Sentiment analysis of New York Times Comments Machine Learning for Natural Language Processing 2020

Ibréhime Traoré
ENSAE Paris
ibrehime.traore@ensae.fr

Merieme Amine
ENSAE Paris
merieme.amine@ensae.fr

Link to the project: [1]

Abstract

The aim of this project is to analyze comments of New York Times readers. We built models to predict the popularity and the sentiment expressed by a comment.

1 Problem Framing

Given the diversity of readers and articles published by New York Times, we start from the fact that the opinions given will allow us to understand what the readers think about current news and what one thinks about others' opinions.
To do that, we firstly labelized sentiment expressed by the comment. Then, we evaluate the popularity of each comment. Finally we built predicting models for the sentiment and the popularity.

2 Experiments Protocol

To carry out our experiments, we choose the comments of April 2018, since they are the latest ones available.
After cleaning the dataset, we used *Sentiwordnet* to calculate the sentiment of each word and then deduce the comment's sentiment.
We defined the popularity of the comment as related to the number of upvotes it gets. We will consider that a comment is popular when the number of upvotes exceeds the median, which is in our case equal to 6 upvotes. We also created another modality for very popular comments, when the number of upvotes exceeds the 80th percentile which is equal to 20 upvotes in out data set.
When conducting the prediction of the popularity of a comment, we compare the "CountVectorizer" and "Word2Vec" methods for the embedding of our tokens. And for the models, we used support vector machine and Random Forest for each outcome variable. As for the sentiment analysis, we try the two previous methods as well as the sentiment classifier model based on BERT uncased.

3 Results

Sentiment expressed by the comment Our models (SVM and Random Forest) give us predictions which are in this case fairly good for the negative and postive sentiments but less optimal for the neutral comments, due to the low number of observations. We denote that the SVM model is better than Randomforests in sentiment analysis in our case, when comparing evaluation metrics (precision, recall and f1-score).
By using word2vec embedding method trained on political text (analyzing the tokens shows that the context of the comments is political), instead of the first method, we get worse predictions.
The sentiment classifier model based on BERT uncased lead us to poor accuracy on test set.

Comment's popularity Our models (SVM and Random Forest) predict well neutral (nonpopular) comments. For the others groups (popular and very popular), predictions are

[1] https://github.com/ibrim1997/nlp-project-ensae2020

not very correct. By changing the embedding method, we can't notice any improvement in predictions. The popularity of a comment is difficult to predict.One other method would be working with only two labels to have more observations for the popular modality and thus better outcome. When it comes to the BERT model, the accuracy is low and the loss function does not seem to decrease with the number of iterations which indicates that this model is not optimal for the sentiment analysis in our case.

4 Discussion/Conclusion

In this project, we tried to predict sentiment and popularity of a given comment. As means of improvement, we could have worked with more comments, regroup some classes (very popular and popular for instance) or evaluate the popularity in a different way (using the number of recommendations for example).

.