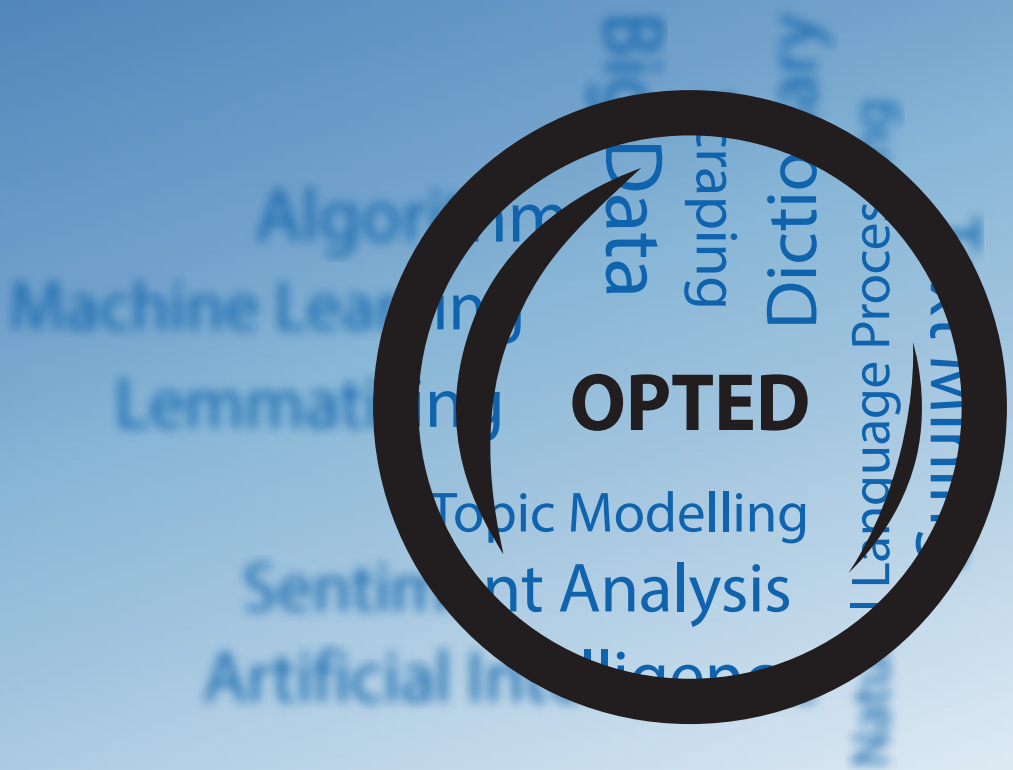
OPTED

Case Study: Named Entity Networks in Alternative Media

Deliverable 3.5

Paul Balluff, Marvin Stecker, Hajo G. Boomgaarden, and Annie Waldherr

University of Vienna



Disclaimer

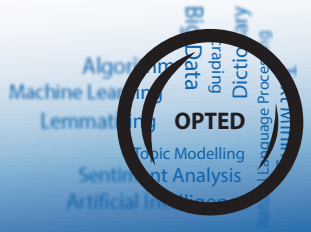
This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



OPTED
Observatory for Political Texts in European Democracies:
Designing a European research infrastructure

Case Study: Named Entity Networks in Alternative Media

Deliverable 3.5

Authors: Paul Balluff, Marvin Stecker, Hajo G. Boomgaarden, and Annie Waldherr

University of Vienna

Due date: May 2023

Contents

Executive Summary

In this case study, we showcase how a larger research infrastructure can assist in an automated analysis of journalistic texts. First, we use *Meteor* (D3.2) to inform our data collection by deciding which news sources we include in our study. We also use the registry to find the right tools for data pre-processing and analysis steps. We try different data access strategies and evaluate their advantages and disadvantages. Furthermore, we use AmCAT 4 (D7.1) to store and explore the large amount of collected text data.

This exploratory study compares the named entity networks of legacy and alternative media on the gas-pipeline Nord Stream 2 in German-speaking online media. Previous studies suggest that alternative media have fundamentally different content features compared to legacy media, such as their preference for referencing different named entities and having simpler networks of actors. The study aims to show that alternative media's named-entity networks exhibit distinctive patterns that can serve as a classification method based on textual features. The research questions focus on the extent to which named entities and their network configurations differ across news sources. The findings are expected to contribute to developing a definition for alternative media and identifying subtypes of alternative media.

1 Introduction

Alternative media are an elusive and hard to define phenomenon (Holtz-Bacha, 1999; Holt, 2018). They are very heterogeneous, and the term is "infuriatingly vague" (Atton, 2008). Contemporary forms of alternative media are even more diverse than 20 years ago (Atton, 2002; Holt et al., 2019), and finding common characteristics is dependent on cultural circumstances as well as their political beliefs (Atton, 2002; Holt et al., 2019). Alternative media are often understood as a form of journalism that is outside the "mainstream" media (Atton & Wickenden, 2005). They typically do not subscribe to the same professional practices as their "legacy" counterparts (Harcup, 2011; Holt et al., 2019). The people authoring stories for alternative media are rarely members of press associations, and seldom join press conferences, raising the question whether they should be called "journalists" at all (Anderson & Schudson, 2019).

Yet, alternative media have become increasingly relevant in a high-choice media environment, because they cater to audiences who have lost trust in legacy media and consequently also influence public opinion (Aelst et al., 2017; Holt, 2018). Moreover, some forms of alternative media have contributed towards the erosion of factual knowledge by creating or disseminating mis-information and dis-information (Aelst et al., 2017; Bennett & Livingston, 2018; McDowell-Naylor et al., 2021). Therefore, it is highly relevant to find a way to comprehensively identify and classify these types of media.

