Understanding Characteristics of Biased Sentences in News Articles

Sora Lim
Kyoto University
Kyoto, Japan
lim.sora.88u@st.kyoto-u.ac.jp

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Masatoshi Yoshikawa
Kyoto University
Kyoto, Japan
yoshikawa@i.kyoto-u.ac.jp

Abstract

Providing balanced and good quality news articles to readers is an important challenge in news recommendation. Often, readers tend to select and read articles which confirm their social environment and their political beliefs. This issue is also known as *filter bubble*. As a remedy, initial approaches towards automatically detecting bias in news articles have been developed. Obtaining a suitable ground truth for such a task is however difficult. In this paper, we describe ground truth dataset created with the help of crowd-sourcing for fostering research on bias detection and removal from news content. We then analyze the characteristics of the user annotations, in particular concerning bias-inducing words. Our results indicate that determining bias-induced words is subjective to certain degree and that a high agreement on all bias-inducing words of all readers is hard to obtain. We also study the discriminative characteristics of biased content and find that linguistic features, such as negative words, tend to be indicative for bias.

1 Introduction

In news reporting it is important for both authors and readers to maintain high fairness, accuracy, and to keep balance between different view points. However, bias in news articles has become a major issue [GM05, Ben16] even though many news outlets claim to have dedicated policy to assure the objectiveness in their articles. Different news sources may have their own views towards the society, politics and other topics. Furthermore, they need to attract readers to make their businesses profitable. This frequently leads to the potentially harmful reporting style resulting in biased news.

To overcome news bias, as a remedy, users often try to choose news articles from news sources (outlets) which are known to be relatively unbiased. Ideally, this should be performed by corresponding recommender systems. However, bias-free article recommendations are still not feasible given the state-of-the-art. Furthermore, the recommendations might not be trusted by users, as readers often need concrete evidence of bias in the form of bias-inducing words and similar aspects.

In this paper, we focus on understanding news bias and on developing a high-quality gold standard for fostering bias-detection studies on the sentence and word levels. We assume here that word choices made by articles' authors might reflect some bias in terms of their viewpoint. For example, the phrases "illegal immigrants" and "undocumented immigrants" chosen by news reporters to refer to immigrants in relation to Donald Trump's decision to rescind Deferred Action for Childhood Arrivals may be considered as case where the choice of words can result in a bias. Here, the use of the word "illegal" degrades the immigrants by inducing more negative value than in the case of using the adjective "undocumented". By such nuanced word choices, news authors may imply their stance on the news event and deliver biased view to the readers.

It is, however, challenging to identify words that cause the article to have biased points of view [BEQ+15]. The bias inherent in news articles tend to be subtle and intricate. In this research, we construct a comparable news dataset which consists of news articles reporting the same news event. The objective is to help designing methods to detect bias triggers[1] and

[1] https://github.com/skymoonlight/newsdata-bias

shed new light on the way in which users recognize bias in news articles. To the best of our knowledge, this is the first dataset with annotated bias words in news articles. In the following, we describe the design of the crowd-sourcing task to obtain the bias labels for the news words and we subsequently analyze the characteristics of detected biased content in news.

2 Related Works

Several prior works have focused on media bias in general and news bias in particular. Generally, according to D'Allessio and Allen [DA00], media bias can be divided into three different types: (1) *gatekeeping*, (2) *coverage* and (3) *statement bias*. *Gatekeeping bias* is a selection of stories out of the potential stories; *coverage bias* expresses how much space specific positions receive in media; *statement bias*, in contrast, denotes how an author's own opinion is woven into a text. Similarly, Alsem *et al.* [ABHK08] divide news bias into *ideology* and *spin*. *Ideology* reflects news outlets' desire to affect readers' opinions in a particular direction. *Spin* reflects the outlet's attempt to simply create a memorable story. Given these distinctions, we consider the bias type tackled in this paper as *statement bias* w.r.t. [DA00] and as *spin bias* according to [ABHK08].

