Sentiment Analysis on Progressive Issues Dataset Machine Learning for Natural Language Processing 2020

NOKRI Amale
ENSAE
amale.nokri@ensae.fr

RAHIS Erwan
ENSAE
erwan.rahis@ensae.fr

Abstract

With the democratisation of social networks, the opinions and positions of web users are becoming increasingly important. Thus, several methods to analyse this data have emerged. We used Word2Vec and Bert to do sentiment and aspect classification. We carried out this work using a database of tweets on social issues.[1]

1 Problem Framing

1.1 Context and motivation

Sentiment analysis, in particular, is used to detect the presence of sentiment in texts on various levels. BERT developed by Google have been extensively used in the past few years to automatically detect sentiment but also topics in textual data. Given the dramatic increase of tweets posted regarding political topics, it is useful to apply those kinds of methods in order to sense the global sentiment of the internet users.

1.2 Dataset

The Progressive Issues dataset contains the results of questions asked to contributors after reading 1 159 tweets in english regarding a variety of left-leaning issues like legalization of abortion, feminism, Hillary Clinton, etc. They classified if the tweets were for, against or neutral on the issue (or none of the above). Our goal is to use this database as a training for an algortihm able to automatically deduce the aspect and the sentiment of a given tweet.

2 Experiments Protocol

2.1 Data cleaning

Before building our models, we started by cleaning up our text data. We removed elements that did not provide us with any information: hashags, URLs, put the tweets in lower case. We also detected multilingual expressions, such as Hillary Clinton, which cannot be separated.

2.2 Word2Vec

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. We used one of the most famous embedding methods: Word2Vec. The Word2vec algorithm uses a neural network model to learn word associations from a corpus of text. Once trained, this model can detect synonymous words or suggest additional words for a partial sentence. Word2vec represents each distinct word with vector.
Then, we used three models: logistic regression, random forest and KNN. We selected the model with the highest accuracy to analyse the confusion matrix. We did this for sentiment, and aspect.

2.3 Bert

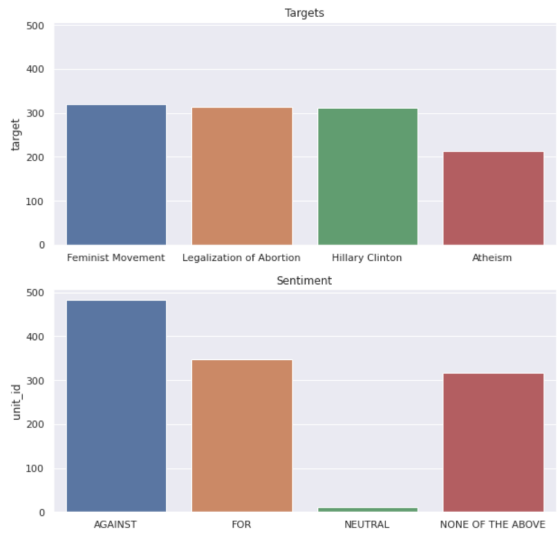
Given a tweet about a (progressive) political issue, it is interesting to perform an analysis of the sentiment and the topic of the tweet. This will be done using the BERT algorithm developed by the Google Teams. Two BERTS algorithms can be ran against a database that already contains the labels both for sentiments and aspects. BERT is a model for representing texts written in natural language. The representation made by BERT has the particularity of being contextual. In addition, BERT's context is bidirectional, i.e. the representation of a word involves both the words that precede it and the words that follow it in a sentence. We used Bert embedding, a neural network with a Bert layer and a linear layer.

[1] https://github.com/amalenkr/NLP–Progressive-issues-sentiment-

3 Results

3.1 Descriptive analysis

The word cloud allows us to see the most used vocabulary in the tweets. Some words refer to politics in the US ('Trump'), others to abortion ('prochoice').



The most talked about topics are: the feminist movement, legalized abortion and Hillary Clinton. The majority of tweets indicate a pro, followed by a con opinion. Very few individuals were neutral.

3.2 Sentiment classification

F1 score	**Word2Vec**	**Bert**
Against	49%	64%
For	43%	27%
Neutral	0%	0%
None of the above	33%	38%

For both methods, we look at the F1 score for each sentiment. The F1 score is 0% for the Neutral, which includes only 11 data. The F1 score is higher for 'Against' with Bert, and for 'For' with Word2Vec.

F1 score	**Word2Vec**	**Bert**
Atheism	52%	41%
Feminism Movement	51%	54%
Hillary Clinton	54%	6%
Legalization of Abortion	38%	31%

3.3 Aspect classification

The F1-score is better with the Word2Vec.

3.4 Qualitative analysis

We analysed the aspects and sentiments of two tweets. The first tweet : "the hijab ban in France is more than enough evidence that if your feminism isn't DECOLONIAL, it is also a tool for imperialist oppression" This tweet is about feminism, the person is for it. According to Bert's model, this tweet is about the feminist movement, and is against it. According to Word2Vec, the tweet is about Hillary Clinton and issues an opinion against. Both predictions are wrong, but Bert is getting closer to the truth.

"Cameras, chokehold bans, "retraining" funds, and similar reform measures do not ultimately solve what is a systemic problem. That system will find a way - killings happen on camera, people are killed in other ways, retraining grows $ while often substituting for deeper measures."

4 Discussion/Conclusion

This work was very interesting, but we encountered some limitations. It would have been better to use a database with more tweets. Also, instead of using Bert for the classification of sentiment and aspect, we could have done a Aspect Based Sentiment Analysis, which would have allowed us to have these two pieces of information, and to optimise our algorithm. Humans answer questionnaires on tweets, the possible answers to the possible feeling neither for, nor against, nor neutral, so none of the above is mutual. If a tweet is neither for nor against, it means it is neutral.

References

[1] Crowdflower. Progressive issues sentiment dataset. https://data.world/crowdflower/progressive-issues-sentiment, 2015.

[2] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv:1903.09588 [cs]*, 2019.