To date, identifying and classifying alternative media is mostly accomplished by examining their self-ascribed labels and looking for keywords where the medium claims to provide alternative news and views in opposition to the mainstream media (Heft et al., 2020; Schwaiger, 2022). However, many forms of alternative media make efforts to mimic legacy media in order to piggyback on their credibility (Holt et al., 2019; Mayerhöffer & Heft, 2021). These deceptive news sources would not label themselves as alternative media.

Therefore we argue that alternative media should also be studied at the content level because, in certain aspects, they exhibit fundamentally different content features compared to legacy media. Previous studies suggest that alternative media tend to prefer referencing different named entities compared to traditional media (Figenschou & Ihlebæk, 2019). For example, alternative media typically prefer to reference elites that are aligned with their own political views and agenda (Atton & Wickenden, 2005; Figenschou & Ihlebæk, 2019). Unlike mainstream media, alternative media make references to mainstream media, but not vice versa (Holt et al., 2019). This means that a traditional newspaper would refrain from using an alternative news outlet as its source. But alternative media often quote legacy media, often to confirm their ideological stance or to directly criticize a mainstream article (Figenschou & Ihlebæk, 2019; Holt et al., 2019). They also tend to have simpler and more focused networks of named entities, meaning that they feature a smaller number of actors who frequently co-occur (Shahsavari et al., 2020; Tangherlini et al., 2020).

We conduct an exploratory case study in which we compare the networks of named entities in German speaking online media. We specifically focus on the coverage in legacy and alternative media on the gas-pipeline *Nord Stream 2* since its early-planning in 2011. We choose this case because there are several conflicting interests surrounding the pipeline (Fischer, 2016), meaning we should be able to observe a great variety of mentioned named entities. Second, the political significance of the pipeline has changed drastically over time, which allows us to cover a long time span and study changes in the named entity networks (Lang & Westphal, 2017). Third, in September 2022 explosions occurred on the pipelines under the Baltic Sea (Reed, 2022), leading to many speculations on their causes. A prior study has shown that alternative media, in general, publish more speculative claims (Holt et al., 2019). Fourth, the infrastructure project has a clear and unique name that allows sampling relevant news articles reliably with little interference of non-relevant articles.

Table 1: Phases of Nord Stream 2

Period	Dates	Description
1	2011 – 2018	Planning of Nord Stream 2
2	2018 – Jan 2022	Construction & sanctions
3	Feb 2022 – Sep 2022	Russia attacks Ukraine
4	Sep 2022 – Dec 2022	Explosions and aftermath

We have the following research question for our study:

> **Research Question:** To what extent do the named entity networks differ across news sources?

We expect to find different network configurations in legacy media compared to alternative media. While the networks featured in legacy media are expected to have a larger number of nodes and a high modularity, the networks in alternative media tend to have fewer nodes and a low modularity (Tangherlini et al., 2020). Previous research suggests, that alternative media focus on a specific set of actors who fall into the categories of "enemies" and "allies" (Mayerhöffer & Heft, 2021; Heft et al., 2020). Especially, right-wing

alternative media tend to paint simplified pictures of complex events (Mayerhöffer & Heft, 2021). Of course, this feature mainly applies to one incarnation of alternative media, but it could also be the case for other forms. Additionally, alternative media tend to have much smaller teams of content creators who often do also have other jobs (Harcup, 2011). Their resources possibly limit their focus on a smaller set of actors and issues. Therefore, the observable implication of our research question is that the named entity networks in alternative media should be distinct from their legacy counterparts. That means, if the labelling of alternative media based on their self-understanding is appropriate, then they should "naturally" form clusters that can be separated from the legacy media.

Putting this case study in the context of a research infrastructure, we show how *Meteor* can serve to assist in automated content analysis. We use *Meteor* to inform our the data collection and to find the right tools for our data pre-processing and analysis steps. Furthermore, we also show different data access strategies and evaluate their advantages and disadvantages.

2 Literature Review

Alternative media have been the subject of research for several decades (Holtz-Bacha, 1999; Atton, 2008). They are understood as a relational phenomenon, because they primarily constitute themselves by *not* being like "mainstream" media (Atton, 2008).

Although alternative media have changed throughout the years, there are certain characteristics from earlier forms that are still relevant today and help us to better understand contemporary forms that have emerged in the past decade.

Earlier forms of alternative media emerged in the 1970s until the early 2000s and can be placed in the context of feminist media (Byerly, 2016) and in the left-wing, progressive, or environmentalist spectrum (Holtz-Bacha, 1999; Harcup, 2013). The producers of these forms of alternative media would not necessarily see themselves as journalists, but rather as activists, reformers, or campaigners (Steiner, 2019). They would also have fewer resources at their disposal and often rejected the practices of traditional journalists as well as business models that depended on advertising revenue (Holtz-Bacha, 1999). But they should not be placed in the sphere of "political resistance media", because they have always been understood and self-described as not being organs of political parties or highly organized political movements (Harcup, 2013).

Another earlier form of alternative could be labelled as "ethnic" or "community" media, meaning that they are part of subcultures which have different media needs and smaller niche outlets aim to cater to these needs (Couldry & Hepp, 2012). They can also be highly political in their contents and often provide a "counter narrative" (Poole, 2019). They consciously produce these narrative in a general sense that they want to "correct" stereotypes, but also for specific events. For example, to have alternative accounts when there is reporting on civil unrest or protests (Poole, 2019).

The forms of alternative media mentioned above have different news values as opposed to mainstream media. They particularly politicized the "repression of events", and care less about the prospective commercial success of news stories (Atton, 2002). By employing a different set of news values, alternative media also actively criticize the selection of news in mainstream media (O'Neil & Harcup, 2019).

