyzing text data over time. However, these tools are often scattered across different packages and libraries, making it difficult for researchers to use them in a consistent and reproducible way.

The `ttta` package is supposed to serve as a collection of tools for analyzing text data over time and can be accessed using its GitHub repository.

Summary

In its current state, the `ttta` package includes diachronic embeddings, dynamic topic modeling, and document scaling. These tools can be used to track changes in language use, identify emerging topics, and explore how the meaning of words and phrases has evolved over time.

Our dynamic topic model approach is based on the model RollingLDA (Rieger et al., 2021), which is a modification of the classic Latent Dirichlet Allocation (Blei et al., 2003), that allows for the estimation of topics over time using a rolling window approach. We additionally implemented the model LDAPrototype (Rieger et al., 2020), serving as a more consistent foundation for RollingLDA than a common LDA. With these models, users can uncover and analyze topics of discussion in temporal data sets and track even rapid changes, which other dynamic topic models struggle with. This ability to track rapid changes in topics is further used in the Topical Changes model put forth by Rieger et al. (2022) and Lange et al. (2022) that identifies change points in the word topic distribution of RollingLDA. Figure 1 visualizes the changes observed by the Topical Changes model in speeches from the German Bundestag (Lange & Jentsch, 2023), which can be analyzed further using leave-one-out word impacts provided by the model or, as Lange et al. (2025) proposed, by asking Large Language Models to interpret the change and relate it to a possible narrative shift.

The Journal of Open Source Software

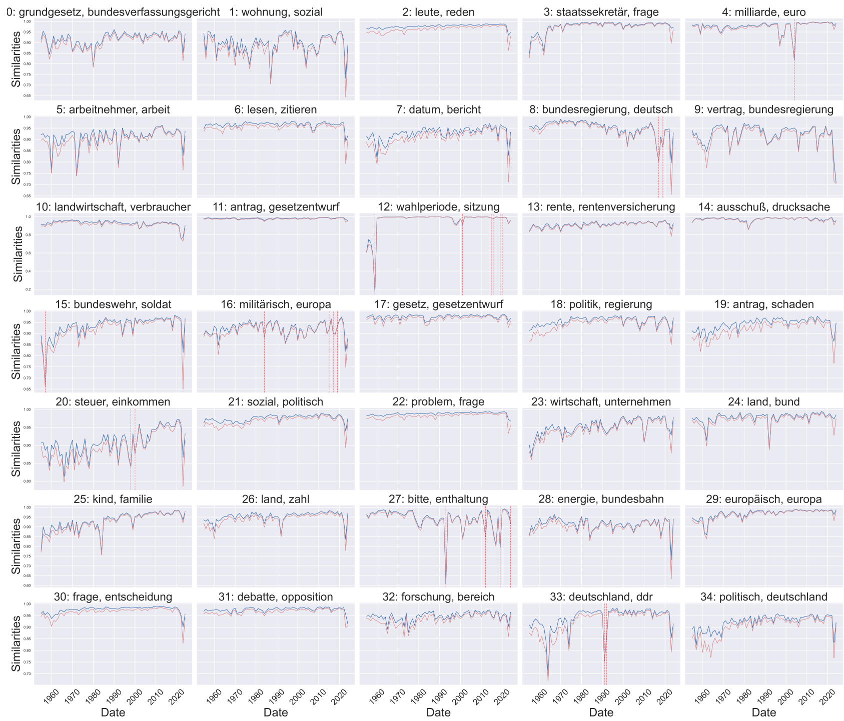


Figure 1: Changes observed by the Topical Changes Model in a corpus of speeches held in the German Bundestag between 1949 and 2023. There is one plot for each topic, with the topic's most defining words over the time frame provided as a title for easier interpretation. Each plot shows the stability of the topic over time (blue line) as well as a threshold calculated with a monitoring procedure (orange line). A change is detected, when the observed stability falls below the threshold, indicated by red vertical lines.

The first diachronic embedding model, originally introduced by Hamilton et al. (2016), builds on the static word embedding model Word2Vec (Mikolov et al., 2013). It enables the estimation of word embeddings over time by aligning Word2Vec vector spaces across different time chunks using a rotation matrix. The second diachronic embedding model is based on the work of Hu et al. (2019), who leveraged BERT's contextual language understanding to associate word usage in a sentence with a specific word sense, thus enabling users to track shifts in word meanings over time.

An example of the evolution of the static diachronic embedding of the word Ukraine in the German Bundestag from 2004 to 2024 is shown in Figure 2. The plot shows the nearest neighbors of Ukraine in the respective embedding spaces, enabling users to observe the trajectory of the target word Ukraine across time chunks in a low-dimensional representation. This visualization highlights the potential of using diachronic embeddings for the interpretation and detection of word context change, as it reflects Ukraine's contextual shift from being closely associated with Russia and China, to aligning more with Europe, and ultimately moving into a different context centered on war.

The Journal of Open Source Software

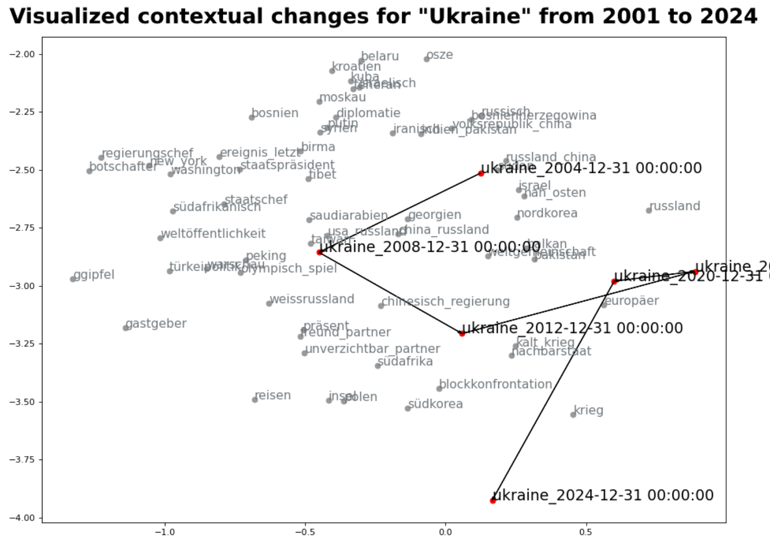


Figure 2: Development of the diachronic embedding of the word "ukraine" from 2004 to 2024 in the German Bundestag. Along with the word itself, its closest neighbors to visualize the target word's track across time. The dimension of the embeddings has been lowered using TSNE.

The Poisson Reduced Rank model (Jentsch et al., 2020, 2021) is a document scaling model, which uses a poisson-distribution based time series analysis to model the word usage of different entities (e.g. parties when analyzing party manifestos). With this model, the user are able to analyze, how the entities move in a latent space that is generated using the word usage counts.

In future work, we plan to add more tools for analyzing temporal text data, as we consider the current state of the package to only be the beginning of development.

Acknowledgements

This paper is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/P16-1141

Hu, R., Li, S., & Liang, S. (2019). Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3899–3908. https://doi.org/10.18653/v1/P19-1379

The Journal of Open Source Software

Jentsch, C., Lee, E. R., & Mammen, E. (2020). Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, *142*. https://doi.org/10.1016/j.csda.2019.106813

Jentsch, C., Lee, E. R., & Mammen, E. (2021). Poisson reduced-rank models with an application to political text data. *Biometrika*, *108*(2). https://doi.org/10.1093/biomet/asaa063

Lange, K.-R., & Jentsch, C. (2023). SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments. *Proceedings of the 3nd Workshop on Computational Linguistics for the Political and Social Sciences*. https://aclanthology.org/2023.cpss-1.3/

Lange, K.-R., Rieger, J., Benner, N., & Jentsch, C. (2022). Zeitenwenden: Detecting changes in the German political discourse. *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis*. https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf

Lange, K.-R., Schmidt, T., Reccius, M., Roos, M., Müller, H., & Jentsch, C. (2025). Narrative shift detection: A hybrid approach of dynamic topic models and large language models. *To appear in Proceedings of the Text2Story'25 Workshop*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). http://arxiv.org/abs/1301.3781

Rieger, J., Jentsch, C., & Rahnenführer, J. (2021). RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. *Findings Proceedings of the 2021 EMNLP-Conference*. https://doi.org/10.18653/v1/2021.findings-emnlp.201

Rieger, J., Lange, K.-R., Flossdorf, J., & Jentsch, C. (2022). Dynamic change detection in topics based on rolling LDAs. *Proceedings of the Text2Story'22 Workshop*. https://ceur-ws.org/Vol-3117/paper1.pdf

Rieger, J., Rahnenführer, J., & Jentsch, C. (2020). Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. *Natural Language Processing and Information Systems*, 118–125. https://doi.org/10.1007/978-3-030-51310-8_11

Narrative Shift Detection: A Hybrid Approach of Dynamic Topic Models and Large Language Models

Kai-Robin Lange[1,*], Tobias Schmidt[2], Matthias Reccius[3], Henrik Müller[2], Michael Roos[3] and Carsten Jentsch[1]

[1]*Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany*
[2]*Institute of Journalism, TU Dortmund University, 44221 Dortmund, Germany*
[3]*Faculty of Management and Economics, Ruhr University Bochum, 44780 Bochum, Germany*

Abstract
With rapidly evolving media narratives, it has become increasingly critical to not just extract narratives from a given corpus but rather investigate, how they develop over time. While popular narrative extraction methods such as Large Language Models do well in capturing typical narrative elements or even the complex structure of a narrative, applying them to an entire corpus comes with obstacles, such as a high financial or computational cost. We propose a combination of the language understanding capabilities of Large Language Models with the large scale applicability of topic models to dynamically model narrative shifts across time using the Narrative Policy Framework. We apply a topic model and a corresponding change point detection method to find changes that concern a specific topic of interest. Using this model, we filter our corpus for documents that are particularly representative of that change and feed them into a Large Language Model that interprets the change that happened in an automated fashion and distinguishes between content and narrative shifts. We employ our pipeline on a corpus of The Wall Street Journal news paper articles from 2009 to 2023. Our findings indicate that a Large Language Model can efficiently extract a narrative shift if one exists at a given point in time, but does not perform as well when having to decide whether a shift in content or a narrative shift took place.

Keywords
change point, narrative, story, Large Language Models, Latent Dirichlet Allocation

1. Introduction

With the rise of populism in western democracies, it has become increasingly important to evaluate narratives in politics, economics and other areas of interest that shape our society. Broadly speaking, narratives are linguistic constructs that boil complex connections between events down to an explainable form. Such narratives do not have to fall in line with facts and have become increasingly important in many areas of our society as they can act as a substitute for fact-based decision making. Consequently, methods to extract and analyze such narratives from large corpora have garnered the interest of many researchers. With both constantly changing and also conflicting narratives spread by entities such as policy makers, news outlets or social media personalities, we consider it especially important to observe, how a narrative develops over time instead of just globally extracting narratives from a corpus that spans over a long time period.

Along with the development and improvement of Natural Language Processing (NLP) methods, narrative extraction methods have also improved. The latest big development in NLP has brought a push towards more reliable human-like narrative extraction: the language understanding capabilities of Large Language Models (LLMs) such as ChatGPT [1], Llama [2], and others show the ability to detect narratives with increasingly complex definitions and thus blur the lines between qualitative and quantitative analyses. Given billions of parameters in the neural networks these models are based on,

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Luca (Italy), 10-April-2025
*Corresponding author.
kalange@statistik.tu-dortmund.de (K. Lange); tobias3.schmidt@tu-dortmund.de (T. Schmidt); Matthias.Reccius@ruhr-uni-bochum.de (M. Reccius); henrik.mueller@tu-dortmund.de (H. Müller); michael.roos@ruhr-uni-bochum.de (M. Roos); jentsch@statistik.tu-dortmund.de (C. Jentsch)
0000-0003-1172-9414 (K. Lange); 0000-0001-7330-4943 (T. Schmidt); 0000-0002-0716-0432 (M. Reccius); 0009-0004-1747-3303 (H. Müller); 0000-0002-5465-9893 (M. Roos); 0000-0001-7824-1697 (C. Jentsch)

they can answer questions about almost any input document and the narratives contained within, but this enormous size also comes with drawbacks. Similarly to how expensive it is to pay experts of a certain field of interest to annotate texts, using LLMs for large scale corpora is not feasible for many researchers due to their computational demand or the financial cost of commercial models. Additionally, their size does not allow users to fully train the models on just a single corpus from scratch. Thus, many methods that use training from scratch to model temporal developments in a corpus, such as many dynamic topic models and diachronic word embedding models [3], cannot be directly transferred to LLMs, leaving fewer options for temporal narrative change detection.

We propose a pipeline that combines the best qualities of both approaches, using dynamic topic models and LLMs. Leveraging the approach of [4], we use a bootstrap-based topical change detection on the topics resulting from the dynamic topic model "RollingLDA" [5]. While this model succeeds in providing us with change points that are based on differences in the word count vectors of individual topics, it does not give us any intricate details about the changes themselves except for the topic of the change, the time chunk in which it happened and some key words that are responsible for the detection of the change. Furthermore, when a change is detected, it is not always certain whether it actually signifies a genuine shift in the narrative or in another dimension of the discourse, such as the factual content or the contextual framework provided. We can however use this method to narrow a large corpus down to a small curated number of documents that are suspected to contain information about some discoursive shift. We do this by filtering the documents within the time period in which the change occurred, given the information we are provided about the change. We then use the language understanding capabilities of an LLM by processing these documents to explain the topical change that occurred and to decide whether the change signifies a narrative shift or not. We additionally provide the LLM with information gathered by the topic model to put the change into context. To accurately guide our identification of shifts in narratives, we use the Narrative Policy Framework (NPF) [6, 7], an analytical approach from political science. Our findings indicate that the LLM performs well when explaining a narrative shift, if one exists, but hallucinates when judging whether a detected change is a narrative shift or not, claiming most content shifts to also be narrative shifts.

We evaluate our model on a corpus of news articles of The Wall Street Journal ranging from 2009 to 2023. As the documents are copyright-protected, we do not use a commercial LLM in the cloud, but rather a local instance of the open source LLM Llama 3.1 8B [2].

2. Related Work

In our research, we focus on media narratives stemming from news articles with a business and finance focus. Thus, we base our definition of narratives on the existing literature on narratives from research in the field of economics and political economics.

2.1. Defining Narratives

The study of narratives has recently gained traction in economic research, though scholars have yet to converge on a single, universally accepted definition of the concept. Shiller [8], as an early and influential example, characterizes economic narratives as "stories that offer interpretations of economic events, or morals, or hints of theories about the economy", thus providing a rather vague definition that leaves ample room for interpretation. A more formal modeling strategy was pioneered by Eliaz and Spiegler [9], who adapt concepts from the literature on Bayesian networks [10]. The authors highlight the causal connections among events and economic variables as sufficient for shaping people's economic and political beliefs. While analytically appealing, this conceptualization sets aside value judgments, a crucial aspect of narratives. We hold that these normative implications ground the impact of political and economic narratives by motivating individuals and groups to act on their beliefs [11, 12, 13].

To empirically capture narratives, including their ideological valence, the Narrative Policy Framework (NPF) offers an alternative lens [6, 7, 14]. The NPF distinguishes narratives by their content and form, systemizing the latter through four elements: a setting, certain characters, a plot and a moral of the story. The setting describes the scenery in which the narrative takes effect, such as a presidential election, a military conflict or a time of high inflation. The characters of a narrative can be persons or organizations, but also other entities take actions in the narrative including even non-sentient entities such as a spreading virus. The plot establishes relationships between the characters in space and time. Lastly, the moral of the story acts as a "takeaway" that often includes implicit or explicit calls to action. With these components, the NPF highlights both the role of ideological charge and the centrality of causal attributions, particularly by emphasizing that identifying "who is to blame for the problem" is an essential part of every narrative [15].

Similarly, Müller et al. [16] propose a narrative definition that is based on six key elements. They build their framework around the theory of media frames [17], proposing that narratives provide dynamic, evolving depictions of events. According to this view, a media narrative comprises one or more media frames (all of which are built around four key elements, see [17]) combined with protagonists (e.g., individuals and institutions) and events arranged chronologically and often presented as causally linked.

For a more comprehensive overview of narrative definitions across disciplines, see the overview papers by Roos and Reccius [11] or Santana et al. [18], among others.

Building upon foundational works in narrative theory, our research focuses on detecting narrative structures and their evolution in large text corpora. To achieve this, we draw on methodologies that have proven effective in identifying thematic shifts or *change points* within textual data. These change points often coincide with shifts in journalistic focus, and we hypothesize that such shifts frequently reflect underlying narrative changes. By leveraging established methods that combine topic modeling with change point detection, we aim to capture these transitions, providing valuable insights into how narratives develop and evolve over time.

2.2. Extracting Narratives from Text

Following the identification of discursive change points, the second critical step in our approach involves the automated analysis of the documents associated with these transitions. To date, there is no widely established (language) model specifically optimized for narrative extraction in this context. Instead, a diverse range of methods has emerged, each attempting to identify narratives or their components through diverse techniques ranging from word-count-based analyses [19] to Large Language Model-based methods [20].

A detailed review of existing NLP-methods to identify systematic parts of narratives is outlined by Santana et al. [18]. The authors focus on identifying key components such as events, participants, and temporal and spatial data, and linking these components to form coherent narratives. The methodology discussed includes part-of-speech tagging, event extraction, semantic role labeling, and entity linking, among others. However, this approach using classical NLP-tools struggles with challenges such as narrative complexity and cross-document narratives.

A similar approach, focusing on political and economic narratives, is proposed by Ash et al. [21]. Their methodology *RELATIO* employs semantic role labeling to identify key narrative components such as agents, actions, and patients within sentences, which culminates in the production of interpretable narrative statements. While their approach effectively identifies simple narrative building blocks, it falls short in detecting and extracting more complex narratives that integrate causality and sense-making. Lange et al. [22] advanced this approach by enhancing the existing RELATIO method with additional pre- and post-processing steps. By combining multiple RELATIO-extracted narrative blocks, the authors were able to link related statements and extract complex narrative structures that better align with causality-based definitions of narratives. At the same time, they emphasize that the increased complexity of their pipeline can amplify error cascades, where even minor changes in a longer input sequence may lead to significantly different results.

A recent study leveraging large language models (LLMs) for narrative extraction is presented by Gueta et al. [20]. The authors explore whether LLMs can effectively capture macroeconomic narratives from social media platforms like X (formerly Twitter). However, the study falls short of providing a robust definition of narratives, focusing instead on sentiment analysis and RELATIO-style statements, leaving key aspects of narrative complexity and causality underexplored.

Collectively, these studies demonstrate the evolving landscape of narrative extraction methodologies, highlighting the integration of advanced NLP techniques to unravel complex narrative structures. By employing a hybrid approach, we seek to address the limitations identified in previous studies, particularly concerning the temporal dynamics of narratives and their computational feasibility. Our proposed pipeline, which combines dynamic topic models and LLMs, aims to provide a more comprehensive understanding of narrative changes over time, thereby contributing to the broader discourse on narrative extraction and analysis.

The choice to utilize LLMs for annotating and categorizing economic narratives stems from their demonstrated ability to excel at complex natural language processing tasks. Modern state-of-the-art LLMs, such as OpenAI's GPT-4 [1] or Anthropic's Claude 3.5 Sonnet, have consistently outperformed traditional NLP models in natural language understanding, classification tasks, and information retrieval. These models enable efficient processing of large corpora, reducing the time needed to annotate thousands of documents from weeks or months to just hours [23, 24, 25].

Unlike traditional NLP tools, which often rely on predefined models like sentiment analysis or topic modeling, LLMs can understand nuanced, contextual relationships in text. Previous approaches to analyzing economic narratives frequently employed machine learning pipelines [21], topic modeling [26], or sentiment analysis [27]. While these methods are effective for identifying broad patterns, as discussed earlier, they lack the depth to identify and categorize predefined, backward-looking narratives, particularly when such narratives involve subtle linguistic cues or complex causal relationships. On the other hand, studies in computational social sciences have shown that LLMs can match or even surpass human coders in annotating political, social, and economic texts [28], underscoring their potential for content analysis. For example, Mellon et al. [29] reported that LLMs achieved 95% agreement with expert annotators when analyzing British election statements. Similarly, Gilardi et al. [30] demonstrated that LLMs like GPT-3.5 could classify tweet content, author stances, and narrative frames more accurately than trained crowd workers. Building on this work, we explore the ability of LLMs to identify narratives and narrative-like structures in text. Detailed prompting and the continuous involvement of humans in the loop ensure that LLM annotations align as closely as possible with human intuition.

3. Methodology

To extract change points in our documents from our corpus, we use the Topical Changes method [4, 31], which is based on the models RollingLDA [5] and LDAPrototype [32], that improve the reliability of the classical Latent Dirichlet Allocation [33] and allows us to apply it to temporal data. After extracting the change points, we further analyze our documents using the LLM Llama 3.1 8B [2]. In this section, we detail the models' functionalities and advantages for the task at hand.

3.1. LDAPrototype

Because the modeling of LDAs is inherently non-deterministic due to its sampling and initialization, there is no way of telling if a single given run does well to represent the corpus or whether it creates "bad" topics by chance. To prevent relying on randomness, we use LDAPrototype.

In terms of language modeling, LDAPrototype follows an LDA [33], but instead of training just one LDA, N LDAs are trained from which one model is chosen as a prototypical LDA. To "represent" the N models, the model with the highest average pairwise similarity to every other model is chosen as the prototype. To do this, Rieger et al. [32] proposed a similarity measure to compare LDAs with. All pair-wise combinations of N LDA models are compared by clustering their topics based on the cosine similarity of the topic's top words into clusters of size two. The topic model similarity between models A_1 and A_2 is then given by

$$\frac{\#\text{topics of model } A_1 \text{ that are matched with a topic of model } A_2}{K}$$

with K as the number of topics in both models A_1 and A_2. This similarity can thus be interpreted as the percentage of topics that is "more similar" to a topic of the different model rather than to a topic of the same model, so a form of topic overlap between two models. The chosen prototype will thus be one of the models with the highest average topic overlap and, conversely, the least "unique" topics that have not been generated by other LDAs and can therefore be considered the most stable out of the N generated models.

3.2. RollingLDA

RollingLDA [5] is a dynamic topic model based on LDA and can, by extension, also be used with LDAPrototype as a back-end. The model uses a rolling window approach that does not train an LDA from all time chunks at once, but rather begins to model the first w time chunks and then proceeds to model the remaining time chunks based on the information and topic assignments of the last m time chunks.

As the training of RollingLDA is thus based on the initially trained model, we use w instead of just one time chunk as a warm-up to ensure the model that initializes RollingLDA is properly trained. This parameter should thus be chosen sufficiently large that a proper topic model that covers most important repeating topics. For instance, choosing $w = 12$ when using monthly chunks, enables the model to initialize on the first 12 months of the data without a temporal component, ensuring its initial model has been trained on observing topics that return in yearly trends.

The memory parameter m creates the rolling window effect of the model. In each time chunk, we provide our model with previous topic assignments from the last m time chunks, which are considered while estimating the document-topic and word-topic distributions during the Gibbs-sampling step. This enables RollingLDA to efficiently model trends in temporal copropra, if the memory parameter is tuned accordingly. For instance, when using monthly chunks, $m = 4$ or $m = 3$ can be chosen to generate topics "remembering" quarterly trends, while "forgetting" older information from years prior. Only providing the model with a bit of past information allows it to change flexibly while still forcing it to keep coherent topics over time. This model is specifically designed to model abrupt changes in rapidly changing news media, which sets it apart from other dynamic topic models, such as the original dynamic topic model [34] and other early iteration of the idea [35, 36, 37].

3.3. Topical Changes

The topical changes model [4, 31] detects change points within topics put out by RollingLDA by comparing the development of word-topic vectors over time. For this, we observe both the current and previous count vectors of word assignments and compare the resulting similarity scores. We then perform a Bootstrap-based monitoring procedure that tests whether a change occurred.

To construct these word count vectors, we count the number of occurrences of each word in that topic over the last z time chunks (or since the last detected change, if it occurred less than z time chunks ago). Thus, z is the maximum number of time chunks the change detection can "look back" to, enabling a rolling window based change detection similar to RollingLDA's rolling window based topic modeling. z can be tuned to, for instance, stop the detection from capturing repeatedly appearing trend effects by setting it as the assumed length of the trend. As the model focuses on finding abrupt changes rather than the slow natural development of language over time, a "mixture" parameter controls, how much language change the user expects from one time chunk to the next, which alters the look-back word topic vectors by mixing them with the current word vector to a certain degree.

The change detection is then performed by sequentially performing a bootstrap-based test with a significance level α for each time chunk in each topic. In this test, the cosine similarities between the word topic distribution in the current time chunks and B Bootstrap samples of the look-back word topic vector are compared. If the cosine distance of the observed look-back word topic vector to the current word topic vector is larger than $B * (1 - \alpha)$ of the bootstrap samples, a change is detected.

This detection does however not give indication about which tokens have actually caused the change.

Rieger et al. [4] propose to use words with high leave-one-out word impacts, that is words for which the cosine distance is reduced the most when leaving the word out of both word-topic vectors that are compared during the detection step. These words can thus be interpreted as the main causes for the detected change, as they had the highest impact on the drop of similarity.

3.4. Llama as a change interpreter

With the topical changes model, we are able to identify not only points in time where a change point in the topics of our corpus occurred, but also which topics are affected by this change. However, as topics are abstract constructs consisting of word distributions, the model can only give us abstract information about what changed. To gather more information, we further analyze our documents using a large language instruction model with great language understanding and summarization abilities. As we are handling copyrighted texts, we will not use a commercial model such as GPT [1], but rather an open source model that runs on a local machine. We use Llama 3.1 8B model, a instruction model designed by Meta with 8 billion parameters [2].

We first narrow down the number of potential documents to give our LLM as an input. Given a change point in topic k between the time chunks t and $t-1$, we employ a filtering strategy that looks for change-related documents to feed into our LLM.

For this, we use the leave-one-out word impacts native to the Topical Changes model as a foundation. We count the number of occurrences of the words found to be significant due to the leave-one-out measure in each document in time chunk t. We then select those 5 documents with the highest count of these words and feed them into our model, letting it compare them to the topic of the previous time chunk, judging from 10 top words. We do not include documents from time chunk $t-1$ here, as preliminary experiments showed that this "confused" the model. Leveraging from the Narrative Policy Framework, we give our LLM the following instruction, alongside the 5 chosen documents:

> ## You are an expert journalist. You will be asked to explain, why a topical change in a corpus of news articles has has been found and what the change consists of. To fulfill this task, you will be provided information from other text analysis models such as parts of the output of a RollingLDA topic model.
> ## Whenever you are asked to analyze a "narrative", assume the definition of a narrative that is laid out in the paper "The Narrative Policy Framework: A Traveler's Guide to Policy Stories". Specifically, respect and apply the following definitory aspects of a narrative: "The NPF posits that while the content of narratives may vary across contexts, structural elements are generalizable. For example, the content of a story about fracking told by a Scottish environmentalist is certainly different from the story told by a right-wing populist who attacks a public agency in Switzerland. However, these stories share common structural elements: They take place in a setting, contain characters, have a plot, and often champion a moral." Keep in mind that a moral must feature a value judgement. When asked to specifiy a moral of a narratives, you must refer to this value judgement or note that there is no moral and thus no narrative! A narrative change must satisfy the four structural criteria, while a content change can simply be caused by an event that shifts the focus of the topic without a clear narrative. Your goal is to determine if a narrative change occurred or if it was a mere content change.
> ## Please explain an apparent change within a RollingLDA topic that has occurred in [date]
> ## The following topic top words might give you an idea of what the topic was about before the change: [10 top words of the topic in chunk $t-1$]
> ## The following topic top words might give you an idea of what the topic was about after the change: [10 top words of the topic in chunk t]
> ## The following words were found to be significant to the detected change: [leave-one-out word impacts]
> ## The following are those articles from the period that make the most use of the words

> found to be significant to the detected change: [Filtered articles]
> ## Provide your output in a strict JSON format. First, summarize each article in one sentence: {“summaries”: [{“article_1”: ...}, {“article_2”: ...}, ...]}. Then formulate what the topic was about before and after the change based on the topic top words, emphasizing the changes induced to the topic, judged by the articles and the change words: “topic_change”: ... Explain how this change in topic indicates a shift in narrative. How did the narrative shift? “narrative_before”: “Before the change, the narrative centered around ...”, “narrative_after”: “After the change, the narrative centers around ...”. Finally, walk through the four structural criteria that true narratives must satisfy according to the Narrative Policy Framework and confirm or disconfirm their existence in the narrative after the break by briefly naming what they are in the texts provided {“narrative_criteria”: [{“setting”: ...}, {“characters”: ...}, {“plot”: ...}, {“moral”: ...}]}. Make sure to specify the exact source of the moral judgement that you may have found. Lastly, make a final judgement if there is a narrative shift to be found with {“true narrative”: True/False}. Do not answer in anything but JSON.

This filtering strategy enables us to specifically observe documents that are found by the model to be significant to the change. It does, however, not capture a larger picture of the topic itself and might lead the model to focus too much on a few significant words. This can happen if more than a few words are significant to the change with similar intensity, not just those that were captured by the leave-one-out word impacts. This might indicate an abrupt and broad topic change that shook the entire word topic distribution.

We also tested out different filtering strategies, such as providing the LLM with documents that are particularly representative of the topic k (i.e. documents with the highest topic share of k) in both time frames $t-1$ and t, but they generally yielded worse results. We aim to further improve our prompting to further optimize our pipeline in the future.

4. Evaluation

We use Python 3.9 to run our scripts. We will publish the code corresponding to this paper as a part of the Tools for Temporal Text Analysis (ttta) Python package[38].

We evaluate our model on a The Wall Street Journal data set containing 795,800 articles dating from 01/01/2009 to 12/31/2023. This high count of documents allows us to use small time chunks for our RollingLDA analysis. We choose monthly time chunks to enable a fine-grained analysis while ensuring that the number of detected changes remains manageable for annotators. We also choose our memory to check the last four months $m = 4$, thus enabling RollingLDA to remember quarterly trends in the data., To provide a stable initial LDA model and to incorporate all yearly trends into that initial model, we choose a warm-up period of $w = 12$ months. As a trade-off between computational efficiency and model reliability, we conduct 10 LDA runs to determine an LDAPrototype. We generated the initial model multiple times with $\{20, 30, ..., 100\}$ topics and decided to choose $K = 50$ topics. Our Topical Changes model is then performed with a look-back-window of 4 months (also to remember for quarterly trends). We use a mixture parameter of 95% and evaluate the similarities to a significance level of $\alpha = 0.01$ to control the severity of changes that are observed to be major changes. We choose $B = 500$ Bootstrap samples to generate the bootstrap percentiles. For our LLM we set the temperature parameter to 0 in the hopes of minimizing hallucinations.

Our Topical Changes model found a total of 68 changes across 156 time chunks. In Figure 1 each topic is represented with its overall top words across the entire corpus in its title for interpretability. Each line plot includes a blue line, representing the word vector similarity, going from one chunk to the next while the orange line signalizes the dynamic threshold calculated using bootstrap samples. A change is detected when the blue line falls below the orange line, resulting in a vertical red line. At these change points, we filter the documents for candidate documents to feed into our LLM to check for a narrative shift.

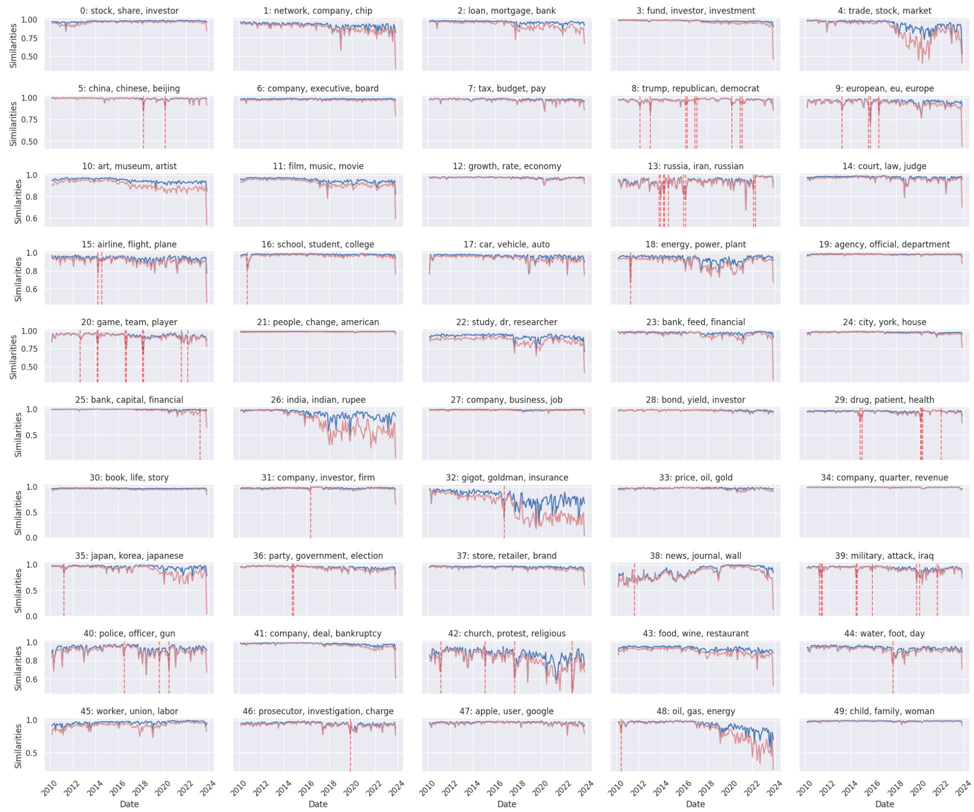


Figure 1: Results of the topical changes model with 50 topics on The Wall Street Journal corpus from 2010 to 2023. Vertical red lines indicate a detected change in the topic at that time chunk.

Three expert annotators discussed each change and annotated them according to the Narrative Policy Framework and afterwards classified the answers of the Large Language Model. They found 37 of them to contain narrative rather than mere content shifts. To exemplify Llama's performance, we show three samples of detected changes as well as their narrative shift evaluation of the LLM in the appendix. We will release the list of annotated narratives upon publication of this paper in a GitHub repository.

We split our evaluation into two parts: We see the binary classification between a narrative shift and a content shift as a first step. In the second step, we check if Llama correctly categorized aspects of the existing narratives according to the NPF. The results show that the LLM does not perform particularly well in classifying the changes into content shifts and narrative shifts, as it finds a narrative in 60, so all but eight cases. This results in a accuracy score of 57.35%, and an f1 score of 0.7010. Since the task involves Llama reproducing all 4 aspects of narratives laid out by NPF, this misclassification is likely caused by hallucinatory behavior, which is a well-known tendency of LLMs [39]. The LLM hallucinates trying to give a satisfactory output – ie. provide the user with aspects of narratives according to the NPF – at all times, resulting in a large false positive rate.

The model does, however, perform better when explaining an existing narrative shift. In those cases the model accurately defines the narrative in 31 out of 37, so 83,78%, of the cases.

Overall, the LLM performs well when a narrative shift exists for a given change point, accurately applying the NPF definition to capture the narrative. However, it does not perform as well when a mere content change occurs, stretching the definitional aspects of the NPF, thus hallucinating narratives. An improved prompt or an additional filtering step could help to solve this issue in future research.

5. Summary

When interpreting political or economic events, people align the corresponding information with their internal world view, combining the two into a narrative. Media narratives have become a big research topic in recent years due to the rise of spreaders of “simple”, often populist narratives. While recent advances in Natural Language Processing, namely the emergence of LLMs, have resulted in improvements in the task of narrative extraction from texts due to their language understanding capabilities, these models are resource intensive. Thus, using them to label narrative in large corpora is often not always feasible.

We introduce a novel pipeline that combines the scalability of the dynamic topic model RollingLDA and its extension Topical Changes with the language understanding of the LLM Llama 3.1 8B. The topic model is used to detect changes within a corpus of The Wall Street Journal dating from 2009 to 2023 in 50 topics. We use two filtering strategies to identify documents that contain information about the nature of the detected change or the before-and-after of the topic. These texts are then fed to Llama to analyze the change according to the Narrative Policy Framework and detect whether a mere shift in content or in the narrative took place.

We processed the articles in monthly time chunks and detected 68 changes within 13 years, as one year served as a warm up period. After manually labeling the changes, we find that 37 of those 68 changes show signs of narrative shifts. While our LLM managed to distinguish content shifts from narrative shifts only 57.35% of the time due to hallucinatory behavior, it correctly explained a narrative shift, if one exists, 83.78% of the times. The incorrect narrative shift detections stem from the model showing a preference to report a narrative shift for each input rather than considering the option of a change that is not caused by a narrative shift. While this result might improve with more careful prompting, it indicates that the language understanding capabilities of an LLM are functional enough to properly fit an existing narrative shift into a given complex definition, yet may fail not because of insufficient language understanding, but rather due to an inclination to oversatisfy the user’s prompt.

While we observed narratives as purely binary cases in this paper, we plan to perform a more nuanced evaluation of narrative extraction techniques in the future, considering a wider array of subjectivity at this complex task.

Limitations

While we demonstrate an exemplary usage of our pipeline here, parameter choices such as the size of time chunks for RollingLDA and the mixture parameter for the Topical Changes can impact the number and granularity of changes detected, which could also alter the results we observed in our study. Furthermore, we didn’t observe cases in which the Topical Changes model incorrectly detected a change, we therefore cannot tell, how the LLM would react to such a case.

Acknowledgments

This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University. The work was also part of the Narrative Economic Alliance Ruhr (NEAR) project, supported by the Mercator Research Center Ruhr (MERCUR) with project number Ko-2022-0015.

References

[1] OpenAI, J. Achiam, S. Adler, S. Agarwal, ..., B. Zoph, GPT-4 Technical Report, 2024. ArXiv:2303.08774 [cs].

[2] A. Dubey, A. Jauhri, A. Pandey, ..., Z. Zhao, The Llama 3 Herd of Models, 2024.

