NLP on URLs
Machine Learning for Natural Language Processing 2020

Mathias Andler
ENSAE
mathias.andler@ensae.fr

Etienne Apers
ENSAE
etienne.apers@ensae.fr

Abstract

With the rise of internet and our connected world, we have access to an increasingly large number of news source from all around the globe. At the same time it is often said that we tend to only read news that confirm our own opinion. Taking steps to mitigate this is bias aren't necessarly easy when we often only read a few lines from an article before switching our attention, or when we only read a title. Our hypothesis is that by using only urls linking to a news article we would be able to situate politically the news source. Colab link

1 Problem Framing

We will use the following database :Millions of News Article URLs: 2.3 million URLs for news articles from the frontpage of over 950 English-language news outlets in the six month period between October 2014 and April 2015.

The objective is to build a Sequence Classification based on Sentiment analysis. The context and sentiment of the headlines will help us identify the political views of the newspapers in question.

2 Experiments Protocol

As a first step, the database was cleaned up. We had URLs that were not usable in their current state and had to format them in order to extract lists of words that made up the titles of the articles.

We then proceeded to embed the words obtained in order to group the articles into clusters and form thematic groups of articles.

Finally, we studied the way in which these subjects are treated by different websites. For this purpose, the pre-trained database proposed by NLTK allowed us to analyse the trends of the articles.

3 Results

A first result we obtained was that the largest sites offered a range of articles that generally covered most of the topics addressed by all the articles in the database. In other words, the largest sites are the generalist sites.

We were then able to group the articles into 40 clusters, some of which are relevant and show, via Sentiment Analysis, that the sites are quite divided on certain topics. Indeed, when we look at the proportion of articles with a positive tendency on a given subject per website, we sometimes see two fairly strong peaks around 0% and 100%.

Finally, we do not obtain very significant results regarding the strong ideological orientation of certain websites. This is mainly due to the fact that the database contains many articles from newspapers such as the New York Times or The Guardian, which are quite moderate and produce relatively heterogeneous articles on most subjects.

4 Discussion/Conclusion

Further interesting work could be pursued by using sentence embedding instead of word embedding techniques, such as those based on BERT. Such methods would probably improve results compared to the average token technique we used in our work.

Also given the poor data quality within URLs We could have obtained better data for each link by using a scraping package like Newspaper3k (https://newspaper.readthedocs.io/en/latest/). But this would have defeated our initial objective of using only URLs in addition of being much more computationally complex.

Finally, hierarchical clustering techniques could have yielded a finer clustering showing the topics/sub-topics/sub-sub-topics nature of newspaper articles.