Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska,
Malte Ostendorff, Julián Moreno-Schneider, Georg Rehm
DFKI GmbH, Berlin, Germany
firstname.lastname@dfki.de

Abstract

We[1] present a data set consisting of German news articles labeled for political bias on a five-point scale in a semi-supervised way. While earlier work on hyperpartisan news detection uses binary classification (i. e., hyperpartisan or not) and English data, we argue for a more fine-grained classification, covering the full political spectrum (i. e., far-left, left, centre, right, far-right) and for extending research to German data. Understanding political bias helps in accurately detecting hate speech and online abuse. We experiment with different classification methods for political bias detection. Their comparatively low performance (a macro-F_1 of 43 for our best setup, compared to a macro-F_1 of 79 for the binary classification task) underlines the need for more (balanced) data annotated in a fine-grained way.

1 Introduction

The social web and social media networks have received an ever-increasing amount of attention since their emergence 15-20 years ago. Their popularity among billions of users has had a significant effect on the way people consume information in general, and news in particular (Newman et al., 2016). This development is accompanied by a number of challenges, which resulted in various NLP tasks that deal with information quality (Derczynski and Bontcheva, 2014; Dale, 2017; Saquete et al., 2020). Due to the data-driven nature of these tasks, they are often evaluated under the umbrella of (un)shared tasks, on topics such as rumour detection or verification (Derczynski et al., 2017; Gorrell et al., 2019), offensive language and hate speech detection (Zampieri et al., 2019; Basile et al., 2019; Struß et al., 2019; Waseem et al., 2017; Fišer et al., 2018; Roberts et al., 2019; Akiwowo et al., 2020) or fake news and fact-checking (Hanselowski et al., 2018; Thorne et al., 2019; Mihaylova et al., 2019).

Several shared tasks concentrate on stance (Mohammad et al., 2016) and hyper-partisan news detection (Kiesel et al., 2019), which predict either the stance of the author towards the topic of a news piece, or whether or not they exhibit allegiance to a particular party or cause. We argue that transparency and de-centralisation (i. e., moving away from a single, objective "truth" and a single institution, organisation or algorithm that decides on this) are essential in the analysis and dissemination of online information (Rehm, 2018). The prediction of political bias was recently examined by the 2019 Hyperpartisan News Detection task (Kiesel et al., 2019) with 42 teams submitting valid runs, resulting in over 30 publications. This task's test/evaluation data comprised English news articles and used labels obtained by Vincent and Mestre (2018), but their five-point scale was binarised so the challenge was to label articles as being either *hyperpartisan* or *not hyperpartisan*.

We follow Wich et al. (2020) in claiming that, in order to better understand online abuse and hate speech, biases in data sets and trained classifiers should be made transparent, as what can be considered hateful or abusive depends on many factors (relating to both sender and recipient), including race (Vidgen et al., 2020; Davidson et al., 2019), gender (Brooke, 2019; Clarke and Grieve, 2017), and political orientation (Vidgen and Derczynski, 2021; Jiang et al., 2020). This paper contributes to the detection of online abuse by attempting to uncover political bias in content.

We describe the creation of a new data set of German news articles labeled for political bias. For annotation, we adopt the semi-supervised strategy of Kiesel et al. (2019) who label (English) articles

[1]This work was done while all co-authors were at DFKI. The new affiliations of the first two authors are ambeRoad Tech GmbH, Aachen, Germany (dmitrii@amberoad.de) and Morningsun Technology GmbH, Saarbrücken, Germany (peter.bourgonje@morningsun-technology.com).

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 121–131

according to their publisher. In addition to opening up this line of research to a new language, we use a more fine-grained set of labels. We argue that, in addition to knowing whether content is hyperpartisan, the *direction* of bias (i. e., left-wing or right-wing) is important for end user transparency and overall credibility assessment. As our labels are not just about hyperpartisanism as a binary feature, we refer to this task as *political bias classification*. We apply and evaluate various classification models to the data set. We also provide suggestions for improving performance on this challenging task. The rest of this paper is structured as follows. Section 2 discusses related work on bias and hyperpartisanism. Section 3 describes the data set and provides basic statistics. Section 4 explains the methods we apply to the 2019 Hyperpartisan News Detection task data (for evaluation and benchmarking purposes) and to our own data set. Sections 5 and 6 evaluate and discuss the results. Section 7 sums up our main findings.

