Bias in Newspaper Headlines

Constantin Brîncoveanu
Timo van Niedek

***Abstract*—The aim of this project is to detect bias in newspaper headlines using machine learning. A dataset containing world news headlines was obtained using the Reddit API. Bias detection was achieved using several approaches. Firstly, Sentiment Analysis was applied on the headlines. Mean sentiment values can provide information about possible biases. Topic Detection was performed to categorize the headlines, and then Sentiment Analysis was performed on those topic classes, as well as Flavor Analysis. Statistical tests yielded significant biases for a few domains. Lastly, we employed domain classification to gain a different perspective on the biases.**

I. Introduction

Although newspapers tend to claim neutrality, it is almost impossible to find a source that does not contain any stereotypes or biases in some form. Biases in news headlines pose a danger because they can propagate stereotypes to the general public. The uninterested or time-pressed reader will skim the headlines of newspaper articles or online news collections, often missing important nuances. It is therefore important to be able to detect biased language not only in the body of a news article, but also in the headline by itself. An automatic bias detection system could help guide readers, and expose the hidden biases that are present in news headlines.

Detecting biases manually requires an enormous amount of effort to code news headlines and apply theoretical bias frameworks. It is infeasible to apply this process to the vast amounts of news stories that are released every day. An alternative approach that has gained interest recently is to detect biases automatically using machine learning techniques. These systems employ natural language processing techniques to detect one of the many ways in which bias can be present in texts. One such method is sentiment analysis, in which the polarity of the text is determined, either positive, negative or neutral. Sentiment analysis is typically applied to highly subjective texts for which the polarity is very clear and has high variance across difference texts. The difficulty with using sentiment analysis on news headlines to detect bias is that the polarity scores lie closely together, and therefore do not clearly indicate a bias on its own. News outlets report negative events such as war, death or protests more often than positive events, which makes it difficult to compare polarity scores. Additionally, bias detection using polarity as a measure is not enough since negative scores appear often, even unbiased reports. We therefore propose a technique that investigates the difference between the headline sentiment and the mean of the polarity scores for a specific topic.

Cognitive Computational Modeling of Language and Web Interaction,
SOW-MKI61-2016-SEM2-V, 13th July 2017, Dr. G.E. Kachergis.
email: c.brincoveanu@student.ru.nl,timo.niedek@student.ru.nl

Another method of investigating news bias is to measure the amount of flavor words used by a specific news source. These flavor words are adjectives, adverbs, comparatives and superlatives. The motivation behind this method is that these words can carry bias, and a truly objective report would need less of them. Our method compares the amount of flavor in a headline with the mean of the flavor across one topic to identify which news sources use significantly more flavor as an indicator of bias.

Our third and final approach is to classify news headlines into their respective sources, which will indicate how distinct a news source is. An unbiased news outlet would be very difficult to classify, whereas a news outlet that consistently uses biased language can easily be classified.

Our methods can be applied in real-world, on-line settings as a tool to make readers aware of the presence of bias. Creating awareness is an important first step in preventing the reinforcement and spreading of unwanted stereotypes. Our proposed methods are applicable to topics and news sources of any kind.

The rest of this paper is structured as follows. In Section II we provide an overview of related work, as well as giving the definition of bias that we used to define our work. Section III describes the technical details of our work, where Section III.B describes the methods that we used to detect bias and Section III.C details how we validated the performance of our techniques. We provide the results in Section IV and conclude with Section V.

II. Background

In cognitive science and sociology, biases in news reports have been studied since well before automatic detection was possible. The first crucial step for research into this topic is to find a clear, measurable definition of bias. One such definition with respect to news reports is

> "[...] the inappropriate intrusion of subjective opinion into an otherwise factual account".[9]

In general, cognitive biases are often defined as a deviation from a norm. We therefore combine this definition with the one above, to define a biased headline as a headline that deviates significantly in some dimension from the average value.

There are several other works in which bias detection is performed on news texts. [10] investigates two linguistic biases, namely framing and epistemological bias, and detects them using heuristic methods.