With the rise of the internet, there was also a sudden proliferation of alternative media (Hellman & Riegert, 2012). The costs of publishing news decreased, while the potential reach increased. While earlier forms of alternative media featured content from non-professionals who were personal acquaintances of the editors (Holtz-Bacha, 1999), the digital incarnations allowed for anonymous and rich contributions by a variety of users (Hellman & Riegert, 2012). The rise of user-generated content also caused a "struggle over the journalistic jurisdiction" (Anderson & Schudson, 2019). It blurred the boundaries between occasional blogging and professional journalism. Not only the internet, but especially smart devices with cameras accelerated this trend (Anderson & Schudson, 2019). This development also brought a differentiation of non-professional journalistic formats, such as citizen journalism. These formats are very similar to alternative media and sometimes almost impossible to distinguish (Allan & Hintz, 2019)

The strongest trend among alternative media in the past two decades was the emergence of right-wing media (Wahl-Jorgensen & Hanitzsch, 2019). As mentioned, the majority of alternative media would have been placed in the spectrum of the political left and in progressive forces which rejected ideas such as commercial news production. Right-wing alternative media share many qualities with alternative media from the other ends of the political spectrum. For example, they also highlight the repression of events by their legacy counterparts and question their news values (Holt et al., 2019). In more general terms, they share a dissatisfaction with mainstream media (Atton, 2008), but the reasons differ. Right-wing alternative media often criticise mainstream media for being biased in favor of liberal or leftist perspectives (Holt, 2018).

Another larger contrast between left-wing and right-wing alternative media are their commercial activities. The non-commercial qualities of the earlier left-wing progressive media no longer apply to what we can observe today: right-wing alternative media often rely on ad revenues (Mayerhöffer & Heft, 2021). Through the data on *Meteor* we can confirm that 49 out of 95 alternative news websites in Germany contain ads (Balluff et al., 2022). And we also have indications that alternative media rarely use pay-wall models (6 out of 95). They rather depend on donations or paid extra content such as their podcasts (see also: Heft et al., 2020).

Modern alternative media sometimes also exhibit higher degrees of professionalization by following a journalistic logic in their reference practices (Mayerhöffer & Heft, 2021). Also the separation from political parties is becoming less and less strict (Atton, 2008; Holt, 2018). Some right-wing alternative media have shown to have ties to political parties that are placed in the same ideological spectrum (Mayerhöffer & Heft, 2021; Heft et al., 2020)

We can see that alternative media have increased their heterogeneity in the past decades, which is mainly due to a richer ideological spectrum. They still have in common that they understand themselves as being in opposition to the dominant discourse in the mainstream media (Holt, 2018). But this relays the problem of "defining alternative media" to "defining mainstream media". Therefore, Holt et al. (2019) suggests to study alternative media as a continuum rather than "absolutely opposed categories".

In most studies, alternative media were classified depending on their self-understanding. When a news source claims to be "alternative", "against the mainstream", "revolutionary" or similarly, then they would be put into the category of alternative media (Mayerhöffer & Heft, 2021; Schwaiger, 2022). From such studies, we could also

learn about content characteristics of alternative media. This approach also has the advantage of relaying the classification problem to the media themselves. For example, it also allows to easily distinguish between alternative media and citizen journalism. However, as mentioned above, there are also deceptive forms of alternative media, that would not label themselves as such, but engage in the dissemination of mis-information and dis-information (Aelst et al., 2017; Heft et al., 2020).

Similar to their legacy counterparts, alternative media source news rather from elites than "ordinary citizens" (Atton & Wickenden, 2005). However, alternative media typically prefer to reference elites that are aligned with their own political views and agenda. Unlike mainstream media, alternative media make references to mainstream media, but not vice versa (Holt et al., 2019). This means that a traditional newspaper would refrain from using an alternative news outlet as its source; but alternative media tend to quote legacy media, often to confirm their ideological stance or to directly criticize a mainstream article (Figenschou & Ihlebæk, 2019; Holt et al., 2019).

Previous studies looking into the actor networks featured in alternative media have mainly focused on dissecting conspiracy narratives (Tangherlini et al., 2020; Shahsavari et al., 2020). They have shown that the actor networks in alternative media tend to have different topologies as compared to legacy media. These networks have a lower number of nodes and also have a low modularity. They focus on a few specific actors and these are all interconnected, instead of covering a large number of actors that form communities (Tangherlini et al., 2020). For example, conspiracy narratives as constructed by an alternative media regarding COVID-19 feature Bill Gates as a central actor who has connections with all other actors in the network (Boberg et al., 2020; Shahsavari et al., 2020).

While these examples specifically focused on conspiracies in social media and alternative media, they provide preliminary findings of named entity networks in alternative media. So far, it was not investigated, whether this pattern holds up to several forms of alternative media. But given the literature, we have plausible reasons to believe so. First, alternative media are highly political and feature ideologically informed content. Therefore, the observable implication is that alternative media focus on a specific set of actors who fall into the categories of "enemies" and "allies" (Mayerhöffer & Heft, 2021; Heft et al., 2020). Second, right-wing alternative media tend to paint simplified pictures of complex events (Mayerhöffer & Heft, 2021). Of course, this feature mainly applies to one incarnation of alternative media, but it could also be the case for other forms. Third, alternative media tend to have much smaller teams of content creators who often do also have other jobs (Harcup, 2011). Their resources possibly limit their focus on a smaller set of actors and issues.

3 Data Collection

The goal of this study is to investigate networks of named entities in the German news coverage about *Nord Stream 2* since its planning in 2011 until the end of 2022. Therefore, we needed to acquire a corpus that covers a time-span of more than 10 years and also includes a variety of news sources. We inform our data collection with *Meteor*, which lists 828 news sources in Germany, including alternative media. The labelling scheme for "alternative media" in *Meteor* follows previous research (e.g., Schwaiger, 2022), meaning

that the news sources are marked as "alternative" according to their self-understanding.