2 Related Work

2.1 Data sets

For benchmarking purposes, we run our system on the data from Kiesel et al. (2019). They introduce a small number of articles (1,273) manually labeled by content, and a large number of articles (754,000) labeled by publisher via distant supervision, using labels from BuzzFeed news[2] and Media Bias Fact Check[3]. Due to the lack of article-level labels for German media, we adopt the strategy of labeling articles by publisher.

Several studies use the data from *allsides.com*[4], which provides annotations on political ideology for individual articles in English. Using this data, Baly et al. (2020) introduce adversarial domain adaptation and triplet loss pre-training that prevents over-fitting to the style of a specific news medium, Kulkarni et al. (2018) demonstrate the importance of the article's title and link structure for bias prediction and Li and Goldwasser (2019) explore how social content can be used to improve bias prediction by leveraging Graph Convolutional Networks to encode a social network graph.

Zhou et al. (2021) analysed several unreliable news data sets and showed that heterogeneity of the news sources is crucial for the prevention of source-related bias. We adopt their strategy of splitting the sources into two disjoint sets used for building train and test data sets respectively.

Gangula et al. (2019) work on detecting bias in news articles in the Indian language Telugu. They annotate 1,329 articles concentrating on headlines, which they find to be indicative of political bias. In contrast to Kiesel et al. (2019), but similar to our approach, Gangula et al. (2019) treat bias detection as a multi-class classification problem. They use the five main political parties present in the Telugu-speaking region as their classification labels, but do not position these parties on the political spectrum.

Taking into account the political orientation of the author, SemEval 2016 Task 6 (Mohammad et al., 2016) worked on stance detection, where sub-task A comprised a set of tweets, the target entity or issue (e. g., "Hillary Clinton", or "Climate Change") and a label (one of *favour*, *against*, *neither*). The tweet-target-stance triples were split into training and test data. Sub-task B had a similar setup, but covered a target not included in the targets of task A, and presented the tweet-target-stance triples as test data only (i. e., without any training data for this target). While (political) stance of the author is at the core of this challenge, it differs from the problems we tackle in two important ways: 1) The task dealt with tweets, whereas we process news articles, which are considerably longer (on average 650 words per text for both corpora combined, see Section 3, compared to the 140-character limit[5] enforced by Twitter) and are written by professional authors and edited before posted. And 2) unlike the shared task setup, we have no target entity or issue and aim to predict the political stance, bias or orientation (in the context of this paper, we consider these three words synonymous and use the phrase *political bias* throughout the rest of this paper) from the text, irrespective of a particular topic, entity or issue.

One of the key challenges acknowledged in the literature is cross-target or cross-topic performance of stance detection systems (Küçük and Can, 2020). Trained for a specific target or topic (Sobhani et al., 2017), performance is considerably lower when these systems are applied to new targets. Vamvas and Sennrich (2020) address this issue by annotating and publishing a multilingual (standard Swiss

[2] https://github.com/BuzzFeedNews/2017-08-partisan-sites-and-facebook-pages

[3] https://mediabiasfactcheck.com

[4] https://www.allsides.com/media-bias

[5] The shared task took place before Twitter increased the character limit of one tweet from 140 to 280 in 2017.

German, French, Italian) stance detection corpus that covers a considerably higher number of targets (over 150, compared to six in Mohammad et al., 2016). Vamvas and Sennrich (2020) work with comments, which are longer than tweets (on average 26 words), but still shorter than our news articles. Similar to Mohammad et al. (2016) but unlike our approach, the data is annotated for stance toward a particular target.