III. Project

Our project consists of several approaches which let us detect biases and analyze newspapers from different perspectives. In this section, we are going to provide information about the dataset we used, as well as describe our

approaches, ranging from general sentiment analysis, topic detection and sentiment and flavor analysis for the specific topics, to domain classification.

A. Data

Firstly, we looked at the Kaggle News Aggregator Dataset[1]. The advantage of this dataset is that it contains over 400k news stories from many different domains. Often, the same event exists in the dataset multiple times, covered by different news sources. This dataset did not fulfill our requirements, because it only covered a quite small timespan in 2014 and it had a limited number of content categories, and no politic content.

Therefore, we started looking for alternative datasets, which led us to the discovery of the Reddit API. We selected newspaper headlines from "r/worldnews", which exhibits roughly 90k headlines per year. Each headline was obtained with a corresponding timestamp and the domain on which the news story was published. One drawback of our dataset is the fact that each event was usually featured only once in our dataset. If the same event would exist in the dataset several times, covered by different domains, a direct comparison between those domains would have been much easier. We scraped headlines starting from 2014, but for our analysis we mostly focused on the timespan between 2016 and 2017, since the full dataset would make Topic Detection too computationally expensive. Furthermore, the number of topics would probably increase, and some topics would only occur during a short timespan.

Since many domains in the dataset only occurred a few times, and the number of headlines per domain would be further reduced because of the Topic Detection, we selected the big domains to make sure that at least a few hundred headlines per domain were available.

B. Method

We employed several methods to detect biases. As a first glance at a possible bias, we used Sentiment Analysis. Later, we decided to refine our results by utilizing Topic Detection, Flavor Analysis, as well as a domain classifier.

B.1 General Sentiment Analysis

One of our first results was a general Sentiment Analysis. We aggregated the results by averaging the sentiments of the headlines for each domain. Those average values ranged from -0.3 to -0.1. This revealed some interesting differences between the domains. It became clear that some domains had very negative average sentiments, because they focused on negative topics such as war and general world news, whereas others have relatively positive average sentiments, because they focused on economics or science news.

B.2 Topic Detection

As a more fine-grained measure of bias, we use sentiment analysis per topic. To extract the topics from our data set, we use the biterm topic model (BTM) [3], which is especially suitable for short texts. BTM uses the word co-occurrence patterns (biterms) to create a model over the entire corpus, thus solving the problem of sparse word co-occurrences within one document. The model parameters are estimated using Gibbs sampling. We create a vocabulary of the top 5000 most occurring words in the corpus, and run the Gibbs sampling algorithm to model $T = 20$ topics. For the Dirichlet priors, we use $\alpha = 50/T = 2.5$ as recommended by [5] and $\beta = 0.01$ in order to make the topics contain a small number of highly distinct words. Due to limitations in computational power and the large size of the data set, we were limited to 250 iterations of Gibbs sampling, which took approximately nine hours to complete. We use a modified version of the BTM implementation found at [8].

B.3 Topic Sentiment Analysis

We determine the topic for all the headlines in the 2017 Reddit data set using the maximum topic probability. We extract the headlines for the selected topics in Table I and compute the polarity score for every headline. We use the VADER sentiment analysis framework [6] since it is has been shown to achieve good performance on short social media text. VADER is able to detect the polarity of complicated sentences with human-like accuracy.

For every topic, we take the mean over the polarity scores of all headlines within the topic, which we call the *mean topic sentiment*. We then calculate the mean over the polarity scores within one topic for every news source separately. The difference between the mean sentiment per news source and the mean topic sentiment is an indicator of bias, using the definition of bias given in Section II.

B.4 Flavor Analysis

As an alternative metric, we introduce the concept of flavor. The amount of flavor in a headline is determined by the amount of words that can potentially carry subjectivity, namely adjectives, adverbs, superlative and comparatives. Intuitively, objective records of news events require less flavor words, while flavorful headlines are more likely to contain bias. We determine the flavor words using the part-of-speech tagger from NLTK [7]. Since longer sentences are more likely to contain flavor words, we define a normalized flavor metric as the number of flavor words, divided by the total number of words in a sentence.

