Analyzing New York Times's Jounalists' Popularity based on the articles comments.
Machine Learning for Natural Language Processing 2020

Naïla El Haouari
ENSAE Paris
naila.el.haouari@ensae.fr

Félix Pasquier
ENSAE Paris
felix.pasquier@ensae.fr

Abstract

Nowadays, with the expansion of access to information, journalists are in constant competition to satisfy their audience. Especialy with the development of social media, traditional newspapers need to reinvent themselves to stay attractive to their public, and in particular to the youngest generations. With the digitalization of their articles, the New York Times can get the reviews from their readers, and adapt their offer to them. Our goal in this project consists in detecting the most popular journalists, based on the readers' comments. Using the NYT comments dataset, we train a sentiment analysis model based on BERT on the Twitter US Airline sentiment dataset, and apply it on the comments' sentences containing the journalist's name – to have the feeling towards the journalist, and not towards to article. Our code (Colab) is accessible on Github[1].

1 Problem Framing

Nowadays, with the expansion of social media and the access to a large amount of information, being a journalist is not an easy jobs. The profession has to re-invent itself to keep attracting its audience. In this project, we aim at analyzing the journalists' popularity among their audience by analyzing the readers' comments for the articles.We are particularly interested in the work of the journalists. In democracies such as the United States or France, they are very important as they set the tone for the public debate. Our goal consists in having some indicators about the journalists' writing: how effective are they in providing food for thoughts? Are people happy/unhappy with their way of seeing social issues and/or with their answers? We are mostly interested in two indicators : the average number of comments a journalist receives per article and the average share of positive/negative comments concerning the journalist's articles. To compute this latter indicator, we use a sentiment analysis model based on BERT to analyze the readers' sentiments about the journalists who write editorials.

2 Experiments Protocol

Data (train and evaluation): We use the New York Times Comments dataset [2], which contains more than 2 million comments, from 9000 articles of the New York Times in 2017-2018. For this analysis, we choose to focus on Editorials and Open Editorials, and to consider only the journalists that wrote more than 10 articles during this period. Since we consider the journalists independently, we duplicate the articles written by several journalists, so that they count for each of them. From there, it is easy to compute our first indicator. Then, we limit our analysis on comments that contain the journalist's name, and for each comment, we extract the sentences that contain the name of the journalist. Indeed, we want to capture the reader's opinion about the journalist's work, so we need to make sure that we are not capturing the reader's opinion about the topic of the article – a reader can express anger towards the reported news in the editorial, but still agree with the journalist and appreciate her work. However, restricting to the sentences might not be enough to capture only the sentiment towards the journalist – but we will discuss this choice later. Our final cleaned data set therefore contains 80000 comments, on 916 articles from 21 journalists.
Since the NYT comments data set is not labeled, we need to use another data set with labels to

[1] https://github.com/Naila-elh/NLP_NYT_comments/tree/master

[2] https://www.kaggle.com/aashita/nyt-comments

train our model. We chose to use the Twitter US Airline sentiment data set[3], since it contains three different types of labels (”positive”, ”negative”, ”neutral”), and it is not as specific as other labeled data sets (like reviews of movies, etc.). We test our model on the NYT Comments data set, and we manually label 300 comments randomly chosen, to see the accuracy of our model.

Model used: We use BertForSequenceClassification, the pre-trained BERT model with a single linear layer on top for classification. We do a multi-class classification, since we want to predict if the comment is positive, negative, or neutral about the journalist's work.

Model training: We train our model on the Twitter US Airline sentiment data set. We divided the data set between a training set (90% of observations) and a validation test (10%), using 2 epochs. We notice that the validation loss is going up with each epoch while the training loss is going down, which means our model is over-fitting on the training data.That's why we decided to train the model on 2 epochs only.

3 Results

The results we obtain on our labeled test of 300 comments set are not that good. We fail at predicting the positive and neutral comments, but our prediction of negative comments is quite fair. When we predict positive comments, we can be pretty sure that they are indeed positive. This is the main point we can use in our analysis

We have some explanations for our bad predictions. First, when we manually labeled the 300 comments on which we test the model and build the accuracy score, some sentences were quite hard to label. Since the readers discuss the quality of articles and of journalists' work, they are often nuanced in their comments, and often highligh both the pros and cons of the articles. In many cases, a single sentence isn't enough to get the global reader's opinion about the journalist's work. Besides, even if we choose to focus only on the sentences of the comments where the journalist's name appear, the reader is sometimes expressing an opinion on the article in the same sentence (they sometimes say they agree with the journalist because they share their bad opinion on a certain topic, in the same sentence). This can also bias our predictions.

Another reason for these bad predictions is that our training dataset – sentiment tweets on US airlines – may not really adequate to our problem, but we haven't found another dataset that was more appropriate, especially with three classes of labels. The other explanation for our performance is that our model overfit, which means it is too close to the data it has trained on. Since our dataset is quite different than the training dataset, the prediction errors might be amplified.

4 Discussion/Conclusion

About our first indicator, we noticed that some journalists have a lot of comments on their editorials, sometimes twice as much as their colleagues. We can expect these journalists to be better at triggering some debate, maybe by picking the right and hot topics.

Concerning our second indicator, if we stand with our analysis, we can identify the most ”preferred” journalists : those who have more positive comments. We cannot see a correlation between the number of articles or comments and the proportion of positive or negative comments. However, there is a correlation between positive and negative comments – journalists that have a lot of positive comments have few negative comments. We can expect those journalists to be better at being controversial and at stirring some debate.

We also notice that there are a lot of negative comments, which might confirm the results obtained on our small labeled data set : our model does not predict well the negative and neutral comments, so many comments are wrongly labeled ”negative”.

Some further interesting research would be to understand what are the topics that foster a lot of comments and debate. Moreover, once we have determined who are the good journalists, it would be fascinating to find some explanations, in their writing style for example, that makes them special.

[3]https://www.kaggle.com/crowdflower/twitter-airline-sentiment

.

References

Chris McCormick and Nick Ryan. Bert fine-tuning tutorial with pytorch.