Bias in Newspaper Headlines

Timo van Niedek, *Student, IEEE*
Constantin Brîncoveanu, *Student, IEEE*

***Abstract*—TODO**

I. Introduction

Although newspapers tend to claim neutrality, it is almost impossible to find a source that does not contain stereotypes or biases in some form. Biases in news headlines pose a danger in that they can propagate stereotypes to the general public. 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.

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.

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

A. Definition of Bias

"Bias in cognitive science is generally defined as a deviation from a norm, deviation from some true or objective value." We examined deviations from the norm by calculating the mean topic sentiments, as well as the mean topic flavours, and then looking at the differences of the respective domains from those mean values.

A second way to detect biases was domain classification.

II. Background

TODO

III. Project

TODO

A. Data

We got data from Newspaper APIs. Firstly, we looked at the Kaggle News Aggregator Dataset. 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" ranging from 2016 to 2017. From those headlines, we further selected the ones that came from the most occurring domains.

B. Design

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 consider the sentiment analysis per topic. To extract the topics from our data set, we use the biterm topic model (BTM) [1], 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 [3] 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.

The

C. Evaluation

TODO

C.1 Topic Detection

C.2 Sentiment Analysis

C.3 Flavour

C.4 Domain Classification

We implemented a Bag of Words model with a Random Forest Classifier. For this task, we handpicked six domains (the five most occurring domains, and Foxnews).

IV. Results

TODO

V. Summary

TODO

VI. Conclusions

There are measurable biases in newspaper headlines. Classification yielded surprisingly good results. The detected differences between the respective domains are statistically significant.

Sentiment Analysis is not accurate enough to detect the true sentiment value for each headline. 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.

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

References

[1] 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.

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

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

Appendix

Topics