Similar to the topic sentiment analysis, we compare the flavor per news source and topic with the average flavor taken over all headlines in one topic, the *mean topic flavor*. It is important to note that a deviation from the mean topic flavor does not indicate that a news source is negatively or positively biased for one topic. Instead, such a deviation means that a news source uses significantly more subjective words than others, which we consider an indicator of bias using the definition in Section I.

B.5 Domain Classification

We implemented a Bag of Words model with a Random Forest Classifier to detect the domain for a given headline. For this task, we handpicked six domains. Those

were the five most occurring domains: BBC, CNN, Independent, Reuters, and The Guardian. We additionally selected Foxnews as a sixth domain, hoping for it to be a contrast to the other five ones. The number of headlines per domain varies a lot, ranging from just 1,418 headlines (Foxnews) to over 11,000 headlines (BBC). Due to this big imbalance, we applied undersampling to create a balanced dataset, containing 8,508 total headlines. Then we split the data into a training set (80%) and a test set (20%).

The Bag of Words model was created by firstly removing stopwords from the headlines and then transforming them into vectors with maximum 5000 features, using a CountVectorizer. Secondly, a Random Forest Classifier was used to predict the domain for the headlines. We found, by trial and error, that 50 is a good number estimators for this task.

C. Evaluation

In this section, we will describe the evaluation metrics used to validate our systems. Since there is no good quantitative measure to evaluate the quality of the detected topics, we instead validated them using a small-scale study in which three participants were asked to define a topic from the top twenty most distinct keywords as found by the BTM, and give a rating on a scale from 1 to 7 for how coherent the topic was. A large-scale study was beyond the scope of this project due to time limitations. We selected a subset of the twenty topics based on average coherence rating and potential for controversy to perform sentiment analysis over. The selected topics can be found in Table I. After extraction of these selected topics, the data set was severely reduced in size. The distribution of headlines per topic is shown in Fig. 1, the number of headlines per news source after selecting topics is shown in Fig. 2.

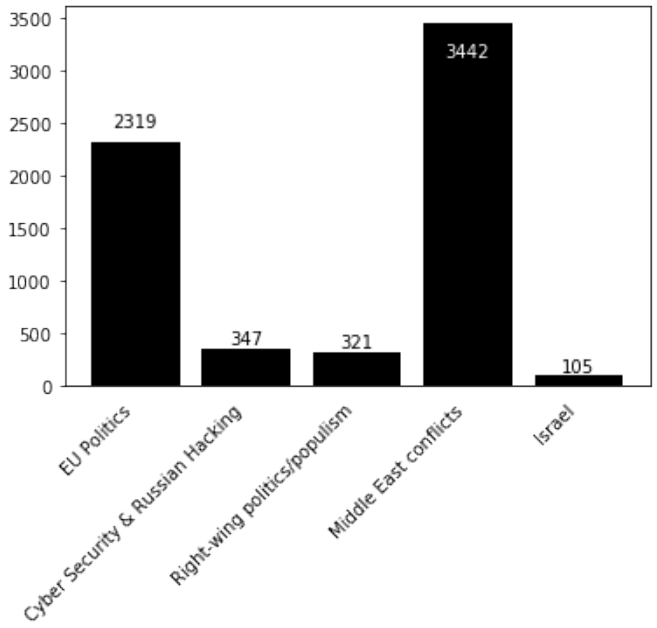

Fig. 1. Number of headlines per topic.

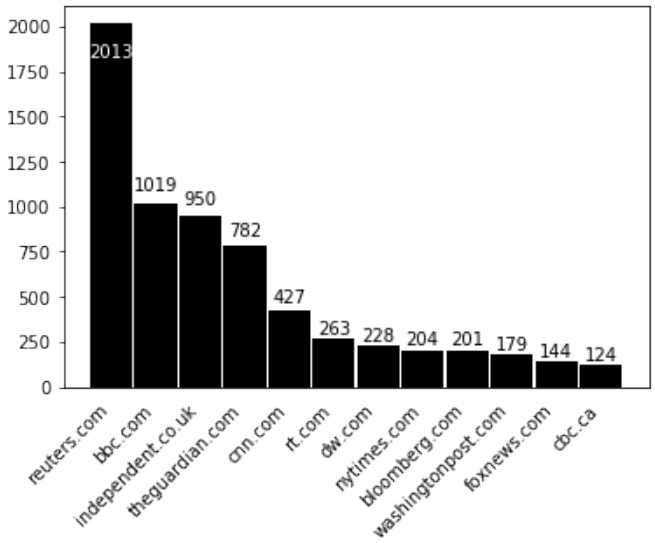

Fig. 2. Number of headlines per news source.

