Information Processing and Management 58 (2021) 102505

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Automated identification of bias inducing words in news articles using linguistic and context-oriented features

Timo Spinde [a,b,*], Lada Rudnitckaia [a], Jelena Mitrović [c], Felix Hamborg [a,e], Michael Granitzer [c], Bela Gipp [b,e], Karsten Donnay [d,e,**]

[a] University of Konstanz, Universitätsstraße 10, DE-78464 Konstanz, Germany
[b] University of Wuppertal, Gaußstraße 20, DE-42119 Wuppertal, Germany
[c] University of Passau, Innstraße 41, DE-94032 Passau, Germany
[d] University of Zurich, Rämistrasse 71 CH-8006 Zürich, Switzerland
[e] Heidelberg Academy of Sciences and Humanities, Germany

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Media bias
Feature engineering
Text analysis
Context analysis
News analysis
Bias data set

ABSTRACT

Media has a substantial impact on public perception of events, and, accordingly, the way media presents events can potentially alter the beliefs and views of the public. One of the ways in which bias in news articles can be introduced is by altering word choice. Such a form of bias is very challenging to identify automatically due to the high context-dependence and the lack of a large-scale gold-standard data set. In this paper, we present a prototypical yet robust and diverse data set for media bias research. It consists of 1,700 statements representing various media bias instances and contains labels for media bias identification on the word and sentence level. In contrast to existing research, our data incorporate background information on the participants' demographics, political ideology, and their opinion about media in general. Based on our data, we also present a way to detect bias-inducing words in news articles automatically. Our approach is feature-oriented, which provides a strong descriptive and explanatory power compared to deep learning techniques. We identify and engineer various linguistic, lexical, and syntactic features that can potentially be media bias indicators. Our resource collection is the most complete within the media bias research area to the best of our knowledge. We evaluate all of our features in various combinations and retrieve their possible importance both for future research and for the task in general. We also evaluate various possible Machine Learning approaches with all of our features. XGBoost, a decision tree implementation, yields the best results. Our approach achieves an F_1-score of 0.43, a precision of 0.29, a recall of 0.77, and a ROC AUC of 0.79, which outperforms current media bias detection methods based on features. We propose future improvements, discuss the perspectives of the feature-based approach and a combination of neural networks and deep learning with our current system.

* Corresponding author at: University of Konstanz, Universitätsstraße 10, DE-78464 Konstanz, Germany.
** Corresponding author at: University of Zurich, Rämistrasse 71 CH-8006 Zürich, Switzerland.
E-mail addresses: Timo.Spinde@uni-konstanz.de (T. Spinde), Lada.Rudnitckaia@uni-konstanz.de (L. Rudnitckaia), Jelena.Mitrovic@uni-passau.de (J. Mitrović), Felix.Hamborg@uni-konstanz.de (F. Hamborg), Michael.Granitzer@uni-passau.de (M. Granitzer), Gipp@uni-wuppertal.de (B. Gipp), Donnay@ipz.uzh.ch (K. Donnay).

https://doi.org/10.1016/j.ipm.2021.102505
Received 31 October 2020; Received in revised form 17 December 2020; Accepted 8 January 2021
Available online 11 February 2021

1. Introduction

News articles in online newspapers are considered a crucial information source that replace traditional media such as television or radio broadcasts and print media, and new sources of information such as social media (Dallmann, Lemmerich, Zoller, & Hotho, 2015). Many people consider such articles a reliable source of information about current events, even though it is also broadly believed and academically confirmed that news outlets are biased (Wolton, 2017). Given the trust readers put into news articles and the significant influence of media outlets on society and public opinion, media bias may potentially lead to the adoption of biased views by readers (Spinde, Hamborg, Donnay, Becerra, & Gipp, 2020a). However, "unrestricted access to unbiased information is crucial for forming a well-balanced understanding of current events" (Hamborg, Donnay, & Gipp, 2018).

Highlighting media bias instances may have many positive implications and can mitigate the effects of such biases (Baumer, Elovic, Qin, Polletta, & Gay, 2015). While complete elimination of bias might be an unrealistic goal, drawing attention to its existence can not only warn readers that content is biased but also allow journalists and publishers to assess their work objectively (Dallmann et al., 2015). Furthermore, such insights could be very interesting for research projects, e.g., in social science.

We want to point out that it is uncertain if and how actual news consumers would like to obtain such information. Only a few systems are already employed to help readers mitigate the consequences of media bias impact. Most of them focus on the aggregation of articles about the same event from various news sources to provide different perspectives (Lim, Jatowt, & Yoshikawa, 2018a). For example, the news aggregator Allsides[1] allows readers to compare articles on the same topic from media outlets known to have different political views. Various media bias charts, such as the Allsides media bias chart,[2] or the Ad Fontes media bias chart[3] provide up-to-date information on media outlets' political slants.

The main objective of this paper is to present a prototypical system for the automated identification of bias-inducing words in news articles. In the following chapter, we will give an overview of research on the topic and show major currently existing drawbacks on the issue. They lead us to more fine-grained research contributions. Mainly, we will:

1. create a labeled data set for media bias analysis on different levels;
2. analyze and engineer features potentially indicating biased language;
3. train a classifier to detect bias-inducing words and
4. evaluate the performance.

This study holds both theoretical and practical significance. We summarize all existing research to give a full overview of possible classification features for media bias. We also show the relevance of all these features. We provide a data set of media bias annotations. It is the first such data set in the field, reporting word and sentence level annotations and detailed information on annotator characteristics and background. Our current data set already significantly extends available data in this domain, providing a unique and more reliable insight into bias perception. It also offers grounds for future extension. Lastly, we train and present a classifier for biased words that outperforms existing feature-based classifiers for bias.

The rest of the paper is organized as follows. Section 2 presents the literature review on media bias and its automated detection. Section 3 details the methodology, and Section 4 presents the results of this study. Finally, in Section 5 we present a discussion of the project and an outlook on future work.

2. Related work

2.1. Media bias

Media bias is defined by researchers as slanted news coverage or internal bias, reflected in news articles (Hamborg et al., 2018). By definition, remarkable media bias is deliberate, intentional, and has a particular purpose and tendency towards a particular perspective, ideology, or result (Williams, 1975). On the other hand, bias can also be unintentional and even unconscious (Baumer et al., 2015; Williams, 1975). Different news production process stages introduce various forms of media bias. In this project, we will focus on the bias that arises when journalists or, more generally, text content producers label the same concepts differently and choose different words to refer to the same concept, namely, bias caused by word choice. Depending on which words journalists select to describe an event, inflammatory or neutral, a reader can perceive the information differently. In turn, an author can manipulate a reader's perception by implying a particular opinion or perspective or inducing positive or negative emotions. The following two examples present instances of media bias by word choice, respectively:

> *1. Practicing **pro-life** litigators know that Trump judges are saving lives by permitting restrictions on abortion to go into effect.*[4]

*2. Tens of millions of children under 12 months are potentially at risk for diseases such as diphtheria and polio as the **Chinese** coronavirus pandemic interrupts routine vaccinations, according to data published by global public health experts on Friday.*[5]

[1] https://www.allsides.com/unbiased-balanced-news, accessed on 2020-10-31.
[2] https://www.allsides.com/media-bias/media-bias-chart, accessed on 2020-10-31.
[3] https://www.adfontesmedia.com, accessed on 2020-10-31.
[4] https://thefederalist.com/2020/04/24/david-french-needs-to-stop-slandering-trump-supporting-christians, accessed on 2020-10-31.
[5] https://www.breitbart.com/health/2020/05/22/report-over-80-million-children-at-risk-as-coronavirus-disrupts-vaccination-schedules/, accessed on 2020-10-31.

In the first example, the author chooses the vaguer word "pro-life" to describe the very concrete "anti-abortion" position as highly positive. In the second example, labeling the coronavirus pandemic with the word "Chinese" implies China's fault in the pandemic.