[3] A. Kutuzov, L. Øvrelid, T. Szymanski, E. Velldal, Diachronic word embeddings and semantic shifts: a survey, in: Proceedings of the 27th International Conference on Computational Linguistics,

Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1384–1397. URL: https://aclanthology.org/C18-1117.
[4] J. Rieger, K.-R. Lange, J. Flossdorf, C. Jentsch, Dynamic change detection in topics based on rolling LDAs, Proceedings of the Text2Story'22 Workshop CEUR-WS 3117 (2022) 5–13.
[5] J. Rieger, C. Jentsch, J. Rahnenführer, RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2337–2347. doi:10.18653/v1/2021.findings-emnlp.201.
[6] M. D. Jones, M. K. McBeth, A narrative policy framework: Clear enough to be wrong?, Policy studies journal 38 (2010) 329–353.
[7] E. A. Shanahan, M. D. Jones, M. K. McBeth, C. M. Radaelli, The narrative policy framework, in: Theories of the policy process, Routledge, 2018, pp. 173–213.
[8] R. J. Shiller, Narrative economics, American economic review 107 (2017) 967–1004.
[9] K. Eliaz, R. Spiegler, A model of competing narratives, American Economic Review 110 (2020) 3786–3816.
[10] J. Pearl, Causality, Cambridge University Press, 2009. URL: https://www.ebook.de/de/product/20478918/judea_pearl_causality.html.
[11] M. Roos, M. Reccius, Narratives in economics, Journal of Economic Surveys 38 (2024) 303–341.
[12] S. R. Shenhav, Political narratives and political reality, International Political Science Review 27 (2006) 245–262. doi:10.1177/0192512106064474.
[13] K. Eliaz, S. Galperti, R. Spiegler, False narratives and political mobilization (2024). doi:10.1093/jeea/jvae047.
[14] C. Schlaufer, J. Kuenzler, M. D. Jones, E. A. Shanahan, The narrative policy framework: a traveler's guide to policy stories, Politische Vierteljahresschrift 63 (2022) 249–273.
[15] D. Crow, M. Jones, Narratives as tools for influencing policy change, Policy & Politics 46 (2018) 217–234.
[16] H. Müller, G. von Nordheim, K. Boczek, L. Koppers, J. Rahnenführer, Der wert der worte–wie digitale methoden helfen, kommunikations-und wirtschaftswissenschaft zu verknüpfen, Publizistik 63 (2018) 557–582.
[17] R. M. Entman, Framing: Toward clarification of a fractured paradigm, Journal of communication 43 (1993) 51–58.
[18] B. Santana, R. Campos, E. Amorim, A. Jorge, P. Silvano, S. Nunes, A survey on narrative extraction from textual data, Artificial Intelligence Review 56 (2023) 8393–8435.
[19] N. Benner, K.-R. Lange, C. Jentsch, Named Entity Narratives, Working Paper 962, Ruhr Economic Papers, 2022. doi:10.4419/96973126.
[20] A. Gueta, A. Feder, Z. Gekhman, A. Goldstein, R. Reichart, Can llms learn macroeconomic narratives from social media?, arXiv preprint arXiv:2406.12109 (2024).
[21] E. Ash, G. Gauthier, P. Widmer, Relatio: Text semantics capture political and economic narratives, Political Analysis 32 (2024) 115–132.
[22] K.-R. Lange, M. Reccius, T. Schmidt, Towards Extracting Collective Economic Narratives from Texts, Ruhr Economic Papers (2022). doi:https://doi.org/10.4419/96973127.
[23] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology (2023).
[24] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, arXiv preprint 2303.18223 (2023).
[25] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv preprint 2303.12712 (2023).
[26] A. Macaulay, W. Song, News media, inflation, and sentiment, in: AEA Papers and Proceedings, volume 113, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2023,

pp. 172–176.
[27] S. Tilly, G. Livan, Predicting market inflation expectations with news topics and sentiment, 2021. ArXiv:2107.07155 [cs].
[28] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science?, Computational Linguistics 50 (2024) 237–291.
[29] J. Mellon, J. Bailey, R. Scott, J. Breckwoldt, M. Miori, P. Schmedeman, Do ais know what the most important issue is? using language models to code open-text social survey responses at scale, Research & Politics 11 (2024) 20531680241231468.
[30] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023). doi:10.1073/pnas.2305016120.
[31] K.-R. Lange, J. Rieger, N. Benner, C. Jentsch, Zeitenwenden: Detecting changes in the German political discourse, Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis@KONVENS2022 (2022). URL: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf.
[32] J. Rieger, C. Jentsch, J. Rahnenführer, LDAPrototype: A model selection algorithm to improve reliability of latent Dirichlet allocation, PeerJ Computer Science 10 (2024) e2279. doi:10.7717/peerj-cs.2279.
[33] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[34] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 113–120. doi:10.1145/1143844.1143859.
[35] X. Song, C.-Y. Lin, B. L. Tseng, M.-T. Sun, Modeling and predicting personal information dissemination behavior, in: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, KDD '05, Association for Computing Machinery, New York, NY, USA, 2005, pp. 479–488. doi:10.1145/1081870.1081925.
[36] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 424–433. doi:10.1145/1150402.1150450.
[37] C. Wang, D. Blei, D. Heckerman, Continuous time dynamic topic models, in: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI'08, AUAI Press, Arlington, Virginia, USA, 2008, pp. 579–586.
[38] K.-R. Lange, N. Benner, L. Grönberg, A. Hachcham, I. Kolli, J. Rieger, C. Jentsch, ttta: Tools for temporal text analysis, 2025. URL: https://arxiv.org/abs/2503.02625. arXiv:2503.02625.
[39] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

Appendix

Sample 1

Topic: 29 (healthcare)
Date: 10/2014
Leave-one-out word impacts: ebola, drug, africa, worker, dallas
Narrative shift?: Yes
Hand-coded narrative shift: After a small ebola outbreak in the U.S. in the month prior and an appropriate response that condemned that outbreak, the focus of the topic shifts. Instead of the fear of an outbreak in the U.S., the articles focus on drug tests that give hope and also shift the focus back on the African continent.
Response of the LLM:
{

summaries: [
{article_1: “Pharmaceutical companies and health officials are racing to develop and test treatments for the Ebola virus, with several experimental drugs showing promise in early trials.”},
{article_2: “The US response to the Ebola outbreak has improved significantly since the first case in Dallas, with officials learning from past mistakes and taking steps to contain the virus.”},
{article_3: “The number of people potentially exposed to Ebola in the US has dropped to 50 from 100, and officials are working to reassure the public that the virus can be contained.”},
{article_4: “A nurse in Texas has tested positive for Ebola, highlighting the need for hospitals to be more vigilant in their infection-control procedures and raising concerns about the preparedness of US hospitals for the disease.”},
{article_5: “The Ebola outbreak in West Africa is being hindered by a lack of qualified staff, with many health workers in Liberia and other affected countries lacking proper training and equipment to combat the virus.”}],
topic_change: “The topic shifted from a focus on the global response to the Ebola outbreak, including the development of treatments and the US response, to a focus on the challenges of containing the outbreak in West Africa, particularly the lack of qualified staff and the need for improved infection-control procedures.”,
narrative_before: “Before the change, the narrative centered around the global response to the Ebola outbreak, with a focus on the development of treatments and the US response, emphasizing the sense of urgency and the need for action.”,
narrative_after: “After the change, the narrative centers around the challenges of containing the outbreak in West Africa, highlighting the difficulties of providing adequate care and the need for improved infection-control procedures, with a focus on the human cost of the outbreak and the need for more effective solutions.”,
narrative_criteria: [
{**setting**: “The setting of the narrative is the Ebola outbreak in West Africa, particularly in Liberia, where the lack of qualified staff and inadequate infection-control procedures are exacerbating the crisis.”},
{**characters**: “The characters in the narrative include health workers, patients, and officials, who are struggling to contain the outbreak and provide adequate care.”},
{**plot**: “The plot of the narrative revolves around the challenges of containing the outbreak, including the lack of qualified staff, inadequate infection-control procedures, and the need for improved solutions.”},
{**moral**: “The moral of the narrative is that the Ebola outbreak highlights the need for improved global health infrastructure and the importance of prioritizing the health and safety of healthcare workers, as emphasized by Dr. Frieden, the CDC Director, who stated that ‘preventing transmission requires scrupulous attention to infection control’.”}
],
true_narrative: true }
LLM evaluation: LLM detected the fact that a narrative shift correctly. It correctly identified that the focus shifted back to the African continent and condemning the disease there. It did not focus on the emotional aspect of “hope vs. fear”, which might be the result of not having been provided from the month prior to the change, resulting in a lack of context.

Sample 2

Topic: 36 (Elections and Polls)
Date: 10/2014
Leave-one-out word impacts: scotland, scottish, independence, neves, scot
Narrative shift?: No
Hand-coded narrative shift: The protagonists (and settings) in this topic have hardly anything in common. Documents deal with Catalonia’s independence movement, Brazil’s presidential election, and Scotland’s decision to remain in the U.K., among others. There is no such thing as a common narrative here.

Response of the LLM:
{ **summaries**: [
{article_1: “Catalonia’s independence movement remains strong despite a court injunction, with activists planning to hold a nonbinding referendum on November 9.”},
{article_2: “Aécio Neves, the opposition candidate in Brazil’s presidential election, is gaining momentum against incumbent President Dilma Rousseff, with polls showing him within striking distance of a runoff.”},
{article_3: “Marina Silva, the third-place finisher in Brazil’s presidential election, is considering endorsing Aécio Neves, the opposition candidate, in a bid to defeat President Dilma Rousseff in a runoff.”},
{article_4: “The Brazilian presidential election is too close to call, with Aécio Neves and Dilma Rousseff neck and neck in the polls, with Neves focusing on the economy and Rousseff maintaining a loyal base of support.”},
{article_5: “The Bank of England had planned to flood the financial system with cash and hold emergency auctions to secure banks’ funding if Scotland had voted for independence in its referendum.”}],
topic_change: “The topic shifted from a discussion of Scottish independence to a discussion of Brazilian politics and the presidential election, with a focus on the opposition candidate Aécio Neves and his chances of defeating incumbent President Dilma Rousseff.”,
narrative_before: “Before the change, the narrative centered around the Scottish independence movement, with a focus on the potential consequences of a ’yes’ vote and the plans of the Bank of England to maintain financial stability.”,
narrative_after: “After the change, the narrative centers around the Brazilian presidential election, with a focus on the opposition candidate Aécio Neves and his chances of defeating incumbent President Dilma Rousseff.”,
narrative_criteria: [
{**setting**: “The setting is the countries of Scotland and Brazil, with a focus on the political and economic systems of each country.”},
{**characters**: “The main characters are the politicians and activists involved in the Scottish independence movement and the Brazilian presidential election, including Aécio Neves, Dilma Rousseff, and Marina Silva.”},
{**plot**: “The plot revolves around the potential consequences of a ’yes’ vote in the Scottish referendum and the outcome of the Brazilian presidential election, with a focus on the opposition candidate Aécio Neves and his chances of defeating incumbent President Dilma Rousseff.”},
{**moral**: “There is no clear moral judgment in the narrative, but the narrative does suggest that the opposition candidate Aécio Neves is a more viable option for Brazil’s economy and that the incumbent President Dilma Rousseff’s policies have been ineffective in addressing the country’s economic problems.”}]
true_narrative: true }
LLM evaluation: The model correctly detected some of the main characters. However, it did not come to the conclusion that, due to the different settings, the provided documents are more or less incomparable. The suggested moral may fit to a single document, but does certainly not consider all provided texts adequately.

Sample 3

Topic: 8 (US Elections)
Date: 01/2016
Leave-one-out word impacts: trump, house, iowa, cruz, sander
Narrative shift?: Yes
Hand-coded narrative shift: Just before the start of the 2016 Republican primaries, Donald Trump’s campaign has gained momentum, overtaking his more mainstream competitors through his charisma and aggressive attacks on competitors, such as Ted Cruz and Marco Rubio, and despite questions surrounding his character and previous business dealings.

Response of the LLM:
{ **summaries**: [
{article_1: "The article discusses Donald Trump's financial struggles in the 1990s, including his debt restructuring and bankruptcy, and how he managed to come out of it by taking advantage of his creditors and using his charisma to negotiate deals."},
{article_2: "The article discusses the 2016 presidential election and how Donald Trump's campaign is gaining momentum, despite his high disapproval ratings and the objections of some voters, and how he may be able to overcome these obstacles by confronting his opponents and using his charisma to win over voters."},
{article_3: "The article discusses the 2016 presidential election and how Ted Cruz and Donald Trump are battling for the lead in Iowa, with Cruz's campaign accusing Trump of being too egotistical and Trump questioning Cruz's eligibility to be president."},
{article_4: "The article discusses the 2016 presidential election and how Donald Trump is targeting Ted Cruz, questioning his eligibility to be president and his evangelical faith, and how Cruz is handling the attacks with humor and poise."},
{article_5: "The article discusses the final debate before the Iowa caucuses, where the candidates, minus Donald Trump, sparred over policy differences and personal attacks, with Ted Cruz and Marco Rubio engaging in a heated exchange over immigration policy."}],
topic_change: "The topic shifted from discussing Donald Trump's financial struggles and personal characteristics to focusing on his presidential campaign and the 2016 election, with a particular emphasis on his rivalry with Ted Cruz and the other Republican candidates.",
narrative_before: "Before the change, the narrative centered around Donald Trump's personal characteristics, such as his financial struggles, his ability to negotiate deals, and his charisma, and how these traits might make him a successful businessman and president.",
narrative_after: "After the change, the narrative centers around Trump's presidential campaign and his rivalry with other candidates, particularly Ted Cruz, and how he is using his charisma and tactics to win over voters and gain momentum in the election.",
narrative_criteria: [
{**setting**: "The setting of the narrative is the 2016 presidential election, with the context of the Republican primary and the Iowa caucuses."},
{**characters**: "The main characters in the narrative are Donald Trump, Ted Cruz, and the other Republican candidates, who are portrayed as rivals and opponents in the election."},
{**plot**: "The plot of the narrative is the struggle for the Republican nomination, with Trump and Cruz engaging in a heated rivalry and the other candidates trying to position themselves as alternatives to the two front-runners."},
{**moral**: "There is no clear moral judgment in the narrative, but the story implies that Trump's ability to win over voters and gain momentum in the election is a testament to his charisma and tactics, rather than any inherent qualities or values."}]
true_narrative: true }
LLM evaluation: The model correctly identified the main characters, the setting and the plot of Donald Trump taking the leading role in the primary field. While the model somewhat overstated the importance of the "moral" implications of this development, all aspects of the NPF were correctly recognized and summarized.

**Identifying economic narratives in large text corpora –
An integrated approach using Large Language Models**

Tobias Schmidt[1] (TU Dortmund University), Kai-Robin Lange[1] (TU Dortmund University), Matthias Reccius[1] (Ruhr-University Bochum), Henrik Müller (TU Dortmund University), Michael Roos (Ruhr-University Bochum) and Carsten Jentsch (TU Dortmund University)

Abstract

As interest in economic narratives has grown in recent years, so has the number of pipelines dedicated to extracting such narratives from texts. Pipelines often employ a mix of state-of-the-art natural language processing techniques, such as BERT, to tackle this task. While effective on foundational linguistic operations essential for narrative extraction, such models lack the deeper semantic understanding required to distinguish extracting economic narratives from merely conducting classic tasks like Semantic Role Labeling. Instead of relying on complex model pipelines, we evaluate the benefits of Large Language Models (LLMs) by analyzing a corpus of Wall Street Journal and New York Times newspaper articles about inflation. We apply a rigorous narrative definition and compare `GPT-4o` outputs to gold-standard narratives produced by expert annotators. Our results suggests that `GPT-4o` is capable of extracting valid economic narratives in a structured format, but still falls short of expert-level performance when handling complex documents and narratives. Given the novelty of LLMs in economic research, we also provide guidance for future work in economics and the social sciences that employs LLMs to pursue similar objectives.

JEL-Code: C18, C55, C87, E70

Key words: economic narratives, natural language processing, large language models

[1]: equal contribution

Acknowledgments: This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University and the Narrative Economic Alliance Ruhr (NEAR) project, supported by the Mercator Research Center Ruhr (MERCUR) with project number Ko-2022-0015. It was also partially funded by the Reality Check incubator project at the Research Center for Trustworthy Data Science and Security.

1. Introduction

Macroeconomic policy decisions are not made in a vacuum. More and more, central banks and other institutions recognize the role that public discourse and widely shared narratives play in shaping expectations and thus economic behavior. The central bank communications literature in particular has described the active role of policy makers in explaining their policies to great detail (Gorodnichenko et al., 2023; Gurkaynak et al., 2005; Hansen et al., 2017, 2019). But neither fiscal nor monetary policy discourses are top-down processes. People's economic views and expectations are increasingly shaped by the media (Fiore et al., 2025). Particularly in times of uncertainty, stories about rising inflation, looming recessions, or job market disruptions spread rapidly and uncontrollably.

As economic dynamics are driven by the beliefs and expectations of households and firms, monitoring the narratives that dominate public discourse is crucial. Quantifying economic narratives in mass media can be challenging, however, as the amount of information transmitted is vast and it's rhetorical packaging diverse. Many researchers perform qualitative analyses, which require expert evaluations of texts. While such procedures produce reliable results, they are not scalable to large text corpora. Increasing numbers of articles, reports, opinion pieces and even comments on social media are released every day, so purely relying on qualitative research is not feasible when analyzing the narratives that circulate in an economy. Therefore, economists need efficient, quantitative methods of extracting and presenting narratives from texts.

State-of-the-art language models such as BERT (Devlin et al., 2019) and its variations have shown considerable success in tasks like sentiment analysis, named entity recognition, and semantic role labeling, which are seen by many as foundational to the identification of economic narratives. However, while these models excel at narrow linguistic tasks, they often fall short when tasked with the deeper, more nuanced understanding required to distinguish between complex economic narratives and simpler textual structures. Pipeline approaches that mix different methods also introduce compounding error risks due to their multiple processing stages. Hence, traditional NLP models address the scalability problem but lack the integrated language comprehension needed to fully capture the contextual meaning and economic relevance of narratives.

This gap presents an opportunity for exploring alternative approaches that can offer greater contextual understanding. The advent of Large Language Models (LLMs), with both commercial models such as GPT (Brown et al., 2020), Claude Sonnet and Google Gemini and open source models like Llama (Dubey et al., 2024), has opened up entirely new possibilities for narrative extraction. LLMs have demonstrated remarkable proficiency in a wide array of tasks, excelling especially in those that require a deeper comprehension of language and context (Shahriar et al., 2024). These models are trained on vast amounts of data and can process complex narratives with a level of understanding closer to that of human readers. But how should economists leverage LLMs to accurately measure complex phenomena such as narratives? And can LLMs match expert annotators in identifying and extracting economic narratives from newspaper articles?

In this paper, we use a set of complex instructions and expert-labeled examples to evaluate the proficiency of LLMs at extracting economic narratives. We utilize `GPT-4o` (OpenAI et al., 2024b), the leading commercial LLM at the time of analysis, for the extraction of inflation narratives from a corpus of Wall Street Journal and New York Times newspaper articles. We further demonstrate how complex concepts like economic narratives can be operationalized for human coding, and then adapted for coding by LLMs. While such models enable new avenues for quantitative economic research, the litmus test for their usefulness lies in the rigorous validation of their outputs. Hence, we compare the results of the LLMs' analysis against a gold-standard set of narratives, representing the consensus of three trained annotators with domain expertise. A detailed and rigorous annotator codebook was developed through multiple iterations and refinements. We find that expert annotators reach notably different results extracting narratives, pointing to the inherent subjectivity involved in narrative sense-making. To refine and adjust our evaluation of GPTs' performance to the subjectivity of the task, we not only consider the gold-standard data set as a set of optimal answers, but also the deviations made by the expert annotators, which we call "expected" deviations. We also compare the results attainable through a set of different LLM prompting strategies recently found to be the most effective by research in Artificial Intelligence (AI).