Earlier work on political stance is represented by Thomas et al. (2006), who work on a corpus of US congressional debates, which is labeled for stance with regard to a particular issue (i. e., a proposed legislation) and which uses binary labels for supporting or opposing the proposed legislation. From this, political bias could potentially be deduced, if information on the party of the person that proposed the legislation is available. However, first of all this correlation is not necessarily present, and second, it results in a binary (republican vs. democratic) labeling scheme, whereas we use a larger set of labels covering the political spectrum from left-wing to right-wing (see Section 3).

A comprehensive review of media bias in news articles, especially attempting to cover insights from social sciences (representing a more theoretical, rational approach) and computer science (representing a more practical, empiricist approach), is provided by Hamborg et al. (2018). The authors observe a lack of inter-disciplinary work, and although our work is mainly empirical, we agree that using a more diverse range of corpora and languages is one way to move away from "too simplistic (models)" (Hamborg et al., 2018, p. 410) that are currently in use. In this respect, we would like to stress that, unlike Kulkarni et al. (2018); Baly et al. (2020); Li and Goldwasser (2019), who all either work on or contribute data sets (or both) to political bias classification in English, we strongly believe that a sub-discipline dealing with bias detection benefits especially from a wide range of different data sets, ideally from as many different languages and cultural backgrounds as possible. We contribute to this cause by publishing and working with a German data set.

2.2 Models

With regard to the system architecture, Bießmann (2016) use similar techniques as we do (bag-of-words and a Logistic Regression classifier, though we do not use these two in combination), but work on the domain of German parliament speeches, attempting to predict the speaker's affiliation based on their speech. Iyyer et al. (2014) use a bag-of-words and Logistic Regression system as well, but improve over this with a Recursive Neural Network setup, working on the Convote data set (Thomas et al., 2006) and the Ideological Book Corpus[6]. Hamborg et al. (2020) use BERT for sentiment analysis after finding Named Entities first, in order to find descriptions of entities that suggest either a left-wing or a right-wing bias (e. g., using either "freedom fighters" or "terrorists" to denote the same target entity or group). Salminen et al. (2020) work on hate speech classification. We adopt their idea of evaluating several methods (features and models, see Sections 4.1 and 4.2) on the same data and also adopt their strategy of integrating BERT representations with different classification algorithms.

3 Data Collection and Processing

We obtain our German data through two different crawling processes, described in Sections 3.1 and 3.2, which also explain how we assign labels that reflect the political bias of the crawled, German news articles. Since the 2019 shared task data which we use for benchmarking purposes is downloaded and used as is, we refer to Kiesel et al. (2019) for more information on this data set.

3.1 News-Streaming Data

This work on political bias classification is carried out in the context of a project on content curation (Rehm et al., 2020).[7] One of the project partners[8] provided us with access to a news streaming service that delivers a cleaned and augmented stream of content from a wide range of media outlets, containing the text of the web page (without advertisements, HTML elements or other non-informative pieces of text) and various metadata, such as publisher, publication date, recognised named entities and sentiment value. We collected German news articles published between February 2020 and August 2020. Filtering these for publishers for which we have a label (Section 3.4) resulted in 28,954 articles from 35 publishers. The average length of an article is 741 words, compared to 618 words for the 2019 Hyperpartisan News Detection shared task data (for the by-publisher data set).

[6] https://people.cs.umass.edu/~miyyer/ibc/index.html
[7] https://qurator.ai
[8] https://www.ubermetrics-technologies.com

Data set	Type	Far-left	Centre-left	Centre	Centre-right	Far-right	General	Regional	Overall
Training	Num. publishers	2	3	11	8	2	23	3	26
	Num. articles	1,146	11,958	11,714	15,624	1,772	41,175	1,039	42,214
Test	Num. publishers	1	3	3	2	1	8	2	10
	Num. articles	215	1,159	1,349	1,754	671	3,597	1,551	5,148

Table 1: Basic statistics of our data set.

3.2 Crawled Data

To further augment the data set described in Section 3.1, we used the open-source news crawler news-please[9]. Given a root URL, the crawler extracts text from a website, together with metadata such as author name, title and publication date.