Several researches made efforts to provide effective means for solving the news bias problem. However, most of them have focused on the news diversification according to the content similarity and the political stance of news outlets. Park *et al.* [PKCS09], for instance, have developed a news diversification system, named *NewsCube*, to mitigate the bias problem by providing diverse information to the users. Hambourg *et al.* [HMG17] presented a matrix-based news analysis to display various perspectives for the same news topic in a two-dimensional matrix. An *et al.* [ACG+12] revealed skewness of news outlets by analyzing their news contents spread throughout tweets.

Alonso *et al.* [ADS17] focused on omissions between news statements which are similar but not identical. The omission occupies one category in news bias in that it is a means of statement bias [GS06]. Ogawa *et al.* [OMY11] attempted to describe the relationship between main participants in news articles to detect news bias. To catch describing way of the relationship, they expanded sentiment words in SentiWordNet [BES10].

Other works focused on linguistic analysis for bias detection on text data. Recasens *et al.* [RDJ13] targeted detecting bias words from the revised sentence history in Wikipedia. They utilized NPOV tags for bias labels, and linguistically categorized resources for the bias feature. Baumer *et al.* [BEQ+15] used Recasens *et al.*'s linguistic features to identify biased language in political news as well as features from theoretical literature on framing.

Table 1: Statistics of Labeled Sentences

Total number of news articles	88
Total number of sentences	1,235
Average tagged sentences per a news article	73.48%
No. of sentences including tagged words	826 (66.88%)
No. of tagged sentences on agreement level 2	431 (34.90%)
No. of tagged sentences on agreement level 3	173 (14.01%)
No. of tagged sentences on agreement level 4	42 (3.40%)
No. of tagged sentences on agreement level 5	7 (0.57%)

3 Annotating Bias in News Articles

3.1 Dataset

To detect the subtle differences which cause bias, one way is to compare words across the content of different news articles which are reporting the same news event. This should allow for pinpointing differences in the subtle use of words by different authors from diverse media outlets to describe the same event. Although, many news datasets were created for news analysis, to the best of our knowledge, none focused on a single event while, at the same time, covering many news articles from various news outlets from a short time range.

We selected the news event titled "Black men arrested in Starbucks" which has caused controversial discussions on racism. The event happened on April 12, 2018. We focused on news articles written on April 15, 2018 as the event was widely reported in different news on that day.

For collecting news articles from various news outlets we used Google News[2]. Google News is a convenient source for our case as it already clusters news articles concerning the same event coming from various sources. We first crawled all news articles available online that described the aforementioned event. Based on manual inspection, we then verified whether all articles are about the same news event. We next extracted the titles and text content from the crawled pages ignoring pages which covered only pictures or contained only a single sentence. In the end, our dataset consists of 89 news articles with 1,235 sentences and 2,542 unique words from 83 news outlets. Articles contain on average 14 paragraphs.

3.2 Bias Labeling via Crowd-Sourcing

To overcome scalability issue in annotations, crowdsourcing has been widely used [FMK+10, ZLP+15]. We also use crowdsourcing to collect bias labels and

[2] https://news.google.com/?hl=en-US&gl=US&ceid=US:en

we choose Figure Eight[3] as our platform. Figure Eight (called CrowdFlower until March 2018) has been used in a variety of annotation tasks and is especially suitable for our purposes due to the focus on producing high-quality annotations. We note that it is difficult to obtain bias-related label information such as binary judgements on each sentence of news articles, as the bias may depend on the news event and its context. To design the bias labeling task, we divided the news dataset into one *reference* news article[4] and 88 *target* news articles. Having a reference news article, users could first get familiar with the overall event. Furthermore, the motivation was to have some reference text which being relatively bias-free allows for detecting bias content in a target article. Our reference article has been selected after being manually judged as relatively unbiased according to several annotators.

