Function Words as Stable Features for German Opinion Articles Classification

Amelie Schmidt-Colberg[1], Simon Burkard[1], Anne Grohnert[1], Michael John[1]
[1]Fraunhofer Institute for Open Communication Systems (FOKUS)
Correspondence: <firstname>.<lastname>@fokus.fraunhofer.de

Abstract

The paper addresses the challenge of stable classification of neutral news reports and opinion articles under shifts in publisher and domain. We apply experiments using different lexical and grammatical word categories on a hand curated data set of German online news and opinion articles that are labeled in terms of topic and publisher trustworthiness. The findings demonstrate that function words, typically used as stop words, are effective markers for opinion article classification that provide stable predictions under domain shifts and across different publishers. The results can help with the development of suitable feature sets to distinguish between opinion-free news articles and opinion articles, for example as a tool to flag online articles with a high level of subjectivity.

1 Introduction

In the modern information society, it is sometimes difficult for consumers to assess the level of subjectivity of journalistic content. That's why a missing labeling of opinion articles can lead to disinformation (Molina et al., 2021). Analyzing the level of objectivity and subjectivity of a text is therefore a major challenge for reception of news articles. Tools for the automatic classification of online news articles can serve as an essential aid (Alhindi et al., 2020).

Regarding the objectivity of the articles, news pieces can be roughly divided into two types: 1) factual, opinion-neutral news reports and 2) subjectively composed opinion articles. News reports can be either - often short and precisely written - pieces from news agencies that are distributed to many publishers or news stories written by journalists especially for the respective publisher. Opinion articles, on the other hand, subjectively present the author's point of view trying to convince the reader of certain stance (e.g. editorials, columns or guest comments). Since both types of articles (news reports and opinion articles) have different functions (fact presentation vs. persuasion), they also differ in their linguistic form (McCabe and Heilman, 2007).

Related research on the classification between opinion pieces and factually neutral articles usually uses sets of various and complex linguistic features (Krüger et al., 2017). However, these approaches typically show a loss in accuracy when subject to publishers- or topic-related distribution shifts. Beyond the analysis of content words, which obviously are markers for different text types, our assumption is that opinion pieces and factually neutral articles also vary in terms of the sole use of content-independent function words, e.g. prepositions, conjunctions and auxiliary verbs.

We thus hypothesize that using function words as classification features we can achieve stable classification distinguishing neutral news articles from opinion articles with respect to changes in publisher and topic. Specifically, the main research question of this paper is: Can function words effectively be used as markers to distinguish between journalistic text types with high and low level of subjectivity, i.e. opinion articles and news reports?

In this paper, we therefore present the experimental results on the classification of online news articles using features constructed from function words compared to features from common part-of-speech tags like adverbs, adjectives, verbs or nouns. We apply the experiments on a hand curated data set of 600 German online news articles which were labeled regarding the reliability of the publisher (reliable vs. non-reliable), the article type (opinion vs. news report) and whether the article originates from a news agency. Hence, the contributions of our work are two-fold:

- Curation of a German-language dataset of 600 labelled news and opinion articles, from both reliable and unreliable news publishers

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 101–111, 2025

- Demonstration that function words can be used as markers to differentiate between factual news reports and opinion pieces, with stable predictions under domain shifts and across different data subsets depending on publisher reliability and news agencies

The results of this work should help to develop a better understanding of how objective and subjective news articles can be distinguished from each other. On this basis, algorithms for the stable recognition of opinions and factual news articles can be developed more effectively, for example as a tool for news consumers to specifically flag articles with a high level of subjectivity.

2 Related Work

Related work to our research task particularly includes other linguistic approaches to classifying news genres as well as related classification approaches primarily based on function words. Also, similar existing German-language news datasets will be briefly mentioned and separated from the dataset we have manually compiled.

2.1 News Genre Classification

As mentioned earlier, one factor that differentiates opinion pieces from more factual news articles is their respective subjectivity level (Krüger et al., 2017; Alhindi et al., 2020). Thus, a common approach for news genre classification has been the sentiment analysis as a sub-task of subjectivity detection. In the context of news genre classification, this has been done on a sentence-level (Wiebe et al., 2004; Yu and Hatzivassiloglou, 2003; Toprak and Gurevych, 2009; Krüger et al., 2017) or on a document-level (Kessler et al., 1997; Freund et al., 2006; Sharoff et al., 2010; Krüger et al., 2017). For sentence-level classification, simple features such as n-grams (Toprak and Gurevych, 2009; Wiebe et al., 2004), TF-IDF constructed from lemma features (Toprak and Gurevych, 2009) or lexicon features (Wilson et al., 2005) were used to train classifiers such as Naive Bayes and support-vector machine (SVM) classifiers.