We used the 40 German news outlets for which we have bias labels (Section 3.4) as root URLs to extract news articles. We applied regular expression patterns to skip sections of websites unlikely to contain indications of political bias[10]. This resulted in over 60,000 articles from 15 different publishers.

3.3 Data Cleaning

After collecting the data, we filtered and cleaned the two data sets. First, we removed duplicates in each collection. Because the two crawling methods start from different perspectives – with the first one collecting large volumes and filtering for particular publishers later, and the second one targeting these particular publishers right from the beginning – but overlap temporally, we also checked for duplicates in the two collections. While we found no exact duplicates (probably due to differences in the implementation of the crawlers), we checked articles with identical headlines and manually examined the text, to find irrelevant crawling output.

Second, we removed non-news articles (e. g., personal pages of authors, pages related to legal or contact information, or lists of headlines). This step was mostly based on article headlines and URLs. Because the vast majority of data collected was published after 2018, we filtered out all texts published earlier, fearing too severe data sparsity issues with the older articles. Due to the low number of articles, a model may associate particular events that happened before 2018 with a specific label only because this was the only available label for articles covering that specific event.

Finally, we inspected our collection trying to detect and delete pieces of texts that are not part of the articles (such as imprints, advertisements or subscription requests). This process was based on keyword search, after which particular articles or sections of articles were removed manually.

This procedure resulted in 26,235 articles from 34 publishers and 21,127 articles from 15 publishers[11] in our two collections respectively. We combined these collections, resulting in a set of 47,362 articles from 34 different publishers. For our experiments on this data, we created a 90-10 training-test data split. Because initial experiments showed that models quickly over-fit on publisher identity (through section names, stylistic features or other implicit identity-related information left after cleaning), we ensured that none of the publishers in the test set appear in the training data. Due to the low number of publishers for certain classes, this requirement could not be met in combination with 10-fold cross-validation, which is why we refrain from 10-fold cross-validation and use a single, static training and test data split (see Table 1).

3.4 Label Assignment

To assign political bias labels to our news articles, we follow the semi-supervised strategy of Kiesel et al. (2019), who use the identity of the publisher to label (the largest part of) their data set. The values for our labels are based on a survey carried out by Medienkompass.org, in which subjects were asked to rate 40 different German media outlets on a scale of partiality and quality. For partiality, a range from 1 to 7 was used with the following labels: *1 – left-wing extremism (fake news and conspiracy theories)*, *2 – left-wing mission (questionable journalistic values)*, *3 – tendentiously left*, *4 – minimal partisan tendency*, *5 – tendentiously right*, *6 – right-wing mission (questionable journalistic values)*, *7 – right-wing extremism (fake news and conspiracy theories)*. For quality, a range from 1 to

[9] https://github.com/fhamborg/news-please

[10] For some websites, the URL was indicative of the category, like domain.com/politics/ or domain.com/sports/. These are filtered out through regular expressions.

[11] For 25 out of the 40 root URLs, we have been unable to extract anything using the news-please crawler.

5 was used: *1 – click bait*, *2 – basic information*, *3 – meets high standards*, *4 – analytical*, *5 – complex*.

A total of 1,065 respondents positioned these 40 news outlets between (an averaged) 2.1 (indymedia) and 5.9 (Compact) for partiality, and between 1.3 (BILD) and 3.5 (Die Zeit, Deutschlandfunk) for quality. We used the result of this survey, available online[12], to filter and annotate our news articles for political bias based on their publisher. In this paper we use the bias labels for classification and leave quality classification for further research.

Because 60-way classification for partiality (1 to 7 with decimals coming from averaging respondents' answers) results in very sparsely populated (or even empty) classes for many labels, and even rounding off to the nearest natural number (i. e., 7-way classification) leads to some empty classes, we converted the 7-point scale to a 5-point scale, using the following boundaries: 1-2.5 – far-left, 2.5-3.5 – centre-left, 3.5-4.5 – centre, 4.5-5.5 – centre-right, 5.5-7 – far-right. We favoured this equal distribution over the scale of the survey over class size balance (there are more far-right articles than far-left articles, for example). The distribution of our data over this 5-point scale is shown in Table 1.

