Project Machine Learning for Natural Language Processing : Sentiment analysis of climate change tweets

HEDFI Houeida

MEFENZA NOUNTU Thierry

April 24, 2021

1 Introduction and problematic

In this project we analyse sentiment analysis of tweets based on climate change. First of all, we will explore the corpus, we will perform preprocessing of our dataset (cleaning, tokenizing, ...). Second we will use Word2Vec Our project will be organized as follows. First, we will do some preprocessing on the dataset (cleaning, encoding, tokenization, etc) and explore the dataset. Second we will use Word2Vec to perform our embedding and construct models (SVM, RandomForest and KNN) to predict tweets sentiments. Third, we will use Bert architecture based on Hugging Face to perform embedding as well as predictions. Afterwords, We evaluate the performances of our models (quantitatively and qualitatively) and we compare our results. Fourth, we explore the aspect based sentiment analysis. We end this project by a conclusion and some possible future work.

2 Data presentation and preprocessing

2.1 Presentation

The dataset is composed by three columns sentiment, message (tweet) and tweetid. It is composed by 41033 unique tweets and each tweet is characterized by a sentiment towards the climate change. The labels -1, 1, 0, 2 are associated respectively with anti, pro, neutral and news sentiments.

2.2 Preprocessing

We have splitted the tweets into a list of tokens using five different tokenizers, we have performed some descriptive statistics and we have kept TweetTokenizer. Then we have built our cleaning function (removing special characters, stop words, hashtags, url, numbers, etc). The cleaning function transforms every tweet in the corpus into a list of cleaned tokens.

3 Embedding and sentiment prediction

We splitted our dataset intro three dataset, one for training (60%), one for validation (20%) and one for test (20%).

3.1 Sentiment prediction based on Word2Vec

Each tweet in all our datasets is transformed into a vector as follows: Each token is transformed into a vector of numbers of size 300 and each tweet is represented by a vector that constitutes the average of all tokens' vectors.

We have trained three models, SVM, RF and KNN. We have obtained, on the validation dataset, an accuracy (F1 score) of 0.58 (0.60) for SVM, 0.63 (0.58) for Random forest, and 0.60 for KNN. The best accuracy is

obtained with RandomForest, but the latter overfits.
The performances of those models are pretty close and not so high. We have analysed SVM model qualitatively. The tweets that we have observed seemed very ambiguous even for a human being. Our predictor couldn't assimilate the complexity of tweets understanding. This could in part explain the bad scores that we obtained. The other reason could be the fact that our embedding method do not take into consideration all the tokens vectors contained in a tweet, but uses only the mean, with possibly many zeros for tokens that do not belong to the vocabulary.
In the following we will try another strategy using Bert architecture.

3.2 Sentiment prediction using BERT

In this part we have encoded our tweets and we have used BertForSequenceClassification model which is obtained from the pretrained model bert-base-uncased with an additional linear layer. For our training we have used two epochs with a batch of size 32. We have obtained, on the test dataset, an accuracy (F1 score) of 0.74 (0.73). We can see that BERT is a more appropriate model for our data, as we have obtained better scores that those of the previous part. However, the results of the qualitative evaluation is similar to those obtained in the previous part.

4 Aspect based sentiments

Analyse sentiments on tweets is not only about predicting sentiments but also about getting some insights of aspects that the sentiments are related to.
In this section we do an aspect-based sentiment analysis in order We first clean our tweets and see it as a sentence. Then, we join consecutive word in a sentence and do Pos-tagging. Afterwords, we get relationships between words using stanza. The, we create a feature list that could possibly contain our sentiments and aspects. Then we find the relations between words in the feature list and we select the aspect and the sentiment from those relations.
However, our model did not gave us the best aspect and sentiment expressed in the tweets.

5 Conclusion

In this project, we first predicted sentiments of tweets on climate change. Our best performance is obtained with Bert. Second, we did an aspect-based sentiment analysis.
To conclude, our prediction as well as our aspect-based sentiment analysis does not perform very well in general. This could be due partly to the ambiguity of tweets, the quality of tweets and their cleaning. We lacked time to tackle these issues and we leave it as a future work.