We focus our sample on digital news sources, because news is nowadays present on a variety of digital channels and audiences tend to increasingly use social media and websites for retrieving news (Scharkow et al., 2020; Forman-Katz & Matsa, 2022). Moreover, the majority of modern alternative media rely on digital channels to reach their audiences. *Meteor* lists 119 printed news sources in Germany, and only 14 of them were labelled as alternative media. In contrast, out of the 212 websites listed in Germany, 95 of them were labelled as alternative media. Therefore, we mainly focus our study on drawing texts from online news sources (websites).

Our initial attempt to retrieve a large collection of news articles was by extracting them from the *Common Crawl*.[1] The *Common Crawl* is an extremely large data collection project which continuously downloads and archives all sorts of websites. The data from the project can be accessed with the tool provided by Hamborg et al. (2017), which also provides structured output of news websites. The tool automatically downloads chunks of the *Common Crawl*, filters the chunk according to a set of domains, and then extracts the full-text and meta-data from the news articles. We used all 212 domains listed in *Meteor* for filtering the *Common Crawl* data; subsequent data extraction took approximately 18 days. It eventually resulted in over 16 million news articles from 200 different domains, which is about 56 GB of data. We ingested the massive dataset into AmCAT4 (see D7.1), where we could inspect the quality of the data and perform some simple aggregations.

The yield of the *Common Crawl* is impressive at first sight, but has two major shortcomings. First, the *Common Crawl* has only a limited historical coverage. The data starts only in 2016 and reaches a peak in 2018, and has fewer articles for 2020, and 2021. This makes the data not suitable for our longitudinal study. The second shortcoming is that news articles extracted from pay-walled websites are incomplete. Since 78 of the 212 websites listed in *Meteor* use some form of a pay-wall, we deem the data not usable for such cases, because there is at least a third of the data missing.

Given the non-satisfying results, we employed different data collection strategies. In order to retrieve full-text data from legacy media, we queried the *LexisNexis* archive with a search string approach[2]. The *LexisNexis* archive yielded 65,992 articles reaching back to 2008 with news articles from 90 sources, where 66 are print news sources, 19 websites, and 5 are transcripts of TV shows.

As mentioned above, digital channels have made it easier for alternative media to distribute news. Many news websites rely on content management software, where the most popular one is *WordPress* (see D3.4). Therefore, we leveraged the WordPress API which is activated by default, and most website administrators do not deactivate it. We tested every of the 212 websites whether they have a WordPress API endpoint available, and 70 of them returned positive. 56 out of the 70 were labelled as alternative media in *Meteor*. The API-based approach resulted in 503,211 news articles spanning from 2005 until the end of 2022. When we retrieved the articles from each news source's API, we did not filter accoring to any search strings or other criteria.

The third strategy is the API made available by *Tagesschau*, a German public news

[1] https://Common Crawl.org/

[2] We used the keywords: "nord stream", "nord-stream", and "nordstream". All search on *LexisNexis* is case-insensitive.

Table 2: Summary of retrieved news articles by data source.

Data Source	Before filtering	After filtering
LexisNexis	65,992	65,150
Tagesschau API	1,088	645
WordPress API	503,211	4,466
Total	570,291	70,261

show that broadcasts on a daily basis and also maintains a news website that is continuously updated. The website offers a public API[3] where news articles can be queried according to search strings. We retrieved 1,088 news articles from the API by using the same search terms as we used for the *LexisNexis* query.

Considering that the articles retrieved from the WordPress API were not filtered and that *LexisNexis* tends to yield broad results, we used *Corpustools* (Welbers & van Atteveldt, 2022) to filter all news articles based on keywords[4]. As Table 2 shows, this reduced the number of articles significantly.

4 Data Analysis Procedure

The unit of analysis in this study are single documents, which in our case are articles collected from news sources, where we extracted all named entities and investigated their co-occurrence. For that purpose, we ran a series of pre-processing steps. Again, we used tools listed in *Meteor* to process our corpus. However, we also used a custom fine-tuned large language model for named entity recognition (see below).

4.1 Pre-processing

4.1.1 Data Cleaning

We performed a number of simple data cleaning steps to remove as much unwanted noise as possible. First, some articles retrieved from *LexisNexis* had inconsistent character encoding. We used *ftfy* (Speer, 2019) which can automatically detect such inconsistencies and convert the data to clean UTF-8 encoded text. Next, we harmonized all variants of quotation marks[5], and normalized the spacing between punctuation marks.

[3] Website of the broadcaster is https://www.tagesschau.de/ and the API documentation is available here: https://tagesschau.api.bund.dev/

[4] *Corpustools* allows for wildcard endings, as well as proximity matching of keywords: `“nord stream*”~4 OR nordstream* OR ns2 OR “nord-stream*”`. The first part of the searchstring `“nord stream*”~4` finds all occurrences of `nord` and `stream*` within a four word range, and also considered all possible endings for `stream*`, such as `streams`.

[5] For example, we replaced Swiss-style guillemets «`Word`» with a normalized double-quotation mark “ `Word` ” and additional spacing.

Table 3: Comparison of model performances with composite corpus.

Model	F1
GottBERT	0.79
GermanBERT	0.80
XLM-RoBERTa	0.84
DeBERTaV3	0.87

4.1.2 Sentence splitting

Extracting named entities tends to work better with smaller text units, because most LLMs cannot handle more than 512 tokens at once. Sentence splitting is a non-trivial task in German. While it is possible to achieve decent results with regular expressions, there are just too many edge cases where the end of a sentence is not captured properly with such a technique. Therefore, we used spaCy (Honnibal et al., 2020) for sentence splitting, which offers a pre-trained neural network for this particular task that was trained with German news data[6]. After splitting, the corpus consisted of roughly 2.6 million sentences, with 36 sentence per article on average.