We validated the per-topic sentiment and flavor analysis using a two-sided paired T-test. For every news source, the measurements were tested against the rest of the population excluding the news source. We did not assume homogeneity of variance across samples. The results of these tests will be shown in Section IV.

We evaluated the classifier by taking its F1-Score into account, but also looking at other influencial factors, such as the number of classes. A confusion matrix (see Fig. 5) helped us to detect domains which were identical to others. For example, `bbc.co.uk` was often confused with `bbc.com`. Since headlines of both of those domains come from BBC, we aggregated them into a single class. The same was done for Reuters, CNN and The Guardian. Furthermore, we eliminated domains that regularly post headlines that came from other domains. One example for that is `yahoo.com`, which takes news stories from various sources and publishes them on their domain. The classification revealed those sources accordingly.

IV. Results

A. Sentiment and Flavor Analysis

The results of the sentiment analysis described in Section III.B.3 evaluated using a paired T-test are shown in Fig. 3.

B. Domain Classification

The domain classification for our six selected domains yielded an F1-Score of around 0.4842. The confusion matrix for the classification is shown in Fig. 5.

Additionally, we took a look at the feature importances and we examined the words with the highest prediction probabilities for a certain domain. Table II shows a list of some of the most indicative words.

	EU Politics	Cyber Security & Russian Hacking	Right-wing politics/populism	Middle East conflicts	Israel
Mean topic sentiment	-0.041	-0.111	-0.111	-0.128	0.215
washingtonpost.com	0.048	-0.021	0.150	0.063	0.093
rt.com	-0.035	0.009	-0.030	0.013	-0.281
nytimes.com	0.039	-0.084	0.226	-0.032	0.100
theguardian.com	-0.009	-0.055	-0.075	-0.071	-0.008
independent.co.uk	-0.064	0.063	-0.105	-0.082	-0.089
dw.com	-0.039	-0.045	0.038	0.020	0.030
cbc.ca	0.019	-0.089	0.109	-0.062	0.047
cnn.com	-0.029	0.020	-0.160	0.038	0.043
reuters.com	0.088	0.035	0.120	0.079	0.089
bloomberg.com	0.121	-0.011	-0.031	-0.070	0.034
bbc.com	-0.039	-0.004	-0.038	-0.070	-0.033
foxnews.com	-0.043	0.076	-0.132	-0.026	-0.217

Fig. 3. Deviation from mean topic sentiment per domain. The top row shows the mean topic sentiment in blue. Significant results with $p < 0.05$ are shown in red.

	EU Politics	Cyber Security & Russian Hacking	Right-wing politics/populism	Middle East conflicts	Israel
Mean topic flavor	0.099	0.099	0.095	0.095	0.118
washingtonpost.com	0.011	-0.026	0.040	0.015	0.132
rt.com	0.028	0.040	-0.014	0.014	-0.028
nytimes.com	-0.051	-0.062	-0.073	-0.060	0.007
theguardian.com	0.009	-0.003	-0.016	-0.001	-0.002
independent.co.uk	-0.002	0.018	0.024	0.011	0.001
dw.com	0.026	-0.000	0.027	0.014	-0.037
cbc.ca	0.021	0.050	0.003	0.007	-0.060
cnn.com	-0.007	0.009	-0.020	-0.006	0.124
reuters.com	0.018	0.010	0.015	0.006	-0.032
bloomberg.com	-0.050	-0.080	-0.077	-0.052	-0.087
bbc.com	-0.018	-0.009	-0.013	-0.003	0.001
foxnews.com	0.023	0.044	0.031	0.005	-0.119

Fig. 4. Deviation from mean topic flavor per domain. The top row shows the mean topic flavor in blue. Significant results with $p < 0.05$ are shown in red.