2.2. Automated identification of media bias by word choice

To the best of our knowledge, there are no tools or systems for automatic identification of media bias based on word choice. The task is a challenging one for several reasons. Firstly, while a vast amount of text data from the news is available, articles naturally are created without such sophisticated labels that could allow us to detect bias inducing words; therefore, existing data are unlabeled. Existing annotated data sets are very small. That means that currently, there is no large-scale gold standard data set for the identification of media bias by word choice (Hamborg et al., 2018). Furthermore, it has been proven that bias identification is a non-trivial task for non-experts (Lim et al., 2018a; Recasens, Danescu-Niculescu-Mizil, & Jurafsky, 2013), which can cause problems while creating such a data set. Secondly, bias words are highly dependent on the context. Therefore, simple presence of specific words is not a reliable indicator of bias (Hube & Fetahu, 2018).

1. *An abortion is the* ***murder*** *of a human baby embryo or fetus from the uterus, resulting in or caused by its death (Hube & Fetahu, 2018).*
2. *In 2008, he was convicted of* ***murder*** *(Hube & Fetahu, 2018).*

In the first example, the word "murder" describes abortion as something highly negative. In the second example, "murder" is used to describe a pure fact of murder. In the first case, the word induces bias, whereas, in the second, it does not. Another good example of context dependence of bias is the bidirectionality of the epistemological bias, i.e., this bias can occur in two cases: when a truthful proposition is questioned or when a false or controversial proposition is presupposed or implicated. The word that will not cause bias in the first case, e.g., factive verb, will cause it in the second, and vice versa (Recasens et al., 2013).

Finally, media bias and bias in reference works are subtle, implicit, and intricate since, in general, the news is expected to be factual and impartial (Baumer et al., 2015; Hube & Fetahu, 2018; Lim et al., 2018a; Recasens et al., 2013). As fairly noticed by Williams (1975), remarkable bias is not extreme but rather reasonable and plausible: extreme bias is obvious and does not threaten values or institutions.

Despite the challenging nature of the task, several researchers attempted to annotate media bias by word choice automatically. We also derive valuable insights for this project from the research attempting to identify the biased language in the related field — reference sources such as encyclopedias, where neutral language is also desired and required.

Lim et al. (2018a) use crowdsourcing to construct a data set consisting of 1235 sentences from various news articles reporting on the same event, namely, the arrest of a black man. The data set provides labels on the articles and word level. The authors then train a Support Vector Machine on the Part of Speech (POS) tags and various handcrafted linguistic features to classify bias on the sentence level, achieving the accuracy of 70%.

In related work, Lim, Jatowt, Färber, and Yoshikawa (2020) propose another media bias data set consisting of 966 sentences and containing labels on the sentence level. The data set covers various news about four different events: Donald Trump's statement about protesting athletes, Facebook data misuse, negotiations with North Korea, and a lawmaker's suicide.

Baumer et al. (2015) focus on the automated identification of framing in political news and construct a data set of 74 news articles from various US news outlets covering diverse political issues and events. They then train Naïve Bayes on handcrafted features to identify whether a word is related to framing, and achieve 61% accuracy, 34% precision, 70% recall, 0.45–0.46 F_1-score.

Hamborg, Zhukova and Gipp (2019) constructed a data set using content analysis. They created a codebook describing frame properties, coding rules, and examples. The data set consists of 50 news articles from various US news outlets and covers ten political events. The authors distinguish the target concept and phrases framing this concept, and define a number of framing properties, e.g., "affection", "aggression", "other bias", etc. The authors then automatically extract the candidates for target concepts and identify frames by looking for words semantically similar to the previously defined framing properties via exploiting word embeddings properties. Then, identified framing properties are assigned to the candidates via dependency parsing. The authors achieve an F1-score of 45.7%.

Fan et al. (2019) create the data set BASIL, annotated by two experts, covering diverse events and containing lexical and informational bias. The data set allows analysis at the token level and relatively to the target, but only 448 sentences are available for lexical bias. Then, they employ BERT lexical sequence tagger to identify lexical and informational bias at the token level and achieve an F_1-score of 25.98%.

Chen, Al-Khatib, Wachsmuth, and Stein (2020) create a data set of 6964 articles covering various topics and news outlets containing political bias, unfairness, and non-objectivity labels at the article level. They then train the recurrent neural network to classify articles according to these labels. Finally, the authors conduct a reverse feature analysis and find that, at the word level, political bias correlates with such LIWC categories (Pennebaker, Boyd, Jordan, & Blackburn, 2015) as negative emotion, anger, and affect.

Recasens et al. (2013) create static bias lexica based on Wikipedia bias-driven edits due to NPOV (Neutral Point of View) violations.[6] The bias lexicon and a set of various linguistic features are then fed into the logistic regression classifier to predict

[6] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view, accessed on 2020-10-31.

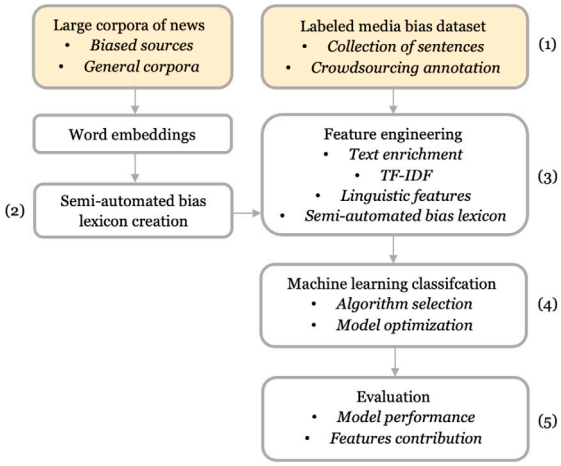

Fig. 1. General workflow of the prototypical system for automated identification of bias inducing words in news articles.

which words in the sentences are bias-inducing. The authors reached 34.35% to 58.70% accuracy for predicting 1 to 3 potential bias-inducing words in a sentence, respectively.

Hube and Fetahu (2018) propose the semi-automated approach to extract domain-related bias words lexicon, based on the word embeddings properties. The authors then feed obtained bias words and other linguistic features into a random forest classifier to detect language bias in Wikipedia at the sentence level. The authors achieve 73% accuracy, 74% precision, 66% recall, and an F_1-score 0.69 on the newly created ground truth based on Conservapedia,[7] and state that the approach is generalizable for Wikipedia with a precision of 66%.

In their later work, Hube and Fetahu (2019) train a recurrent neural network on a combination of word embeddings and a few handcrafted features to classify bias in Wikipedia at the sentence level and achieve 81.9% accuracy, 91.7% precision, 66.8% recall, and 0.773 F_1-score.

Spinde, Hamborg, and Gipp (2020b, 2020c) analyze media bias in German news articles covering the refugee crisis. The three components: an IDF component, a combined dictionary-based component, and a component based on a semantically created bias dictionary, are analyzed to identify bias on the word level. The combination of the dictionary component and the topic-dependent bias word dictionary achieves an F_1-score of 0.31, precision of 0.43, and recall of 0.26. The authors point out that considering adjectives separately increased the performance.

3. Methodology

3.1. Workflow overview

The general workflow of a prototypical system for automated identification of bias-inducing words in news articles is presented in Fig. 1. In our work, we start from collecting the sentences and gathering annotations via a crowdsourcing process (1). We then obtain various features (3) described in more detail in Section 3.5. One of the features is a bias lexicon built semi-automatically by computing words similar to potential bias words using outlet-specific word embeddings (2). We then train a supervised classifier on our engineered features and annotated labels (4). After the best model is selected and optimized, we evaluate the performance of the feature-based approach for detection of media bias. Furthermore, we evaluate all features individually (5).

3.2. Data set creation

One of the challenges in the automated detection of media bias is the lack of a gold standard large-scale data set with labeled media bias instances. The existing data sets described in Section 2.2 either do not allow for the analysis of media bias on the word level or can induce drawbacks due to the following limitations: (1) they only include a few topics (Lim et al., 2020, 2018a), (2) they mostly focus exclusively on framing (Baumer et al., 2015; Hamborg, Zhukova et al., 2019), (3) annotations are target-oriented (Fan et al., 2019; Hamborg, Zhukova et al., 2019), (4) annotations are not on the word level (Lim et al., 2018a), or (5) training data are too small (Fan et al., 2019). Therefore, we decided to create a diverse, robust, and extendable data set to identify media bias. We hand-picked 1.700 sentences from around 1.000 articles. According to our collection strategy, most of the sentences should contain media bias instances, while the smaller number of sentences should be neutral. However, the final annotations are made by

[7] https://conservapedia.com/Main_Page, accessed on 2020-10-31.

the crowd-source annotators. The sentences equally represent the full political spectrum since we used the articles from the major left and right-wing outlets, classified by their political ideology within the media bias ratings of www.allsides.com. We covered 14 different topics (selected randomly out of a variety of possible topics), from very contentious (e.g., abortion, elections) to less contentious topics (e.g., student debt, sport).