4.2 Named Entity Recognition

We ran several tests for determining the optimal model and parameters to perform named entity recognition in German. We chose four state-of-the-art German or multilingual models from the BERT family: German-BERT[7], GottBERT (Scheible et al., 2020), XLM-RoBERTa (Conneau et al., 2019), and DeBERTaV3 (He et al., 2021). We used a composite corpus that we generated using two German corpora and one English corpus for fine-tuning the models and for validation. The models' inherent multilingual capabilities allowed us to use training datasets from different languages to improve the models performances (Balluff et al., 2023). We used the German CoNLL corpus (Tjong Kim Sang & De Meulder, 2003), the GermEval corpus (Benikova et al., 2014), and the English part of the OntoNotes 5.0 corpus (Weischedel et al., 2013). All corpora consist mainly of news texts, so they are suitable choices for our application.

We ran several experiments where we systematically varied different hyperparameters (i.e., batch size, training epochs, and learning rate) and kept note of each model's best results (Table 3).

The pre-tests show that the newest model, DeBERTaV3, performs best. Therefore, we used this model for our named entity recognition task[8]. We also validated the model against the designated validation sets for each corpus separately (Table 4).

Using this model, we extracted more than 4 million named entities, where 942,938 are persons, 974,968 are organisations, and more than 2 million are locations.

[6] We used the model `de_core_news_lg` (version 3.5.0) which was trained on the TIGER corpus (Brants et al., 2004) and achieved a F1-measure of 0.95.

[7] see: https://huggingface.co/bert-base-german-cased

[8] Optimal results were achieved using 2 epochs, a batch size of 8, and a learning rate of 3e-5.

Table 4: Performance of DeBERTaV3 against individual corpora.

Corpus	F1
CoNLL2003 (German)	0.75
GermEval	0.85
OntoNotes (English)	0.91

4.3 Named Entity Reconciliation

Next, we harmonized the extracted entities. For example, one document might mention "Olaf Scholz" and another one might mention "Chancellor Scholz". Both instances refer to the same individual but use different phrases. There are computer-assisted approaches available for named-entity reconciliation that query the Wikidata API and try to guess whether two different phrases actually refer to the same actor (e.g., Poschmann & Goldenstein, 2022).

We use the WikiData reconciliation API by Delpeuch (2020) which allows for fuzzy matching of search terms against unique WikiData IDs. The API accepts queries with additional filters. For example, to narrow down the search for a person, a phrase can be matched for entities of the category "human". It returns a list of matches, where each match has a score for best fit. The API was designed for *OpenRefine* (Delpeuch et al., 2023) which offers a graphical user interface and requires human interaction. However, we extracted 148,750 unique unreconciled named entities, which were too many for *OpenRefine* with semi-automated reconciliation. Instead, we used a Python client to programmatically interface with the reconciliation service and only retain the best matches (Balluff, 2023). To enhance the success of automated reconciliation, we used spaCy again but this time for lemmatization of the named entity phrases. The Python client uses the lemmatized phrase first, and if no result is returned from the API, it retries with the original phrase as a fallback query.

Using automated reconciliation, we could automatically disambiguate over 98\% of all named entity phrases, which resulted in around 67 thousand unique named entities. This left 27,360 unreconciled named entities, of which we manually reviewed all phrases that occurred more than 5 times in the corpus. The manual consolidation yielded another 395 unique named entities. The most frequent named entities per category are shown in Table 5. Finally, we only kept the most frequent named entities, which are all in the third quartile. This corresponds to named entities that are mentioned at least 8 times in the entire corpus.

4.4 Network Generation

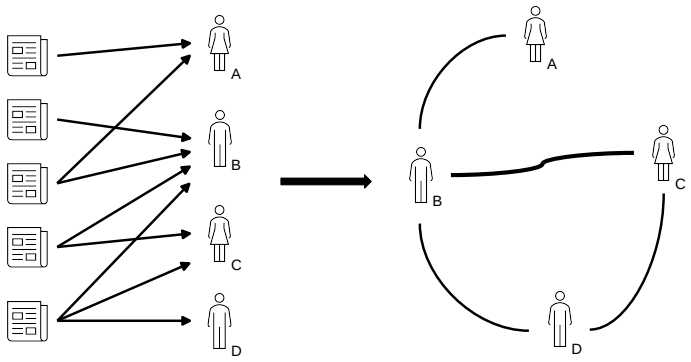
With the consolidated dataset we constructed weighted two-mode networks (incidence matrices) for each news source. The two node types are documents and named entities, and the two nodes have a link between them, when a document mentions a named entity (Figure 1). The weight between the nodes is the number of times a document mentions a named entity. The advantage of this representation is that we can subset the data based on date periods and study how the network changes over time.

Table 5: Most frequent named entities after reconciliation

Persons	N	Locations	N	Organizations	N
Wladimir Putin	95,841	Russia	393,200	European Union	93,844
Angela Merkel	48,470	Germany	265,527	Gazprom	43,999
Donald Trump	42,268	Ukraine	194,508	NATO	30,384
Olaf Scholz	39,242	USA	139,117	SPD (political party)	24,599
Gerhard Schröder	37,505	Europe	115,085	Greens (political party)	19,859
Joe Biden	32,208	Moscow	45,644	DPA (news source)	19,159
Robert Habeck	20,408	Berlin	44,776	Uniper	15,897
Annalena Baerbock	16,498	Poland	44,120	Die Welt (news source)	15,125
Wolodymyr Selenskyj	15,211	PRC China	29,532	European Commission	11,401
Manuela Schwesig	14,640	France	21,032	US-Government	11,375

We projected the two-mode networks to weighted co-occurrence networks of named entities (adjacency matrices), where the nodes are only named entities and the links between represent that the named entities were mentioned together in documents. The edge weight between the named entities count the number of times that the entities co-occurred.