The rest of the paper is structured as follows. In Section 2, we shed light on previous research that has aimed to define economic narratives or automatically extract them from texts. We also provide a short overview of inflation-related literature, as inflation is the center of our evaluation strategy. In Section 3, we introduce the data as well as the LLM we use in our experiments. Section 4 covers our narrative extraction codebook as well as a description of the process of creating a gold-standard data set. In Section 5 we explain our prompting strategy and compare it to other approaches used in the field of AI to maximize LLM performance. The results of the experiments are then displayed and interpreted in Section 6 where we also provide an outlook for future research. Finally, we conclude in Section 7.

2. Related Work

2.1. Defining economic narratives

The field of narrative economics has gained remarkable prominence in recent years (Roos and Reccius, 2024). Only eight years after the well-known paper by Shiller (2017) on this topic, it has become widely accepted in the economic literature that individuals do not simply react to economic data, but interpret the world through personal or publicly shared interpretations of what is going on. Such stories, or narratives, are expected to have the potential to shape personal expectations and even guide collective behavior, when widely shared (Bénabou et al., 2018; Flynn and Sastry, 2024; Larsen and Thorsrud, 2019). This relationship appears particularly relevant in times of uncertainty, when traditional expectation formation models struggle to capture real-world behavior. As King and Kay (2020) argue, under conditions of radical uncertainty, economic agents are unable to maximize utility in the conventional sense. Instead, they turn to shared sense-making structures to reduce complexity and navigate decision-making.

Despite the growing recognition that narratives constitute a promising field of research, the definition of what precisely qualifies as an "economic narrative" remains fragmented, with researchers highlighting different aspects depending on their analytical objectives. Early contributions, most notably Shiller (2017), describe economic narratives as broad stories that convey interpretations of economic events, morals, or simplified theories. While Shiller's intuitive approach to narratives captures their communicative power, high-level concepts like "moral" or "interpretation" resist the consistent formal structure required for empirical analysis. More recent studies, therefore, aim to articulate definitions that are both theoretically rigorous and operationalizable.

A central contribution to the formalization of economic narratives comes from Eliaz and Spiegler (2020, 2024), who conceptualize narratives as simplified causal models represented by directed acyclic graphs (DAGs). Drawing on Bayesian network theory, their 2020 model shows how individuals adopt competing narrative-policy pairs that interpret long-run correlations to maximize anticipatory utility. Narratives, in their framework, are not neutral representations of data but selective causal stories that distort objective relationships by omitting relevant variables or misattributing causal directions. In their 2024 extension, the authors apply this logic to media markets, showing how media platforms strategically supply both biased information and empowering narratives to boost consumer engagement. The empirical study by Andre et al. (2024) follows a similar logic. Specifically, the authors define narratives as backward-looking causal accounts of recent events, or explanations that people construct in order to make sense of what has happened, and that in turn shape their forward-looking expectations. The DAG-structure of narratives that is pervasive in this line of research can be viewed as a key contributor to the simplifying and organizing function narratives fulfill for individuals trying to make sense of information. Roos and Reccius (2024) synthesize these and other perspectives to propose a definition of collective economic narratives. They argue that economically relevant narratives are not just individual stories, but shared sense-making structures that arise in a social context, explain economic events, and suggest collective action. For a more detailed overview over different definitions of economic narratives, we refer readers to this work.

A recurring insight across the literature is that causality is the core ingredient of any narrative. It is what distinguishes a narrative from adjacent concepts such as topics or frames. While a topic identifies what is being discussed, a narrative links who did what to whom, and with what consequence. It attributes responsibility, defines trajectories, and frames problems in a way that invites action. Strikingly, "who is to blame for the problem" (Crow and Jones, 2018) is central to understanding how narratives influence public opinion and policymaking. Against this background, the present study puts particular emphasis on recognizing this specific property of narratives. In order to automatically identify narratives in large text data, detecting causal connections between events—implicit as well as explicit connections—is paramount.

Building on this literature, we define economic narratives as causal connections between two temporally and semantically distinct events, formulated in the structure *A causes B* or *A is caused by B*, that reflect an interpretative framing of economic developments. For example:

1. the prices for cucumbers rose by 100% this month - causes - people stop buying cucumbers
2. increasing money supply - causes - prices for real estate go up
3. the FOMC raised the policy rate - causes - turmoil on Wall Street

Crucially, not every causal claim qualifies as a narrative in our framework. Narratives, as we understand them, are more than just descriptions of factually accurate chains of events; they often imply a perspective, or a selective emphasis that helps make sense of economic developments. To this end, for example, we distinguish between entities that qualify as genuine *events* (such as "inflation is on the rise") and those that do not (such as "high prices", which has no temporal dimension). For a detailed operationalization of our definition, see Section 4.

2.2. Extracting economic narratives from texts

Historically, narrative extraction has been performed qualitatively. Hoping to create a more scalable solution, researchers have switched their focus to quantitative NLP methods to extract narratives from texts automatically. To be able to extract complex narratives however, classic NLP tasks such as topic modeling or sentiment analysis do not suffice. Each of these methods extracts information that is only a small part of most narrative definitions. For instance, topic modeling extracts latent topics from texts, enabling researchers to make an educated guess as to what the documents are about. This can be interpreted as the setting for a story, containing all necessary places and characters, but falls short of connecting these individual parts to a coherent sense-making story.

As a result, researchers have increasingly turned to pipeline-based approaches. Ash et al. (2024) propose such a pipeline, named RELATIO, that heavily relies on semantic role labeling to extract narrative actors. They define narratives as a connection between so called "agents" and "patients", where agents actively perform an action that affects patients. Notably, the causal component inherent in most theoretical approaches to narratives is not considered. For example, Ash et al. (2024) consider "ECB raise interest rate" to be a narrative, with "ECB" as the agent, "interest rate" as the patient and "raise" as the corresponding verb that defines the connection between the two. Lange et al. (2022a) extend this concept by incorporating additional NLP tasks into the pipeline, such as coreference resolution and causal discovery, with the aim of connecting agent-patient pairs to form a sense-making story. However, they also acknowledge that complex pipelines with many interdependent components can cause even minor errors to cascade into serious downstream failures. This sensitivity has motivated the search for integrated models capable of handling these tasks within a unified framework.

Fixing this major issue of narrative extraction techniques, while simultaneously not watering down the definition, requires NLP models of a different architecture. We therefore propose to use LLMs to extract economic narratives from texts. For conceptual clarity, we use the term LLM only when referring to generative language models that have undergone instruction tuning. This limits the scope of the term to models like `GPT-4o`, while excluding families of discriminative language models like BERT. Given the "world knowledge" and language understanding capabilities encoded in LLMs, it is possible to prompt a model with the definition of an economic narrative and extract those narratives directly without the need to run the text through an error prone pipeline. Previous works have shown that LLMs are capable of solving related tasks, such as summarization, speaker attribution and Retrieval Augmented Generation (Bornheim et al., 2024; Fan et al., 2024; Zhang et al., 2025).

Recent research further underscores the potential of LLMs for content analysis. Studies in computational social science have shown that LLMs can match or even surpass human coders in annotating political, social, and economic texts (Ziems et al., 2024). Mellon et al. (2024) reports that LLMs achieved 95% agreement with expert annotators when analyzing British election statements. Similarly, Gilardi et al. (2023) demonstrates that LLMs like `GPT-3.5` could classify tweet content, author stances, and narrative frames more accurately than trained crowd workers. The applicability of LLMs to economic narratives is highlighted by Gueta et al. (2025), who use `GPT-3.5` to extract macroeconomic narratives from Twitter data. Although their operationalization of narratives relies heavily on sentiment analysis and topic modeling, their work demonstrates that LLMs are capable of handling unstructured textual data effectively. Schmidt (2025) makes use of the large language model (LLM) `Claude 3.5 Sonnet` to identify predefined inflation narratives in the German media coverage in 2022. Using a detailed codebook and exhaustive prompting techniques, the author was able to detect the same inflation narratives proposed by Andre et al. (2024) and track their appearance in coverage, demonstrating that LLM prompting is a promising approach to extract backward-looking "blame" narratives from text. Lange et al. (2025) showed that combining scalable methods such as topic modeling with LLMs can lead to additional benefits and can be utilized to extract shifts in complex economic narratives over time. Instead of analyzing every available document with an LLM, the authors propose to only extract narratives at points in time in which a change in narrative is suspected. They detect change points in the word-topic distributions of the dynamic topic model RollingLDA (Rieger et al., 2021) using bootstrap percentile tests (Lange et al., 2022b; Rieger et al., 2022). Afterwards, the authors utilize the LLM `Llama 3.1 8B` (Dubey et al., 2024) to categorize the changes according to the Narrative Policy Framework from political science (Jones and McBeth, 2010; Schlaufer et al., 2022; Shanahan et al., 2018). Such an analysis can be adapted to work with other change detection methods (e.g. Benner et al., 2022), allowing researchers to use a detection method suited for their purposes. This approach of Lange et al. (2025) does, however, also reveal the limitations of `Llama 3.1 8B`, as it infers narratives from most inputs, even when none are present. In our approach, which covers a more complex notion of an economic narrative, we therefore opt for a high-performing model to observe what kind of results a state of the art model can produce.

Given the increasing feasibility of narrative extraction through LLMs, a natural next step is to explore substantive domains where narratives matter most. The topic of inflation is particularly instructive in this regard, as few economic issues are more closely tied to public sentiment, expectation formation, and media framing.

2.3. Inflation-related literature

Inflation is particularly interesting in the context of narrative extraction. First and foremost, the topic is widely researched due to its formal relation to personal beliefs. In standard New Keynesian frameworks, expectations are not merely passive reflections of current conditions but actively shape future inflation trajectories (Werning, 2022). Agents form beliefs about inflation that, in turn, influence their consumption, pricing, and wage-setting behavior (Bachmann et al., 2015; Burch and Werneke, 1975; Juster et al., 1972). In this context, economic narratives play an important role, as narratives help individuals understand the world around them and make decisions. This tight linkage between belief formation and macroeconomic outcomes is one reason why inflation narratives have attracted more research attention than, say, labor market narratives. Central banks in particular have shown strong interest in the topic. In the modern forward guidance era, they aim to actively shape the inflation expectations of households and firms (Blinder et al., 2008). Given this policy regime, it is unsurprising that a growing number of central bank researchers are engaged in studying the role of narratives in the expectation formation process (Kalamara et al., 2020; Nyman et al., 2021; Ter Ellen et al., 2022; Tuckett et al., 2020).

One central finding from inflation expectation research is that media coverage appears to play an important role in expectation formation processes. As Conrad et al. (2022) show in the German context, traditional media consumption is associated with more accurate perceptions of past and expected inflation, particularly when media coverage is intensive. Similarly, Lamla and Lein (2014) find that intensive inflation reporting increases the forecast accuracy of households, while negative or sensationalist framing can lead to exaggerated inflation perceptions. Ter Ellen et al. (2022) further show that narrative monetary policy shocks (i.e., stories about interest rate decisions) can affect real macroeconomic variables if successfully disseminated.

Another central contribution in this domain is provided by Andre et al. (2024), who analyze the narratives individuals construct to explain the recent surge in inflation. Drawing on open-ended survey responses from thousands of U.S. households, the authors find that while expert narratives predominantly attribute inflation to demand-side factors (e.g., fiscal and monetary expansion), household and manager narratives are more heterogeneous and often emphasize supply-side shocks or political mismanagement. Through randomized priming experiments, the authors show that media narrative exposure causally shifts beliefs, highlighting the importance of inflation narratives in expectation formation.

Other lines of research suggests, however, that responsiveness of expectations to media coverage varies over time. Building on the rational inattention framework (Reis, 2006; Sims, 2003), Coibion and Gorodnichenko (2015) and Bracha and Tang (2025) provide evidence that attention to inflation is cyclical. As the authors show, media attention to inflation spikes in periods of high inflation and declines otherwise. Recent research by Schmidt et al. (2023) references to the same effect by analyzing how varying levels of inflation reporting affect inflation expectations. The authors make use of the Inflation Perception Indicator (IPI), which tracks thematic shifts in German inflation coverage, to analyze narrative shifts in German inflation reporting. Employing a threshold VAR framework, the authors show that the influence of media coverage on inflation expectations is regime-dependent: only during high-inflation periods does narrative intensity significantly affect expectations and real variables.

In what follows, we put these insights into practice. Using inflation as a well-studied and socially salient topic, we apply an LLM-based narrative extraction approach to a corpus of media reports. Although inflation serves as the empirical domain of this paper, our methodology is generalizable to other economic issues. Our goal is not only to assess the performance of LLMs in a high-stakes context, but also to demonstrate how economic narratives can be systematically identified in large text corpora.

3. Model and data

In our experiment, we aim to get the best performance possible given the current generation of LLMs. In this section, we argue why the model `GPT-4o` (OpenAI et al., 2024b) is our model of choice to accomplish this and describe how we have created the data set for our experiment.

3.1. Choosing an LLM

At the time of analysis, `GPT-4o` (OpenAI et al., 2024b) is the newest iteration in the 4th generation of the Generative Pre-trained Transformer family of models by OpenAI. We chose this particular LLM—model snapshot `gpt-4o-2024-11-20` —over other commercial models, such as `GPT-4` (OpenAI et al., 2024a), as it offers comparable or superior performance while being more scalable due to lower financial costs. Its performance in human reasoning and language understanding tasks is especially impressive (Shahriar et al., 2024), as these are crucial components of narrative extraction, which requires deep understanding of complex economic circumstances. In particular, we choose to use `GPT-4o` over the more recent class of reasoning models, such as `OpenAI-o1` (OpenAI, 2024), as our prompt requires Chain-of-Thought-prompting (Wei et al., 2022) tailored specifically to our narrative definition. As recent research shows that prompting reasoning models with additional Chain-of-Thought sequences does not improve performance (DeepSeek-AI, 2025; Wang et al., 2024), we opt for our own specialized prompting over the generic reasoning offered by models like `OpenAI-o1`. Our prompting strategy is described in detail in Section 5.

While analyzing the potential of large commercial models to extract economic narratives is valuable, their use also entails notable downsides. For one, the cost of using such models on large data sets is very high and might not be feasible for universities, research centers or companies with low financial backing. Additionally, since these models are not publicly available and operate on proprietary hardware, their long-term availability is uncertain, as companies like OpenAI are likely to prioritize profit over maintaining access to older models. Therefore, all evaluations performed by older models will ultimately lose their reproducibility, undermining a cornerstone of good scientific practice. For this reason, the use of the commercial model `GPT-4o` in this paper should be understood as a theoretical example of what state-of-the-art LLMs are currently capable of, and what open-source models may achieve in the future as their development progresses.

3.2. The text corpus

We evaluate the proficiency of `GPT-4o` on the task of economic narrative extraction using a curated corpus of Wall Street Journal and New York Times news paper articles published between 01-01-1985 and 27-09-2023. We filtered for articles containing the words "inflation" or "price (increase|hike|surge)". This time period was selected because it includes both high- and low-inflation phases. Focusing on inflation-related documents allows us to work with a thematically coherent corpus, enabling aggregation and comparison of narratives, even within a relatively small dataset.

To further narrow down the potential economic narratives in these documents, we provide the model with a short excerpt rather than the full article. Each excerpt includes the sentence containing the filter terms, along with the two preceding and two following sentences. This allows us to focus only on the topic of inflation while still providing enough context to interpret potential narratives, including those that span multiple sentences.

From this corpus, we randomly sample 100 documents to create gold-standard responses for the LLM. While a larger annotated dataset is always desirable, we prioritize a thorough annotation of a smaller subset over a broader but less precise coverage. The annotation task is complex, demands sustained focus, and is prone to errors when done hastily or without sufficient care.

4. The codebook and the gold-standard

When trying to operationalize rich concepts like economic narratives for empirical research, we often struggle to translate verbal definitions into a rigorous and workable codebook. Aspects of narratives that seem sensible enough in a verbal explanation are sometimes hard or even impossible to identify in an empirical setting. This problem gets exacerbated when a codebook is required to apply to LLM-based extraction in addition to human coders. While LLMs are trained to appear as human-like partners in interactions, they differ substantially from humans in the way they process information when performing tasks (Mondorf and Plank, 2024).

Therefore, we will proceed by clarifying what defining aspects of economic narratives we consider to be important in an empirical setting. Based on these aspects, we create a codebook designed for human coders to create a consensus gold-standard data-set. We then translate the human codebook into a prompt designed to maximize the LLM's performance, accounting for the model's peculiarities relative to human annotators.

4.1. A narrative codebook

Translating any of the narrative conceptualizations presented in Section 2 to empirical work entails some fundamental challenges. These definitions operate with high-level concepts such as *event*, *story* and *moral*. As a result, when two experts are given only a definition to guide text annotation, their results are likely to differ significantly. Because a narrative is tied to language, domain understanding, and the reader's interpretation of the author's intent, no narrative extraction can be truly objective. Additionally, from a purely linguistic perspective, narrative extraction is not a simple span detection task in which one specific part of a text must be marked. Instead, key contextual information given in the text must be extracted and synthesized into a coherent structure. These structures must also be amenable to aggregation when comparing narratives across large corpora.

We propose a codebook that both narrows down the type of narratives that are to be annotated and outlines a clear structure for the annotation process. The codebook is the end product of multiple workshops that were held with experts and non-experts. These sessions aimed to develop clear guidelines for extracting economic narratives, minimizing common coding errors and sources of ambiguity wherever possible. The entire codebook can be found in Section A in the appendix. The most important aspects of the codebook can be summarized as follows.

- **Goal**: The main task is to accurately extract narratives from newspaper excerpts, defined as a causal link between two consecutive events. Multiple researchers should code the same texts to minimize subjective bias.

- **Target Form**: Each narrative must be reduced to one of two forms: *event 1 - causes - event 2* or *event 1 - is caused by - event 2*. To preserve the natural order of the text, the first event in the source should remain as *event 1* in the coded narrative.

- **Event Types**: Permissible events include occurrences, activities, conditions, future events, plans and policies. The coded event should remain as close as possible to the original wording, avoiding synonyms or abstractions.

- **Coreference Resolution**: When entities are referred to indirectly (e.g., with the use of pronouns), the coder is to replace them with the correct entity to maintain clarity (e.g., "{He|President Biden}").

- **Statements**: When an event consists of a statement made by an individual, coders must distinguish whether the causal link to another event pertains to the content of the statement or to the statement itself. Typically, the focus should be on the content. However, if the statement itself triggers a reaction, it should be coded as the event. Typical examples include forward-looking statements in corporate disclosures or forward guidance issued by central banks.

- **Embedded Clauses**: Coders should ignore non-essential elements within events (e.g., appositions, parentheses) unless they contribute meaningfully to the narrative.

- **Chained narratives**: Each event in a chained narrative (e.g. "event 1 causes event 2; event 2 causes event 3") should be coded separately.

- **Narrative forks**: When an event is caused by multiple events or an event causes multiple other events, each causal connection has to be noted as a separate narrative unless the combination of events is integral to the narrative.