To gather annotations of the sentences and the words, we developed our own survey platform, combining classical survey functionality with text annotations, and hired participants via Amazon Mechanical Turk to complete microtasks. Annotation quality ensured by experts is often preferable, but we expressively wanted to collect a large number of annotations from non-experts. It has been shown that many complex problems can be resolved successfully through crowdsourcing if the existing crowdsourcing platforms are used in combination with appropriate management techniques and quality control (Mitrović, 2013; Mladenović, Mitrović, & Krstev, 2016)[8].

Seven hundred eighty-four annotators participated in the survey, all located in the United States. The vast majority (97.1%) of the annotators were native English speakers, 2.8% were near-native speakers. The annotators from diverse age groups participated in the survey; people from 20 to 40 years old (67.4%) prevail over other age groups. The annotators' gender is balanced between females (42.5%) and males (56.5%). The annotators have a diverse educational background; more than half have higher education. The political orientation is not well balanced: liberal annotators prevail (44.3%) over conservative annotators (26.7%) and annotators from the center (29.1%). The vast majority of the annotators read the news sometimes, more than half — one (46.4%) or more (23.1%) times per day. Each annotator received 20 randomly reshuffled sentences. We showed each sentence to ten annotators.

Within our platform,[9] we first instruct participants about the general goals of the study. We explain the tasks in detail and ask them to leave aside their personal views. We also give them a few examples of bias and ask a control question to check whether participants understood media bias's general concept. If the control question was not answered correctly, participants had to reread the instructions. Within the annotation task itself, we provide detailed instructions on the workflow. We then ask each annotator to highlight words or phrases that induce bias according to the provided instructions. After that, we ask them to annotate the whole sentence as biased or impartial, and whether they would describe it as opinionated, factual, or mixed.

To the best of our knowledge, our data set is the first in the research area to collect detailed background demographic information about the annotators, such as gender, age, education, English proficiency, but also information on political affiliation and news consumption. Overall, our data set allows for performing three different tasks: bias identification on the word level, sentence level, and a classification of the sentence as being opinionated, factual, or a mixture of both. We discuss the results of the annotation in Section 4.1.

3.3. Biased words lexicon creation

As one of our features, we present a lexicon of biased words, built explicitly for the news domain (Hube & Fetahu, 2018). Interestingly, although such a lexicon cannot serve as an exhaustive indicator of media bias due to high context-dependence (Hube & Fetahu, 2019), it can potentially serve as a useful feature of a more complex media bias detection system. To extract a biased word lexicon of high quality, we replicate the method proposed by Hube and Fetahu (2018). The authors proposed a semi-automated way to automatically extract biased words from corpora of interest using word embeddings. We present the whole pipeline of the approach in Fig. 2.

We first manually create a list of words that describe contentious issues and concepts. Then, we use this list to manually select "seed" biased words in the two separate word embedding spaces trained on news articles potentially containing a high number of biased words. We select seed biased words among the words that have high cosine similarity to the words describing contentious issues. We publish our list of seed biased words at https://bit.ly/36guHdu.[10]

We assumed that news outlets with presumably stronger political ideology would use bias words when describing contentious issues with a higher likelihood than neutral mediums. To capture both liberal and conservative biases, we train word embeddings separately on the corpora of news articles from HuffPost and Breitbart, respectively. In the choice of the outlets, we relied on the information provided by Allsides: both outlets are presented at the media bias chart,[11] and for both outlets, the confidence level of the assigned ratings is high.[12] Noteworthy, these two sources are also ones of the most popular media sources that left- and right-leaning communities share respectively in Soliman, Hafer, and Lemmerich (2019). The articles from both sources, published from 2010 to 2020, are scraped from Common Crawl[13] using NewsPlease (Hamborg, Meuschke, Breitinger and Gipp, 2017). We split the initial text into lower-cased tokens, remove punctuation marks and numbers, and train Word2Vec word embeddings (Mikolov, Chen, Corrado, & Dean, 2013). The chosen hyper-parameters are summarized in Table 1.

Since evaluation of such an unsupervised task as word embeddings creation is quite challenging (Bakarov, 2018), we choose the hyper-parameters based on the existing research (Spinde, Rudnitckaia, Hamborg, & Gipp, 2021a). The number of dimensions is set to 300 and is not increased further due to the scarcity of the training data (Mikolov et al., 2013). The window size is set to 8 since

[8] The data set described in this paper is based on a recent poster publication (Spinde, Rudnitckaia, Sinha, Gipp, & Donnay, 2021b)

[9] An anonymized version of our platform including all instructions and questions are public on http://tassy.blind-review.org/. Access to the non-anonymized platform will be granted in case of approval.

[10] The complete and non-anonymous GitHub repository will be made available in case of acceptance.

[11] https://www.allsides.com/media-bias/media-bias-chart, accessed on 2020-10-31.

[12] https://www.allsides.com/news-source/huffpost-media-bias, https://www.allsides.com/news-source/breitbart, both accessed on 2020-10-31.

[13] https://commoncrawl.org, accessed on 2020-10-31.

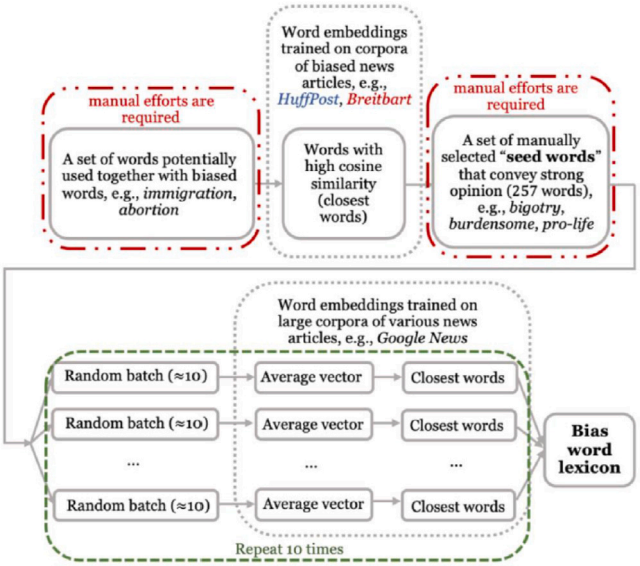

Fig. 2. Pipeline for building bias lexicon semi-automatically.

Table 1
Hyper-parameters for training word embeddings on HuffPost and Breitbart.

Hyper-parameter	Value
Dimensionality	300
Window size	8
Subsampling rate	10^{-5}
# of iterations	10
Maximum token length	28
n-grams threshold (1 pass)	90
n-grams threshold (2 pass)	120
Minimum frequency	25

larger window size can capture broad topical content (Levy, Goldberg, & Dagan, 2015). We increase the number of iterations to 10 since the size of training data is small and cannot be increased. In the scope of this project, it is important to avoid unstable low-qualitative vectors, therefore words appearing less than 25 times are excluded. Finally, treating n-grams as single units may lead to better training of a given model (Camacho-Collados & Pilehvar, 2018). We use the default scoring for n-grams generation and run two passes over training data. The thresholds for n-grams inclusion are based on manual analysis of the generated n-grams. The rest of hyper-parameters are set to the default values.

As a next step, we divide the set of seed biased words into random batches consisting of ten words and repeat this process ten times to create batches with various combinations of words. Then, for each batch, the average vector in the word embedding space trained on a 100 billion Google news data set[14] is calculated. For each average vector, we extract the top 100 words close to this average vector. Hube and Fetahu (2018) do not reshuffle words in batches and extract the top 1000 words. However, the average cosine similarity of farthest words among the top 1000 is 0.47, whereas the average cosine similarity of farthest words among the top 100 is 0.52. Besides, extracting the top 1000 words introduces noise. Finally, we add extracted words to the resulting bias word lexicon, and remove duplicates.