Figure 1: Two-mode network of documents mentioning named entities and its projection as a co-occurrence network.



4.5 Network Analysis

We want to understand the changes in the networks according to different time periods. Therefore, we subset the data into four time periods (see Table 1). We then projected the co-occurrence networks for each news source at every time period. We used the *igraph* (Csardi & Nepusz, 2006) package to calculate basic metrics of every network: density, diameter, transitivity, and growth rate of number of unique nodes. Additionally, we ran the Louvain community detection algorithm to calculate each network's modularity (Blondel et al., 2008).

Finally, we performed k-means clustering to group networks with similar topologies

Table 6: Mean metrics of network topologies

Metric	Alternative Media	Legacy Media
Network Size	344.60	655.83
Network Growth	21.59	7.57
Network Density*	0.40	0.19
Components	1.12	1.12
Diameter*	24.50	43.95
Transitivity*	0.68	0.48
Communities*	4.04	6.07
Modularity*	0.17	0.25
N Networks	104	329

*Difference between groups is statistically significant using Wilcoxon Signed-Rank Test

for each time period. We ran a silhouette analysis to determine the optimal number of clusters for each time period.

5 Results

If we work with the assumption that the self-ascribed labels for alternative media is indeed correct, we can see in Table 6 that their network topologies are different from legacy media. Alternative media appear to have significant differences for network density, diameter, transitivity, and modularity. Our results appear to agree with previous findings [Tangherlini et al. (2020); Shahsavari2020].

Taking into account that the self-ascribed labels for alternative media are not always accurate, we get a less clear picture. The results for the cluster analysis as presented in the Figures 2a–2d. Comparing the changes of the clusters across time periods, it appears that each time period has at least one cluster for news sources that did not report much or nothing at all about *Nord Stream 2*. These are the clusters at the bottom of each plot where the news sources are all closely located to the cluster's center. The number of clusters is a bit unstable and ranges from 6 to 9 clusters, where the highest number is in period 2 and the lowest in period 4.

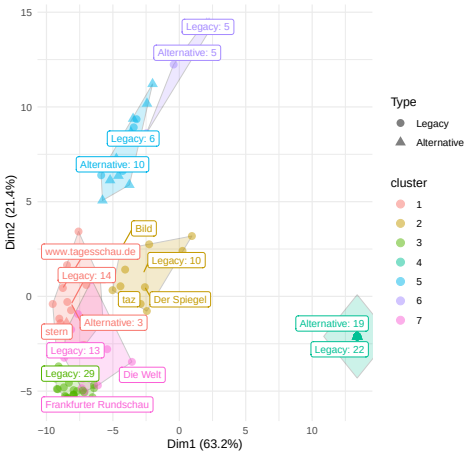
The relative positions of the most influential legacy media (see Weischenberg et al., 2006) provide insights to the face validity of the cluster analysis. The influential legacy media are mostly positioned close to each other, but also in their own distinct clusters.

Periods 1, 3, and 4 each have at least one larger cluster that exclusively contains legacy media. These clusters could be referred to as the "mainstream". However, the alternative media are distributed across all other clusters, and it appears that their network topologies share many characteristics with legacy media.

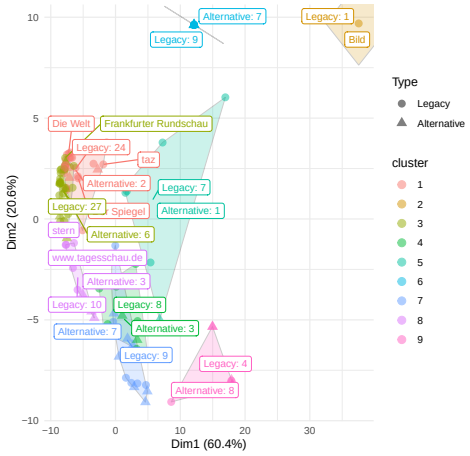
Figure 2: Result of K-Means Clustering.

Each plot shows the position of a news outlet in a latent space and their membership in clusters. Legacy media are marked with triangles and alternative media with circles. The annotations show the total number of legacy and alternative media respectively. Additionally, leading media are highlighted.

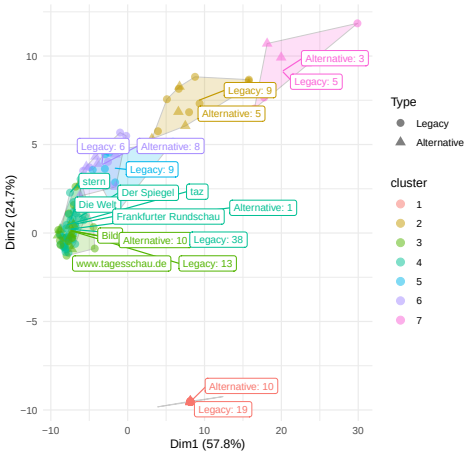
(a) Period 1: 2011 – 2018



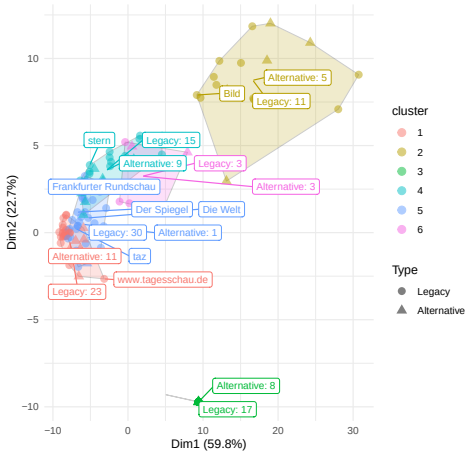
(b) Period 2: 2018 – Feb 2022



(c) Period 3: Feb 2022 – Sep 2022



(d) Period 4: Sep 2022 – 2022 end



6 Discussion

In this study, we investigated the named entity networks in different news sources, based on their co-occurrences across four time periods. The descriptive statistics of the network topologies (Table 6) appear to agree with previous findings (Shahsavari et al., 2020; Tangherlini et al., 2020). On average, the networks of alternative media are indeed smaller, have a lower modularity, and have a higher density[^statisticaltests].