3.5 Topic Detection

To get an overview of the topics and domains covered in the data set, we applied a topic detection model, which was trained on a multilingual data set for stance detection (Vamvas and Sennrich, 2020) where, in addition to stance, items are classified as belonging to one of 12 different news topics. We trained a multinomial Naive Bayes model on the BOW representation of all German items (just under 50k in total) in this multilingual data set, achieving an accuracy of 79% and a macro-averaged F_1-score of 78. We applied this model to our own data set. The results are shown in Table 2. Note that this is just to provide an impression of the distribution and variance of topics. Vamvas and Sennrich (2020) work on question-answer/comment pairs, and the extent to which a topic detection model trained on such answers or comments is eligible for transfer to pure news articles is a question we leave for future work.

Since the majority of articles was published in 2020, a year massively impacted by the COVID-19 pandemic, we applied simple keyword-based heuristics, resulting in the estimate that approximately 40% of all articles are about COVID-19, as illustrated in the bottom rows of Table 2.

Topic	Training set	Test set
Digitisation	53	6
Economy	4,843	628
Education	1,379	126
Finances	1,309	79
Foreign Policy	8,638	969
Healthcare	925	79
Immigration	3,881	455
Infrastructure & Environment	3,132	473
Political System	5,087	563
Security	7,175	883
Society	4,077	709
Welfare	1,715	178
About COVID-19	16,994	2,414
Not about COVID-19	25,220	2,734

Table 2: Predicted topics of the articles

We publish the data set as a list of URLs and corresponding labels. Due to copyright issues, we are unable to make available the full texts.

4 Methodology

In this section we describe the different (feature) representations of the data we use to train different classification models on as well as our attempts to alleviate the class imbalance problem (Table 1).

4.1 Features

Bag-Of-Words Bag-of-Words (BOW) represents the text sequence as a vector of $|V|$ features with V being the vocabulary size. Each feature value contains the frequency of the word associated with the position in the vector in the input text. The vocabulary is based on the training data.

TF-IDF Term-Frequency times Inverse-Document-Frequency (TF-IDF) differs from BOW in that it takes into account the frequency of terms in the entire corpus (the training data, in our case). In addition to its popularity in all kinds of IR and NLP tasks, TF-IDF has recently been used in hate speech detection tasks (Salminen et al., 2019).

BERT Since its introduction, BERT (Devlin et al., 2019), has been used in many NLP tasks. We use the German BERT base model from the Hugging Face Transformers library[13]. We adopt the fine-tuning strategy from (Salminen et al., 2020): first, we fine-tune the BertForSequenceClassification model, consisting of BERT's model and a linear softmax activation layer. After training, we

[12] https://medienkompass.org/deutsche-medienlandschaft/

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

drop the softmax activation layer and use BERT's hidden state as the feature vector, which we then use as input for different classification algorithms.

4.2 Models

Logistic Regression We use logistic regression as our first and relatively straightforward method, motivated by its popularity for text classification. We add L2 regularization to the cross-Entropy loss and optimize it using Stochastic Average Gradient (SAGA) (Defazio et al., 2014).

Naive Bayes Equally popular in text classification, Naive Bayes is based on the conditional independence assumption. We model BOW and TF-IDF features as random variables distributed according to the multinomial distribution with Lidstone smoothing. BERT features are modeled as Gaussian random variables.

Random Forest Random Forest is an ensemble algorithm using decision tree models. The random selection of features and instances allows reduction of the model's variance and co-adaptation of the models. To handle class imbalance we use the Weighted Random Forest method (Chen and Breiman, 2004). This changes the weights assigned to each class when calculating the impurity score at the split point, penalises mis-classification of the minority classes and reduces the majority bias.

EasyEnsemble EasyEnsemble is another ensemble method targeting the class imbalance problem (Liu et al., 2009). It creates balanced training samples by taking all examples from the minority class and randomly selecting examples from the majority class, after which AdaBoost (Schapire, 1999) is applied to the re-sampled data.

