Sentiment analysis on Donald Trump NYT comments
Machine Learning for Natural Language Processing 2020

Géraldine Houatra
ENSAE ParisTech
geraldine.houatra@ensae.fr

Benjamin Joly
ENSAE ParisTech
benjamin.joly@ensae.fr

Abstract

Donald Trump is quite a controversial politician. As a consequence, many articles about him are published in newspapers. Through our work, we tried to figure out what is the opinion about him emerging from the New York Times newspaper. Fort that, we performed a sequence classification task by training a BERT model. From the results, we can say with quite a good accuracy that the New York Times is far from being a pro-Trump newspaper.

1 Problem Framing

During the 2016 U.S. presidential election, many polls incorrectly predicted Hillary Clinton's victory over Donald Trump. Since its election for American presidency, the latter has always been a very controversial political personality.
The New York Times (NYT), one of the most famous American newspapers, mainly deals with political subjects. As a result, a large part of the articles published are related to Donald Trump. Through our project, we want to analyze thanks to a sequence classification task whether NYT articles convey a certain image of the president and if that image is highlighted in the comments or not.

2 Experiments Protocol

We exploited the NYT data from March 2018[1] and decided to build a train database composed of 50k comments. Among those comments, 20% are from articles dealing specifically with Donald Trump and 80% are from articles dealing with any other topic. We then divided it to build a 10k comments validation set. We also built two different test data sets : one composed of around 2k headlines and another one of 5k comments from articles dealing with Trump.

Our first work has been to label those comments ourselves as there were no label variable in the initial database. We first planed to finetune BERT on another database already labeled and dealing with the same topic as ours so that we could then apply it to our own database. Unfortunately we couldn't find any suitable source database for that. We then consider hand-labeling with Pigeon, but it turned out to be very time consuming given that NYT comments can be very long and dealing with complex topics. Moreover, we were willing to enrich our base with some comments from pro-Trump newspaper but we couldn't find such content. We thus finally decided to use AFINN lexicon in order to label and then train a BERT model.

This model seemed very appropriate for our sequence classification task given its bidirectional and attention based approaches, thus allowing to capture long-term dependencies quite naturally. This is an advantage over other models usable for that task such as LSTM which proceeds much more sequentially. We used the BERT base uncased version.
We first trained the BERT model from the fifth lab session. For that we made a three level labelling with AFINN : negative, neutral and positive[2]. We observed a mean loss decrease during the epochs on the validation set, but the mean accuracy was strangely always the same. Moreover, the accuracy on the test set was about 43% which is not really good.
We thus decided to train BERT once again, this

[1] data available at https://www.kaggle.com/aashita/nyt-comments

[2] see the 'Labellisation avec le lexique AFINN' part at https://github.com/geraldine-ht/NLP/blob/master/NLP_project_Joly_Houatra.ipynb

time based on a github notebook(**?**) using BERT for binary classification of IMBD movie reviews. For that, we thus made a two level labelling with AFINN : positive and negative.[2]
About three-thirds of the comments composing our training set have a length lower than 130 tokens so we decided to fix the maximum sentence length to 128 for computational resources reasons. For the modelling parameters, we set a batch size of 32, ten epochs with 500 steps for each and a learning rate of 2e-5.

3 Results

The previously described BERT model training led to a 85% accuracy on the comments test set. As we can see in the confusion matrix (Figure 1), only about 700 comments out of 5000 are wrongly classified. Moreover, by comparing those results with the one of an opinion lexicon such as BING, we can note the difference in performances. Indeed, the exploitation of BING[3] led to an accuracy of only 55%, with a lot of fake negative classified comments.
Qualitatively, our model is also pretty good. Indeed, our classification predictions on new sentences seem consistent with how an human being could interpret them.

We tested our model on 5k comments from more than 2k articles concerning Donald Trump. About 80% of those articles' headlines are negative according to our AFINN labeling. Concerning the comments, about 60% of them are negative. The NYT is thus definitely not a pro-Trump newspaper.

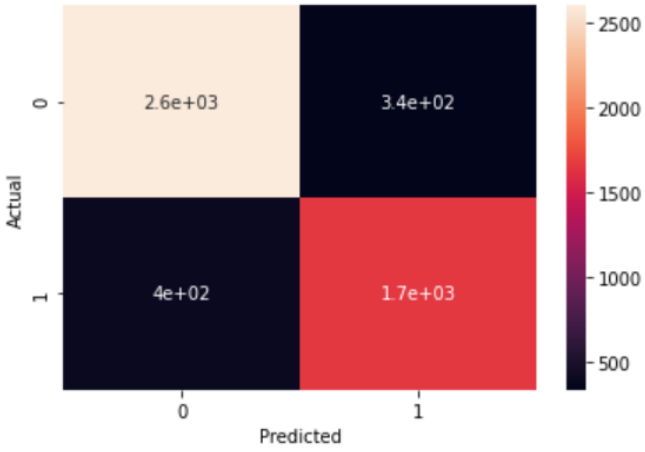

Figure 1: BERT confusion matrix on the test dataset

4 Discussion/Conclusion

According to our analysis, it seems that the NYT is a newspaper conveying a quite negative image of the American president through its articles. Those negative headlines may negatively influence NYT readers. Actually, most of the comments from those articles are negative too. We thus can conclude that NYT readers are rather Trump detractors than supporters.

A first discussion can be about the labels. Indeed, we mainly exploited binary labelling, but the exploitation of a neutral label could be interesting. Some informative and factual comments can be considered as neutral. This would allow to reconsider the balance between positive and negative comments. According to our analysis, we suspect that it would consequently decrease the share of positive comments, which is already not high. On the other and, we also suspect to be more difficult to get a good accuracy with a third label.
Moreover, it could be interesting to compare our BERT model results with those of an LSTM one, especially because our database is composed of quite long comments (up to some hundreds of tokens). Indeed, there is a maximum sentence length constraint of 512 tokens in BERT model, while there is not for LSTM. As a results, perhaps an LSTM may outperform BERT for such long comments.

[3] see 'Evaluation Quantitative - BING' part at https://github.com/geraldine-ht/NLP/blob/master/NLP_project_Joly_Houatra.ipynb

References

- https://lesdieuxducode.com/blog/2019/4/bert--le-transformer-model-qui-sentraine-et-qui-represente

- https://colah.github.io/posts/2015-08-Understanding-LSTMs

- https://mccormickml.com/2019/07/22/BERT-fine-tuning/?fbclid=IwAR1d1b2jyp8I6nZ3tKRryTzPDf9Sqff-HX7fzd3af6jX2NWUSWauRMl95DY

- https://www.tidytextmining.com/sentiment.html

- https://arxiv.org/pdf/1810.04805.pdf

- https://medium.com/@himanshu_23732/sentiment-analysis-with-afinn-lexicon-930533dfe75b

- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

- https://github.com/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb