

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>

This Version is available at:

<https://hdl.handle.net/10419/264398>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



# RUHR

## ECONOMIC PAPERS

Niklas Benner  
Kai-Robin Lange  
Carsten Jentsch

### Named Entity Narratives

## 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

Prof. Dr. Thomas K. Bauer

RUB, Department of Economics, Empirical Economics

Phone: +49 (0) 234/3 22 83 41, e-mail: [thomas.bauer@rub.de](mailto:thomas.bauer@rub.de)

Prof. Dr. Ludger Linnemann

Technische Universität Dortmund, Department of Business and Economics

Economics – Applied Economics

Phone: +49 (0) 231/7 55-3102, e-mail: [Ludger.Linnemann@tu-dortmund.de](mailto:Ludger.Linnemann@tu-dortmund.de)

Prof. Dr. Volker Clausen

University of Duisburg-Essen, Department of Economics

International Economics

Phone: +49 (0) 201/1 83-3655, e-mail: [vclausen@vwl.uni-due.de](mailto:vclausen@vwl.uni-due.de)

Prof. Dr. Ronald Bachmann, Prof. Dr. Manuel Frondel, Prof. Dr. Torsten Schmidt,

Prof. Dr. Ansgar Wübker

RWI, Phone: +49 (0) 201/81 49-213, e-mail: [presse@rwi-essen.de](mailto:presse@rwi-essen.de)

### Editorial Office

Sabine Weiler

RWI, Phone: +49 (0) 201/81 49-213, e-mail: [sabine.weiler@rwi-essen.de](mailto:sabine.weiler@rwi-essen.de)

### Ruhr Economic Papers #962

Responsible Editor: Torsten Schmidt

All rights reserved. Essen, Germany, 2022

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

The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors.

---

## **Ruhr Economic Papers #962**

Niklas Benner, Kai-Robin Lange, and Carsten Jentsch

# **Named Entity Narratives**

## **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<sup>1</sup>

# 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

---

<sup>1</sup> 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 (LiD03), 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 Harris 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<sup>1</sup>. 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.

---

<sup>1</sup><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

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

where  $\pi$  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  $\rho$  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

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

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 occurred 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<sup>2</sup>, these together form an event.

---

<sup>2</sup>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

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

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  $\beta$  days. A structural break differs from an event due to the next  $\gamma$  days, in which the incident still affects the narrative. This time line is shown in Figure 1

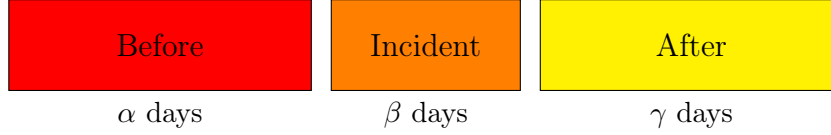


Figure 1: Life span of a narrative defined by a structural break. After the incident, the narrative still affects the word frequencies for  $\gamma$  days.

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

$$\Delta_t = S_C(\text{event}, \text{after}) - S_C(\text{event}, \text{before}) \quad (2)$$

$$= S_C\left(\sum_{i=t}^{t+\beta-1} \Gamma_i, \sum_{i=t+\beta}^{t+\beta+\gamma-1} \Gamma_i\right) - S_C\left(\sum_{i=t}^{t+\beta-1} \Gamma_i, \sum_{i=t-\alpha}^{t-1} \Gamma_i\right) \quad (3)$$

where  $\Gamma_i$  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.

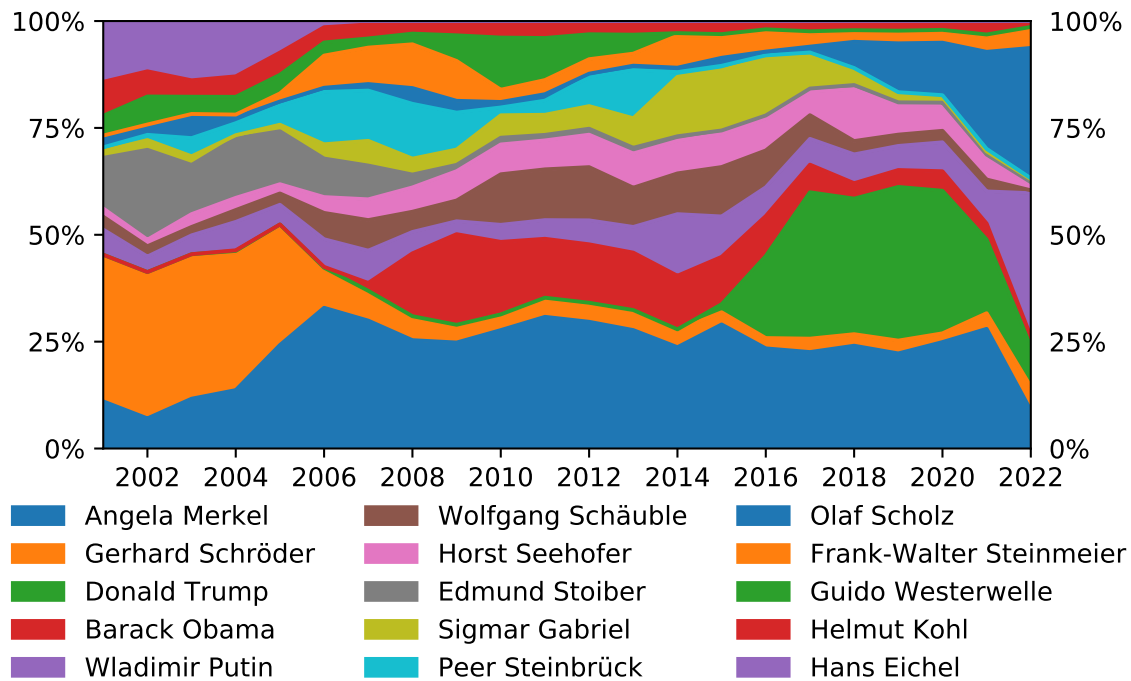


Figure 2: Relative occurrences of the 15 most frequent named entities over the entire data set.

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 least 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).

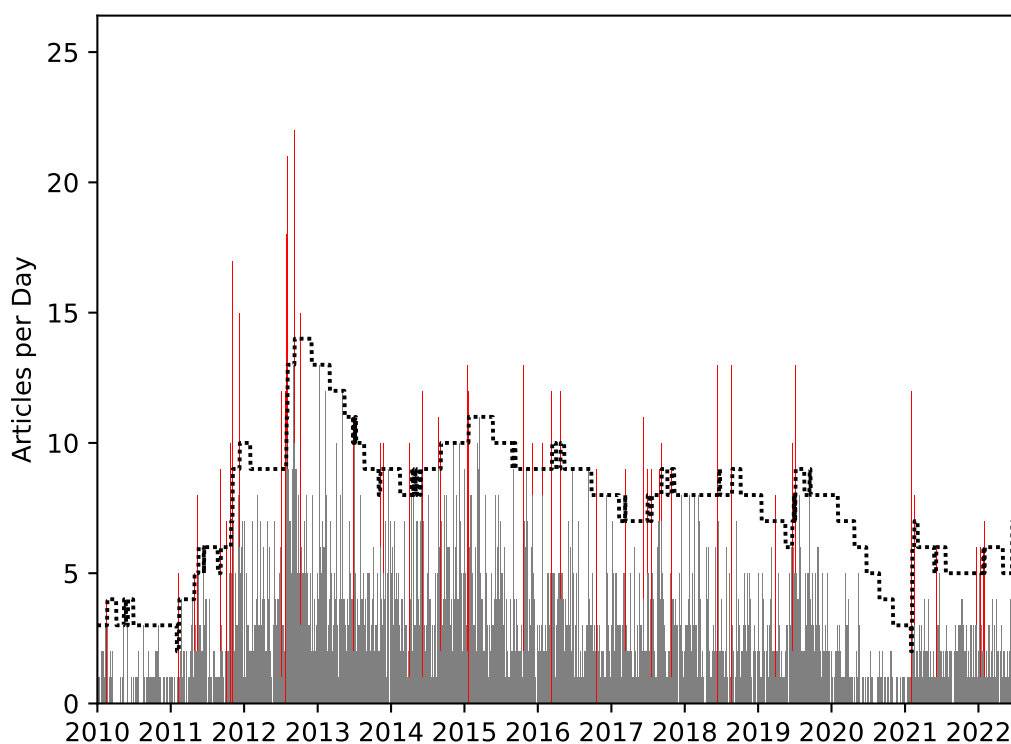


Figure 3: Number of articles mentioning "Mario Draghi" per day

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  $\Delta_t$  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.