It was found that unigrams generally outperform more complex n-grams whereas PoS features perform worse than others and word features outperform all other classes (Krüger et al., 2017). In document-level classification, PoS tags (Karlgren and Cutting, 1994), lexical features (Kessler et al., 1997) or n-grams (Freund et al., 2006; Sharoff et al., 2010) were used. Lexical features and PoS tags were found to be robust against topic shifts whereas n-grams were particularity vulnerable (Petrenz and Webber, 2011) (Krüger et al., 2017).

It should be noted, that the subjectivity-level does not fully match to the journalistic categories of news and opinion, where these categories may display an overlap and other genres can display similar subjectivity levels.

More recently, on a document-level, Krüger et al. (2017) extracted 74.000 reports and 2000 opinion pieces from the BLIIP WSJ Datasets (Charniak et al., 2000) and showed that using their linguistic features (derived from sentence- and token-length, negation, punctuation connectives, citations, tenses, modalities, pronouns, PoS-tags and sentiment), news and opinion articles can successfully be classified by a SVM classifier. They used the NLTK Brown Corpus (Bird et al., 2009) to test their model on one unseen publisher and texts from the New York Times Annotated Corpus of the Linguistic Data Consortium (NYT) (Sandhaus, 2008) to test on unseen topics. When confronted with topic-related distribution shifts, their linguistic features outperformed PoS tags and bag-of-lemma approaches.

Alhindi et al. (2020) expanded on this work and used an unspecified model to estimate argumentative features on a sentence level for 1751 news and 1751 opinion articles taken from the WSJ dataset. They used the NYT dataset to create two balanced sets of 2000 articles on two separate topics, defence (NYT-Defence) and medicine (NYT-Medicine) to evaluate model performance on topic-related distribution shifts. They also constructed another balanced dataset containing 7092 articles from multiple US publishers to evaluate model performance under publisher distribution shift. Argumentative features were more robust than linguistic features or BERT embeddings (Devlin et al., 2019) when confronted with topic-related distribution shifts, achieving an f1 score of 90 % and 88 % on NYT-Defense and NYT-Medicine, respectively. Similar results from Blackledge and Atapour-Abarghouei (2021) show, that BERT models trained on either the ISOT dataset (Ahmed et al., 2017) or the Combined corpus dataset (Khan et al., 2021) are not robust when evaluated on previously unseen topics during training.

Building upon the foundational work of Alhindi et al. (2020), Lin et al. (2023) focused on understanding the perception of news consumers. They

curated a unique corpus comprising various Dutch media formats, including news articles, TV talk shows, TV news, podcasts, and satire. This corpus consisted of 113,427 sentences extracted from 1,607 distinct items, annotated for factuality and formality by human reviewers. They trained a BERT model for sentence-level classification using this dataset. Their findings revealed distinct patterns in the data: news articles consistently exhibited high levels of formality and factuality. Interestingly, opinion articles from news outlets also demonstrated high formality and notably higher factuality than contributions from other outlets. This trend was likely influenced by including non-news-related outlets in the study, which typically feature less formal and factual content. The research focused on a publisher split to evaluate the model but did not include topic-related splits, leaving the model's performance under topic-related distribution shifts untested.

Although some approaches have proven to be more robust than others when confronted with domain shifts, all experienced a non-negligible performance loss in their reported metrics. This performance loss is difficult to estimate as it varies depending on the topic shift. The topic splits in the reviewed work (Alhindi et al., 2020; Devlin et al., 2019; Krüger et al., 2017) investigated a specific distribution shift, comparing performance between two topic groups; articles during training belonged to one topic group, and articles during testing belonged to a topic group where the topic label definitions were broad. Therefore, the results only partially represent a model's performance under topic shift. Furthermore, none of the works was tested on German online news datasets and, to our knowledge, none of the datasets intentionally included unreliable publishers. Also, the effect of separating news agency articles from news reports was not investigated.

2.2 Text Classification using Function Words

As mentioned, various linguistic feature sets can be used to classify news articles, mostly derived from both content words (e.g. nouns, verbs, adjectives) and function words including prepositions, conjunctions, articles, specific adverbs and auxiliary verbs. Function words are also known as linking words or connectors in the clause structure. They usually fulfill one or more grammatical functions within the sentence. In contrast to content words, the independent lexical meaning of function words is only slightly pronounced. Often, function words are regarded as stop words being removed during pre-processing steps in NLP tasks. However, function words are elementary components of a sentence in written languages by connecting content words with each other and creating the meaningful relationships between them through their syntactic or semantic functions (grammis, 2024a).