We let the workers make judgements on each target news article (using also the reference news article). Each article has been independently annotated by 5 workers. In order to ensure a high-quality labeling, we produced various test questions to filter out low quality answers. To create reliable answers to our test questions, we conducted a preliminary labeling task on a set of five randomly selected news articles from the same news collection, plus the same reference news article used for comparison. Nine graduate students (male: 6, female: 3) labeled bias-inducing words in these news articles. The words which have been labeled as "bias-inducing" by at least two people were considered as "biased" in general and served as ground truth for our test questions.

The instructions and main questions given to the workers in the crowdsourcing tasks and to annotators in the preliminary task can be summarized as follows:

1. Read the target news article and the reference news article.
2. Check the degree of bias of the target news article by comparing with the reference news article.
   - not at all biased, slightly biased, fairly biased, strongly biased.
3. Select and submit words or phrases which cause the bias, compared to the reference news article.
   - Submit words or phrases with the line identifier.
   - Try to submit as short as possible content and don't submit whole paragraphs.
   - If no bias inducing words are found, submit "none".
4. Select your level of understanding of the news story
   - four scale ratings from "I didn't understand at all." to "I understood well."

In total, 60 workers participated in the task. We only used the answers from 25 reliable workers who passed at least 50% of test questions. Overall, for 88 documents, we collected 2,982 bias words (1,647 unique words) covered by 1,546 non-overlapping annotations.

3.3 Analysis of Perceived News Bias

We next analyze what kind of words are tagged as bias triggers by the workers. First, we analyze the phrases annotated as biased in terms of the word length. Each annotation consists of four words on average (examples being "did absolutely nothing wrong", "putting them in handcuffs", "racism and racial profiling", "merely for their race", and "Starbucks manager was white"). Most answers submitted by workers are, however, single words, for example, "accuse", "absurd", "boycott", "discrimination", and "outrage". These examples also show a tendency of negative sentiment and that rather extreme, emotion-related words are annotated, which could be extracted almost without considering the context. As second most frequent phrase pattern, three words in a sentence have been annotated, such as "absolutely nothing wrong", "accusations of racism", "black men arrested", "who is black", and "other white ppl". These are typical combinations of sentiment words and modifiers or intensifiers. These sentiment words (with positive or negative polarity) are typically associated with the overall topic or event and can also be considered as outstanding or salient to some degree.

We aggregated the answers of the crowd-workers on the sentence level assuming that if a sentence includes any word annotated as biased, the sentence itself is biased. Note that the information on sentence level bias might be enough for the purpose of automatic bias detection. However, we let users annotate the specific bias-inducing phrases, since this lets us gain a fine-grained insight in the actual thoughts of users and allows to choose appropriate machine learning features for bias-detection algorithms, as well as to show concrete evidence of bias-inducing aspects in the texts to users. Table 1 shows the statistics of the dataset and labeled results. Agreement level n denotes that only annotations tagged by at least n people are considered. When we only consider the unique, i.e., fusioned answers from the workers, among 1,235 sentences in the whole data set, 826 sentences (66.88%) included bias-annotated words. On average, 73.48% of the sentences would be then considered potentially biased in an article. Yet, assuming an agreement of 2 workers the average number of biased sentences is 34.9%, while for $n = 3$ the corresponding number is 14.01%. These statistics reveal that people consider different words as representing biased content through different words.

Inter-rater agreement. We next investigated the inter-rater agreement among the five workers' answers

[3] https://www.figure-eight.com/.
[4] https://reut.rs/2ve3rMz

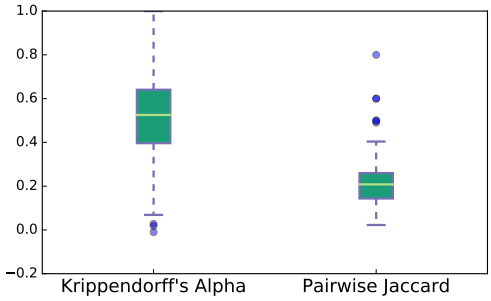

Figure 1: Inter-rater reliability on the Crowdsourcing result: (a) Krippendorff's alpha (b) Pairwise Jaccard.