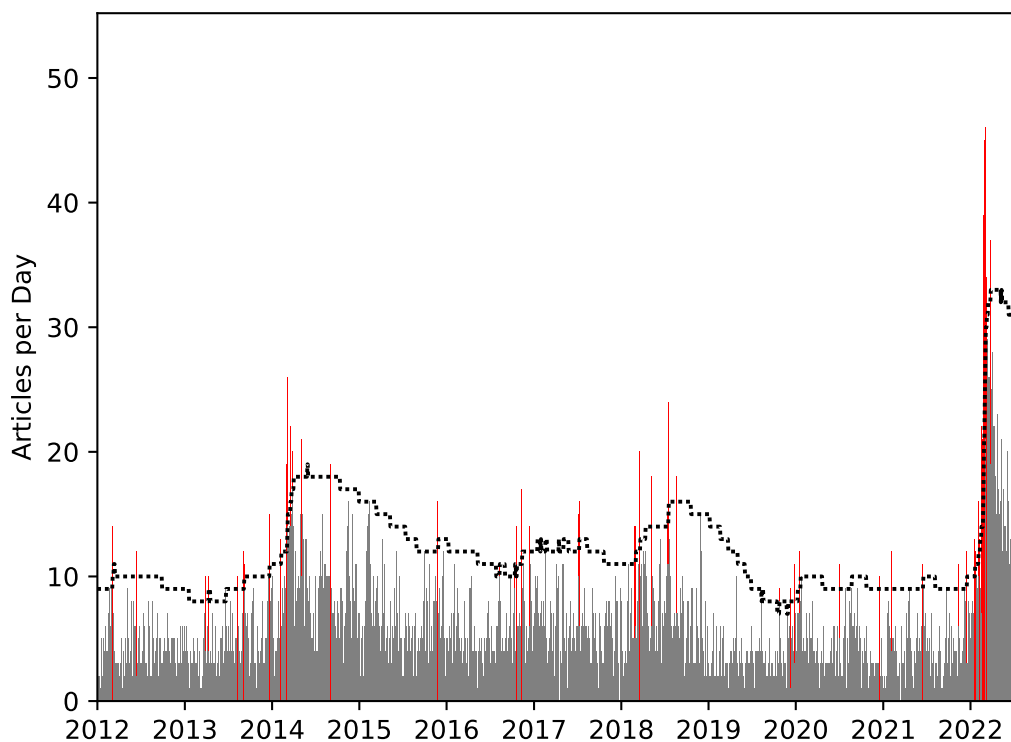


Figure 4: Number of articles mentioning "Wladimir Putin" per day

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 occurred 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

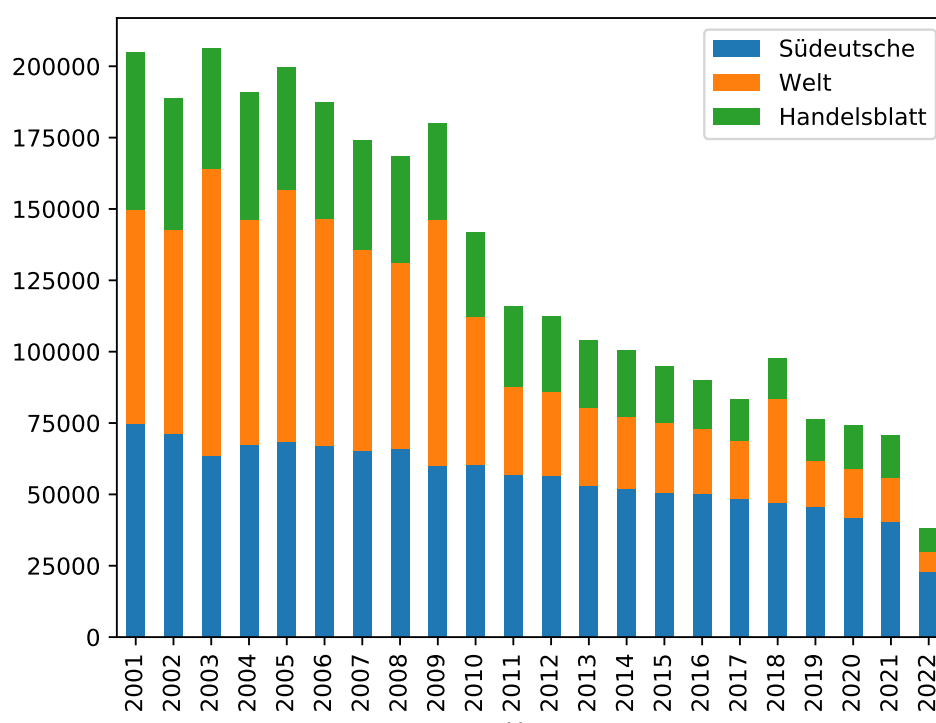


Figure 5: Number of articles published by "Süddeutsche", "Welt" and "Handelsblatt".