Recent work has shown that function word based features can successfully be used for text classification. For example, Halvani et al. (2020) use topic-agnostic features sets derived from function words for effective authorship verification using similarity based clustering. Venglařová and Matlach (2024) show that function words can be used stand-alone to distinguish between different types of texts, e.g. novels, poems and academic articles. The approach of primarily using function words in classification tasks has been less researched. To our knowledge, there is no known work that predominantly examines the use of function words to differentiate between news and opinion articles.

2.3 German News Datasets

The above-mentioned existing works on news genre classification use extensive non-German article datasets, for example the NYT Annotated Corpus or the WSJ Dataset comprising news and opinion pieces of a single publisher.

Unfortunately, German-language data sets of labelled news articles from different publishers are less common. A data set called *One Million Posts Corpus* is based on an extensive collection of articles from a single Austrian daily newspaper (Schabus et al., 2017). This data set categorizes news articles according to rather broad topic categories, but does not explicitly distinguish between opinion and news articles. The project *Deutscher Wortschatz* also provides news corpora from various German-language news portals, but only on the basis of individual sentences and also without the option of separating news reports and opinion items (Goldhahn et al., 2012). Another related data set was created for a news genre classification task as part of the *International Workshop on Semantic Evaluation* (Piskorski et al., 2023). Although this data set contains some labeled news reports and opinion pieces in German, its size is very small and not balanced (27 news and 86 opinion pieces). We therefore have decided to create our own data set.

Figure 1: Word cloud representing the frequency of different tags appearing in the topic label keywords

3 Dataset

The hand curated dataset contains 600 articles from different online news platforms published in the period between August 2020 and April 2024. It comprises 300 news reports and 300 opinion articles (see Table 1) collected from 34 different reliable and 14 different non-reliable publishers. Of these publishers, 8 reliable and 8 unreliable publishers published articles labelled as opinions.

The classification into reputable and untrustworthy publishers was based on the NewsGuard rating system, which assesses the reliability of a news outlet on a scale from 0 to 100 (Lühring et al., 2025). Publishers with ratings above 90 were classified as reliable, and publishers below 60 were classified as non-reliable. Publishers without a rating available or with ratings between 60 and 90 were not considered for the data set.

The articles were manually extracted from german online news portals. This ensured that the texts only contained the mere body texts of the articles without other disruptive elements (captions, advertising, other article references).

The following general conditions were taken into account when collecting the articles: as many different publishers as possible were considered within the corpus in a balanced way (balance of publishers). Also, matching counterparts from at least one other reliable or non-reliable publisher were included for the topic of each news report (thematic balance). Opinion articles were initially found by the publisher categorizing the article as a "comment" or "opinion" within their news page structure. They were compiled evenly over different publication periods, ensuring a chronological and, thus, thematic balance.

		Reliable		
		False	True	
Opinion	False	123	177	300
	True	146	154	300
		269	331	600

Table 1: Number of articles in dataset (news and opinion pieces) from reliable and non-reliable publishers

All articles collected were assigned by hand to a specific topic. The high granularity of the 299 assigned different topics in total is illustrated in Figure 1). This implies that the topics diversity in our data set is more refined than in data sets used in related work (Alhindi et al., 2020; Devlin et al., 2019; Krüger et al., 2017). Especially the news reports were labeled in terms of text type (opinion, news report) and publisher trustworthiness (reliable, non-reliable).

For gaining a high annotation quality all news articles were read in their entirety by at least two annotators with scientific background to ensure their informative and opinion-neutral character. Differences in annotation results were discussed. If no agreement could be reached among the annotators regarding classification, the articles were removed from the dataset. In the case of news reports, it has also been flagged whether the news was written by the news platform's own editorial team or whether it represents a news agency report. Usually, this is indicated by the publishers by means of corresponding abbreviations at the beginning or end of the news item.

Although the data set may not be very extensive, it was intentionally designed to be as complex and balanced as possible, in particular with regard to the choice of topics and reliability of publishers, making it as representative as possible.

