Trumpster: A web scraping tool with sentiment analysis

Sigurd Thomassen[*]

University of Tromsø
Institute of Informatics
sigurd14@gmail.com

April 19, 2017

Abstract

This paper presents Trumpster, a web scraping tool with sentiment analysis. It allows the user to quickly retrieve articles from specific online newspapers and score it with sentiment analysis, to see how the Norwegian media is reporting positively or negatively on a case. It also supports a wide spectrum of languages, so it can be configured to work on online newspapers from other countries. The Trumpster shows that it can score articles based on sentiment simlilarly to a human.

I. Introduction

This paper presents a web scraping tool, that retrieves articles from some specific online newspapers and does sentiment analysis on them. It aims to analyze how Norwegian media is reporting on different topics in the news. Given a keyword, the web scraping tool can extract the related articles, and score it on sentiment either positively or negatively. Alongside the sentiment analysis, the Trumpster can also translate the articles. This is mostly for the analysis part, as it is done in English, but it is also a way to read articles in foreign languages, translated. Using the Trumpster, we will also evaluate the relevance of the sentiment analysis by reading the articles ourselves, and scoring them thereafter.

In summary, we make the following contributions:

- A web scraping tool available for all, that can extract articles and analyse them on sentiment.
- An evaluation on the analysis of articles based on the keyword "trump".

II. Scraping like no tomorrow

The Trumpster is built with the Python programming language, and its existing libraries. Using a library called BeautifulSoup we were able to parse html from the websites we targeted. Before analyzing however, one needs to find the links to the different articles on the specific online newspapers. It starts by giving the Trumpster a link to the page where it should start looking for links. It then grabs everything on the page with BeautifulSoup, and grabs all the links in the soup. Then it refines the links, and makes them ready to be explored further. Depending on which online newspaper it is, the links may look different and might be refined differently. When the links are ready, the Trumpster follows them, and extracts the whole article they lead to. It then takes the title and all the paragraphs in the article, and combines them to one text. We now have a clean article with a title and a body. This is then sent to translation, which uses

[*] Thanks to Google for all the good answers, and Camilla Stormoen for the great name.

Table 1: *Article Trumpster score*

Article Id	Sentiment Score
1	−0.0111
2	−0.0151
3	−0.1166
4	−0.055
5	0.0439
6	0.1588
7	0.0323
8	−0.0156
9	0.0108
10	−0.0675
11	0.16
12	−0.2587
13	0.1799
14	0.0431
15	0.0747

Table 2: *Article human score*

Article Id	Sentiment Score
1	−0.15
2	−0.1
3	0.0
4	−0.23
5	−0.3
6	0.2
7	0.05
8	−0.2
9	−0.2
10	−0.5
11	−0.2
12	−0.4
13	0.2
14	0.0
15	0.0

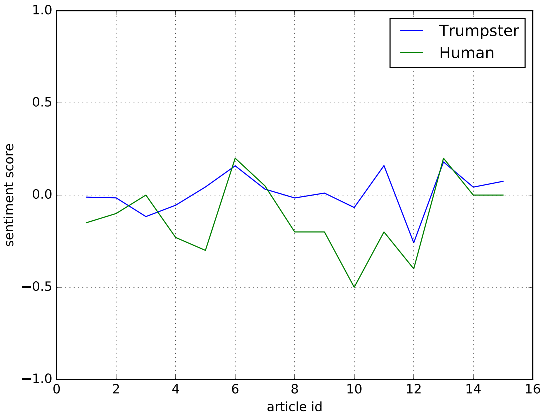
an api from *mymemory* [1] to translate the body. The returned response is the translated string, which is then analyzed on sentiment with a library called vaderSentiment, which uses a library of words loaded with weights. Each sentence is analyzed individually, and when it has analyzed the whole article, a compound score is calculated from the weights, which lie between -1 and 1. After the Trumpster has finished analyzing an article, it moves on to the next link and repeats, until there are no more links left.

III. Results

When executing the Trumpster with the keyword *trump* on the online Norwegian newspapers Aftenposten and Dagbladet, we got a few articles back. They were also scored from -1 to 1, and the article id alongside the score is represented in table 1. An anonmyous colleague read the same articles and scored the articles based on sentiment as well. This is represented in table 2.

In figure 1 we can see the tables plotted as graphs.

Figure 1: *Comparison of human and Trumpster*



IV. Discussion

What we can see from figure 1 is that Trumpster and the human analyser has some similarities. They both agree to the sentiment in some extent. Though the human analyzer has more diverse answers, there is however still a trend in the graph.

The diversity in the human's answers might come from political biases of the person doing the analysing. Since they are still similar

[1] A translator api, www.mymemory.translated.net

though, one can conclude that Trumpster can within a certain degree score articles based on sentiment.

As there has been a lot of talk recently about the media not doing their job, and fake news, we wanted to figure out for ourselves how the Norwegian newspapers Aftenposten and Dagbladet wrote their articles. Specificly articles about Donald Trump. The idea was that Norwegian newspapers in general are negative towards the american president, and that would be reflected in their journalism.

From the results in figure 1 we can see that both The Trumpster and the human analyser gave similar judgment to the articles in context of the sentiment score metric. No articles are over 0.2 in sentiment score. So from that figure, we can clearly see that out of the 15 articles analysed, none of them are positively angled. The human scored most articles neutral trending towards negative, while The Trumpster scored the articles mostly neutral.

What would be interesting is to check this up with more newspaper from both sides of the political spectrum. It is hoped that this study will stimulate further investigations in that regard. Assuming the accuracy of The Trumpster is correct, The Trumpster could be used to do this.

The need to differentiate between fake and real news is immense. A way to start doing this is by weeding out articles that are clearly biased. If an article is scored to be neutral it is a step in the right direction, though it is not guaranteed that it is real news. The Trumpster will help to get this done, but there is still more work needed to completely weed out fake news.

V. Conclusion

With the amount of fake news and criticism of the media these days, the need for tools to differentiate between news and false news is immense. The Trumpster will help to some degree in this regard, by weeding out very biased news based on a sentiment score. There is still a lot more work to be done to completely get rid of fake news, but this is a tool that can help getting there.