We used k-means clustering to determine, whether alternative media would form distinct clusters or fall into similar clusters as their legacy counterparts. Our results suggest that there are some distinct clusters for only legacy media, but none that only contains alternative media. Contrary to our expectations, the alternative media share clusters with the legacy media. Previous research indicates that alternative media make references to legacy media, but not vice versa (Figenschou & Ihlebæk, 2019). Looking at the neighborhoods of alternative media in our cluster analysis, one could argue that the authors of alternative media regularly read mainstream media which in turn might inform their own writing and news stories. This would align with previous research which suggests that alternative media make references to mainstream media, but not vice versa (Holt et al., 2019). However, with our current approach, we cannot show the direction of media referencing.

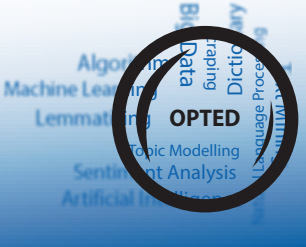
With regard to the building blocks of the OPTED infrastructure, we focus on the German media ecosystem for this deliverable. However, the presented workflow and the analysis, are reproducible for every country covered by *Meteor* and could also be studied in comparative perspective. For the data selection, the repository allows to draw different samples, e.g. a comparison of legacy media outlets in Scandinavia, or only weekly publications that are ad-supported in Southern Europe. The country filters are fully integrated with all other meta-information on outlets, meaning a URL list for filtering the *Common Crawl* or testing with the *WordPress* API can easily be obtained. Full names of the outlets can instead be used to query databases such as *Factiva* or *LexisNexis*.

Using the web-interface of *Meteor* also provides a first, basic overview of the sample composition. Before data collection, one can inspect the various sub-setting criteria that are of possible interest in a study and see whether each of them contains enough outlets for the proposed analysis. If the goal is to compare news sources with foci on the subnational, national or multinational level, using *Meteor's* filters allows researchers to see how the outlets of interest are distributed across these conditions. If, for instance, it turns out that there are too few outlets with a multinational focus, the research strategy can be adjusted before data collection has begun.

For the analysis, *Meteor* provides important meta-information to compare content across different publishers, again for every country covered in the inventory so far. The focus of this study lies on the distinction between self-described "alternative" news media, but we could also easily subset our entire corpus according to other types of information, such as the ownership structure or the geographic focus of a news outlet. The main limitation of *Meteor* is currently its coverage. We still do not have a comprehensive overview of available news sources for many countries. Especially, alternative media outside of the German speaking context are not well covered by *Meteor* yet.

Nevertheless, *Meteor* was useful for choosing a sample of media outlets to include in the study. The platform also helped in finding the right software tools to perform different

pre-processing steps and text analysis.



References

Aelst, P. V., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

Allan, S., & Hintz, A. (2019). Citizen Journalism and Participation. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed., pp. 435–451). Routledge.

Anderson, C. W., & Schudson, M. (2019). Objectivity, Professionalism, and Truth Seeking. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed.). Routledge.

Atton, C. (2002). *Alternative media*. SAGE Publications.

Atton, C. (2008). *Alternative Journalism*. SAGE Publications.

Atton, C., & Wickenden, E. (2005). Sourcing Routines and Representation in Alternative Journalism: a case study approach. *Journalism Studies*, *6*(3), 347–359. https://doi.org/10.1080/14616700500132008

Balluff, P. (2023). wikibase-reconcile: A concurrent client to reconcile entities against wikidata [PyPI version 0.2.0]. https://pypi.org/project/wikibase-reconcile/

Balluff, P., Boomgaarden, H. G., & Waldherr, A. (2023). *Automatically finding Actors in Texts: A performance review of multilingual named entity recognition tools* [Manuscript submitted for publication].

Balluff, P., Lind, F., Stecker, M., Dinhopl, C., Boomgaarden, H. G., & Waldherr, A. (2022). *Meteor: Inventory of news sources, media organizations, and text analysis tools* (Database). OPTED. https://meteor.opted.eu

Benikova, D., Biemann, C., & Reznicek, M. (2014). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2524–2531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic Populism: Facebook Pages of Alternative News Media and the Corona Crisis - A Computational Content Analysis. *CoRR*, *abs/2004.02566*. https://arxiv.org/abs/2004.02566

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, *2*(4), 597–620. https://doi.org/10.1007/s11168-004-7431-3

Byerly, C. M. (Ed.). (2016). *The Palgrave International Handbook of Women and Journalism*. Palgrave Macmillan.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*. arXiv: 1911.02116.