4 Methodology

4.1 Function Words

The list of German function words was compiled from three sources. Firstly, from the list of function words in the dictionary of connectors of the Leibniz-Institut für Deutsche Sprache (grammis, 2024b). Secondly, from the lemma database of the Digital Dictionary of the German Language (der Wissenschaften, 2024). As a third source we used the function word list from Halvani and Graner (2021) and translated it from English into German. Double entries were deleted. In total, our func-

tion word list comprises 4568 context-independent function words. The function word list will also be made publicly available.

4.2 Opinion Classification

We define our problem statement as follows: let $\mathcal{D}$ denote the domain of all news and opinion articles from both reliable and unreliable publishers. The set of topics within these articles is defined as $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$, where each t_i corresponds to a distinct topic that was identified post-collection, considered its own subdomain $\mathcal{S}_\rangle$ within $\mathcal{D}$. Each subdomain $\mathcal{S}_\rangle$ comprises a collection of articles $\mathcal{A}_\rangle$, such that:

$$\mathcal{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{im}\}$$

where a_{ij} is an individual article labelled either as an opinion or a news article and categorized under the topic t_i after collection. Each article a_{ij} is associated with a label $l_{ij} \in \mathcal{L}$ where $\mathcal{L} = \{opinion, news\}$, and a publisher reliability indicator $r_{ij} \in \mathcal{R}$ where $\mathcal{R} = \{reliable, unreliable\}$. The entire corpus $\mathcal{C}$, consisting of 600 articles, can be defined as:

$$\mathcal{C} = \bigcup_{i=1}^{n} A_i$$

where $|\mathcal{C}| = 600$.

The objective is to train a classifier $f : \mathcal{C} \rightarrow \mathcal{L}$ that predicts the label l_{ij} for any article $a_{ij} \in \mathcal{C}$, based on features extracted from the articles.

To implement our models and preprocessing pipelines, we utilized *sklearn* (Pedregosa et al., 2011) for machine learning algorithms and evaluation metrics, *spacy* (Honnibal et al., 2020) for part-of-speech (PoS) tagging and matching the function words, specifically employing the spacy model pre-trained on the *de_core_news_md*[1] dataset. As preprocessing steps, we removed all contraction apostrophes and transformed all characters to lowercase, for both the input articles and the function word lists. Our features were constructed using the term frequency-inverse document frequency (TF-IDF) method on each article for each word set shown in Table 2.

The word groups represented by the PoS tags can generally be assigned either to the group of function words or to the concept of content words (Haspelmath, 2001). In particular, the word groups of nouns, verbs and adjectives can be considered as content words (lexical features), as they all refer to semantic concepts. On the other hand, function words mainly including the PoS tags of adpositions, auxiliary words, particles or conjunction. Adverbs can be both, content words or function words.

Feature Type	Description
function_words	Only function words contained in the prepared function-word list
all	All words appearing in the text
all_wo_function	All words except the function words were used to construct the tf
PoS tags	Each PoS Tag was considered as an individual word group

Table 2: Description of Text Features

In our experiments we explored various n-gram combinations of the TF-IDF representations, ultimately determining that 1-gram representations yielded the same or better performance as other n-gram combinations. We employed a Support Vector Machine (SVM) as an appropriate, standardized method for classification, regression and outlier detection in high-dimensional feature spaces, which is typical for text classification tasks. Unlike LLMs, SVMs provide explainable, traceable results. We used three different dataset splits to ensure robust evaluation and assess our model's generalisation capabilities: random-, publisher- and topic-split. This approach aimed to test the model's ability to generalise across different publishers and topics. We performed 5-fold cross-validation and used parameter grid search for hyperparameter selection. Since our dataset includes a variety of news sources - encompassing reliable and unreliable news publishers, opinion articles and news agency reports - we investigated the model's performance across these different data subsets. A detailed description of these subsets and our findings is provided in the following sections. For some of these subsets, it was not possible to maintain a balanced class distribution across all folds, so we report a weighted f1 score.

5 Experiments

We trained the SVM classifier on different subsets of the data to investigate for possible confounding

[1] de_core_news_md dataset: https://spacy.io/models/de#de_core_news_md, last visited 11.08.2025

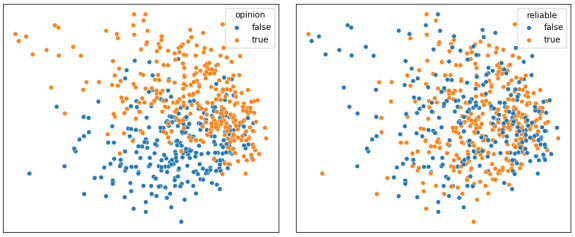
effects of publisher reliability and the presence of news agency articles. This is motivated by the following hypothesis: we expect reliable publishers to adhere more to journalistic standards with correctly labelled opinion and news articles; thus, a classifier should be able to differentiate better between these labels. Similarly, we consider news agency articles to be publisher-independent prime examples of neutral opinion-free news reporting, thus expecting that a classifier can better differentiate between such agency articles and opinion pieces. Following from this, we defined three data subsets:

- All Data: All opinion and news articles
- Opinion & Agency News: All opinion pieces and news articles from news agencies
- Opinion & Non-Agency News: All opinion pieces and news articles without agency news

For each dataset we trained a separate classifier on a random-, publisher- and topic split, using TF-IDF features constructed on function words (*function_words*), all words (*all*) and all words except for the function words (*all_wo_function*). We repeated the experiments for both the reliable and unreliable subgroups. Finally, we also trained a separate model on a random-, publisher- and topic split using TF-IDF features derived from distinct word groups associated with their respective PoS tag. Here, we only looked at the dataset containing all articles. For an initial analysis, we performed dimensionality reduction and plotted the class labels. Due to the sparsity of the TF-IDF feature vector constructed from the function words, we used truncated singular value decomposition (SVD).

6 Results

The SVD scatterplot of the articles marked by opinion label (Figure 2a) and publisher reliability (Figure 2b) show that the articles are separable by the former. The SVM classification results for the different word groups (Table 3) show that function word based features are the only word category which remains stable across the different splits. Although the *function_words* perform slighty worse than *all* and *all_wo_function*, this performance difference becomes negligible for the topic splits, with *function_words* performing equivalently well to *all* and outperforming *all_wo_function*. Some other word categories (adpositions, auxilary words, particles, conjunctions), although with overall worse



(a) Opinion (b) Publisher reliability

Figure 2: Scatterplot for functions words using TF-IDF features for all articles. Articles form two slightly overlapping clusters when categorized by opinion (a), but are not clearly separable when categorized by publisher reliability (b)

performance, also remain stable across all three splits. Notably, all these word categories have significant overlap with the function word list.

Not only do the *function_words* remain stable across splits, but also display stable behaviour over the different data subsets showing with only minor differences when news agency articles are included or excluded. This is in sharp contrast to the feature types *all* and *all_wo_function* that show significant performance loss when agency news are excluded (Table 4). When considering only reliable or unreliable news publishers, however, some performance difference can be observed with articles from reliable publishers leading to better classification results than articles from unreliable news publishers, with one exception being the *Opinion & Agency News* subset (Table 5 and Figure 3). A full list of all conducted experiments can be found in appendix A (Table 6 and Table A).

7 Discussion

Though functions words are commonly used as stop words, our experiments show that they do contain relevant information in the case of news and opinion classification and can be considered as robust features. Our work reveals that the predictions become more unstable under domain and publisher shifts if function words are excluded, supporting the content-independence of function words and the instability of lexical features (e.g. verbs, nouns, adjectives). The strong variation in classification results, i.e. the dependence on publisher reliability, topic and the inclusion or exclusion of agency news, highlights the importance of incorporating such meta information during dataset collection so that these aspects can be considered during model evaluation. The results therefore support the as-

Word Group	Random		Publisher		Topic	
	Mean	Std	Mean	Std	Mean	Std
function_words	0.88	0.02	0.87	0.02	0.87	0.03
all	0.96	0.03	0.94	0.02	0.88	0.04
all_wo_function	0.95	0.00	0.88	0.12	0.80	0.07
ADJ	0.80	0.03	0.67	0.24	0.62	0.16
ADP	0.73	0.02	0.74	0.04	0.72	0.02
ADV	0.84	0.01	0.76	0.12	0.83	0.04
AUX	0.84	0.02	0.83	0.04	0.83	0.04
NOUN	0.87	0.06	0.66	0.06	0.68	0.05
PART	0.64	0.02	0.61	0.06	0.63	0.06
PRON	0.75	0.03	0.72	0.04	0.71	0.05
CCONJ	0.66	0.05	0.66	0.03	0.65	0.03
SCONJ	0.68	0.03	0.75	0.03	0.68	0.03
VERB	0.89	0.07	0.51	0.07	0.47	0.10

Table 3: Weighted f1 score for SVM-classification using TF-IDF features constructed from different word groups on random-, publisher-, and topic-split. A more complete list of results for all word groups can be found in Appendix A

Data	function_words		all		all_wo_function	
	Mean	Std	Mean	Std	Mean	Std
All Data	0.88	0.02	0.96	0.03	0.95	0.00
Opinion & Agency News	0.88	0.07	0.97	0.01	0.92	0.02
Opinion & Non-Agency News	0.86	0.03	0.73	0.10	0.76	0.19

Table 4: Weighted f1 score for SVM-classification on different data subsets (agency news included and excluded) using three different feature types

sumption that function words do not primarily influence the topic of the article, but rather the function of the text, i.e. in our experiments the opinion-free presentation of facts (news reports) and persuasive argumentation (opinion pieces).

