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 one or the other way. Our goal in this project is to apply machine learning to find those biases. We specifically aimed at looking at minorities, such as refugees, and we also examined general politics and anything that might be controversial.

We approached this task by firstly selecting newspaper headlines that belong to a certain topic. This was done by Topic Detection, which will be explained in detail later in this paper. After that selection, a Sentiment Analysis revealed the biases of the individual newspapers.

Another perspective on biases was achieved by implementing a classifier, which revealed further differences between the respective newspapers.

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. Design

TODO

A.1 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. Evaluation

TODO

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

B.3 Sentiment Analysis

B.4 Flavour

B.5 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.

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