3.4. Detection methodology

We define bias-inducing words detection as a binary classification problem where we have only two mutually exclusive classes: whether a word is biased (class 1) or not (class 0). With our binary classifier, and in the context of media bias by word choice, no exhaustive set of precise media bias characteristics exist. Therefore, we combine different linguistic features of biased language proposed by Recasens et al. (2013) and a variety of other syntactic and lexical features (Lim et al., 2020). As the context is highly important when distinguishing between unbiased and biased words, we attempt to capture useful information from context by including collective features adding two previous and two following words into a word's feature vector. We admit that such a way

[14] https://code.google.com/archive/p/word2vec/, accessed on 2020-09-04.

Table 2
The complete set of features used in our approach for detecting biased words.

Feature	Description
POS tags	POS tag indicating the syntactic role of each word, e.g., noun, adverb, etc. Honnibal and Montani (2017).
Syntactic dependencies	Dependencies revealing how words in the text relate to each other, e.g., whether a word is a root, object, or subject (Bird, Loper, & Klein, 2009; Honnibal & Montani, 2017).
Named entity types	Named entities, e.g., persons, organizations, locations, etc. Bird et al. (2009) and Honnibal and Montani (2017).
Word vector norms	Norms of GloVe word embedding vectors pre-trained on the Common Crawl[a] Honnibal and Montani (2017).
TF–IDF	Frequency of the term in a document and in the whole article collection (Lim, Jatowt, & Yoshikawa, 2018b; Pedregosa et al., 2011).
Linguistic features	Word is a report/implicative/assertive/factive/positive/negative word, is strongly or weakly subjective, or a hedge (Recasens et al., 2013).
Additional lexica	Classifications as kill verb (Greene & Resnik, 2009), hyperbolic term (Chakraborty, Paranjape, Kakar, & Ganguly, 2016), boosters, and attitude markers (Hyland, 2019).
LIWC features	LIWC features based on psychological and psychometric analysis (Pennebaker et al., 2015).
Semi-automated bias lexicon	Previously described semi-automatically created bias word lexicon (Hube & Fetahu, 2018).

[a]https://commoncrawl.org, accessed on 2020-10-31.

to account for context is not optimal and requires elaboration in future. We compare different combinations of the features, and also train different machine learning classification algorithms, such as logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), complement Naïve bayes (NB), support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), XGBoost and a simple neural network — multilayer perceptron (MLP). To the best of our knowledge, we present the first detailed comparison of classifiers for word-level bias detection.

3.5. Feature engineering

We present our entire feature list in Table 2. In continuation, we describe the individual features and the intuition behind using them for our task. For POS tags, syntactic dependencies, named entity types, word vector norms, and linguistic features, we refer to the previous work on these topics as described by Hube and Fetahu (2018) and Recasens et al. (2013).

TF–IDF. Lim et al. (2018b) propose detection of bias using inverse document frequency (IDF) as one of the features to detect media bias, under the assumption that more rarely occurring terms are more likely to be extreme in any direction and are hence more likely to induce bias. As our data set consists of sentences where the terms are probably rarely repeated within one sentence, we adjusted Lim et al.'s assumption (Lim et al., 2018b) slightly by calculating the TF–IDF statistic basing on the whole text of articles the sentences were collected from.

LIWC Features. Linguistic Inquiry and Word Count (Pennebaker et al., 2015) is a common approach to analyzing various emotional, cognitive, and structural components in language. It identifies linguistic cues related to psychological processes such as anger, sadness, or social wording. We consider all feature categories from LIWC, as this has shown to be the most effective usage of the resource to identify bias.

Additional lexical features. It is not well known which features are the most efficient indicators of media bias (Baumer et al., 2015). Therefore, we test additional features that have been used by researchers to study similar constructs but have not been applied for the detection of media bias yet.

According to Greene and Resnik (2009), the so-called "kill verbs" together with the relevant grammatical relation (governor or dependent term) cause different sentiment perception. In the following example, the second one is perceived as more negative since it implies an intention (Yano, Resnik, & Smith, 2010):

1. *Millions of people **starved** under Stalin.*
2. *Stalin **starved** millions of people.*

Chakraborty et al. (2016) study click baits in online news and find that click bait headlines usually include hyperbolic words — words with highly positive sentiment. We assume that hyperbolic words used in click bait titles to attract readers' attention can be used to emphasize some concepts and induce bias in news articles. Hyperbolic words are, for example, "absolutely", "brilliant", or "impossibly".

Hyland (2019) introduces linguistic features that help authors to express their views on the discussed proposition. One such subcategory are boosters — words that express certainty about a particular position, e.g., "believed", "always", "no doubt". In some regard, boosters are opposite to hedges, which, on the contrary, reduce the confidence of a statement. Another subcategory are attitude markers — indicators of the author's expression of affective attitude to statements, e.g., "fortunately", "shockingly", "disappointed", etc.

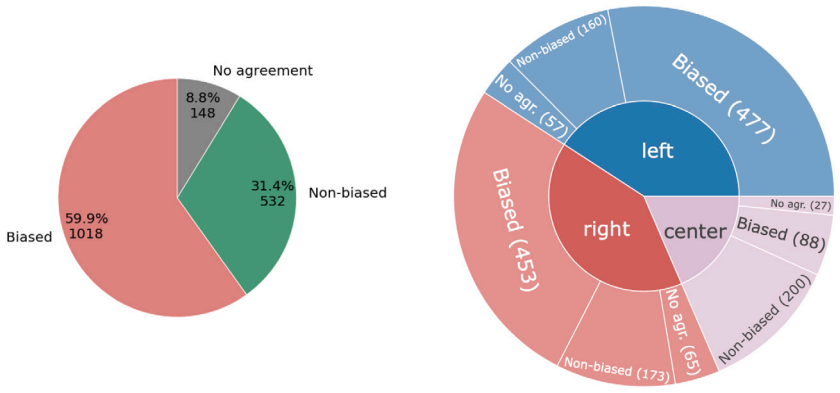

Fig. 3. Distribution of biased and non-biased sentences in the data set: on the left in general, on the right per ideology of media outlets the sentences were collected from.

3.6. Evaluation

In the final data set, classes are highly imbalanced (Section 4.1). Since accuracy does not capture the capability of a model to predict rare class correctly (Branco, Torgo, & Ribeiro, 2015), we focus on such evaluation metrics as confusion matrix, precision, recall, F1-score, and receiver operating characteristic area under curve (ROC AUC).

We compare the performance of our system to several possible baselines:

- Baseline 1 (B1) – a purely random classifier;
- Baseline 2 (B2) – occurrence of a word in the negative sentiment lexicon;
- Baseline 3 (B3) – occurrence of a word in the negative or positive lexicon;
- Baseline 4 (B4) – occurrence of a word in the semi-automated bias lexicon.

In our final data set, each observation corresponds to one word, its feature vector (including the collective context features), and the label. We perform 10-fold cross-validation when compare performance of different classifiers, 5-fold cross-validation when optimize hyper-parameters of the selected model, and finally, estimate the final performance on a test set of words that did not participate in training and manually investigate correctly and wrongly classified instances.

4. Experiments

4.1. Data set and bias perception

To gain insights into the characteristics of our data set, we analyze it quantitatively and qualitatively.

The results of sentences classification are presented in Figs. 3 and 4. The annotation results confirm our data sampling strategy: biased and non-biased statements are not balanced in the data set: biased statements prevail over non-biased statements. Besides, most media bias instances are taken from liberal and conservative news sources, whereas sources from the center were used mainly to retrieve non-biased statements. Note that this does not imply that liberal and conservative news outlets generally experience media bias by word choice and provide opinionated news more often than news outlets from the center. It is valid only due to our data collection scheme.

We assigned a biased or impartial label to a sentence if more than half of respondents annotated a sentence as biased or impartial, respectively. 149 sentences could not be labeled due to a lack of agreement between annotators. Many measures for assessing inter-annotator agreement have been used in similar computational linguistics projects. Based on the way in which we have organized our crowdsourcing workflow, i.e. having 10 annotators per task, we have decided to use Fleiss (1971) to assess the inter-annotator agreement. It represents the task's general difficulty: for example, Hube and Fetahu (2018) reported $\alpha = 0.124$ on word-level bias, and Recasens et al. (2013) reported a 40.73% agreement when looking at only the most biased word in Wikipedia statements. The value of 0.21 that we achieved can be considered as a fair agreement.