Couldry, N., & Hepp, A. (2012). Media Cultures in a Global Age: A Transcultural Approach to an Expanded Spectrum. In I. Volkmer (Ed.), *The Handbook of Global Media Research* (pp. 92–109). Wiley.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. https://igraph.org

Delpeuch, A. (2020). Running a reconciliation service for Wikidata. *19th International Semantic Web Conference*, *2773*(19). https://ceur-ws.org/Vol-2773/paper-17.pdf

Delpeuch, A., Morris, T., Huynh, D., Mazzocchi, S., Jacky, Guidry, T., Stephens, O., elebitzero, Matsunami, I., Sproat, I., Santos, S., allanaaa, kushthedude, Fauconnier, S., Mishra, E., Larsson, A., Beaubien, A., Magdinier, M., Liu, L., … Chandra, L. (2023). *OpenRefine v3.7.2* (Version 3.7.2). Zenodo. https://doi.org/10.5281/zenodo.7803000

Figenschou, T. U., & Ihlebæk, K. A. (2019). Challenging Journalistic Authority. *Journalism Studies*, *20*(9), 1221–1237. https://doi.org/10.1080/1461670X.2018.1500868

Fischer, S. (2016). Nord Stream 2: Trust in Europe. *CSS Policy Perspectives*, *4*. https://doi.org/10.3929/ETHZ-A-010682973

Forman-Katz, N., & Matsa, K. E. (2022, September 20). *News Platform Fact Sheet*. Retrieved October 28, 2022, from https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/

Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (2017). news-please: A Generic News Crawler and Extractor. *Proceedings of the 15th International Symposium of Information Science*, 218–223. https://doi.org/10.5281/zenodo.4120316

Harcup, T. (2011). Alternative journalism as active citizenship. *Journalism*, *12*(1), 15–31. https://doi.org/10.1177/1464884910385191

Harcup, T. (2013). *Alternative Journalism, Alternative Voices*. Routledge.

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Heft, A., Mayerhöffer, E., Reinhardt, S., & Knüpfer, C. (2020). Beyond Breitbart: Comparing Right-Wing Digital News Infrastructures in Six Western Democracies. *Policy & Internet*, *12*(1), 20–45. https://doi.org/10.1002/poi3.219

Hellman, M., & Riegert, K. (2012). Emerging Transnational News Spheres in Global Crisis Reporting A Research Agenda. In I. Volkmer (Ed.), *The Handbook of Global Media Research* (pp. 156–174). Wiley.

Holt, K. (2018). Alternative Media and the Notion of Anti-Systemness: Towards an Analytical Framework. *Media and Communication*, *6*(4), 49–57. https://doi.org/10.17645/mac.v6i4.1467

Holt, K., Figenschou, T. U., & Frischlich, L. (2019). Key Dimensions of Alternative News Media. *Digital Journalism*, *7*(7), 860–869. https://doi.org/10.1080/21670811.2019.1625715

Holtz-Bacha, C. (1999). Alternative Presse. In J. Wilke (Ed.), *Mediengeschichte der Bundesrepublik Deutschland* (pp. 330–349). Böhlau Verlag. https://doi.org/10.7788/9783412328733-015

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Lang, K.-O., & Westphal, K. (2017). *Nord Stream 2: a political and economic contextualisation* (Vol. 3/2017). Stiftung Wissenschaft und Politik. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-51318-5

Mayerhöffer, E., & Heft, A. (2021). Between Journalistic and Movement Logic: Disentangling Referencing Practices of Right-Wing Alternative Online News Media. *Digital Journalism*, 1–22. https://doi.org/10.1080/21670811.2021.1974915

McDowell-Naylor, D., Thomas, R., & Cushion, S. (2021). Alternative online political media. In *The Routledge Companion to Media Disinformation and Populism* (pp. 169–177). Routledge. https://doi.org/10.4324/9781003004431-19

O'Neil, D., & Harcup, T. (2019). News Values and News Selection. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed., pp. 213–228). Routledge.

Poole, E. (2019). Covering Diversity. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed., pp. 496–486). Routledge.

Poschmann, P., & Goldenstein, J. (2022). Disambiguating and Specifying Social Actors in Big Data: Using Wikipedia as a Data Source for Demographic Information. *Sociological Methods & Research*, *51*(2), 887–925. https://doi.org/10.1177/0049124119882481

Reed, S. (2022). Mysterious Blasts and Gas Leaks: What We Know About the Pipeline Breaks in Europe. *The New York Times*. https://www.nytimes.com/2022/09/28/world/europe/nordstream-pipeline-gas-leak-explosions.html

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). GottBERT: a pure German Language Model.

Schwaiger, L. (2022). *Alternative Nachrichtenmedien im deutschsprachigen Raum.* transcript Verlag. https://doi.org/doi:10.1515/9783839461211

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, *3*(2), 279–317. https://doi.org/10.1007/s42001-020-00086-5

Speer, R. (2019). ftfy: Fix Text For You [PyPI version 6.1.1]. https://doi.org/10.5281/ZENODO.2591652

Steiner, L. (2019). Gender, Sex, and Newsroom Culture. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (2nd ed., pp. 452–468). Routledge.

Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. (2020). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLOS ONE*, *15*(6), 1–39. https://doi.org/10.1371/journal.pone.0233879

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. https://aclanthology.org/W03-0419

Wahl-Jorgensen, K., & Hanitzsch, T. (Eds.). (2019). *The Handbook of Journalism Studies* (2nd ed.). Routledge. https://doi.org/10.4324/9781315167497

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). OntoNotes Release 5.0. https://doi.org/10.35111/XMHB-2B84

Weischenberg, S., Malik, M., & Scholl, A. (2006). Journalismus in Deutschland 2005. *Media Perspektiven*, *7*, 346–361. https://www.ard-media.de/media-perspektiven/publikationsarchiv/2006/artikel/journalismus-in-deutschland-2005/

Welbers, K., & van Atteveldt, W. (2022). *corpustools: Managing, Querying and Analyzing Tokenized Text* [R package version 0.4.10]. https://CRAN.R-project.org/package=corpustools