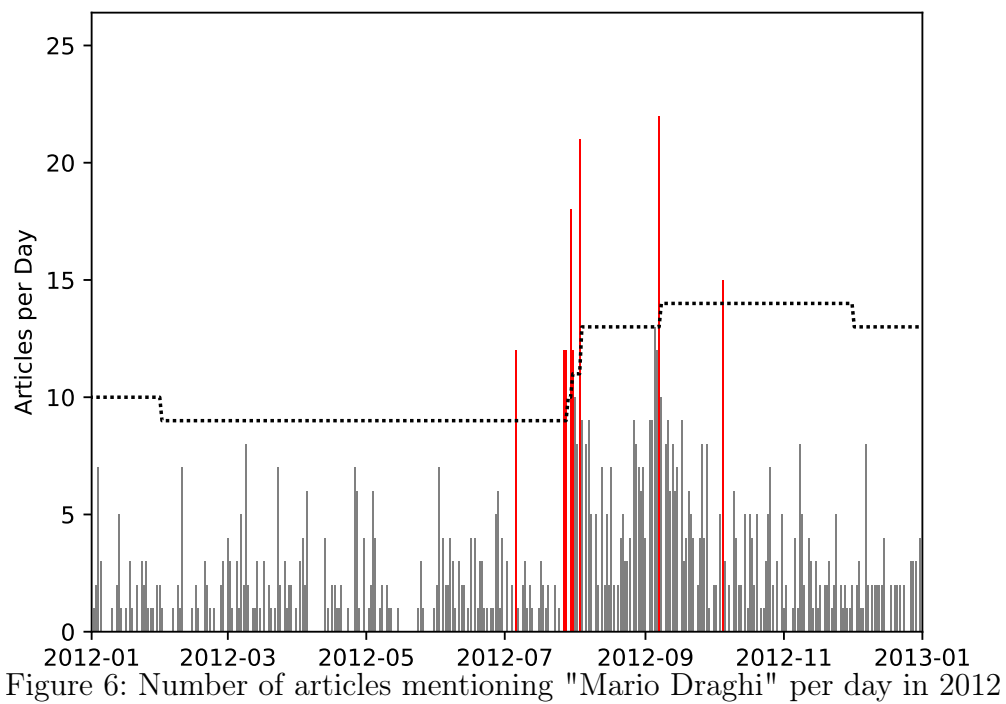


Figure 6: Number of articles mentioning "Mario Draghi" per day in 2012

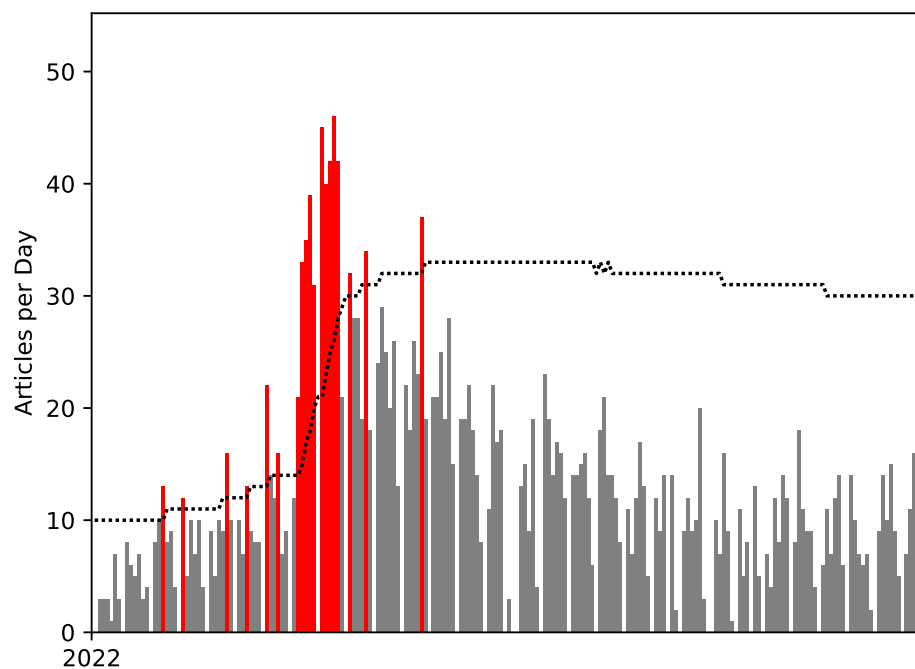


Figure 7: Number of articles mentioning "Wladimir Putin" per day in 2022