for the each target news. We calculated Krippendorff's alpha and pairwise Jaccard similarity coefficients. Krippendorff's alpha are used for quantifying the extent of agreement among multiple raters, and Jaccard similarity is mainly used for comparing the similarity between two sets. Here, we regard each sentence in a target news as item to be measured. The mean scores calculated over all the target articles are 0.513 for Krippendorff', and 0.222 for Jaccard, as shown also in Figure 1. The agreement scores show relatively low tendency which means the answers from the five workers are diverse and with slight agreement. In practice, it is hard to get substantial agreement on news articles in general [NR10]. This may have several reasons in our case: Firstly, the degree of perception concerning bias differs from person to person. Secondly, the answer coverage by people is different and imperfect. For example, some people might feel it is enough to submit around five different answers on a target news article, while others might try to find as many as possible evidences of biased content. It is then hard to decide whether the differences are from insincerity of individuals or the matter of their perception.

Analysis of POS tags. We investigated the part of speech tags included in the sentences. The Stanford POS Tagger [TKMS03] was employed in this process. To that end, we considered different agreement levels, i.e., the minimum number of users who tag words as biased in the same sentences. We conducted the t-test for the bias tagged sentences and non-tagged sentences. Table 2 shows the statistically significant POS tags under the p-value < 0.001.

Analysis of further linguistic features. We also investigate words by using the linguistic categories proposed by [RDJ13], including sentiment, subject/object, verb types, named entity and so on. In Table 3, we observe that the most significant word category is negative subject words in agreement level 1. Also weak subject words and negative words are shown to be significant. We believe this result is because our news event is controversial and related to the arrest, therefore, many negative words affect to the bias cognition of users. Interestingly, factive verbs do not show any significant difference.

Table 2: POS Feature Effects by t-test in Each Agreement Level[5]

Agreement Level	1	3	5
Cardinal number (CC)	5.19	4.0554	
Determiner (DT)	4.87		-4.4403
Existential there (EX)	3.81		-6.9333
Preposition/subordinating participle conjunction (IN)	7.63	3.4378	
Adjective (JJ)	9.2987	3.4507	
Adjective, superative (JJS)			-7.6947
Noun (NN)	7.5422		
Noun, plural (NNS)	5.3969		
Predeterminer (PDT)	3.7788		-8.7549
Adverb	5.3142		
Adverb, superative (RBR)		-3.4822	-3.4797
Particle	5.6674		-11.969
Verb, past tense (VBD)	6.5408		
Verb, gerund/present (VBG)	7.4645	3.3702	
Verb, past participle (VBN)	8.2355	4.0162	-2.6979
Verb, 3rd ps. sing. present (VBZ)	6.1593	3.713	
Wh-pronoun (WP)	5.4197	2.4701	
Wh-adverv (WRB)			-15.243

For the preliminary experiments, we next use the POS tags and the mentioned linguistic features for approaching the task of automatically detecting bias. We employ a standard SVM model and use randomly selected 80% of the sentences for training the model and the remaining 20% of sentences for testing. The classification accuracy is 70%. As our data set is primarily designed for linguistic analysis, larger numbers of train/test examples are needed for obtaining more reliable evaluation results.

Further extensions. We analyzed bias in the news sentences perceived by people using crowdsourcing. In this research, we used a news event that occurred in a short time period. Thus, users do not need to spend much time to understand the context of the news event. However, in case of a long time lasting news event, the news topic tends to be complicated or consists of many sub-events and there might be many aspects to be aware of. For example, politics-related news events, typically have a long time span when they cover elections the reports on actions of candidates appear in the weeks beforehand. For detecting and/or minimizing the news bias under more complex situations, an alternative strategy for obtaining a rea-

[5]Only significant results are shown ($p < 0.001$).

Table 3: Linguistic Feature Effects by t-test in Each Agreement Level[5]

Agreement Level	1	3	5
Factive verb			-10.154
Assertive verb		-3.2339	-4.3784
Implicative verb		-3.7975	
Entailment		-2.7975	
Weak subject word	5.5862	4.917	
Negative word	7.5961	5.6002	
Bias Lexicon		-2.9986	
Named Entity		3.375	
Negative subject words	9.7921	8.2414	

sonable ground truth concerning news bias might be to focus on credibility aspects and to target the recommendation of citations to clearly and formally stated facts and/or events, such as ones in existing knowledge bases.