We assigned an opinionated, factual, or mixed label to a sentence if most respondents annotated a sentence as opinionated, factual, or mixed, respectively. We could not label 174 sentences due to the lack of agreement between annotators. According to our crowdsourced annotations, the data set contains an almost equal number of factual, opinionated, and mixed statements.

The annotation scheme for biased words allowed respondents to highlight not only the words but also short phrases. A word was considered biased if at least four respondents highlighted it as biased. On average, a sentence that contains biased words contains two biased words. Out of 31,794 words for training, only 3018 are biased, which constitutes only 9.5% of our current data. The types of words annotated as biased are presented in Table 3.

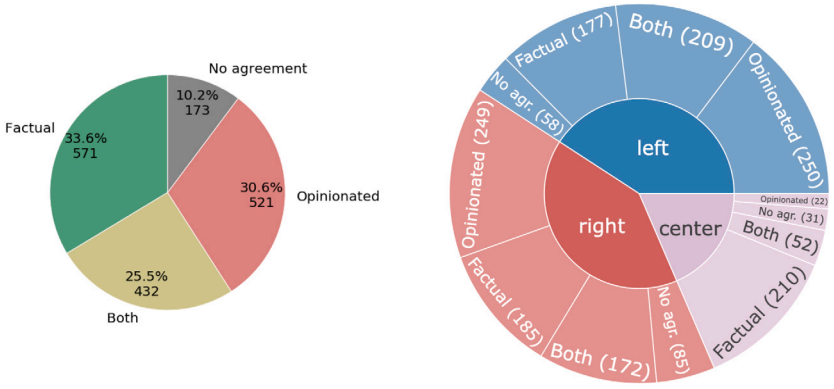

Fig. 4. Distribution of opinionated, factual and mixed sentences in the data set: on the left in general, on the right per ideology of media outlets the sentences were collected from.

Table 3
The characteristics of the words annotated as biased.

Category	Amount	Percentage
NOUN	1053	34.9%
ADJ	962	31.9%
VERB	784	26.0%
ADV	169	5.6%
PROPN	48	1.6%
Named entities	47	1.6%

We observe that annotators select not only extreme and emotional words that can be considered biased even without context, but also *context-dependent* bias words. For instance, while the word "Chinese" is generally not biased it can be in specific contexts, such as "House Democrats' **Chinese** coronavirus relief package bails out coastal millionaires and billionaires while ensuring big businesses are able to freely hire illegal aliens and visa overstayers over unemployed Americans".[15]

Albeit emphasizing in the instructions that words that are connected to very controversial topics or have very negative sentiment are not necessarily biased, some of such words were still annotated as biased. For example, the term "neo-Nazis" in the sentence "For years now, Fox News has been mainstreaming arguments that used to be the province of fringe websites run by **neo-Nazis** and other groups who believe the U.S. is meant to be a country of white people and for white people".[16]

Furthermore, we find that annotators sometimes fail to mark words as biased if a sentence contains clearly extreme and emotional words. For example, a majority of annotators marked "cray-cray" as biased but did not notice "totally" in the sentence "Over the past few decades, RFK Jr.'s famous name has helped him get in the door to talk to important people, and it probably is not long before the person who is all jacked up to meet a Kennedy realizes the guy is totally **cray-cray**".[17]

As expected, we find a positive correlation between marking sentences as biased and opinionated, and factual and non-biased. Furthermore, more controversial topics are annotated as non-biased, on average, 7.4 p.p. less than less controversial topics. Interestingly, in 49.3% of the sentences labeled as non-biased, annotators still labeled some words as biased.

Annotators who estimate themselves as conservative mark 3.76 p.p. more sentences as biased than others who describe themselves as being liberal — except if the sentence stems from a conservative news outlet (Yano et al., 2010). Furthermore, annotators who report that they check news at least sometimes, label sentences as biased 6.85 p.p. more than those who report to check news very rarely, and 19.95 p.p. more than those who report that they never check news.

4.2. Lexicon of biased words

In this section, we first present the characteristics of the articles we used to train our word embedding models and the performance of the trained word embeddings. We also provide the characteristics of the pre-trained Google News embeddings. We measure semantic word similarity and word analogy (Bruni, Tran, & Baroni, 2014; Finkelstein et al., 2001; Mikolov et al., 2013). Table 4 depicts the results of our measures. Two data sets – WordSim-353 (Finkelstein et al., 2001) and MEN (Bruni et al., 2014) – allow to estimate the Pearson correlation for the semantic similarity between pairs of words in respective word embeddings

[15] https://www.alternet.org/2019/07/fox-news-has-gone-so-deep-into-white-nationalism-that-donald-trump-now-believes-its-how-hell-win-in-2020/, accessed on 2020-10-27.

[16] https://www.alternet.org/2019/07/fox-news-has-gone-so-deep-into-white-nationalism-that-donald-trump-now-believes-its-how-hell-win-in-2020/, accessed on 2020-10-27.

[17] https://thefederalist.com/2017/01/12/no-anti-vaxxer-robert-kennedy-jr-wont-trumps-vaccine-czar/, accessed on 2020-10-27.

Table 4
Characteristics and evaluation results of word embedding models.

Corpora	# tokens	Vocab. size	WordSim-353	MEN	Google
HuffPost	68 M	53 K	0.65	0.71	0.50
Breitbart	39 M	36 K	0.57	0.59	0.38
Google News	100 B	3 M	0.62	0.66	0.74

quiddity, transcends, instigate, foolishness, nonracist, overhasty, harangue, similarly, stoically, bigotted, inuendo, gleefully, thoughtless, upbraiding, ahistoric, majoritarianism, bigots, antilabor, nauseating, postmodernists, subterfuge, defeatest, denounces, militarising, marshbaum, disloyalty, pandered, nonrational, mendaciously, blantantly, gutlessness, narrowminded, rawly, necrophiliacs, bsing, oppressive, condescension, dissemblers, brutalises, bureaucratization, scandalizes, solipsistic, delegitimise, hyping, impugns, contumely, totalistic, unwise, bureaucratize, invective, triumphalism, insinuating, mobocracy, bewails, jackassery, rankles, greedy, dishonesties, pathetic, chafes, childish, teapartiers, barbarism, sneery, obamian, resents, immobilism, carped, oppression, vilification, fuzzily, libertines, hogarthian, snub, vapidly, backpedals, incommunicable, particularist, incensed, satans, communitarians, enlightened, yobbishness, naysay, thuggy, credulously, fundamentalism, pffft, demurring, morals, maligns, tactless, scarcely, eggheaded, parvenus, wickedness, bestiality, nutty, smacks, negativists

Fig. 5. Random sample of the semi-automatically extended dictionary of biased words.

and as estimated by human assessors. The Google analogy test set (Mikolov et al., 2013) allows to evaluate accuracy. Even though those evaluation data sets are not perfectly suited for our task, the comparison shows that our data sets are large enough to give comparable results to the full Google News data set. We also manually inspected the embeddings' results and confirmed that they capture bias to a reasonable extent.

Second, we qualitatively investigate the lexicon of biased words resulting from the semi-automated expansion (Section 3.3). We manually inspected a random sample of 100 words and find that the vast majority (Around 69% in a random sample of 100 words) are negatively connotated, are emotional, and convey strong opinion (Fig. 5). Furthermore, the dictionary consists of disproportionally many rather uncommon words (e.g., "teapartiers", "obamian", "eggheaded", "mobocracy"), that are only interpretable when knowing about the circumstances they developed in. We find only one word ("similarly") that cannot directly be related to bias, while we ourselves would classify all 99 other words as being very likely to induce bias. Among 96 words for which POS tag can be identified unambiguously, 41.7% are nouns, 24.0% are verbs, 21.9% are adjectives, and 11.5% are adverbs.