- **Positive Causality Only**: Only positive causal links should be coded, negative ones (e.g., "does not cause") should be excluded.

- **Economic Focus**: Only narratives with an explicit economic context should be coded. Non-economic narratives should be ignored.

- **Maintain Full Meaning**: The full message of the original narrative is to be retained without omitting crucial information, even if some parts seem superfluous.

- **Edge Cases**: When a code is unsure if a sentence contains a causal narrative, they code it anyway and resolve uncertainties later with other coders.

4.2. Creating a gold-standard data set

To annotate our 100 randomly sampled documents, three expert annotators —familiar with the definition of economics narratives, experienced in applying it and proficient in English at a near-native level—were provided with the codebook. Each annotator was tasked with coding all narratives from the 100 documents independently, without discussing their results with the other annotators. To address any remaining language barriers, such as idiomatic expressions or culturally specific references, annotators were given access to a built-in DeepL API, allowing them to translate the text into their native language if needed. In such cases, narratives were still coded based on the original English version.

In regular intervals during the annotation process, the annotators met to revise their understanding of the codebook and discuss edge cases they were uncertain about.

After coding all 100 documents individually, the annotators met in multiple sessions to discuss the results. Each individually coded narrative was discussed by all three annotators. If they unanimously agreed it was correctly coded, the narrative was added to the list of gold-standard narratives for the corresponding document. In this process, it did not matter whether a narrative had been identified by one, two, or all three annotators. Given the subjectivity of the task and the possibility that even experts may overlook narratives without considering multiple perspectives, inclusion of a narrative in the gold standard was based solely on substantive discussion grounded in the codebook. Such procedures have been shown to produce more valid and reliable results than simply applying majority voting (Burla et al., 2008; Neidhardt, 2010).

The individual annotations by each annotator, however, were not discarded. Instead, we use them to compute an *expected deviation* from the gold-standard, representing the level of variation that is acceptable among expert annotators. The output of the LLM is then compared to this baseline. We also document the reasons for each deviation from the gold-standard labels. Suppose an event has multiple consequences—i.e. *Event A* causes three distinct events: *B*, *C* and *D*. Annotator 1 did not split these into three individual narratives but instead coded *Event A* as causing the combined outcome of *B*, *C* and *D*. This would be classified as a *minor deviation*, with the reason noted as "missing multiple consequences". This approach of differentiating major from minor deviations allows us to formalize complex patterns within our codebook while still comparing human and AI performance effectively. *Major deviations* either completely alter the meaning of a narrative, identify a narrative that does not exist or miss an existing narrative. Minor deviations, by contrast, occur when the overarching narrative is correctly identified but is either formally incorrect (e.g. due to missing coreference resolution) or contains subjective elements that differ from the gold-standard annotation.

In total, the gold-standard data set contains 291 narratives in 100 documents.

5. Prompting LLMs to extract narratives

Prompting language models to extract and annotate data in research is a relatively new venture. We feel strongly that LLMs open entirely new avenues for quantitative economic research. Korinek (2023) was among the first to urge economists to explore this potential. However, as of the time of this writing, the major use case for LLMs in economic research is sentiment analysis (Examples include Jeong and Ahn (2025) and Han et al. (2024)), for which high-quality models exist that do not require Generative AI. Due to the novelty of LLMs, no best practices for crafting prompts have been established to date in any of the social sciences. However, recent literature in AI and computational linguistics has introduced a set of general strategies that can enable researchers to make informed choices. We hope that our task can provide a useful example for related studies, and we will cite relevant work throughout. Using `GPT-4o`, we explore multiple approaches to extract economic narratives informally. We then narrow down the more useful principles which we subject to a formal evaluation. We require narratives to be extracted in a consistent format that can be parsed by machines. Therefore, we always use the "JSON" option in the OpenAI API, which guarantees that the model will output consistent Java Script Object Notation (JSON) files.

5.1. Prompt optimization: strategies and challenges

The least data-intensive yet most heuristic prompting strategy is zero-shot learning (Kojima et al., 2022; Wei et al., 2021). Generally speaking, LLM-assistants like ChatGPT are fine-tuned for instruction-following, which is the reason users can interact with them in a way that feels natural and human-like. By using zero-shot prompting, researchers can exploit this property by only supplying the model with the research question and task instruction along with the text to perform inference on. To set a benchmark for comparison to human performance, we use this zero-shot strategy, supplying our entire unabridged codebook as an instruction. The results are very inconsistent, both regarding the form of the outputs and in their faithfulness to the codebook.

We also explore a variation of this strategy, using a condensed version of our codebook. While current LLMs use interpolation techniques (Zhong et al., 2025) to expand hard context windows beyond the length of any common-sense codebook, it remains unclear how well performance is maintained as context size increases. Due to the intransitivity of LLM architectures, few theoretical guarantees can be considered for applied work. However, empirical investigations have documented phenomena such as the lost-in-the-middle effect (Liu et al., 2024), which suggests that increasing the length of inputs tends to result in subtle forms of information loss, mostly in the middle of input sequences. The distribution of information in prompts appears to be a relevant criterion as well. Tian et al. (2024) suggest that a higher relative distance between crucial pieces of information in a prompt adversely affects information retrieval rates, which could compromise tasks like ours. Hence, ceteris paribus, limiting prompt length is advisable. However, the level of detail in our codebook must still cover the intricacies of the coding task. To find the sweet spot regarding this prompt-length trade-off, we condense our codebook for several iterations, emphasizing different aspects in each version. We evaluate by probing the outputs that result from using different codebook variations.

Generally, we find that increasing the information density of the codebook tends to be beneficial for performance. However, a priori, it is not obvious what kinds of information should be compressed and what concepts must be dealt with extensively in the instructions. For example, several passages in the codebook deal with the definition of an event, including detailed elaborations on the types of events we consider to be the most relevant constituents of economic narratives. These include the implementation of a policy or the future-facing intentions of firms or policy makers. We endowed human coders with a great level of detail here partly because issues involving the delineation of events had come up regularly during the codebook workshops. However, as it turns out, `GPT-4o` defaults to a conceptualization of events that closely corresponds to our definition. Hence, simply mentioning that narratives consist of two events turns out to be the optimal level of information we can provide about events. By contrast, the ordering of events in the output, which we require to remain unchanged from the source text, turns out to be a problem that needs explicit mentioning in multiple spots of the instructions.

In general, simply adding more requirements to prompts does not reliably increase LLM-performance. Additional aspects often introduce competing constraints that the model must resolve implicitly (Yang et al., 2025). Researchers should therefore prioritize which concepts to explain in detail. Morover, since an LLM's conceptual understanding is shaped by its training data and the alignment techniques used during training — both of which are unknown to researchers — practitioners should always probe the model's sensitivity to the way information is provided, even for seemingly uncontroversial concepts.

In addition to descriptive instructions, various sections of the codebook (see Section A) contain synthetic examples of full narratives and of narrative components. These examples are meant to highlight specific aspects of the extraction task for human coders to focus on when familiarizing themselves with the coding process. Such partial demonstrations are essential for designing high-quality codebooks, as they have shown to be helpful for human-coders (Saldaña, 2016). Conversely, they do not contribute positively to LLM performance on our task. The model appears to struggle with synthesizing these aspect-driven examples into a consistent, multi-aspect narrative definition, which adversely affects model performance.

5.2. Few-shot Chain-of-Thought prompting

Given the results of the initial explorations, we decide to fully adapt our original codebook for LLM inference. Instead of using instructions that mirror the structure of the original codebook outlined in Section 4, we only provide the model with a succinct description of the key concepts and processes that it contains. Crucially, along with these instructions, we supply hand-labeled input-output examples. This widely used method is called few-shot learning in the AI literature (see Song et al. (2023) for a review). Unlike the supervised learning paradigm that most empirical economists will be familiar with, effective few-shot learning only requires a handful—rather than hundreds—of hand-labeled training examples. In contrast to the strategy outlined in Section 5.1, where distributed examples throughout the codebook each highlight a specific aspect, few-shots are complete input-to-output demonstrations that are provided in a dedicated block after the verbal instructions.

The few-shot paradigm allows the model to infer information about the desired output from full-fledged examples rather than from lengthy prosaic instructions. Consequently, the selection of few-shots and their total number are crucial. Few shots should be representative of the full data-set, ideally without adding redundancies. During this final round of prompt engineering, we use rigorous formal validation: From our set of 100 hand-annotated, gold-standard documents, we randomly select 20 for cross-validation. Any of these 20 examples can be used either as few-shots in prompts or for evaluating the performance of a candidate-prompt. The rest of the gold-standard documents are held out entirely for testing the performance of the final prompt. Given our validation set, we explore using between 1 and 9 few-shots in a single prompt. After evaluating the performance on the remaining validation example, holding constant the rest of the instructions, we settle for using 7 few-shots. Finally, we run evaluations with rotating sets of seven examples to identify the few-shot combination that yields the best evaluation results. Intuitively, the optimal set captures the most relevant narrative characteristics for the model to learn the task effectively.

We combine few-shot learning with the Chain-of-Thought (CoT) prompting paradigm, which means that we specify intermediate steps for the model to take when constructing an answer (Wei et al., 2022), with the few-shot examples mirroring this step-wise format. When dealing with problems that require complex forms of reasoning, dividing up a task into sub-problems turns out to be advantageous. This is because LLMs can only "reason" about a problem insofar as they can produce tokens about it. The CoT-logic of transforming narrative extraction into a set of sequential tasks allows the model to devote more computational power to each step in the reasoning chain. Every step must be framed as a discrete action and output separately by the model before moving on to the next step. In addition to enhancing performance, CoT-prompting also increases transparency, as errors in LLM inference can be traced to particular steps in the chain.

Conceptually, the CoT paradigm is reminiscent of classic pipelines approaches to economic narrative extraction (see Section 2), where sub-tasks are handled by different language models rather than a unified LLM (Ash et al., 2024; Lange et al., 2022a). However, the CoT pipeline merely imitates a full separation into sub-task. The LLM always has access to the prior steps since they are stored in its context. Since our codebook contains interdependent tasks that require economic judgment and adherence to linguistic principles, this feature is highly desirable.

5.3. An integrated narrative extraction pipeline

Our final prompt relies heavily on the few-shot CoT paradigm introduced in the previous section. A detailed description can be found in Section C in the appendix. The prompt is divided into two main parts, *Codebook* and *Examples*, where the former subsumes the written instructions and the latter only encompasses the few-shots.

The first subsection of the codebook, *Basic Idea*, is designed to prime the model by broadly outlining the task and the source medium of the input texts. Subsection 2, *Definition*, briefly states the crucial structural elements needed in every extracted narrative:

An economic narrative consists of exactly two events
and a causal connection that is asserted between those events.

Subsection 3 is called *Event structures*. It introduces three distinct types of linguistic constructions in which causally linked events are typically embedded when they are discussed within the article excerpts. We build on the formalism of graphical models introduced by Pearl (2009), which is widely used in the AI field and in parts of economics, to distinguish *direct causal effects* from causal *chains* and *forks*. A causal chain occurs if an *Event A* causes an *Event B* which in turn causes an *Event C*. A causal fork occurs if an *Event B* is a common cause for two Events, *A* and *C*. Embracing this terminology enables us to represent causal structures as directed acyclic graphs (DAGs), as is common practice in the literature on economic narratives. Using this established formal language also allows us to optimally leverage `GPT4o's` pre-training, because we do not have to explain concepts in the instructions to the LLM as long as they are encapsulated in the model's world knowledge. As stated in Section 4, we require chains and forks to be coded as separate narratives by the model. While some of the few-shots contain examples of such causal constellations, the model turned out to struggle with them, which prompted us to incorporate them into the *Codebook* section. Hence, apart from raising awareness about the aforementioned causal construct, subsection 3 also sets up expectations on how the LLM should deal with these structures when they arise.

Subsections 4 and 5, *Rules* and *Target forms*, precisely establish the set of allowable forms that extracted narratives are expected to conform to. We assert that the expression denoting the causal connection between events in a narrative must be classified into either "causes" or "caused by", depending on the direction of the causal relationship. As laid out in Section 2, causality is a core ingredient of economic narratives, because causal connections enable individuals to make sense of why events have unfolded. Many economic choices by households and firms are based on this sense-making inference. Quite commonly, however, causal connections that appear obvious to human readers are not written down explicitly in the source text. One of the main advantage of using LLMs for our task is their ability to recognize a diverse set of explicit and implicit causal connections from a given context. While extracting implicit causality—when related events are not connected using a specific causal cue—is challenging, it is equally important. For example, when events occur at different points in time, authors often imply causal relationships through temporal cues. Distinguishing whether the author intends to signal causality or merely describes a sequence of independent events often depends on contextual knowledge. LLMs excel at such tasks, which require deep language understanding and nuanced interpretation. This capability sets them apart from dictionary-based methods that rely on explicit cue words to detect causality or language models like BERT (Devlin et al., 2019) that require fine-tuning.

In the final subsection, *Extraction process*, we describe the CoT that the model should produce for every input. Figure 1 displays this procedure for an example text.

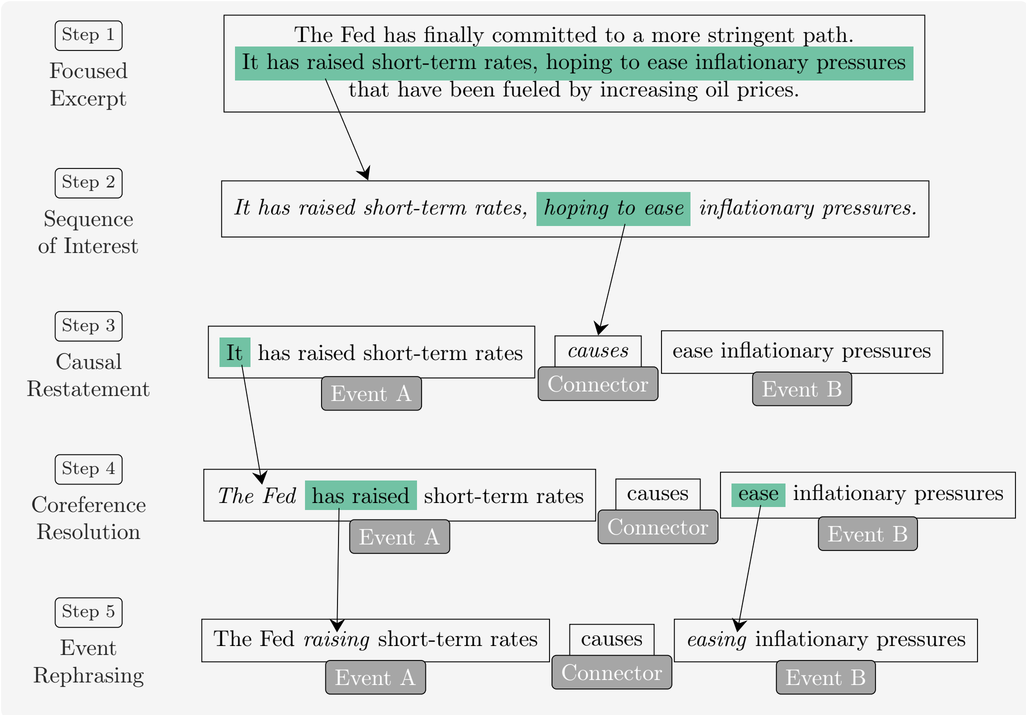


Figure 1: A diagram depicting the Chain-of-Thought transformations that our prompting strategy induces. The chain extracts an economic narrative from an example document and molds it into a standardized form step by step. Green highlighting indicates segments to be changed in the subsequent step, arrows map the source segments to their corresponding results. Results are cursive.

As mentioned before, LLM performance hinges on steering the model's attention mechanism to relevant sections of the inputs. Step 1, therefore, consists of extracting a *focused excerpt* from the input, which is an article fragment that features a term deemed indicative for an inflation discourse. Since the input texts vary significantly in their narrative density, it is helpful to first let the model disregard any sections from it that do not contain any events or causal relationship. This initial recognition of a narrative sequence is already non-trivial. In our example, the events are the Fed's realized policy measure—raising interest rates—and it's intention of easing inflationary pressures in the economy. Their causal connection is hypothetical because the outcome event has not occurred and is not guaranteed to occur at all. These nuances must be recognized by the LLM in the first CoT step, at least to the degree that it will include the segment in the focused excerpt instead of disregarding it altogether.

Step 2 delineates a single, previously unextracted narrative from the focused excerpt by producing a *Sequence of interest.* The task requires the LLM to isolate exactly two events that are causally connected, while disregarding any unrelated elements found in the text. Sequences of interest are essentially full narratives in their rawest form. In our example, the focused excerpt actually features two more events—the Fed committing to a more stringent path and increasing oil prices. While either of those might be important for further narratives, Step 2 requires the LLM to not attend to them for the time being. The *Causal restatement* in Step 3 serves to separate the events from the causal connector and to make the latter explicit. Collectively, these first 3 CoT steps perform the "heavy lifting" of narrative recognition.

Steps 4 and 5 only apply for a subset of extracted narratives. Steps 4 replaces pronouns in the sequence of interest with the entities they refer to. This task is known as *coreference resolution* in the NLP literature and is generally handled with high accuracy by the LLM. If required, this step is crucial because entities that act or are acted upon are essential component of many narratives (Gehring and Grigoletto, 2023). Step 5 consists of some final grammatical adaptations: For the sake of event aggregation, we require events to be grammatically interchangeable. *Event rephrasing* standardizes grammatical structure, allowing for easier extraction, comparison, and inference over events. This step mostly entails nominalization, in which a verb phrase is rephrased as a noun phrase.

After Step 5, we redirect the LLM to the focused excerpt to check for further narratives.

To validate our prompting strategy, we use between 1 and 9 few-shots from the validation sample and evaluate performance on the remaining 11 examples. In our experiments, performance peaks with 7 few-shots. Subsequently, we test and evaluate `GPT-4o` with our chosen prompt on the remaining 80 documents with gold-standard labels. We set the LLMs temperature hyperparameter, which governs the randomness or creativity of its responses, to 0.2. This setting makes responses more deterministic than the default while retaining enough variability to allow the model to explore the space of possible answers. For the purpose of reproducibility, eliminating any randomness by using a temperature setting of 0 would be preferable. However, a fully deterministic setting is known to degrade performance of generative language models, causing what is known as the *likelihood trap* (Huang et al., 2023; Zhang et al., 2021).

5.4. Narrative abstraction and event clustering

After extracting all narratives in the form of *Event A causes Event B* from the 80 test documents, we apply a series of post-processing steps aimed at enabling large-scale abstraction and aggregation of recurring economic patterns. The goal of this procedure is to move beyond individual instances and identify broader narrative structures. The following steps represent an initial attempt to structure and cluster narrative content in a principled way. While the primary focus of this paper lies in the extraction of narrative pairs, we briefly outline this pipeline to illustrate how the extracted data can be further processed and organized for downstream applications.

Event decomposition. First, we split compound events into distinct atomic propositions. Events such as "Energy and food prices are on the rise" were forked into two independent entries, such as "Energy prices are on the rise" and "food prices are on the rise". This decomposition is essential, as early tests revealed that our narrative extraction model occasionally fails to fork such multi-entity constructions in the desired way (see above). Accordingly, this step functions as a correction layer that may become redundant in future implementations if the model's extraction capabilities improve. We implement this step using a few-shot prompt with `GPT-4o-mini`, designed to detect conjunctions and rewrite them into separate, grammatically coherent event statements. This approach ensures that each entry represents a clearly interpretable unit of analysis. By contrast, a simple regex-based "and"-splitting approach would often yield semantically incomplete fragments, such as "Energy" in the upper example, that lack narrative coherence.