Moreover, it is reasonable that using the semantic subclasses of function words could lead to a better classification of the author's subjective or objective perspective within an article. This is not exclusive for the classification of opinion and news articles, as also indicated by the overlap of the two clusters. It could be concluded from the results that news agency articles are useful in particular for subjectivity level learning as the ability to distinguish between opinion pieces and agency news seems to work slightly better than the differentiability of opinion articles and non-agency news (Table 4). Following the goal to train a subjectivity detection system, it is therefore important to also consider the type and the publisher of the news articles included. Future work would be to identify further markers that differentiate opinion and news articles beyond subjectivity, possibly alleviating the overlap between these two classes.

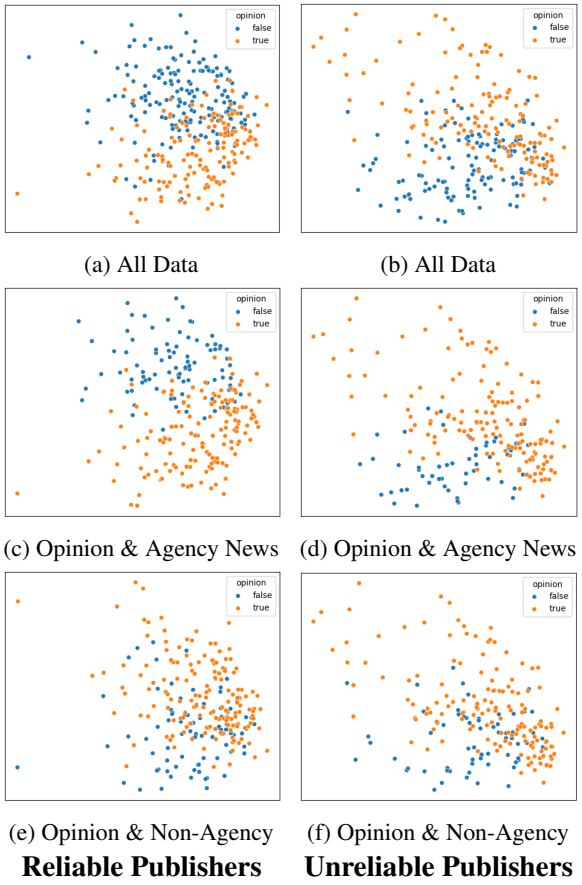
Also, considering that the classification of opinion and news reports originating from unreliable publishers is more difficult, we conclude that the difference between opinion and news pieces is less pronounced with untrustworthy publishers (see *All Data* in Table 5 as well as Figures 3a and 3b) Assuming that reliable publishers are more likely to adhere to journalistic standards in opinion pieces than non-reliable publishers, this might lead to greater homogeneity and a clearer distinction of the text type. This statement is consistent with (Lin et al., 2023), who observed different levels of formality and factuality in opinion articles depending on the outlet. However, this needs further verification utilizing a larger data set.

8 Conclusion and Outlook

In our work, we conducted experiments on the classification of opinion pieces using different lexical and grammatical word categories on the basis of a hand curated data set of German-language news reports and opinion articles originating from various online news publishers. The results demonstrate that function words (e.g. prepositions, conjunctions, pronouns) can effectively be used as markers to differentiate between opinion-free factual news

Data	All publishers		Reliable		Unreliable	
	Mean	Std	Mean	Std	Mean	Std
All Data	0.88	0.02	0.90	0.03	0.83	0.02
Opinion & Agency News	0.88	0.07	0.85	0.18	0.92	0.04
Opinion & Non-Agency News	0.86	0.03	0.88	0.03	0.81	0.01

Table 5: Weighted f1 score for SVM-classification using function word features on different data subsets (agency news included and excluded & reliable and non-reliable publishers)



(a) All Data (b) All Data

(c) Opinion & Agency News (d) Opinion & Agency News

(e) Opinion & Non-Agency (f) Opinion & Non-Agency

Reliable Publishers **Unreliable Publishers**

Figure 3: Scatterplot using functions words for TF-IDF features for three data subsets (all data, opinion & agency news and opinion & non-agency news), separated by reliable (left) and non-reliable publishers (right)

reports and opinion pieces, with stable predictions under domain shifts and across different data subsets depending on publisher reliability and type of news report (agency news or non-agency news).