Finally, we compare the method of batches developed by Hube and Fetahu (2018) to the naive approach where close words are retrieved for a single seed bias word instead of an average of a batch. We find the employed batched extraction to be superior. Specifically, while both approaches yield a high proportion of biased words, naive approach also yields many words that are not biased but co-occur with biased words, such as "abortion", "personhood", "immigrants", "mexicans", etc. Table 5 contrasts extraction for two methods.

So far, the lexicon seems to be valuable especially in finding negatively connotated neologisms and words that convey strong opinion even by themselves. Despite being more efficient than the naive approach, the method of batches, nevertheless, still cannot avoid some degree of noisy words and words falsely included as biased. Among such words are misspellings, abbreviations, and words that describe a contentious or a negative issue or concept, e.g., "xenophobia", "criticize", "anti-Semite", "solipsist". We will further evaluate the resource in future work.

4.3. Detection of biased words

We first train "quick and dirty" models (Gron, 2017) on all available features with default parameters (as implemented in Scikit-Learn (Pedregosa et al., 2011) and XGBoost (Chen & Guestrin, 2016) libraries) and compare the performance based on scores

Table 5
Comparing the method of batches and the naive approach.

Batch: ghastly, deterred, incitement, pains, hyping, unsettling, colossal, **prolife**, unscrupulous, bluntly	Single word: **prolife**	Batch: doubtful, **illegals**, harassment, instigating, unskilled, oppression, outrages, deceptively, troublemaker	Single word: **illegals**
shameful	antiabortion	splittists	illegals
calumnious	prochoice	oppression	undocumented
hyprocrisy	abortion	harassment	noncitizens
dishonest	antichoice	racist	immigrants
immoralities	personhood	tyrannize	mexicans
demonizing	faith2action	islamophobic	hispanics
disconcerts	dfla	racists	illegaly
hypocricy	naral	persecution	otms
grotesque	prolifers	harrassment	immigration
unpatriotism	nrlc	opressed	aliens
gadaon	massresistance	facism	lawbreakers
hyping	grtl	descrimination	wetbacks
sensationalization	bereit	hateful	noncriminals
disgraceful	dannenfelser	bigots	imigration
appalling	mccl	extremisms	guestworker
despicable	evangelical	udbkl	latinos
demonizing	baipa	immiseration	imigrants
hypocritical	paulites	exclusionist	alipac
reprehensible	homosexualist	nonracist	migrants
shameless	lifenews.com	mobocracy	arizonians

Table 6
Performance of algorithms for bias word detection.

Model	ROC AUC (sd)	F_1 (sd)	P (sd)	R (sd)
Logistic regression	.82 (.03)	.38 (.05)	.26 (.05)	.67 (.06)
LDA	.82 (.04)	.41 (.04)	.50 (.08)	.35 (.06)
QDA	.76 (.03)	.19 (.00)	.10 (.00)	**.99 (.02)**
NB	.82 (.03)	.35 (.04)	.23 (.03)	.74 (.05)
KNN	.70 (.03)	.21 (.04)	.45 (.09)	.14 (.03)
DT	.62 (.03)	.31 (.04)	.30 (.05)	.33 (.05)
RF	**.84 (.03)**	.26 (.04)	**.71 (.12)**	.16 (.04)
SVM (linear kernel)	.83 (.02)	.38 (.04)	.26 (.04)	.70 (.06)
SVM (rbf kernel)	.78 (.02)	.35 (.04)	.39 (.06)	.31 (.05)
XGBoost	**.84 (.03)**	**.42 (.04)**	.32 (.04)	.64 (.07)
MLP	.63 (.03)	.34 (.06)	.35 (.07)	.33 (.06)

averaged from ten-fold cross-validation (Pedregosa et al., 2011). We compare F_1-score, precision, recall, and ROC AUC. Since data are imbalanced (only $\approx$ 10% are biased), weighting of classes is employed for all methods (where possible). Table 6 shows that no model yields a high F_1-score. Instead, best performing models yield either high precision or high recall. Since results of our method are, for now, intended to be verified by a user, we prefer recall over precision while still aiming for moderately high F_1-score.

We choose XGBoost for further optimization since it achieved both the highest ROC AUC score and the highest F_1-score. It also has a relatively high recall: the model predicts more True Positives (biased words as biased) than False Negatives (biased words as unbiased). The model suffers from predicting many False Positives (unbiased words as biased) but to the smaller extent than other models with higher recall (Logistic regression, QDA, NB, SVM).

XGBoost is "a scalable end-to-end tree boosting system" (Chen & Guestrin, 2016). The algorithm is based on gradient boosting — an ensemble method that adds predictors sequentially to an ensemble, each new one is fit to the residual errors made by the previous one (Gron, 2017). Thus, the final model – a combination of many weak learners – is itself a strong learner. In addition to the fact that XGBoost already achieved best results on our data set, it provides several advantages: it accounts for sparsity caused by one-hot encoding (Chen & Guestrin, 2016), allows for a fine parameter tuning using a computationally efficient algorithm (Bentéjac, Csörgo, & Martínez-Muñoz, 2019), and allows to estimate feature importance since we do not have reliable prior information about the importance of the features (Baumer et al., 2015).

We fine-tune five hyper-parameters that help to control for overfitting.[18]

For the fine-tuned model, we quantitatively evaluate the performance and feature importance. Table 7 shows that fine-tuning yields an insignificant performance improvement of F_1 = 1p.p. We find that the model suffers from underfitting since performance

[18] We find the following values to be optimal. max-depth = 6, min-child-weight = 18, subsample = 1, colsample-bytree = 1, and eta = .2. Other hyper-parameters are set to the default values. The maximum number of boosting rounds is set to 999, and early stopping is applied if the F_1-score for validation set does not increase within ten rounds. The model is weighting the imbalanced classes. The evaluation metric is the F_1-score averaged on five-fold cross-validation.

Table 7
Excerpt of models and their performance.

Model	ROC AUC	F_1	P	R
Baselines				
B1: random	.50	.17	.10	.52
B2: negative	.69	.40	.35	.47
B3: neg. & pos.	.68	.32	.22	.57
B4: bias lexicon	.56	.20	**.62**	.12
XGBoost				
All features	**.79**	**.43**	.29	**.77**
Importance ≥ 10	.77	.41	.28	.74
Importance ≥ 400	.69	.40	.36	.47
All but TF–IDF	.77	.42	.29	.75
All but enrichment	.75	.41	.29	.67
All but linguistic	.74	.35	.23	.75
All but LIWC2015	.76	.40	.28	.72
All but bias lexicon	.77	.41	.28	.75
All but context	.78	**.43**	.30	.75

is also low on training ($F_1 = 0.51$) and validation sets ($F_1 = 0.50$). Comparing XGBoost performance to the defined baselines, we see that XGBoost significantly outperforms the random baseline (B1) but fails to significantly outperform the naive usage of the negative sentiment lexicon (B2). However, when analyzing results in a confusion matrix, we see that using just the negative dictionary, in fact, predicts 53% of biased words incorrectly as non-biased words whereas XGBoost predicts only 23% incorrectly. High F_1-score and ROC AUC of baseline B2 is mostly due to the low number of False Positives, but this is a behavior close to simply predicting all words as non-biased.

Since we do not have the prior information on which features are the most contributing into media bias detection, we first trained the classifier on the all the available features. When analyzing feature importance, we find that the most important features are the occurrence of a word in a negative lexicon (gain = 1195) and being a proper noun (470). The bias lexicon that we created semi-automatically is among the top 10 important features (gain = 106). Among linguistic features proposed by Recasens et al. (2013) as indicators of bias, only sentiment lexica, subjectivity lexica, and assertive verbs are among the top 30 important features. While report verbs and hedges still have minor importance, factive and implicative verbs have zero importance.

We train several models feeding them with features that have different importance. Excluding features with low feature importance does not improve the performance (Table 7). Besides, we test how the model performs when different feature groups are not included. Thus, we train a model with all features except one particular feature or group of features. We notice that the performance drops significantly only when linguistic features are not used, most likely because of the negative sentiment lexicon's high importance.

Lastly, we qualitatively investigate automatically detected bias candidates. Examples of correctly classified biased words (TPs) include mostly emotional words that can be considered biased even without context. Words that can be described as causing negative emotions occur more often than those causing positive emotions. 12.5% of TPs correctly indicate less obvious bias, which is most likely generally more rare. The following examples illustrate (1) obvious negative bias, (2) obvious positive bias, and (3) slightly more subtle bias among correctly classified words.