Valence and topic assignment. Next, each atomic event is passed through another LLM-based classification step to extract both its semantic *topic* (e.g., "inflation", "interest rates") and its directional *valence* (e.g., "rising", "falling", "high"). The prompt follows a structured format with few-shot examples and requires the model to output a JSON object containing both events A and B and their respective valences, thereby ensuring consistent outputs across all extractions.

Topic normalization and abstraction. Since we prompted our initial narrative extraction model to stick as closely to the original wording of the text as possible during inference, many of the extracted narratives vary lexically while describing the same or similar phenomena (e.g., "interest rates" vs. "borrowing costs"). Therefore, we implement a semi-automated abstraction step to cluster semantically similar topics. First, we embed all extracted topic strings using the `all-MiniLM-L6-v2` sentence-transformer model. Second, we map each topic embedding to a controlled set of embeddings representing different macroeconomic categories (e.g., "government spending", "interest rates", "inflation"). To compare the topic embeddings and the predefined anchor terms, we use cosine similarity metrics, as they are invariant to the high number of embedding dimensions. Previous research has shown that clustering by anchor terms can serve as an alternative to fine-tuning based classification approaches in an unsupervised setting and help to interpret embeddings (Lange et al., 2024; Mathew et al., 2020). We take care to preserve economically meaningful distinctions, e.g., "inflation expectations" and "inflation" remain separate due to their distinct roles in economic theory and policy, and even in public understanding. We then manually refine the mappings by carefully inspecting cluster compositions. In cases where embedding similarity alone appears insufficient, we use domain knowledge and manually add recurrent topic synonyms to the respective topic cluster or remove misclassified terms. Via pattern-based string matching we then replace all topics that were assigned to a certain cluster with the respective topic label. This hybrid approach ensures that topic abstraction remain both semantically robust and economically meaningful. An example of this mapping is provided in Figure 2.

Cluster: Bond Market
- bond market
- bond markets
- Treasury bonds
- ...

Cluster: Consumer Spending
- consumer spending
- consumer demand
- household spending
- ...

Cluster: Energy Prices
- energy prices
- electricity prices
- home energy prices
- wholesale energy prices
- fuel prices
- ...

Figure 2: Example illustrating how events are grouped by a cluster topic.

Finally, we translate all valence expressions indicating an increase or general high level of something with an upwards-pointing arrow (↑) and all expressions indicating low levels or a decrease with a downwards pointing arrow (↓). In cases where implicit negations are involved, these are reversed to ensure that, for example, *rising economic vulnerability* and *rising economic stability* are represented by different arrows, even if the term "rising" is used in both contexts.

We are well aware of the information loss this step inevitably entails and the economic limitations that come with it. For instance, real interest rates differ fundamentally from nominal interest rates, which in turn are not the same as borrowing costs. However, we argue that such simplifications are justified given the inherent nature of narrative formation (i.e., the tendency to reduce complexity) and the context of our corpus, which consists of general-interest newspaper articles. Distinctions that are highly relevant in academic economics (e.g., between monetary aggregates or types of interest rates) may

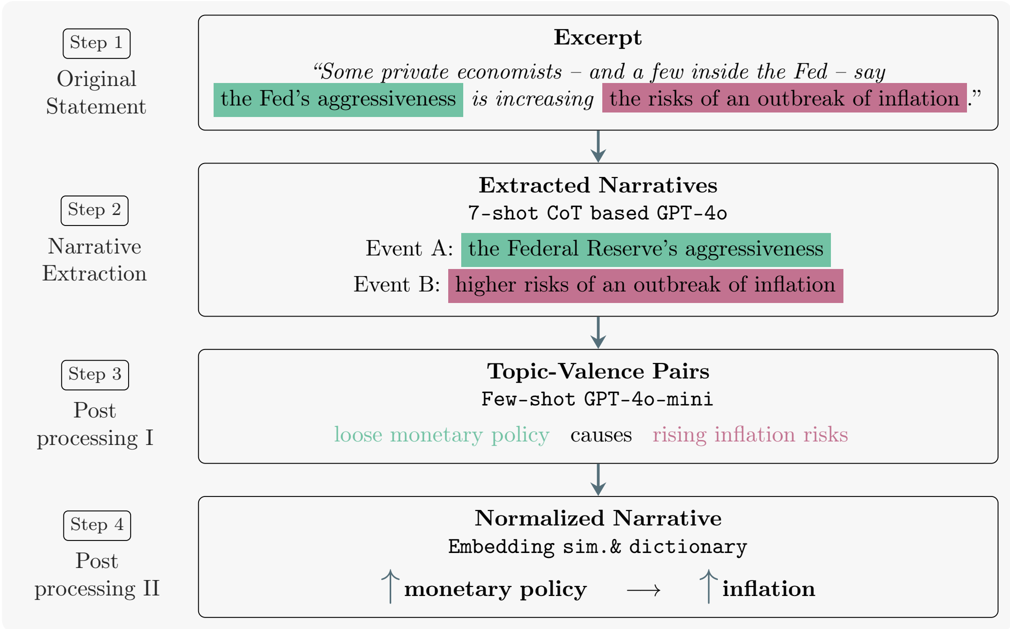


Figure 3: Illustration of our narrative abstraction pipeline using a real example from the corpus.

be less salient or even irrelevant to the typical audience of U.S. dailies. In some cases, however, this loss of specificity results in tautological or arbitrary narrative formulations, such as *rising interest rates cause higher interest rates*. While such artifacts are rare in our dataset, they illustrate one of the limitations of semantic abstraction: when the model reduces different expressions to a shared label, it may inadvertently collapse distinct concepts into self-referential loops. We acknowledge this as a trade-off inherent to our current aggregation strategy, though it affects only a small number of cases and might be mitigated through more refined post-processing in future applications. Figure 3 visualizes the final extraction pipeline using a real example from our corpus.

6. Results

Using the codebook presented earlier, we let GPT-4o process the remaining 80 documents that were neither used for evaluation nor for model prompting. The results offer a nuanced picture of the model's ability to extract economic narratives from text. We will first discuss the results that we have attained through prompting the LLM and then turn to the critical post-processing step of aggregating the extracted narratives. Despite the complexity involved in creating the gold-standard, comparing GPT-4o to this benchmark turned out to be straight-forward, since essentially all disagreements had been resolved beforehand.

The dataset containing test and evaluation documents, including the gold standard as well as the model's and individual experts' annotations, is publicly available and can be found on GitHub.

6.1. Narrative identification

On a structural level, the model performs remarkably well. In almost all cases, GPT-4o precisely follows the formatting instructions and consistently produces valid JSON outputs. In Section B of the appendix, we present three examples comparing the gold standard narratives to those generated by the model, alongside the original excerpt for reference.

The LLM also closely adheres to the CoT-procedure laid out in Figure 1. In Steps 1 and 2, GPT-4o typically paraphrases relevant parts of the source text in a semantically faithful and economically coherent manner. The model also correctly reproduces event order and directions of causality in most narratives. The CoT logic likely aids this process, as it requires the model to state the narrative sequence as continuous text twice before imposing causal structure and determining event order in Step 3. Even when the model misorders events, it typically preserves correct causal directionality, making such errors relatively minor and more grammatical than substantive. Misorderings tend to favor the causal connector *causes* over *caused by*. The reason for this preference is speculative. GPT-4o may favor stating causes first and effects second because such constructions occur more frequently in its training data.

In Table 1, we compare the model's performance to that of our expert annotators' using basic summary statistics. Notably, the LLM identifies a similar number of narratives per document as the human coders, on average. This suggests that the overall level of performance is not compromised by persistent over- or under-identification of narratives. However, a closer look at the distribution reveals important limitations. While human annotators show substantial variance in how many narratives they identify per document (standard deviation of 2.03, on average) the model's output is remarkably stable, averaging 2.32 narratives per document with a standard deviation of only 1.15. This points to a potential inductive bias: the model appears to have an implicit prior about how many narratives it "expects" to find, leading to systematic over-identification in narrative-sparse

Measure	**Expert 1**	**Expert 2**	**Expert 3**	**Gold**	**Model**
Average	2.22	2.36	2.29	2.91	2.32
Standard deviation	1.95	2.12	2.04	2.35	1.15

Table 1: Narrative counts per document for human experts and `GPT-4o`.

texts and under-identification in narrative-dense ones. This pattern is also evident in edge cases. In four documents where the gold standard specifies no narrative, `GPT-4o` nonetheless extracts at least one narrative in each case. This suggests that the model may be prone to overfitting the 7-shot examples provided in the prompt.

To better understand this pattern, it is useful to examine some specific challenges the model faces. The lack of variability in the number of narratives is partly due to the model struggling with causal chains and forks. Even with explicit instructions and a dedicated CoT step in place, the model tends to return most narratives in a non-forked fashion, even in clear-cut cases where all human experts agree. Rather than disentangling forked or chained narratives into multiple atomic phrases, the model frequently compresses them into a single generic statement (see Figure 4). This issue is related to stylistic differences across documents. Some of the articles break down inflation-related phenomena in a technical style, even referencing the channels of monetary policy that are thought to be at work. In such articles, causal chains and forks may intermix, even within a single sentence, yielding multiple narratives from a short text span. When `GPT-4o` struggles with disentangling these structures and coding all narratives separately, the model's narrative count gets notably deflated for the respective article. Conversely, narrative-sparse articles tend to focus less on textbook economic explanations and more on human interest and storytelling. Given the latter, the model tends to infer inaccurate or speculated narratives much more liberally than human coders. This lack of consistency is related to the phenomenon of hallucinations, a term describing the well-known tendency of LLMs to generate plausible but inaccurate answers (Huang et al., 2023).

In general, it appears that the model performs particularly well on short documents with relatively straightforward narratives, while struggling with longer, more complex documents. `GPT-4o` sometimes fails to detect crucial narrative structures, especially when narratives are implicitly stated or span multiple sentences. As the performance metrics in Table 2 illustrate, these characteristics come with substantial downstream effects: `GPT-4o` averages 1.25 major deviations per document overall, more than twice the rate of any expert coder. This translate to an "unexpected major-deviation rate" of 0.91 narratives per document——a metric that captures how many more major deviation the model made per case compared to the human experts. Taken together, these false positives and false neg-

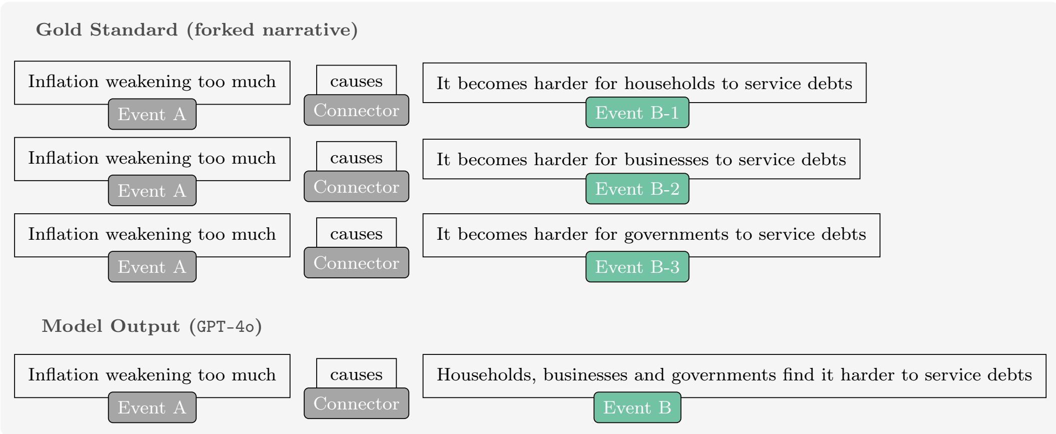


Figure 4: Comparison of a forked narrative (as indicated by the gold standard) and a non-forked narrative (as returned by the LLM).

atives drive down the model's overall accuracy to 44%, compared to the 67–74% accuracy range achieved by individual annotators.

It is important to note that not all errors that we classify as major deviations are created equal. Some of the deviations `GPT4o` produces represent grave misunderstandings and violations of the codebook while others are quite subtle. Occasionally, the economic expertise encoded in the model appears to collide with the aim of faithfully representing the contents of the documents. For example, `GPT-4o` encodes "plans to slow an influx of hard currency that is fueling rapid money-supply growth and pushing up inflation" as a narrative chain, asserting that *money-supply growth* is said to *push up inflation*. While economically plausible, the text clearly asserts the *influx of hard currency* as the cause for inflation. While both variants are related, the model's reading shifts the focal point and chain of causality, resulting in a distorted interpretation of the document.

Measure	**Expert 1**	**Expert 2**	**Expert 3**	**Model**
Major-deviation rate	0.40	0.35	0.49	1.25
Unexpected major deviations	—	—	—	0.91
Accuracy (vs. gold)	0.72	0.74	0.67	0.44
Jaccard similarity	0.59	0.60	0.59	0.40

Table 2: Annotator agreement and deviation metrics for human experts and `GPT-4o`.

In an additional evaluation, we benchmark the narrative extraction capabilities using a Jaccard similarity metric. This merely quantitative approach quantifies lexical overlap between predicted and reference token sets on a scale from 0 (no overlap) to 1 (perfect match). Matching the impressions described earlier, the human experts outperform the model, achieving mean scores across all documents of J = 0.59 or 0.6, whereas the model reaches J = 0.40. This gap again points towards some limitations, despite the measure being rather lenient in nature, as it rewards partial lexical matches and ignores semantic coherence, causal directionality, or narrative plausibility.

6.2. Narrative aggregation

Given the partially promising yet still limited extraction capabilities of `GPT-4o`, any downstream analysis of the extracted narratives must be interpreted with caution. At this stage, results should be seen as indicative rather than conclusive. Nonetheless, as outlined earlier, our goal is to develop a comprehensive narrative extraction framework which also requires us to consider how the extracted data can be processed and aggregated for future applications and econometric analysis. Below, we present the results of this aggregation step, even though its implications remain exploratory at this point.

The abstraction and harmonization of extracted events enable us to compare and group narrative elements across documents. This, in turn, allows for the identification of frequently recurring causal patterns, such as the link between government spending and inflation, or between interest rate hikes and reduced economic growth. Figure 5 illustrates a selection of such simplified narrative arcs that appeared multiple times across our corpus.

One narrative that recurs notably in our sample is the link between government spending and inflation (n=3). This causal association aligns with theoretical arguments suggesting that expansionary fiscal policy—especially if debt-financed—can trigger inflationary pressure, particularly in environments of constrained monetary policy or supply-side rigidity (Sargent, Wallace, et al., 1981). Notably, such narratives are not only central to formal macroeconomic models but also resonate in journalistic discourse (Andre et al., 2024; Schmidt, 2025).

If such narratives appear frequently in the media, they are expected to influence how individuals interpret macroeconomic developments, especially in periods of heightened uncertainty. As discussed in Section 2, narratives are not just post-hoc explanations, but

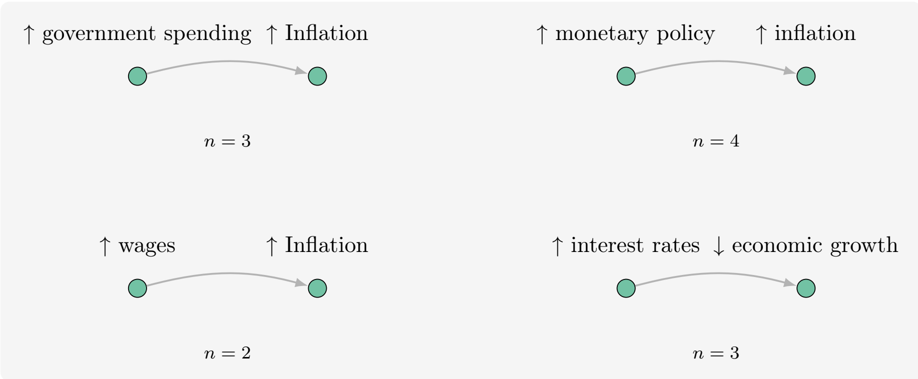


Figure 5: Simplified narratives derived from our LLM based narrative extraction pipeline. "↑" in the context of monetary policy translates to "loose m.p.". n=x indicates how often this narrative can be found in our sample.

sense-making stories that shape how people form expectations and make decisions. This might certainly also apply to the kinds of narratives presented here. For instance, if media audiences are repeatedly exposed to a narrative like "higher wages cause rising inflation" (n=2 in our sample), it is reasonable to assume that such frames may shape their thinking, e.g., influencing how they approach future wage negotiations. Although recent research points in this direction (Andre et al., 2024), it remains unclear how often a reader needs to encounter a given narrative for it to have a measurable effect. With most prominent narrative types in our sample occurring in just 2–4 of 80 articles, the strength and reach of such effects remain speculative and warrant further investigation.

Other frequently recurring narratives include the effect of loose monetary policy on inflation rates (n=4)—which is also a prominent argument in both theory and public discussion—, and interest rate hikes on economic growth (n=3). All four of these narrative arcs point toward classical macroeconomic concerns that have been widely discussed in both academic and policy debates. The most frequently occurring narrative-parts (valence-topic combinations) in our sample are presented in Table 3.

Further analysis of the directional roles played by different events in the extracted narratives reveals additional information on the structure of economic storytelling in the media. For instance, the topic *stock market* appears predominantly as an "effect event" (90% of all instances) suggesting that (business) journalism often frames stock prices as a barometer reacting to other economic developments. In contrast, narratives in which stock market movements are framed as a causal force influencing other domains are relatively rare. Conversely, *government spending* is predominantly discussed as a "cause event" (91% of instances) highlighting its frequent use as an explanatory device in inflation reporting.

Event (example)	**Valence**	**Sym.**	**Topic**	**Count**
higher inflation	rising	↑	Inflation	58
economic stability in the us	positive	↑	Economy	29
higher interest rates	rising	↑	Interest rates	27
the Fed's aggressiveness	loose	↑	Monetary policy	25
the ECB tightening its m.p.	tight	↓	Monetary policy	23
economy plunges toward hard landing	falling	↓	Economy	22
lower inflation rates	falling	↓	Inflation	21
higher gas prices	rising	↑	Energy prices	19
stocks are getting attractive again	rising	↑	Stock Market	15

Table 3: Most frequent valence-topic combinations in extracted narratives. The presented *Events* (left column) are sampled from all events that translate to the respective valence-topic combination.

This asymmetry mirrors typical theoretical priors in economics: while public spending is often viewed as a policy lever that sets economic processes in motion, the stock market tends to be treated as an outcome variable, shaped by shifts in policy, sentiment, or macroeconomic conditions.

However, the aggregation process also reveals important limitations. Despite careful normalization and abstraction, many extracted narratives remain highly context-specific and cannot be assigned to larger clusters. This is partly due to the large combinatorial space of possible valence-topic pairs: the number of possible combinations of clusters A and B and valences A and B is very high, leading to overall rare instances where different texts lead to the same simplified narrative statement. After decomposition and standardization, our 80 evaluation documents yielded a total of 409 distinct narratives, all made up of two valences and two events. Many of them only appear once in the whole dataset, which explains why we count only two or three instances of our most prominent complete narratives.