5 Evaluation

5.1 Hyperpartisan News Detection Data

For benchmarking purposes, we first apply our models to the 2019 Hyperpartisan News Detection task. This data set uses binary labels as opposed to our 5-point scale. Since the 2019 shared task used TIRA (Potthast et al., 2019), the organisers requested submission of functioning code and ran the evaluation on a dedicated machine to which the shared task participants did not have access. The test set used in the shared task was *not* published and even after submission deadline has not been made publicly available. As a consequence, we use the validation set to produce our scores on the data. This renders a direct comparison impossible. To provide an estimate of our performance, we include Table 3, which lists the top 3 systems participating in the task. As illustrated by the row TF-IDF+Naive Bayes (our best-performing setup on this data set), we achieve a considerably lower accuracy score, but a comparable macro F_1-score. The performance of the other setups is shown in Table 3. BERT+Logistic Regression scored just slightly worse than TF-IDF+Naive Bayes, with a precision score that is one point lower.

5.2 German Data Set

We apply the models to our own data. The results are shown in Table 5 for accuracy and in Table 6 for macro-averaged F_1-score. The per-class performance is shown in Table 7, which, in addition, contains performance when binarising our labels (the last three rows) to compare this to the 2019 shared task data and to provide an idea of the difference in performance when using more fine-grained labels. We assume articles with the labels Far-left and Far-right to be hyperpartisan, and label all other articles as non-hyperpartisan. The accuracy for binary classification (not listed in Table 7) was 86%, compared to 43% (Naive Bayes+BOW in Table 5) for 5-class classification.

From the results we can conclude the following. First, class imbalance poses a serious problem, though some setups suffer from this more than others. Linear Regression, on all different features, performed poorly on the Far-left articles. We assume this is due to the small number of Far-left articles (215 in the test set, 1,146 in the training set) and publishers (one in the test set, two in the training set). Despite the high degree of class imbalance, the EasyEnsemble method, designed to target this problem particularly, does not outperform the others with any of the different feature sets. Second, BERT features scored surprisingly low with all classification models. Overall, we can conclude that the two best-performing setups that show both high accuracy and F_1-score are BOW+Naive Bayes and TF-IDF+Random Forest features. Table 7 includes the scores for TF-IDF+Random Forest, our best-performing setup.

6 Discussion

In many NLP tasks, the strategy of using BERT as a language model that is fine-tuned to a specific

Team	Rank	Accuracy	Precision	Recall	F_1
tintin	1	**0.70**	**0.74**	0.63	0.68
joseph-rouletabille	2	0.68	0.64	0.83	**0.72**
brenda-starr	3	0.66	0.63	0.81	0.71
TF-IDF + Naive Bayes (ours)	n. a.	0.58	0.55	**0.84**	0.67

Table 3: Our best performing setup (TF-IDF + Naive Bayes) on the 2019 Hyperpartisan News Detection validation set compared to the top 3 systems of the 2019 Hyperpartisan News Detection task on the by-publisher test set.

Model	Accuracy	Precision	Recall	F_1
BOW + Random Forest	0.51	0.51	0.59	0.55
BOW + Naive Bayes	0.57	0.54	0.81	0.65
TF-IDF + Random Forest	0.52	0.51	0.59	0.55
TF-IDF + Naive Bayes	0.58	0.55	0.85	0.67
BERT + Logistic Regression	0.58	0.55	0.84	0.66
BERT + Logistic Regression (10%)	0.56	0.54	0.85	0.66

Table 4: Results of our setups on the 2019 Hyperpartisan News Detection task (by-publisher validation set).