However, it has to be noted that these indications were based on a carefully curated, but rather small data set. For more meaningful results, the data set should be expanded and other data sets should be incorporated. In future work, also the composition of the function word list could be examined in more detail: semantic groups of most relevant functions words could be detected and separated. Also, function words could be combined with other lexical and argumentative features that have been proven to be stable across topic splits. This would probably result in even more stable feature groups.

Further work would also be beneficial with regard to the analysis of topic granularity. Measuring the kind of distribution shift, one could experiment with various types of topic granularity, eventually defining e.g. far- and near-out of domain data sets. Last but not least, current approaches as LLMs should also be taken into account. This work is still pending.

Limitations

Our work is subject to a few limitations: First of all, it has to be noted that the compiled data set is thoroughly compiled but rather small and might still show some bias, e.g. an unbalanced number of publishers and topics in the set of opinion articles in comparison to data set of factual news reports. It would be desirable to expand the data set and conduct experiments again with additional, larger data sets, possibly also in other languages, to obtain more meaningful findings.

With regard to the labeling of topics, an alternative approach would also be feasible. We attempted to simulate topic distribution shifts by labelling the articles according to topic retrospectively. A more causally thorough approach would be to draw the articles directly from the different topic pools provided by the publisher.

The created function word list is fairly extensive. However, it has not fully been checked manually. It is possible that some function words are missing or that the list contains words whose character as function words is debatable.

Eventually, it should be pointed out that - besides our function word based feature sets - we only tested feature sets based on various PoS tags in our experiments. Other established, more complex feature sets, e.g. the linguistic features of Krüger et al. (2017), were not applied to our data set for comparison.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. Opinion: the Role of Argumentation Features in News Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming fake news: Robust generalisable news classification using transformers. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3960–3968. IEEE.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Berlin-Brandenburgische Akademie der Wissenschaften. 2024. Digitales wörterbuch der deutschen sprache (DWDS).

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, Minneapolis, Minnesota. Association for Computational Linguistics.

L. Freund, C.L.A. Clarke, and E.G. Toms. 2006. Towards genre classification for ir in the workplace. © 2006 ACM. This is an author produced version of a paper subsequently published in Proceedings of the 1st international conference on Information interaction in context. Uploaded in accordance with the publisher's self-archiving policy.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Grammatisches Informationssystem grammis. 2024a. Konnektoren als funktionale Klasse. *Leibniz-Institut für Deutsche Sprache: Systematische Grammatik*.

Grammatisches Informationssystem grammis. 2024b. Leibniz-Institut für Deutsche Sprache: Wörterbuch der Konnektoren.

Oren Halvani and Lukas Graner. 2021. POSNoise: An Effective Countermeasure Against Topic Biases in Authorship Analysis. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–12, Vienna Austria. ACM.

Oren Halvani, Lukas Graner, and Roey Regev. 2020. TAVeer: an interpretable topic-agnostic authorship verification method. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pages 1–10, Virtual Event Ireland. ACM.

M. Haspelmath. 2001. Word Classes and Parts of Speech. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 16538–16545. Elsevier.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain. Association for Computational Linguistics.

Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. 2021. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032.

K. R. Krüger, A. Lukowiak, J. Sonntag, S. Warzecha, and M. Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Zilin Lin, Kasper Welbers, Susan Vermeer, and Damian Trilling. 2023. Beyond Discrete Genres: Mapping News Items onto a Multidimensional Framework of Genre Cues. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:542–553.

Jula Lühring, Hannah Metzler, Ruggero Lazzaroni, Apeksha Shetty, and Jana Lasser. 2025. Best practices for source-based research on misinformation and news trustworthiness using newsguard. *Journal of Quantitative Description: Digital Media*, 5.

Anne McCabe and Karl Heilman. 2007. Textual and interpersonal differences between a news report and an editorial. *Revista Alicantina de Estudios Ingleses*, (20):139.

Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *American Behavioral Scientist*, 65(2):180–212.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philipp Petrenz and Bonnie Webber. 2011. Squibs: Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Cigdem Toprak and Iryna Gurevych. 2009. Document level subjectivity classification experiments in deft'09 challenge. *Actes du cinquième DÉfi Fouille de Textes*, page 91.