This lack of convergence, however, is not considered as a methodological weakness. Instead, it reflects a central feature of narratives as they appear in public discourse, and also in the media. Most newspaper articles tend to convey causal claims that are tailored to specific events, actors, and settings, making generalization inherently difficult. Additionally, our strict definition of narratives as precise, two-event causal claims may contribute to this fragmentation, as it emphasizes nuance and specificity over generalization. It is also important to note that not all causal relationships extracted by the model are economically meaningful. Some narratives form around very specific events, locations or protagonists, which cannot be translated to a broader economic category and may not even be of notable relevance.

All in all, our results highlight both the potential and the limitations of LLM-based narrative aggregation and clustering. While the described framework is capable of detecting economically plausible causal structures and producing interpretable output, further methodological refinement is needed to move from fragmented extractions toward more robust generalizations.

7. Discussion

Our study provides evidence that general-purpose language models can be meaningfully used to extract economic narratives from text. The results achieved by `GPT-4o` on our complex annotation task are promising in that they show considerable improvement compared to earlier attempts using previous generations of LLMs. Identifying economic narratives from text requires abstract reasoning, domain-specific knowledge, and the ability to resolve ambiguity and co-reference. We show that a general-purpose, instruction-tuned LLM is able to approach this task with some consistency without any task-specific fine-tuning. This illustrates meaningful development in the field of narrative economics. The progress we observe here reflects ongoing advances in language model capabilities, particularly in tasks that require instruction-following and context-sensitive reasoning.

We also see considerable promise in combining LLM-based extraction with scalable methods of abstraction and clustering. Our initial experiments show that harmonizing extracted narratives into macroeconomic categories enables the identification of recurring narrative patterns. Although most extracted narratives remain too specific to be aggregated meaningfully, the potential for scalable aggregation is clear. Recent work by Reccius (2025, forthcoming), for example, introduces a polar embedding-based approach that could be used to measure the distance between narratives along theoretically meaningful dimensions. Integrating such techniques with LLM-based extraction appears promising for both theoretical and empirical work on economic narratives.

Our findings also reflect some critical shortcomings of LLMs. While tracing the exact origins of each LLM-misstep is currently not possible, it is worth remembering that LLMs are pre-trained to produce coherent text and then fine-tuned to follow instructions. There is some evidence that, as a side effect of this training, models occasionally try to oversatisfy prompts, placing helpfulness over accuracy and thus avoiding rejections (Chen et al., 2025). This tendency results in the model trying to 'squeeze blood from a stone', yielding outputs that are unfaithful to instructions. These drawbacks are especially problematic in research settings and must be addressed.

Importantly, our evaluation approach also highlights the challenges of defining and operationalizing the concept of an economic narrative in practice. Even among expert annotators, we observed substantial variation in annotator opinion about which narratives were identified, how events were segmented, and how causal relations were interpreted. This variability stresses the subjective nature of the sense-making process that occurs when people consume narratives through media. The extraction of economic narratives from text is inherently interpretive, even when formal definitions and structured codebooks are applied. Consequently, fully automating the process remains an ambitious goal.

Looking ahead, we envision a future in which economic narrative research will increasingly be shaped by hybrid workflows—pairing the scalability of language models with the interpretive capacity of human experts. Rather than replacing researchers, LLMs will act as amplifiers of their judgment. Human coders will remain central in defining what constitutes a narrative in the first place, developing precise annotation frameworks, and crafting high-quality prompts that translate conceptual definitions into machine-readable tasks. One key component of this collaboration will be the creation of exemplary few-shot examples, which LLMs rely on for task adaptation.

Accordingly, the time saved by automating the annotation of thousands of documents will not come without cost. It will need to be reinvested in the evaluation process itself. Unlike more standardized tasks such as sentiment classification, narrative extraction requires a deep understanding of semantics, context, and domain-specific framing to evaluate if a models' results are valid. That is, quick eyeballing or resorting to simple accuracy scores will not suffice. Instead, each output must be assessed on its interpretive merit, demanding a level of scrutiny that (for now) only human coders can provide. In that sense, our study underscores the continued importance of human involvement in content analysis and social science research (Haim et al., 2023). LLMs may extend what is possible; but only when embedded in workflows that preserve interpretive oversight and methodological rigor.

To conclude, our results suggest that while narrative extraction via LLMs is still in its early stages, it offers potential for scalable, interpretable, and theoretically informed analysis of economic discourse. With further refinement, including improved prompts, fine-tuned models, and downstream validation, the approach has the potential to contribute meaningfully to the emerging field of narrative economics.

7.1. Outlook

While LLMs will further develop and excel even more at language understanding in the future, we believe that the performance of the current generation of models can be improved on with more specialized training, such as (parameter-efficient) fine-tuning. As best practices for prompting evolve alongside model architectures, we see a need to continuously refine both our extraction and aggregation techniques.

The aggregation framework presented in this paper reflects only an initial stage. Future work should explore more advanced methods of narrative clustering and abstraction. A promising direction is offered by Reccius (2025, forthcoming), who proposes a polar embedding-based approach to improve the alignment of semantically similar narrative elements. Incorporating such techniques may allow us to systematically group related narratives across documents, even when surface-level variation is high. Once greater reliability in extraction and aggregation is achieved, a natural next step is to scale the method to larger corpora. This would enable a more systematic assessment of narrative prevalence and diffusion over time and across domains.

However, while the performance of `GPT-4o` on our task showed that Large Language Models have great potential for narrative extraction, using them requires good hardware or financial backing. In addition to that, the environmental impact of running such models should not be understated. We therefore also aim to develop methods that cleverly combine scalable NLP methods with LLMs, enabling a thorough analysis of large corpora with a minimal amount of resources.

Ultimately, this paper focused on establishing the feasibility of LLM-based narrative extraction. Future research should move toward substantive applications. In particular, linking extracted narratives to external economic indicators—such as inflation expectations, consumer sentiment, or financial market volatility—could offer novel insights into the role of narrative framing in macroeconomic dynamics. Such connections would not only extend the methodological contribution of this paper, but also advance our theoretical understanding of how economic stories shape the economy itself.

8. Conclusion

Narratives play an crucial role in everyday economic decision-making. All actors in an economy, regardless of their level of sophistication or macroeconomic importance, are influenced by the economic narratives circulating in their environment. Extracting such narratives from mass media enables us to trace their origins, understand how they spread through society, and examine their development over time. For economists, this is a crucial yet complex task.

To enable a quantitative analysis of economic narratives, we present an annotation codebook as a stepping stone into extracting economic narratives from a corpus of texts. We utilize this codebook to let three expert annotators code narratives from a total of 100 documents and form a gold-standard data set out of these coded narratives. We then prompt `GPT-4o` to perform the same task and evaluate it on our gold-standard annotations.

We categorize deviations from the gold-standard annotations into minor and major deviations, where minor deviations cover formal or small subjective deviations and major deviations cover cases in which a narrative has been coded incorrectly. We also compare the resulting deviations to the ones that resulted from our expert annotators.

Our results indicate that the level of language comprehension and economic expertise encoded in contemporary Large Language Models (LLMs) enables them to extract economic narratives from newspaper articles effectively. We further find that few-shot Chain-of-Thought prompting—a state-of-the-art AI strategy—is well-suited for LLM-based text retrieval tasks in economics, as it combines high performance with strong transparency in complex annotation tasks.

Our findings also underscore some limitations. Most notably, `GPT-4o` exhibits biases in narrative density, at least when prompted as presented, under-identifying narratives in some contexts while over-identifying them in others. The model struggles with structurally complex narratives, particularly those involving nuanced causal structures like forks or chains. It also shows limited adaptability when encountering challenging edge cases or hidden and implicit causal links between events that may span multiple sentences.

These shortcomings suggest that, while LLMs possess strong general capabilities, economic narrative extraction is a specialized task where expert judgment cannot (and should not) be fully replaced by general-purpose LLMs. Achieving human-level reliability will require more capable foundational models and methodological refinements.

Acknowledgments

This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University and the Narrative Economic Alliance Ruhr (NEAR) project, supported by the Mercator Research Center Ruhr (MERCUR) with project number Ko-2022-0015. It was also partially funded by the Reality Check incubator project at the Research Center for Trustworthy Data Science and Security.

References

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart (2024). *Narratives about the Macroeconomy.* SAFE Working Paper 426. DOI: 10.2139/ssrn.4947636.

Ash, Elliott, Germain Gauthier, and Philine Widmer (2024). "Relatio: Text Semantics Capture Political and Economic Narratives". In: *Political Analysis* 32.1, pp. 115–132. DOI: 10.1017/pan.2023.8.

Bachmann, Rüdiger, Tim O Berg, and Eric R Sims (2015). "Inflation expectations and readiness to spend: Cross-sectional evidence". In: *American Economic Journal: Economic Policy* 7.1, pp. 1–35. DOI: 10.1257/pol.20130292.

Bénabou, Roland, Armin Falk, and Jean Tirole (2018). *Narratives, Imperatives, and Moral Reasoning.* NBER Working Paper 24798. DOI: 10.3386/w24798.

Benner, Niklas, Kai-Robin Lange, and Carsten Jentsch (2022). *Named Entity Narratives.* Ruhr Economic Papers 962. DOI: 10.4419/96973126.

Blinder, Alan S, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen (2008). "Central bank communication and monetary policy: A survey of theory and evidence". In: *Journal of economic literature* 46.4, pp. 910–45. DOI: 10.1257/jel.46.4.910.

Bornheim, Tobias, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski (2024). "Speaker Attribution in German Parliamentary Debates with QLoRA-adapted Large Language Models". In: *Journal for Language Technology and Computational Linguistics* 37.1, pp. 1–13. DOI: 10.21248/jlcl.37.2024.244.

Bracha, Anat and Jenny Tang (2025). "Inflation levels and (in) attention". In: *Review of Economic Studies* 92.3, pp. 1564–1594. DOI: 10.1093/restud/rdae063.

Brown, Tom B. et al. (2020). "Language models are few-shot learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Curran Associates Inc. ISBN: 9781713829546. DOI: 10.5555/3495724.3495883.

Burch, Susan W and Diane Werneke (1975). "The stock of consumer durables, inflation, and personal saving decisions". In: *The Review of Economics and Statistics*, pp. 141–154.

Burla, Laila, Birte Knierim, Jurgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel (2008). "From text to codings: intercoder reliability assessment in qualitative content analysis". In: *Nursing research* 57.2, pp. 113–117. DOI: 10.1097/01.NNR.0000313482.33917.7d.

Chen, Shan, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman (2025). "When Helpfulness Backfires: LLMs and the Risk of Misinformation Due to Sycophantic Behavior". In: *Research Square*, rs–3.

Coibion, Olivier and Yuriy Gorodnichenko (2015). "Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation". In: *American Economic Journal: Macroeconomics* 7.1, pp. 197–232. DOI: 10.1257/mac.20130306.

Conrad, Christian, Zeno Enders, and Alexander Glas (2022). "The role of information and experience for households' inflation expectations". In: *European Economic Review* 143, p. 104015. DOI: 10.1016/j.euroecorev.2021.104015.

Crow, Deserai and Michael Jones (2018). "Narratives as tools for influencing policy change". In: *Policy & Politics* 46.2, pp. 217–234. DOI: 10.1332/030557318X15230061022899.

DeepSeek-AI (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dubey, Abhimanyu et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783v2.

Eliaz, Kfir and Ran Spiegler (2020). "A model of competing narratives". In: *American Economic Review* 110.12, pp. 3786–3816. DOI: 10.1257/aer.20191099.

Eliaz, Kfir and Ran Spiegler (2024). "News media as suppliers of narratives (and information)". In: arXiv: 2403.09155 [econ.TH].

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li (2024). *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models*. arXiv: 2405.06211 [cs.CL].

Fiore, Fiorella De, Alexis Maurin, Andrej Mijakovic, and Damiano Sandri (2025). *Monetary policy in the news: The FOMC's media coverage and inflation expectations*. URL: https://cepr.org/voxeu/columns/monetary-policy-news-fomcs-media-coverage-and-inflation-expectations (visited on 06/02/2025).

Flynn, Joel P. and Karthik A. Sastry (2024). *The Macroeconomics of Narratives*. NBER Working Paper 32602. National Bureau of Economic Research. DOI: 10.3386/w32602.

Gehring, Kai and Matteo Grigoletto (2023). *Analyzing Climate Change Policy Narratives with the Character-Role Narrative Framework*. CESifo Working Paper 10429. DOI: 10.2139/ssrn.4456361.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences* 120.30. ISSN: 1091-6490. DOI: 10.1073/pnas.2305016120.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (Feb. 2023). "The Voice of Monetary Policy". In: *American Economic Review* 113.2, pp. 548–584. DOI: 10.1257/aer.20220129.

Gueta, Almog, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart (2025). "Can LLMs Learn Macroeconomic Narratives from Social Media?" In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, pp. 57–78. ISBN: 979-8-89176-195-7. URL: https://aclanthology.org/2025.findings-naacl.4/.

Gurkaynak, Refet S., Brian P. Sack, and Eric T. Swanson (2005). "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements". In: *International Journal of Central Banking* 1, pp. 55–93. DOI: 10.2139/ssrn.633281.

Haim, Mario, Valerie Hase, Johanna Schindler, Marko Bachl, and Emese Domahidi (2023). "(Re) Establishing quality criteria for content analysis: A critical perspective on the field's core method". In: *Studies in Communication and Media (SCM)* 12, pp. 277–288.

Han, Di, Wei Guo, Han Chen, Bocheng Wang, and Zikun Guo (2024). "LEST: Large language models and spatio-temporal data analysis for enhanced Sino-US exchange rate forecasting". In: *International Review of Economics & Finance* 96, p. 103508.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2017). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach". In: *The Quarterly Journal of Economics* 133.2, pp. 801–870. DOI: 10.1093/qje/qjx045.

Hansen, Stephen, Michael McMahon, and Matthew Tong (2019). "The long-run information effect of central bank communication". In: *Journal of Monetary Economics* 108, pp. 185–202. DOI: 10.1016/j.jmoneco.2019.09.002.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu (Nov. 9, 2023). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2, pp. 1–55. ISSN: 1558-2868. DOI: 10.1145/3703155. arXiv: 2311.05232 [cs.CL].

Jeong, Minhyuk and Kwangwon Ahn (2025). "Energy organization sentiment and oil return forecast". In: *Energy Economics* 141, p. 108105.

Jones, Michael D. and Mark K. McBeth (2010). "A Narrative Policy Framework: Clear Enough to Be Wrong?" In: *Policy Studies Journal* 38.2, pp. 329–353. DOI: 10.1111/j.1541-0072.2010.00364.x.

Juster, F Thomas, Paul Wachtel, Saul Hymans, and James Duesenberry (1972). "Inflation and the Consumer". In: *Brookings Papers on Economic Activity* 1972.1, pp. 71–121. DOI: 10.2307/2534115.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2020). "Making text count: economic forecasting using newspaper text". In: *Journal of Applied Econometrics*.

King, Mervyn and John Kay (2020). *Radical uncertainty: Decision-making for an unknowable future*. Hachette UK.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). “Large language models are zero-shot reasoners”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3601883.

Korinek, Anton (2023). “Generative AI for Economic Research: Use Cases and Implications for Economists”. In: *Journal of Economic Literature* 61.4, pp. 1281–1317. ISSN: 0022-0515. DOI: 10.1257/jel.20231736.

Lamla, Michael J and Sarah M Lein (2014). “The role of media for consumers’ inflation expectation formation”. In: *Journal of Economic Behavior & Organization* 106, pp. 62–77.

Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch (2022a). *Towards Extracting Collective Economic Narratives from Texts*. Ruhr Economic Papers 963. DOI: 10.4419/96973127.

Lange, Kai-Robin, Jonas Rieger, Niklas Benner, and Carsten Jentsch (2022b). “Zeitenwenden: Detecting changes in the German political discourse”. In: *Proceedings of the 2nd Workshop on Computational Linguistics for the Political and Social Sciences*, pp. 47–53. URL: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf.

Lange, Kai-Robin, Jonas Rieger, and Carsten Jentsch (2024). “Lex2Sent: A bagging approach to unsupervised sentiment analysis”. In: *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*. Association for Computational Linguistics, pp. 281–291. URL: https://aclanthology.org/2024.konvens-main.28/.

Lange, Kai-Robin, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025). “Narrative Shift detection: A hybrid approach of Dynamic Topic Models and Large Language Models”. In: *Proceedings of the Text2Story’25 Workshop*. URL: https://www.di.ubi.pt/~jpaulo/Text2Story2025/paper6.pdf.

Larsen, Vegard H and Leif Anders Thorsrud (2019). *Business cycle narratives*. CESifo Working Paper 7468. DOI: 10.2139/ssrn.3338822.

OpenAI (2024). URL: https://openai.com/index/learning-to-reason-with-llms/ (visited on 05/07/2025).

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). “Lost in the Middle: How Language Models Use Long Contexts”. In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173. DOI: 10.1162/tacl_a_00638. URL: https://aclanthology.org/2024.tacl-1.9/.

Mathew, Binny, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier (2020). “The POLAR Framework: Polar Opposites Enable Interpretability of Pre-Trained Word Embeddings”. In: *Proceedings of The Web Conference 2020*. WWW ’20. Taipei, Taiwan:

Association for Computing Machinery, pp. 1548–1558. ISBN: 9781450370233. DOI: 10.1145/3366423.3380227.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman (2024). "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale". In: *Research & Politics* 11.1. DOI: 10.1177/20531680241231468.

Mondorf, Philipp and Barbara Plank (2024). "Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 9370–9402. DOI: 10.18653/v1/2024.acl-long.508.

Neidhardt, Friedhelm (2010). "Selbststeuerung der Wissenschaft: Peer Review". In: *Handbuch Wissenschaftspolitik*, pp. 280–292. DOI: 10.1007/978-3-531-91993-5_19.

Nyman, Rickard, Sujit Kapadia, and David Tuckett (2021). "News and narratives in financial systems: exploiting big data for systemic risk assessment". In: *Journal of Economic Dynamics and Control* 127, p. 104119. DOI: 10.1016/j.jedc.2021.104119.

OpenAI et al. (2024a). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

OpenAI et al. (2024b). *GPT-4o System Card*. arXiv: 2410.21276 [cs.CL].

Pearl, Judea (2009). *Causality*. Cambridge University Press. ISBN: 9781139632997.

Reis, Ricardo (2006). "Inattentive producers". In: *The Review of Economic Studies* 73.3, pp. 793–821. DOI: 10.1111/j.1467-937X.2006.00396.x.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2021). "RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, pp. 2337–2347. DOI: 10.18653/v1/2021.findings-emnlp.201.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). "Dynamic change detection in topics based on rolling LDAs". In: *Proceedings of the Text2Story'22 Workshop*, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/paper1.pdf.

Roos, Michael and Matthias Reccius (2024). "Narratives in Economics". In: *Journal of Economic Surveys* 38.2, pp. 303–341. DOI: 10.1111/joes.12576.

Saldaña, Johnny (2016). *The coding manual for qualitative researchers*. 3E. SAGE. ISBN: 9781473902480.

