A Proposed Framework for Acessing Bias on English Newspapers

Curado, Antonio & Dahl, Morten

Masters in Advanced Analytics @ Nova IMS

Abstract

This project proposes a framework for the detection fake news by analysing bias in english newspapers. The articles have been grouped by topic and analysed to detect the bias and tendencies. It results in a visualization, which shows the media bias per newspaper, topic and keyword.

s

Motivation

The motivation for this project lays in the recent doubts on media neutrality. People tend to distrust the media and call serious newspapers "fake news", whereas dubious newspapers gain popularity. So this project aims to provide the following:

- provide support to detect biaseness and tendious media coverage
- create transparency which newspapers tend to publish more tendentious articles
- initial notions towards an automated biasness detection in media coverage

s

Methods

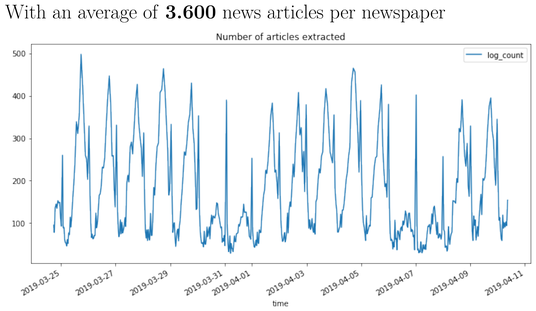
The methodology followed in this project is as followed: As a first step, data was acquired from a **web scrapping** aproach from which news articles are downloaded from the web representations of various well-known newspapers. The data was limited to the politics section of each newspaper. It was tried to pick a balanced selection of different politically located newspapers through a qualitative assessment. These articles were then filtered on keywords [e.g. "Trump", "Syria", "Brexit"] and used to train a **LDA** model for topic detection. Simultaneously a **sentiment analysis** was carried out to measure the tendentious nature of each article. By calculating the amount of words frequently associated with negative and positive sentiments in each article. Finally the results of both analysis were mapped and visualised in an interactive scatterplot.

s

Data Extraction

As a way to make the analysis relevant and up to date with the most current news topics, it has been developed a new news dataset with the following properties:

- Built a dataset with over **70.000** news articles
- Scraped over **19** newspapers for over **2** weeks
- With an average of **3.600** news articles per newspaper



The dataset was build only using the newspapers3k python package.

s

LDA Analysis

The LDA (Latent Dirichlet Allocation)[1] analysis aims to detect topics within all articles. We trained a model for each partition of the dataset. The dataset was partitioned based on a specific keyword (e.g. Trump) in order to have a meaningful filter. Otherwise the dataset would be too big to analyse each detected topic. This algorithm takes three *hyperparameters*: number of topics, alpha and eta. All of these parameters were finetuned for each of the three trained models. The coherence score served as a measure for this assessment.

KeyWord	Topics	Alpha	Eta
"Trump"	14	0.01	0.01
"Brexit"	6	0.01	0.01
"Syria"	6	0.6	0.2

Table: Parameters for LDA models

The result of this analysis is a mapping for each article in order to make them comparable. Each topic is represented as a concatination of words assigned with a probability. In the example topic shown below one can understand by examining the words that the topic is about Trump and the wall to Mexico.

0.033"border" + 0.016*"mexico" + 0.016*"trump" + 0.010*"immigration" + 0.008*"states" + 0.008*"united" + 0.008*"president" + 0.007*"illegal" + 0.007*"migrants" + 0.006*"people"*

s

Sentiment Analysis

Another way to search for bias is to analyse the sentiment which each newspaper looks at a topic. The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.
Valence Aware Dictionary for sEntiment Reasoning technique[2] has been used to achieve this task.

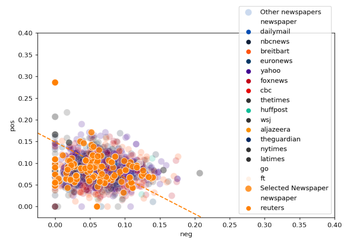
Sentence1 = "*In November, Al-Qahtani and 16 others were designated by the State Department under the Global Magnitsky Act along.*"
Sentence2 = "*Congressman Adam Schiff, who spent two years knowingly and unlawfully lying and leaking, should be forced to resign from Congress!*"
Sentence3 = "*Optimism that China and the United States will eventually hammer out a trade deal has helped global markets rack up huge gains so far this year.*"

	Positivity(pos)	Neutrality(neu)	Negativity(neg)
Sentence1	0.0	1.0	0.0
Sentence2	0.0	0.652	0.348
Sentence3	0.344	0.656	0.0

Table: Sentence Polarity

s

Trump

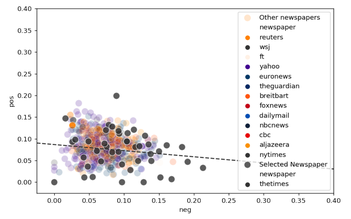


This is an example plot of the analysis of a topic containing the keyword "Trump". The topic covers the investigation done by Robert Mueller regarding the involvement of Russia in the presidential election in 2016. The news agency **reuters** apparently covered this topic in a slight negative manner, but still more or less balanced. Each point in the scatter represent an article, whereas the x-axis shows the negative score and the y-axis the positive score. In this framework a rather lower score on both axis means less usage of *tendencious* words.

s

Brexit

This plot shows the coverage of the topic that the Brexit will impact the global trade. The newspaper **the times** covers this topic apparently more ambivalent. The sparse distribution of points suggest that the coverage of the different articles have different tones.



s

Syria

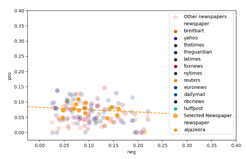


Figure: Al Jazeera

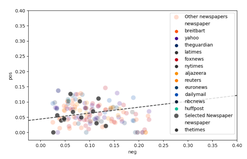


Figure: The Times

The two plots show how articles from the **al jazeera** and **the times** newspapers compare to each other, on a positivity/negativity scale. From this, one could possibly infer that when Al Jazeera publishes on the topic, it takes a more serious and possible more negative approach to it.
Never the less, these are only possible assumptions that the framework helps to clarify. It would be needed to read each article in detail to derive further conclusions.

s

Conclusion

Measuring bias is a very subjective task. Usually computers have a hard time executing subjective tasks. The framework developed does not aim to automate bias detection but rather to empower humans with summarization capabilites otherwise not possible. From the three example analysis above shown, it is possible to validate the usefullness of the framework proposed on a hard topic as bias detection. As news generation gets automated, as is the case of fakenews, the methods of flaging them also require modern digital frameworks. This analysis and framework takes another step in this direction.

s

Future Work

As previously mentioned this project aims to set a framework for discusion on automated detection of bias. As so there is much further work to be done, mainly on 5 different areas:

- Increase the number of newspapers and find a method to select the same number of articles for each one
- Extrapolate the work to different regions and languages
- Automate topic detection
- Build a ML/Rule based algorithm to flag possible high bias on articles
- Apply sliding window mechanimns to track changes in sentiment towards selected topics

s

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation.*", Journal of machine Learning research 3, Jan 2003

[2] Hutto, C.J. and Gilbert, Eric, *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, 2015.

The project was possible with the great work done in the newspaper3k, nltk and gensim python libraries
All code can be easily accessed in github.com/morten-novaims/Text_Mining_HW

s