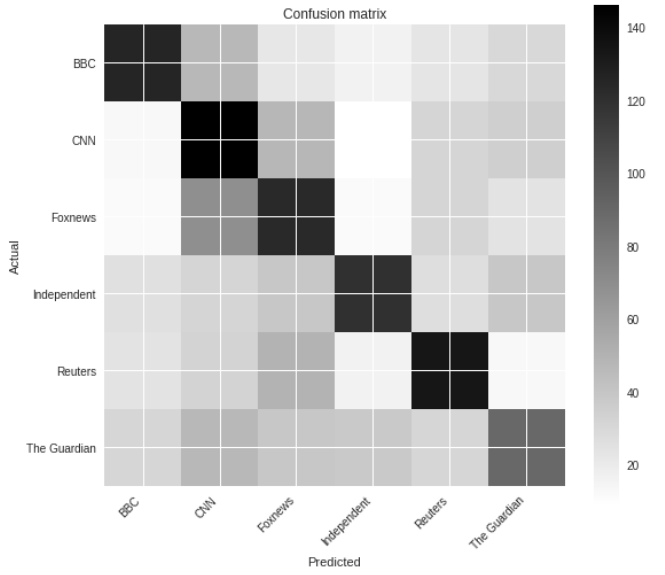

Fig. 5. Confusion matrix for the domain classification.

V. Conclusions

Detecting biases in newspapers is an important task. It helps guide readers and exposes hidden biases that are present in news headlines. There are many possible approaches to implement an automatic bias detection system using machine learning. One such approach is Sentiment Analysis, which we used as well in combination with Topic Detection. Another method that we used is measuring the amount of flavor words, i.e. adjectives and adverbs. Our third and final method is to classify news headlines into their respective sources.

Merely applying Sentiment Analysis to a single headline is not accurate enough to detect true sentiments or biases. In fact, the individual values often appear disappointingly inaccurate. However, if the number of headlines is big enough, the aggregated mean values per domain show significant differences from the general mean. This proves that there are indeed great differences between the domains, but that could also be because each domain focuses on different topics. A topic might carry negative or positive sentiment because of its nature, regardless of who talks about it. Therefore, we decided to examine sentiments per topic to find out whether different domains talk about the same topic differently.

Since big deviations are expected due to the relative inaccuracy of the Sentiment Analysis, statistical tests were performed to detect whether those deviations are significant. Looking at the results of the sentiment analysis per topic, we first note that the distribution of sentiment values is informative, since a random distribution would only yield three significant deviations on average. The mean topic sentiment values are negative for all topics except the last topic ("Israel"), which confirms that news reports often yield a negative sentiment value overall. The most interesting result is the large negative relative sentiment value for the topic "Israel" for `rt.com` (Russia Today).

The domain classification provided a different perspective on biases in news headlines. In contrast to Sentiment Analysis, it cannot tell us anything about positivity or negativity of a domain towards a specific topic, but it can detect a general bias. A higher classification score means that there are existing biases in the domains. If a domain is being classified well, it means that it has distinct features that other domains do not have. For six handpicked domains, there was an F1-Score of 0.4842, which clearly proves a bias. For 16 domains, the F1-Score was still at 0.2913. Here, it becomes clear that some domains can be classified better than others.

Bias might be the explanation for some of those differences. However, it is not a reliable indicator, since a high or a low accuracy for a domain can have many reasons.

For example, the classifier performs very poorly for Yahoo. This can be explained by the fact that Yahoo usually doesn't publish news itself, but rather takes news stories from other news sources and publishes them on Yahoo. There are other possible explanations for a high or low accuracy which have nothing to do with bias, such as journalistic style.

To conclude, we found that there are measurable biases in newspaper headlines. Classification yielded surprisingly good results. The detected differences between the respective domains are statistically significant.