4 Conclusions and Future Works

Detecting news bias is a challenging task for computer science as well as linguistics and media research areas due to the subtle nature and heterogeneous, diverse kinds of biases. In this paper, we set up a crowdsourcing task to annotate news articles concerning bias-inducing words. We then analyzed features concerning the annotated words based on different user agreement levels. Based on the results, we make the following conclusions:

1. Generally, it is hard to reach an agreement among users concerning biased words or sentences.
2. According to results, it is reasonable to focus on linguistic features, such as negative words, negative subjective words, etc. for detecting bias on a word level. This also means that for detecting bias, capturing the context, such as having semantically-structured representations of statements or sentences might not be needed for a shallow bias detection.
3. Our experiments on the characteristics of bias-inducing words indicate that presenting the readers with bias-inducing words (e.g., by highlighting them in the text) is still worthwhile to be pursued in the future.
4. A deeper analysis of bias in the news is needed. Current efforts, such as the *SemEval 2019 Task 4* ("Hyperpartisan News Detection")[6], can be seen as first steps in this direction. More generally, we argue that we need novel ways to measure the actual bias of news (and other texts). This could be achieved by measuring the effect of article reading by not only asking readers before and after the reading about their opinion on topic/event, but also by correlating the read news with actions, such as the votes of readers in upcoming elections.

[6] `https://pan.webis.de/semeval19/semeval19-web/`

Acknowledgments This research was supported in part by MEXT grants (#17H01828; #18K19841; #18H03243).

References

[ABHK08] Karel Jan Alsem, Steven Brakman, Lex Hoogduin, and Gerard Kuper. The impact of newspapers on consumer confidence: does spin bias exist? *Applied Economics*, 40(5):531–539, 2008.

[ACG+12] Jisun An, Meeyoung Cha, Krishna P Gummadi, Jon Crowcroft, and Daniele Quercia. Visualizing media bias through Twitter. In *Proc. of ICWSM SocMedNews Workshop*, 2012.

[ADS17] Héctor Martínez Alonso, Amaury Delamaire, and Benoît Sagot. Annotating omission in statement pairs. In *Proc. of LAW@EACL 2017*, pages 41–45, 2017.

[Ben16] W Lance Bennett. *News: The politics of illusion*. University of Chicago Press, 2016.

[BEQ+15] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proc. of NAACL HLT 2015*, pages 1472–1482, 2015.

[BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proc of LREC 2010*, 2010.

[DA00] Dave D'Alessio and Mike Allen. Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50(4):133–156, 2000.

[FMK+10] Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proc. of CSLDAMT'10*, pages 80–88, 2010.

[GM05] Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237, 2005.

[GS06] Matthew Gentzkow and Jesse M Shapiro. Media bias and reputation. *Journal of political Economy*, 114(2):280–316, 2006.

[HMG17] Felix Hamborg, Norman Meuschke, and Bela Gipp. Matrix-Based News Aggregation: Exploring Different News Perspectives. In *Proc. of JCDL 2017*, pages 69–78, 2017.

[NR10] Stefanie Nowak and Stefan M. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proc. of MIR 2010*, pages 557–566, 2010.

[OMY11] Tatsuya Ogawa, Qiang Ma, and Masatoshi Yoshikawa. News bias analysis based on stakeholder mining. *IEICE Transactions*, 94-D(3):578–586, 2011.

[PKCS09] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proc. of SIGCHI on Human Factors in Computing Systems*, pages 443–452, 2009.

[RDJ13] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proc. of ACL 2013*, volume 1, pages 1650–1659, 2013.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proc. of HLT-NAACL 2003*, pages 173–180, 2003.

[ZLP+15] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proc. of WWW 2015*, pages 347–353, 2015.