Model	BOW	TF-IDF	BERT
Logistic Regression	0.4289	**0.4472**	0.4202
Naive Bayes	**0.4304**	0.4021	0.4188
Random Forest	0.3980	0.4258	**0.4320**
EasyEnsemble	0.3811	0.3798	0.3646

Table 5: Accuracy for different features and classification methods

Model	BOW	TF-IDF	BERT
Logistic Regression	0.3132	0.2621	0.3389
Naive Bayes	**0.4243**	0.2234	0.3637
Random Forest	0.4007	**0.4303**	**0.3836**
EasyEnsemble	0.4197	0.4070	0.3432

Table 6: Macro-averaged F_1-measure for different features and classification methods

Class	Precision	Recall	F_1	Support
Far-left	0.59	0.40	0.48	215
Centre-left	0.34	0.38	0.36	1,159
Centre	0.31	0.23	0.27	1,349
Centre-right	0.51	0.55	0.53	1,754
Far-right	0.46	0.58	0.51	671
Total	**0.44**	**0.43**	**0.43**	**5,148**
Hyperpartisan	0.56	0.81	0.66	886
Non-hyperpartisan	0.96	0.87	0.87	4262
Total	**0.76**	**0.84**	**0.79**	**5,148**

Table 7: Experimental results for TF-IDF+Random Forest, per class for political bias and hyperpartisan classification.

task, has recently been shown to exhibit significant improvements over previously used methods and models, such as Naive Bayes and Random Forest. To determine why our BERT-based setups did not outperform the others, we investigated the impact of training data volume. We trained the BERT+Logistic Regression setup on only 10% of the original training data of the 2019 setup explained earlier and evaluated it on the same test setup (i. e., the validation set in the 2019 shared task). As illustrated by the last row in Table 4, the accuracy dropped by only 2% and F_1-score remained the same, suggesting that data volume has relatively little impact.

To further analyse our results, we examined the attention scores of the first BERT layer and selected the ten tokens BERT paid most attention to for every article. We then combined adjacent tokens and finished non-complete words (with their most likely candidate) to determine the key phrases of the text that the model used for classification. We repeated this procedure on all hyperpartisan articles (i. e., Far-left and Far-right) and derived a list of words and phrases that the model paid most attention to. The result is shown in Table 8.

The question whether or not attention can be used for explaining a model's prediction is still under discussion (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Note that with Table 8, we attempt to gain insight into how words are used to construct BERT embeddings, and not necessarily which words are used for prediction.

The lists of words show that the majority of words for the Far-left classification are neither exclusively nor mainly used by left-wing news media in general, e. g., *wirkt* (works), *seither* (since) or *Geliebte* (beloved, lover). An exception is *antisemi-*

Far-left	Far-right
wirkt	Checklisten
neunziger	*Willkommenskultur*
Hungernden	*Wohlverhaltensvorschriften*
antisemitische	Alltagsgebrauch
Seither	*Tichys [Einblick]*
Geliebte	Witz
Plausch	*Islam*
biologistischen	*Gutmenschen*
Sahelzone	korrekte
undurchsichtige	*Diversity*

Table 8: The top ten words most indicative of Far-left or Far-right content according to BERT's attention scores.

tische (anti-semitic), with anti-semitism in society being a common topic in left-wing media. Other highlighted words are likely to be related to the topic of refugee migration and its causes, such as *Hungernden* (hungry people) and *Sahelzone* (Sahel), an area known for its conflicts and current societal challenges. In contrast to the words we identified for the Far-left, we found most of the words we identified for the Far-right to be more descriptive of this side of the political spectrum. Nearly all words listed under Far-right in Table 8 are typically either used sarcastically or in a highly critical manner in typical right-wing media outlets. For example, *Willkommenskultur* (welcoming culture) is a German compound describing a welcoming and positive attitude towards immigrants, which is often mocked and criticised by the far right. Another example is *Gutmensch* (of which *Gutmenschen* is the plural), a term mainly used by the right as an ironic or contemptuous denigration of individuals or groups that strive to be 'politically correct'. Another word in the right column of Table 8 is *Tichys*, referring to the blog and print magazine *Tichys Einblick*. This news magazine calls itself a platform for authors of the liberal and conservative spectrum but is considered by some observers to be a highly controversial right-wing magazine with neo-liberal tendencies.[14] Since we made sure that the training data publishers and test data publishers are disjoint sets, this cannot be a case of publisher identity still being present in the text and the model over-fitting to this. Upon closer investigation, we found[15] that indeed, many other publishers refer to *Tichy's Einblick*, and these were predominantly publishers with the Far-right label.