During this project we deepened our understanding of Sentiment Analysis, Natural Language Processing, Topic Detection and Text Classification. We practically applied all of those methods to obtain significant results and to detect biases in newspaper headlines. We are generally happy with the way we conducted this project.

However, there are a few things that we would have done differently. We would have definitely benefitted from a larger, more complete dataset. If in our data, a single event would have been covered by multiple news sources, we could have performed a direct comparison per event. The fact that the dataset does not include many headlines probably raised the deviations.

Therefore, we had to give up on a few methods that would have worked on a complete dataset. Research should focus on a single method generally, but in our case it was probably good to try different methods and to find out which ones work well.

There are numerous possible approaches to detecting newspaper biases, and we only scratched the surface of this broad topic. Further research is definitely going to reveal more.

References

[1] Lichman, M. (2013). *UCI Machine Learning Repository* `http://archive.ics.uci.edu/ml`. Irvine, CA: University of California, School of Information and Computer Science.

[2] Caverni J.-P., Fabre J. M., Gonzales, M. (1990): Cognitive Biases: Their Contribution for Understanding Human Cognitive Processes. Amsterdam: *Elsevier Science Publishers B. V.*

[3] Xiaohui Yan, Jiafeng Guo, Yanyan Lan and Xueqi Cheng, "A Biterm Topic Model for Short Texts", *Proceedings of the 22nd international conference on World Wide Web*, ACM, 2013, pages 1445–1456.

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation" *Journal of machine Learning research*, Jan. (2003), pp. 993–1022.

[5] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics", *Proceedings of the National academy of Sciences*, vol. 101, suppl 1, 2004 pages 5228–5235.

[6] Clayton J. Hutto and Eric Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text.", *Eighth international AAAI conference on weblogs and social media*, Ann Arbor, MI, USA, June 2014.

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

[8] Joan Capdevila Pujol (Universitat Polytècnica de Catalunya), "jcapde/Biterm: Biterm topic model", *GitHub*, `https://github.com/jcapde/Biterm`, posted October 14, 2015, accessed May 24, 2017.

[9] S. Holly Stocking and Paget H. Gross, *How Do Journalists Think? A proposal for the study of cognitive bias in newsmaking*, ERIC Clearinghouse on Reading and Communication Skills, 2805 E. 10th St., Smith Research Center, Suite 150, Bloomington, IN 47405, 1989.

[10] Recasens, Marta, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky, "Linguistic Models for Analyzing and Detecting Biased Language", *ACL (1)*. 2013.

[11] Gupta, Sonal. "Finding bias in political news and blog websites", 2009.

Appendix

TABLE I
Selected topics within the 2017 Reddit data set

Keywords	Topic	Average Coherence
election, EU, Brexit, vote, party, president, minister, UK, parliament, presidential, Theresa, May, says, prime, European, Le, Pen, government, would, French	EU Politics	5.67
Russian, Russia, US, intelligence, Trump, hacking, hackers, attack, cyber, CIA, FBI, security, government, former, ex, spy, data, says, officials	Cyber security & Russian hacking	6.00
Germany, party, Trump, right, says, German, Nazi, world, anti, social, pope, new, election, far, Brexit, leader, president, French, speech, saying	Right-wing politics & populism	4.33
Syria, Russia, US, north, Korea, Trump, military, missile, UN, war, Israel, turkey, Iran, state, said, Syrian, says, united, forces	Middle East conflicts	5
Israel, Trump, Israeli, president, Netanyahu, house, white, US, Palestinian, Donald, says, minister, Jerusalem, bank, peace, Palestinians, visit, settlements, west, two	Israel	5.66

TABLE II
Most indicative words for the domain classification

Word	Highest probability	Predicted domain
mosul	0.94	CNN
us	0.84	CNN
sues	0.815	BBC
aleppo	0.8	CNN
reportedly	0.8	Foxnews
japan	0.78	CNN
turns	0.77	BBC
investigates	0.76	BBC
cnn	0.76	CNN
pepe	0.747	BBC
cyclone	0.745	BBC
bbc	0.727	BBC
erdo	0.725	The Guardian
billion	0.72	Reuters
indonesian	0.72	Foxnews
football	0.718	BBC