Sargent, Thomas J, Neil Wallace, et al. (1981). "Some unpleasant monetarist arithmetic". In: *Federal reserve bank of minneapolis quarterly review* 5.3, pp. 1–17.

Schlaufer, Caroline, Johanna Kuenzler, Michael D Jones, and Elizabeth A Shanahan (2022). "The narrative policy framework: a traveler's guide to policy stories". In: *Politische Vierteljahresschrift* 63.2, pp. 249–273. DOI: 10.1007/s11615-022-00379-6.

Schmidt, Tobias (2025). *Narrating inflation: How German economic journalists explain post-covid price rises.* DoCMA Working Paper 14. DOI: 10.17877/DE290R-25380.

Schmidt, Torsten, Henrik Müller, Jonas Rieger, Tobias Schmidt, and Carsten Jentsch (2023). *Inflation perception and the formation of inflation expectations.* Ruhr Economic Papers 1025. DOI: 10.4419/96973191.

Shahriar, Sakib, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool (2024). *Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency.* arXiv: 2407.09519v1.

Shanahan, Elizabeth A, Michael D Jones, Mark K McBeth, and Claudio M Radaelli (2018). "The narrative policy framework". In: *Theories of the policy process.* Routledge, pp. 173–213. DOI: 10.4324/9780429494284.

Shiller, Robert J. (2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004. ISSN: 0002-8282. DOI: 10.1257/aer.107.4.967.

Sims, Christopher A (2003). "Implications of rational inattention". In: *Journal of monetary Economics* 50.3, pp. 665–690. DOI: 10.1016/S0304-3932(03)00029-1.

Song, Yisheng, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo (July 2023). "A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities". In: *ACM Computing Surveys* 55.13s, pp. 1–40. ISSN: 1557-7341. DOI: 10.1145/3582688.

Ter Ellen, Saskia, Vegard H Larsen, and Leif Anders Thorsrud (2022). "Narrative monetary policy surprises and the media". In: *Journal of Money, Credit and Banking* 54.5, pp. 1525–1549.

Tian, Runchu, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang, and Xiaojiang Liu (2024). *Distance between Relevant Information Pieces Causes Bias in Long-Context LLMs.* arXiv: 2410.14641 [cs.CL].

Tuckett, David, Douglas Holmes, Alice Pearson, and Graeme Chaplin (2020). *Monetary policy and the management of uncertainty: a narrative approach.* Bank of England working papers 870. Bank of England. URL: https://ideas.repec.org/p/boe/boeewp/0870.html.

Wang, Guoqing, Zeyu Sun, Zhihao Gong, Sixiang Ye, Yizhou Chen, Yifan Zhao, Qingyuan Liang, and Dan Hao (2024). "Do advanced language models eliminate the need for prompt engineering in software engineering?" In: arXiv: 2411.02093 [cs.SE].

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (Sept. 3, 2021). *Finetuned Language Models Are Zero-Shot Learners.* Tech. rep. arXiv: 2109.01652 [cs.CL].

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-thought prompting elicits reasoning in large language models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3602070.

Werning, Iván (2022). *Expectations and the Rate of Inflation*. NBER Working Paper 30260. DOI: 10.3386/w30260.

Yang, Chenyang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu (2025). *What Prompts Don't Say: Understanding and Managing Underspecification in LLM Prompts*. Tech. rep. arXiv: 2505.13360 [cs.CL]. URL: https://arxiv.org/abs/2505.13360.

Zhang, Haopeng, Philip S. Yu, and Jiawei Zhang (2025). "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models". In: *ACM Comput. Surv.* Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3731445.

Zhang, Hugh, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan (2021). "Trading Off Diversity and Quality in Natural Language Generation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Ed. by Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina. Online: Association for Computational Linguistics, pp. 25–33. URL: https://aclanthology.org/2021.humeval-1.3/.

Zhong, Meizhi, Chen Zhang, Yikun Lei, Xikai Liu, Yan Gao, Yao Hu, Kehai Chen, and Min Zhang (2025). "Understanding the RoPE Extensions of Long-Context LLMs: An Attention Perspective". In: *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 8955–8962. URL: https://aclanthology.org/2025.coling-main.600.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang (2024). "Can Large Language Models Transform Computational Social Science?" In: *Computational Linguistics* 50.1, pp. 237–291. DOI: 10.1162/coli_a_00502.

Appendix

A. Codebook

1. Basic Idea

```
The goal of this coding task is to extract all narratives from the provided
texts as accurately and comprehensively as possible.  The texts are excerpts
from newspaper articles published in the NYT or the Wall Street Journal.[a]
```

We define a narrative as a causal connection between two consecutive events.

Simple examples of a narrative:

- Russia's war on Ukraine - causes - People leaving Ukraine
- rising prices - causes - lifting of global bond yields
- rapid money-supply growth - is caused by - influx of hard currency

These narratives are sometimes explicitly mentioned in the text and easy to identify. Often, however, they require some interpretation and abstraction. That is why it is crucial for our research that multiple coders read and annotate the same texts. This ensures that results are not biased by individual interpretations and that the extracted narratives are recognizable by other coders as well.

2. Target Format

We aim to train a language model to reduce all narratives in a text to one of two possible formats:

- Format A: event 1 - causes - event 2
- Format B: event 1 - is caused by - event 2

Each narrative must follow one of these formats. The distinction between A and B preserves the order of events as they appear in the original text. The event mentioned first must be event 1 in your coded narrative.

Examples:

- *Source text:* The Titanic struck an iceberg on 14 April 1912. As a result, the Titanic sank.
  the Titanic struck an iceberg - causes - the Titanic sank
- *Source text:* People have a hard time finding jobs because the Fed keeps interest rates high.
  people have a hard time finding jobs - is caused by - the Fed keeps interest rates high

These examples show: The narrative does not have to be a coherent sentence. The task is to reduce running text into a "bullet point" format that strips away all information not part of the causal chain.

3. Events

3.1 Types of Events

Events can be categorized into events and activities as well as states and circumstances.

- "the collapse of the Berlin Wall in 1989" (event)
- "Mexico grew rapidly" (activity)
- "Trump signed the bill into law" (activity)

- “inflation” (state)
- “lack of competitive innovation” (circumstance)

States and circumstances can also include properties of people or other entities:

- “Stubborn Indian inflation”
- “The prime minister’s arrogance”

Events and activities can also include future events or plans.

- “The ECB will probably raise rates by next year” (plan)
- “There might be more earthquakes in Germany in the future” (future event)

Policy measures also count as events:

- “The Schuldenbremse”
- “The Affordable Care Act”

3.2 Unchanged Coding of Events

Events should generally be coded in the exact form used in the source text. The closer the extracted narrative is to the source text, the faster the model will learn.

Specifically:

- Do not use synonyms for terms used in the text
- Keep the same verb tense (present stays present, past stays past, etc.)
- Do not abstract semantically or economically (e.g., do not interpret “rising prices” as “inflation”)
- Exception: Explanatory clauses (e.g., appositions) can usually be left out (see 3.5)

3.3 Coreference Resolution

An exception to 3.2 applies when entities (e.g., people, countries) are only indirectly mentioned.

Example:
“She talked a lot about North Korea. ’The country is really poor, therefore many citizens suffer from hunger’, she said.”
-> Instead of coding “The country”, include the resolved entity:

- {The country|North Korea} is really poor - causes - citizens suffer from hunger

Similarly:
“President Biden discussed the matter today. He downplayed the probability of a government shutdown which has calmed the markets.”

- {He|President Biden} downplayed the probability of a government shutdown - causes - the markets were calmed

3.4 Opinions

Sometimes a causal link is part of a third-party opinion or assessment. In most cases, the narrative refers to the content of the opinion, not the act of expressing it.

Example:
"Analysts believe that bond yields might rise even further which would put even more strain on pension funds."

- bond yields might rise even further - causes - more strain on pension funds

But in some cases, the act of expression itself is causal:

"ECB president Lagarde made it clear that no further rate raises were on the horizon. Markets responded calmly."

- ECB president Lagarde made it clear that no further rate raises were on the horizon - causes - Markets responded calmly

3.5 Parentheses and Side Clauses

Explanatory side information set off by commas, dashes, or parentheses should be left out unless essential for understanding the narrative.

Example:
"The highest inflation in decades and the war in Ukraine - the first major military conflict in Europe in over 30 years - raised concerns..."

- highest inflation in decades - causes - concerns about a possible recession
- war in Ukraine - causes - concerns about a possible recession

3.6 Chained Events

Some statements include multiple events connected by "and" or "as well as". These should be split into separate narratives.

Example:
"She wants you to forget that federal spending contributed to soaring prices, as well as to the labor shortages across the economy."

- federal spending - causes - soaring prices
- federal spending - causes - labor shortages across the economy

4. Only Positive Causal Links

Please code only positive causal relationships. Ignore negative causality.

Example:
Putin invading Ukraine - does not cause - President Zelensky fleeing Ukraine *(do not code)*

5. Multiple Narratives per Text

A document may contain multiple narratives. Please write each narrative on a new line. Do not use enumerations.

Correct:
- the prices for cucumbers rose by 100% this month - causes - people stop buying cucumbers
- huge money supply - causes - prices for real estate go up

Wrong:
1. the prices for cucumbers rose ... | 2. huge money supply ...

6. Thematic Scope

6.1 Explicit Economic Reference

Please only code narratives that have an explicit economic relevance. You can ignore all other narratives. If you are unsure, write down the narrative and add a note (uncertain).

Examples:

- Putin invading Ukraine - causes - rising energy prices *(code)*
- Putin invading Ukraine - causes - enormous military support by the United States *(do not code)*

6.2 Focus on Inflation Narratives

We are primarily interested in inflation narratives. Therefore, please mark all narratives without explicit reference to inflation or price changes with an "(x)". As a reminder: narratives that have no economic relevance at all should not be coded in the first place.

Examples:

- Putin invading Ukraine - causes - disruption of agricultural supply chains (x) *(code)*
- Putin invading Ukraine - causes - gas price increases in Germany *(code)*
- Putin invading Ukraine - causes - enormous military support by the United States *(do not code)*

The example from 3.6 should therefore be:

- federal spending - causes - soaring prices
- federal spending - causes - labor shortages across the economy (x)

7. Preserve Content

This point is similar to point 3.2. Sometimes, there is a strong temptation to omit parts of a sentence in order to make the narrative more pointed. In some cases, that's acceptable; however, the rule should be to include all parts of the sentence that are necessary to preserve the core message of the narrative. Please preserve all essential content in the narrative.

Example 1:

- wage inflation going up - causes - the risk of yields on longer-dated bonds rising at a faster pace compared to those on shorter-dated notes (correct)
- wage inflation going up - causes - longer-dated bonds rising faster than shorter-dated notes (incorrect)

Example 2:

- The highest inflation in decades - causes - concerns about a possible recession (correct)
- inflation - causes - recession (incorrect)

8. Edge Cases

If you are unsure whether something really constitutes a narrative or not, code it anyway to be on the safe side. You can later compare all such cases with the other coders or with us.

Example 1:
"Semi"-causal cues (such as "as", "suggesting", ...) can be interpreted as either temporal or causal.

John Sicher, editor and publisher of Beverage Digest, said that soft-drink sales have shown some improvement this year, estimating that they rose about 3.5% in January as price increases stabilized.

In such cases code the narrative just to be safe, and compare your results with other coders afterward.

Example 2:
If you interpret the "as" in the example above as a causal cue, a second ambiguity arises: you could shorten the passage in two ways:

John Sicher, editor and publisher of Beverage Digest, said that soft-drink sales have shown some improvement this year, estimating that they rose about 3.5% in January as price increases stabilized.

- price increases stabilized - causes - soft-drink sales showing some improvement (x)
- price increases stabilized - causes - soft-drink sales rose about 3.5% in January (x)

Again, if you're unsure which version to choose, write down both to be on the safe side.

```
If unsure, include both options.
```

[a]Since the texts are excerpts, some passages may seem out of context. In most cases, however, the meaning should still be clear.

B. Annotated Examples

2004-05-27 | WSJ | China Targets Auto Loans, Speculators

At the same time, overinvestment in the industry has led to rising car inventories. GM says its own inventory levels are healthy and for some models it has waiting lists. Separately, China's foreign-exchange regulator announced plans to slow an influx of hard currency that is fueling rapid money-supply growth and pushing up inflation. Speculators are betting that Beijing will revalue its currency, the yuan, and are parking money in Chinese banks to take advantage of interest rates higher than those in the U.S.The new rules also aim to reduce a run-up in foreign debt, as some borrowers switch to foreign-currency loans that are cheaper than yuan-denominated loans. As part of the tighter measures to come into force on July 1, foreign-invested companies must state a specific use for any funds greater than $200,000 entering China, said a statement posted on the State Administration of Foreign Exchange's Web site.

Narratives

- overinvestment in the industry - causes - rising car inventories (x)
- an influx of hard currency - causes - rapid money-supply growth
- an influx of hard currency - causes - pushing up inflation
- run-up in foreign debt - is caused by - some borrowers switch to foreign-currency loans that are cheaper than yuan-denominated loans
- The new rules - causes - reducing a run-up in foreign debt. (x)

Model Output

- overinvestment in the car industry - causes - rising car inventories
- an influx of hard currency - causes - rapid money-supply growth
- rapid money-supply growth - causes - pushing up inflation
- Speculators are betting that Beijing will revalue its currency, the yuan - causes - speculators parking money in Chinese banks to take advantage of interest rates higher than those in the U.S.
- foreign-currency loans are cheaper than yuan-denominated loans - causes - some borrowers switch to foreign-currency loans

2023-03-03 | WSJ | Business News: Bud Brewer's Volume Fell as Prices Jumped

And people are eating and drinking more at home, a setting that tends to favor beer over other drinks, he said.̈ Beer remains resilient,M̈r. Doukeris said. Ï'll never say that beer is immune to the inflation and everything that is happening out there, but [it] remains a very resilient category.̈ The brewer has for years seen its mainstream brands Bud and Bud Light decline, partly amid a broader shift by American drinkers from beer toward wine and spirits. In response, it has worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois.

Narratives

- people are eating and drinking more at home, a setting that tends to favor beer over other drinks - causes - Beer remains resilient (x)
- Bud and Bud Light decline - is caused by - a broader shift by American drinkers from beer toward wine and spirits (x)
- Bud and Bud Light decline - causes - {it|the brewer} has worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois (x)

Model Output

- people are eating and drinking more at home - causes - beer being favored over other drinks
- a broader shift by American drinkers from beer toward wine and spirits - causes - Bud and Bud Light decline
- Bud and Bud Light decline - causes - the brewer worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois

2023-03-03 | NYT | Market Gains After 3 Days of Losses

Shares of Salesforce soared 11.5 percent for one of the market's biggest gains after it topped forecasts for profit and revenue last quarter. It also gave a stronger-than-expected forecast for upcoming results. Expectations have been coming down recently for profits at big U.S. companies given still-high inflation and interest rates. But several joined Salesforce in rising Thursday after posting encouraging results. Macy's rose 11.1 percent after reporting stronger profit and revenue for the holidays than analysts expected.

Narratives

- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - {it|Salesforce} topped forecasts for profit and revenue last quarter. (x)
- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - {it|Salesforce} gave a stronger-than-expected forecast for upcoming results (x)
- {it|Salesforce} gave a stronger-than-expected forecast for upcoming results (x) - is caused by - {it|Salesforce} topped forecasts for profit and revenue last quarter. (x)
- Expectations have been coming down recently for profits at big U.S. companies - is caused by - still-high inflation
- Expectations have been coming down recently for profits at big U.S. companies - is caused by - still-high interest rates

- {Several|U.S. companies} joined Salesforce in rising Thursday - is caused by - {U.S. companies} posting encouraging results (x)
- Macy's rose 11.1 percent - is caused by - reporting stronger profit and revenue for the holidays than analysts expected (x)

Model Output

- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - Salesforce topped forecasts for profit and revenue last quarter
- Salesforce gave a stronger-than-expected forecast for upcoming results - causes - Shares of Salesforce soared 11.5 percent for one of the market's biggest gains
- Macy's rose 11.1 percent - is caused by - Macy's reported stronger profit and revenue for the holidays than analysts expected

C. Prompt

```
# Codebook:

## Basic Idea:
I will provide you with an excerpt from newspaper articles that have appeared in
the New York Times or The Wall Street Journal.
Your task is to extract all economic narratives that occur in the excerpt.  This
task is a well-defined multi-step extraction process.  The term "narrative"
is frequently used with different meanings and in a variety of contexts.  The
multi-step extraction process is designed to make sure that your work respects
a very specific definition of economic narratives.  For each step that outlined
below, it is imperative that you only make the changes that are indicated.  I will
provide you with this definition next.

## Definition:
An economic narrative consists of exactly two events and a causal connection that
is asserted between those events.

## Event structures:
In the terminology of causal inference, the definition corresponds to finding the
"direct causal effect" of Event A on Event B. The causal pathways you encounter
in the excerpts may be more complex.  Therefore, in addition to direct causal
effects, find and deconstruct the following causal structures as follows:
1) Fork:  If Event A is said to be a common cause of Events B and C, code both
pathways as separate narratives.
2) Chain:  If Event B is said to be a mediator between Event A and Event C, also
code both pathways as separate narratives.
```

```
## Rules:
You also have to follow a couple of hard-and-fast rules:
1) Retain the order of the two events from the source text at all times.  The
event that appears first in the text must be coded as "Event A", the event that
appears second must be coded as "Event B".
2) When you state the narrative in its "Target Form", always represent the causal
connection by using one of the following phrasings:  i) "causes":  Use this causal
connector when Event A is the cause and Event B is an effect of Event A. ii) "is
caused by":  Use this causal connector when Event B is the cause and Event A is an
effect of Event B.

## Target Forms:
{
"Event A":  "[...]",
"causal connector":  "causes",
"Event B":  "[...]"
}
{
"Event A":  "[...]",
"causal connector":  "caused by",
"Event B":  "[...]"
}

## Extraction Process:
1) "Focused Excerpt":  The excerpts vary in length and in their narrative
density.  Parts of each excerpt may obviously not contain any narrative.  To
focus on relevant parts of the excerpt, start by repeating it, but leaving out the
sentences that you are sure no narrative appears in.
2) "Sequence of Interest":  Go through the "Focused Excerpt" and state the first
narrative sequence you find, that is a sequence that contains two events and what
you consider to be a causal connection between them.
3) "Causal Restatement":  Based on the current "Sequence of Interest", restate the
narrative in the appropriate target form.  Make sure to repeat the events verbatim
without rephrasing.
4) "Coreference Resolution":  If necessary, rephrase one or both events to clearly
identify the entities that occur in the narrative.  For the most part, this will
involve replacing personal pronouns with the entity itself.
5) "Event Rephrasing":  If necessary, rephrase one or both events again so that
the narrative uses correct language.
After every narrative, continue by circling back to 1) and restate the "Focused
Excerpt".  Do so even if you did not previously detect more narratives in it.  Use
it to refocus your attention and double-check if more economic narratives occur in
the excerpt.

# Example:
```

```
## Excerpt:
"[...]"
{
"Focused Excerpt": "[...]",
"Sequence of interest": "[...]",
"Causal Restatement":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
},
"Coreference Resolution":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
},
"Event Rephrasing":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
}
},
{
"Focused Excerpt": [...],
[...]
}

# Your Work
## Excerpt:
"[...]"
```