Generally, entries in Table 8 (for both the Far-left and Far-right columns) in italics are those we consider indicative of their particular position on the political spectrum. Some words on the right side are in themselves neutral but often used by right-wing media with a negative connotation, which is why we italicised them, too (e. g., *Islam*, *Diversity*).

7 Conclusion and Future Work

We present a collection of German news articles labeled for political bias in a semi-supervised way, by exploiting the results of a survey on the political affiliation of a list of prominent German news outlets.[16] This data set extends on earlier work on political bias classification by including a more fine-grained set of labels, and by allowing for research on political bias in German articles. We propose various classification setups that we evaluate on existing data for benchmarking purposes, and then apply to our own data set. Our results show that political bias classification is very challenging, especially when assuming a non-binary set of labels. When using a more fine-grained label set, we demonstrate that performance drops by 36 points in accuracy, from 79 in the binary case to 43 in the more fine-grained setup.

Political orientation plays a role in the detection of hate speech and online abuse (along with other dimensions, such as gender and race). By making available more data sets, in different languages, and using as many different publishers as possible (our results validate earlier findings that models quickly over-fit to particular publisher identity features), we contribute to uncovering and making transparent political bias of online content, which in turn contributes to the cause of detecting hate speech and abusive language (Bourgonje et al., 2018).

While labeling articles by publisher has the obvious advantage of producing a larger number of labeled instances more quickly, critical investigation and large-scale labeling of individual articles must be an important direction of future work.

Acknowledgments

This work has received funding from the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR (no. 03WKDA1A, https://qurator.ai) and PANQURA (no. 03COV03E).

[14] https://www.politico.eu/article/new-conservative-magazine-takes-on-angela-merkel-and-the-media-roland-tichy-tichys-einblick/ (last visited: March 21, 2021).

[15] Through simple string search on "Tichy" in the articles.

[16] The URLs of the documents in our data set and the labels can be found at https://github.com/axenov/politik-news.

References

Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors. 2020. *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, Online.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Felix Bießmann. 2016. Automating political bias prediction. *CoRR*, abs/1608.02195.

Peter Bourgonje, Julián Moreno Schneider, and Georg Rehm. 2018. Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication. In *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 180–191, Cham, Switzerland. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Chao Chen and Leo Breiman. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley*.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada. Association for Computational Linguistics.

Robert Dale. 2017. NLP in a post-truth world. *Natural Language Engineering*, 23(2):319–324.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives.

Leon Derczynski and Kalina Bontcheva. 2014. Pheme: Veracity in digital social networks. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

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.

Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors. 2018. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2018. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries (IJDL)*, pages 391–415.

Felix Hamborg, Anastasia Zhukova, Karsten Donnay, and Bela Gipp. 2020. Newsalyze: Enabling news

consumers to understand media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 455–456, New York, NY, USA. Association for Computing Machinery.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020. Reasoning about political bias in content moderation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13669–13672.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1).

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

X. Liu, J. Wu, and Z. Zhou. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Nic Newman, Richard Fletcher, David A. L. Levy, and Rasmus Kleis Nielsen. 2016. Reuters institute digital news report.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA integrated research architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer.

Georg Rehm. 2018. An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena. In *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 216–231, Cham, Switzerland. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Georg Rehm, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezanezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova, and Franziska Heine. 2020. QURATOR: Innovative Technologies for Content and Data Curation. In *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Berlin, Germany. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020.

Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors. 2019. *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy.

Joni Salminen, Hind Almerekhi, Milica Milenkovic, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Jim Jansen. 2019. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media.

Joni Salminen, Maximilian Hopf, S. A. Chowdhury, Soon-Gyo Jung, H. Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943.

Robert E. Schapire. 1999. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg, pages 352 – 363, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *CoRR*, abs/2003.08385.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Emmanuel Vincent and Maria Mestre. 2018. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, volume 2276 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.