1. Large majorities of both parties seem to like the Green New Deal, despite efforts by Fox News to paint it as **disastrous**.[19]
2. Right-wing media sprung into action to try to discredit her, of course, by implying that a woman who graduated summa cum laude with an economics degree is a bimbo and with Twitchy using a screenshot to make the usually **genial** Ocasio-Cortez somehow look like a ballbuster.[20]
3. As leading 2020 Dems advocate spending **big** on the Green New Deal, it turns out most Americans are worried about other issues.[21]

Examples of biased words incorrectly classified as non-biased (FNs) include words that are (1) parts of phrases, (2) ambivalent as to whether they are biased, (3) not generally biased but only in a particular context, (4) mistakes in the annotation, and random misclassifications:

1. By threatening the kids and their families with deportation, the administration's U.S. Citizenship and Immigration Services was effectively delivering death **sentences**.[22]

[19] https://www.alternet.org/2019/04/just-a-cover-for-sexism-and-white-nationalism-paul-krugman-explains-why-the-rights-attacks-on-new-democratic-lawmakers-are-bogus/, accessed on 2020-10-31.
[20] https://www.alternet.org/2019/01/alexandria-ocaio-cortez-is-absolutely-right-there-shouldnt-be-any-billionaires/, accessed on 2020-10-31.
[21] https://fxn.ws/370GuwZ, accessed on 2020-10-31.
[22] https://www.msnbc.com/rachel-maddow-show/trump-admin-backs-plan-deport-critically-ill-children-msna1280326, accessed on 2020-10-31.

2. When the Muslim ban was first enacted, it **triggered** chaos at airports and prompted widespread protest and legal challenges, and it continues to impose devastating costs on families and people who wish to come to the U.S.[23]
3. The **specter** of "abortion regret" has been used by lawmakers and judges alike to impose or uphold rules making it harder for people to get abortions.[24]
4. Gun enthusiasts cannot admit that they like firearms because they fear black **people**.[25]

Examples of non-biased words misclassified by the model as biased (FPs) include words that (1) are ambivalent as to whether they are biased, (2) describe negative or contentious issues, (3) are due to erroneous annotation, and (4) random misclassifications:

1. Justice Sonia Sotomayor, in her dissent, **accused** the majority of weaponizing the First Amendment — an unconscionable position for a person tasked with "faithfully and impartially" discharging the duty to protect the inherent rights of all Americans.[26]
2. He also denounced the policy of Chancellor Angela Merkel and the attitude of the German media, which "are constantly pushing" for Europe to welcome more and more **migrants**, in opposition to the will of the Hungarian people.[27]
3. Michelle Williams won a Golden Globe for her role in "Fosse/Verdon" on Sunday night, but perhaps her **biggest** moment came during her acceptance speech when she defended abortion rights and encouraged women to vote "in your own self-interest".[28]
4. The case was sent back to lower **courts** to determine whether the gun owners may seek damages or press claims that the amended law still infringes their rights.[29]

5. Discussion and conclusion

One of the main contribution of our work is the creation and annotation of a robust and diverse media bias data set. Other than already existing studies on the topic, our data contain background information about the annotators, increasing our results' transparency and reliability. We perform visual analysis and observe the following findings. Topics that are less controversial are annotated as non-biased slightly more often than very controversial topics. Conservative annotators perceive statements as biased slightly more often than liberal annotators, but for both, it is only true unless the statement is from a conservative media outlet. We also find that annotators who read news never or very rarely are less likely to annotate statements as biased. Besides, our annotation results show the connection of bias and opinion.

Even though our data set was developed on a small scale and cannot serve as a gold standard for the field of media bias research, it offers a complete framework for further data extension, especially in combination with our specifically developed survey platform. We plan to extend the data set in the future. While using crowdsourcing gave us an insight into the perception of bias by a broad audience, some related issues could not be resolved, e.g., the submission of random words. Furthermore, even honest workers made mistakes because the identification of media bias is, in general, not a trivial task, especially for non-experts (Lim et al., 2018a; Recasens et al., 2013). We will follow a dual strategy in future work: While extending the existing data set, we will develop an expert annotation guideline and evaluate the same sentences in cooperation with experts in the field. We will also test the difference between the annotation of single, isolated sentences, and sentences within an article's scope. Lastly, we will evaluate from a psychological perspective how different types of questioning affect the perception of bias, and have already collected around 500 varying questions on the issue.

Regarding the quality of the annotations, our main strategy is to exclude noise from the final labels by setting up a threshold of the number of annotators required to label a word as biased or not. However, after manual analysis of final annotations, we find that setting up any strict threshold will introduce some noise and result in some words being omitted. The threshold of four is the most reasonable, but we admit that some words are omitted, and some non-biased words are included. We will experiment with more annotators per sentence to see whether we can reduce the percentage of errors.

Our semi-automatically created bias lexicon is indeed able to find emotional words and words that convey a strong opinion. However, we conclude that while capturing emotional and negative opinionated words, the lexicon is unlikely to be exhaustive. So far, the approach lacks an additional method on how to expand the lexicon without adding non-biased words.

Overall, our prototypical system achieves an F_1-score of 0.43, precision of 0.29, recall of 0.77, and ROC AUC of 0.79. Our data set is the largest and most transparent in the area to date and our classifier is the first built on these data, making a direct comparison to other methods unfeasible. On their respective data sets, researchers who detected media bias on the word level achieved an

[23] https://www.alternet.org/2020/02/conservative-magazine-denounces-trumps-cruel-expansion-of-his-muslim-ban/, accessed on 2020-10-31.

[24] https://www.alternet.org/2020/01/debunking-the-abortion-regret-narrative-data-shows-women-99-percent-of-women-feel-relief-over-their-decision/, accessed on 2020-10-31.

[25] https://www.alternet.org/2019/07/how-far-will-republicans-go-to-destroy-democracy/#.XSY6699P5zg.twitter, accessed on 2020-10-31.

[26] https://thefederalist.com/2018/06/27/was-gorsuch-worth-a-trump-presidency-its-starting-to-look-that-way/, accessed on 2020-10-31.

[27] https://www.breitbart.com/politics/2019/01/10/hungarys-orban-says-he-must-fight-french-president-macron-on-immigration/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+breitbart+%28Breitbart+News%29, accessed on 2020-10-31.

[28] https://www.nbcnews.com/news/us-news/michelle-williams-champions-woman-s-right-choose-globes-acceptance-speech-n1110986?cid=public-rss_20200106, accessed on 2020-10-31.

[29] https://www.reuters.com/article/us-usa-court-guns/u-s-supreme-court-sidesteps-major-gun-rights-ruling-but-more-cases-loom-idUSKCN22920S, accessed on 2020-10-31.

F1-score of 0.26 (Fan et al., 2019) and 0.31 (Spinde et al., 2020b); researchers, who detected framing on the word level, achieved an F1-score of 0.45–0.46 (Baumer et al., 2015; Hamborg, Zhukova et al., 2019).

We present the most complete collection of features for classification to date, extending the work of Hube and Fetahu (2018) and Recasens et al. (2013). Especially Boosters were not used in previous research, but are among the most important features. We will continue our detailed analysis of feature importance for the overall task with our larger crowdsourced data set and the expert data. We will also improve the quality of our features. For example, for implicative verbs, Pavlick and Callison-Burch (2016) introduced a method to automatically predict implicativeness of a verb based on the known constraints on the tense of implicative verbs. We could also expand our sentiment and subjectivity lexicons by using WordNet, a de facto lexico-semantic network (Fellbaum, 1998; Miller, 1995), or SentiWordNet 3.0, a lexical resource that assigns sentiment scores to each synset of WordNet (Baccianella, Esuli, & Sebastiani, 2010).