Klára Venglařová and Vladimír Matlach. 2024. Beyond content: discriminatory power of function words in text type classification. *Digital Scholarship in the Humanities*, page fqae013.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136.

A Detailed Experimental Results

	Random		Publisher		Topic	
	Mean	Std	Mean	Std	Mean	Std
function_words	0.88	0.02	0.87	0.04	0.87	0.03
all	0.96	0.03	0.94	0.02	0.88	0.04
all_wo_function	0.95	0.00	0.88	0.12	0.80	0.07
ADJ	0.80	0.03	0.67	0.24	0.62	0.16
ADP	0.73	0.02	0.74	0.04	0.72	0.02
ADV	0.84	0.01	0.76	0.12	0.83	0.04
AUX	0.84	0.02	0.83	0.04	0.82	0.05
DET	0.72	0.05	0.66	0.06	0.66	0.04
INTJ	0.36	0.01	0.24	0.07	0.30	0.04
NOUN	0.87	0.06	0.68	0.09	0.66	0.05
NUM	0.63	0.08	0.50	0.15	0.56	0.06
PART	0.64	0.02	0.61	0.06	0.63	0.06
PRON	0.75	0.03	0.72	0.04	0.71	0.05
PROPN	0.82	0.08	0.72	0.11	0.63	0.05
PUNCT	0.82	0.03	0.29	0.09	0.33	0.06
CCONJ	0.66	0.05	0.66	0.03	0.65	0.03
SCONJ	0.68	0.03	0.68	0.03	0.68	0.03
VERB	0.89	0.03	0.75	0.24	0.79	0.07
X	0.57	0.07	0.51	0.07	0.47	0.10

Table 6: Comparison of SVM classification using different PoS word groups to construct TF-IDF features across different dataset splits

	Opinion & News						Opinion & Agency News						Opinion & Non Agency News					
	Random		Publisher		Topic		Random		Publisher		Topic		Random		Publisher		Topic	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
function_words																		
accuracy	0.88	0.02	0.87	0.04	0.87	0.03	0.89	0.06	0.81	0.19	0.91	0.03	0.86	0.03	0.85	0.03	0.84	0.05
precision	0.88	0.05	0.88	0.06	0.84	0.05	0.87	0.07	0.66	0.39	0.89	0.04	0.87	0.04	0.86	0.07	0.83	0.06
recall	0.89	0.07	0.88	0.07	0.92	0.02	0.99	0.01	0.75	0.38	0.99	0.01	0.93	0.03	0.94	0.03	0.96	0.05
f1 weighted	0.88	0.02	0.87	0.04	0.87	0.03	0.88	0.07	0.79	0.25	0.91	0.03	0.86	0.03	0.85	0.04	0.83	0.05
all																		
accuracy	0.96	0.03	0.94	0.02	0.88	0.04	0.97	0.01	0.95	0.03	0.92	0.04	0.72	0.11	0.84	0.13	0.78	0.06
precision	0.96	0.05	0.94	0.02	0.83	0.04	0.96	0.02	0.76	0.38	0.90	0.04	0.72	0.11	0.83	0.13	0.75	0.05
recall	0.96	0.02	0.93	0.05	0.97	0.04	1.00	0.01	0.79	0.40	1.00	0.01	1.00	0.01	0.98	0.02	0.99	0.01
f1 weighted	0.96	0.03	0.94	0.02	0.88	0.04	0.97	0.01	0.96	0.02	0.92	0.04	0.61	0.17	0.81	0.19	0.73	0.10
all_wo_function																		
accuracy	0.95	0.00	0.88	0.13	0.80	0.07	0.92	0.01	0.69	0.35	0.71	0.06	0.81	0.13	0.84	0.13	0.72	0.03
precision	0.95	0.01	0.96	0.03	0.76	0.06	0.91	0.02	0.67	0.35	0.70	0.06	0.80	0.12	0.83	0.14	0.71	0.03
recall	0.94	0.01	0.84	0.17	0.89	0.15	0.98	0.02	0.80	0.40	1.00	0.01	0.99	0.01	0.98	0.01	0.99	0.02
f1 weighted	0.95	0.00	0.88	0.12	0.80	0.07	0.92	0.02	0.65	0.35	0.60	0.10	0.76	0.19	0.81	0.19	0.65	0.07

Table 7: Experimental results of SVM classification using different opinion and news article subsets for different feature sets