While recognizing around 77% of biased words correctly, our approach misclassifies around 20% of non-biased words. Due to the classes' imbalance, 20% of the misclassified majority class significantly decreases overall performance. In this section, we discuss the drawbacks of the implementation that lead to low performance. Especially words that are biased only in a particular context are rarely classified correctly, highlighting how media bias is usually very subtle and context-dependent. However, so far, we only accounted for context by using one collective feature for the window of four words surrounding the word. Overall, we believe that the feature-based approach is especially valuable because of its explanatory character, relating bias to specific features, which is impossible with automated feature extraction. It is also not as dependent on the amount of data as a neural network. However, we will integrate deep learning into our approach, as such an architecture especially helps to account for inter-dependencies between words. Both methods can also be combined, with our specific features giving meaning to the identification of words with automatically identified features. We will also investigate whether the classifier works better on the political left, right, or center sources. We will also determine whether we can distinguish bias for any particular ideology in our overall vocabulary of biased words.

In this paper, we propose an approach to identifying media bias using an automatic feature-based classification. To evaluate our method, we also present a 1700-sentence data set of crowdsourced biased word annotations, where we show each sentence to ten survey participants. For the first time in the research area, we report each person's background and make our data more transparent and robust. We extend existing feature sets for the task and especially evaluate each feature in detail. We also experiment with different classifiers, with our final choice returning an F_1-score of 0.43, a precision of 0.29, a recall of 0.77, and a ROC AUC of 0.79. Our results slightly outperform existing methods in the area. To increase our results further, we show how we can improve our data and features in future work and how neural networks could be a suitable option to combine with our method, even though some drawbacks have to be overcome. We publish our code and current system at https://anonymous.4open.science/r/9d305f68-e5c5-4adb-a9fa-61447540069c/.

CRediT authorship contribution statement

Timo Spinde: Initiated the project and implemented/completed all technical components as well as all writing except the discussion. **Lada Rudnitckaia:** Initiated the project and implemented/completed all technical components as well as all writing except the discussion. **Jelena Mitrović:** Helped to implement and completely financed the successful survey execution. **Felix Hamborg:** Wrote the discussion chapter. **Michael Granitzer:** Gave permanent expert feedback on our work. **Bela Gipp:** Gave permanent expert feedback on our work. **Karsten Donnay:** Helped to implement and completely financed the successful survey execution.

Acknowledgments

The Hanns-Seidel-Foundation, Germany supported this work.

SPONSORED BY THE

Federal Ministry of Education and Research

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01—S20049. The author is responsible for the content of this publication. All authors discussed the results and contributed to the final manuscript.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Bakarov, A. (2018). A survey of word embeddings evaluation methods, ArXiv 1801 (09536).

Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Human language technologies: the 2015 annual conference of the north american chapter of the ACL* (pp. 1472–1482).

Bentéjac, C., Csörgo, A., & Martínez-Muñoz, G. (2019). A comparative analysis of xgboost. ArXiv abs/1911.01914.

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with python, O'Reilly Media Inc.

Branco, P., Torgo, & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. ArXiv 1505 (01658).

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. http://dx.doi.org/10.1613/jair.4135.

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In *BlackboxNLP@EMNLP*. http://dx.doi.org/10.18653/v1/W18-5406.

Chakraborty, A., Paranjape, B., Kakar, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 9–16) http://dx.doi.org/10.1109/ASONAM.2016.7752207.

Chen, W.-F., Al-Khatib, K., Wachsmuth, H., & Stein, B. (2020). Analyzing political bias and unfairness in news articles at different levels of granularity. arXiv preprint arXiv:2010.10652.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Dallmann, A., Lemmerich, F., Zoller, D., & Hotho, A. (2015). http://dx.doi.org/10.1145/2700171.2791057.

Fan, L., White, M., Sharma, E., Su, R., Choubey, P., Huang, R., & Wang, L. (2019). In plain sight: Media bias through the lens of factual reporting. In *EMNLP/IJCNLP* http://dx.doi.org/10.18653/v1/D19-1664.

Fellbaum, C. (1998). *WordNet: An electronic lexical database and some of its applications*. MIT press Cambridge.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web* (pp. 406–414). http://dx.doi.org/10.1145/371920.372094.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters.. *Psychological Bulletin*, *76*(5), 378. http://dx.doi.org/10.1037/h0031619.

Greene, S., & Resnik, P. (2009). More than words: Syntactic Packaging and Implicit Sentiment, HLT-NAACL, http://dx.doi.org/10.3115/1620754.1620827.

Gron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems, O'Reilly Media, Inc.

Hamborg, F., Donnay, K., & Gipp, B. (2018). Automated identification of media bias in news articles: an interdisciplinary literature review, International Journal on Digital Libraries, 1–25, http://dx.doi.org/10.1007/s00799-018-0261-y.

Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (2017). news-please: A generic news crawler and extractor. In *Proceedings of the 15th international symposium of information science*, Maria and Trkulja, Violeta and Petra, Vivien, Berlin, Gaede (pp. 218–223).

Hamborg, F., Zhukova, A., & Gipp, B. (2019). Automated identification of media bias by word choice and labeling in news articles. In *Proceedings of the 19th ACM/IEEE-CS joint conference on digital libraries, urbana-champaign*, Illinois, USA, http://dx.doi.org/10.1109/JCDL.2019.00036.

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

Hube, C., & Fetahu, B. (2018).Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018* http://dx.doi.org/10.1145/3184558.3191640.

Hube, C., & Fetahu, B. (2019). Neural based statement classification for biased language. In *Proceedings of the twelfth acm international conference on web search and data mining* http://dx.doi.org/10.1145/3289600.3291018.

Hyland, K. (2019). Metadiscourse: Exploring interaction in writing, Bloomsbury Academic, New York, NY.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. http://dx.doi.org/10.1162/tacl_a_00134.

Lim, S., Jatowt, A., Färber, M., & Yoshikawa, M. (2020). Annotating and analyzing biased sentences in news articles using crowdsourcing, LREC.

Lim, S., Jatowt, A., & Yoshikawa, M. (2018). Understanding characteristics of biased sentences in news articles. In *CIKM workshops*.

Lim, S., Jatowt, A., & Yoshikawa, M. (2018). DEIM Forum 2018 C1-3 towards bias inducing word detection by linguistic cue analysis in news articles.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communations of the ACM*, *38*, 39–41. http://dx.doi.org/10.1145/219717.219748.

Mitrović, J. (2013). Crowdsourcing and its application. *INFOtheca*, *14*(1), 37–46.

Mladenović, M., Mitrović, J., & Krstev, C. (2016). A language-independent model for introducing a new semantic relation between adjectives and nouns in a WordNet. In *Proceedings of 8th global wordnet conference* (pp. 218–225).

Pavlick, E., & Callison-Burch, C. (2016). Tense manages to predict implicative behavior in verbs. In *EMNLP*, http://dx.doi.org/10.18653/v1/D16-1240.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 2825–2830.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. http://dx.doi.org/10.15781/T29G6Z.

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *ACL*.

Soliman, A., Hafer, J., & Lemmerich, F. (2019). A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and social media*.

Spinde, T., Hamborg, F., Donnay, K., Becerra, A., & Gipp, B. (2020). Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020* http://dx.doi.org/10.1145/3383583.3398585.

Spinde, T., Hamborg, F., & Gipp, B. (2020). An integrated approach to detect media bias in german news articles. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020* http://dx.doi.org/10.1145/3383583.3398585.

Spinde, T., Hamborg, F., & Gipp, B. (2020). Media bias in german news articles : A combined approach. In *Proceedings of the 8th international workshop on news recommendation and analytics (INRA 2020)*, Virtual event.

Spinde, T., Rudnitckaia, L., Hamborg, F., & Gipp, B. (2021). Identification of Biased Terms in News Articles by Comparison of Outlet-specific Word Embeddings. In *Proceedings of the iConference 2021*.

Spinde, T., Rudnitckaia, L., Sinha, K., Gipp, B., & Donnay, K. (2021). MBIC - A Media Bias Annotation Dataset Including Annotator Characteristics. In *Proceedings of the iConference 2021*.

Williams, A. (1975). Unbiased study of television news bias. *Journal of Communication*, *25*, 190–199.

Wolton, S. (2017). Are biased media bad for democracy?, Asymmetric & Private Information eJournal, Microeconomics, http://dx.doi.org/10.2139/ssrn.2285854.

Yano, T., Resnik, P., & Smith, N. (2010). Shedding (a thousand points of) light on biased language. In: Mturk@HLT-NAACL.