Prediction of speaker during Democratic Primary Debates
Machine Learning for Natural Language Processing 2020

Robin Beuraud
3A DSBD
robin.beuraud@ensae.fr

Coline Fouché
3A DSBD
coline.fouche@ensae.fr

Link to Google Colab : https://colab.research.google.com/drive/1FVBw7UOb7cDESbyHt0DXoPFqyHafyR0K

Abstract

In order to find out whether the Democratic candidates, although ideologically close, were recognizable by their oratory intervention, we build models that predict the speaker as accurately as possible. We perform Sequence Classification using BERT word embedding and compare its performances with a classification using TF-IDF features from scikit-learn. We find that support-vector machine models predict relatively accurately which candidate is speaking and, in our case, works better than complex neural network method.

1 Problem Framing

The U.S. Democratic presidential primaries are the process, which runs from February to June 2020, by which supporters and members of the Democratic Party nominate their candidate for the 2020 presidential election. Since politics in the United States depend a lot on a candidate's ability to convince others through speech, we analyse the content of speeches from the candidates during 2019-2020 Democratic debates. Do candidates have distinct speaking strategies? In other words, **is it possible to predict the speaker from an oral intervention in a debate?**

2 Experiments Protocol

- Data (train and evaluation)

We use data from https://www.kaggle.com/brandenciranni/democratic-debate-transcripts-2020. Our dataset contains text transcripts of mutliple Democrat debates between June 2019 and Feb. 2020, with one row everytime a different person speaks. We clean and tokenize the speeches using nltk's TreeBankTokenizer, and apply a filter to keep only interventions from the five main candidates. This leaves us with 2071 observations. We randomly slice the dataset into 60% training, 20% validation, and 20% test. Our classes are balanced.

- Model used

BERT (Bidirectional Encoder Representations from Transformers) is a LSTM (long short term memory) neural network. We compare it with an SVM classification over TD-IDF features which had the best performance among a set of models.

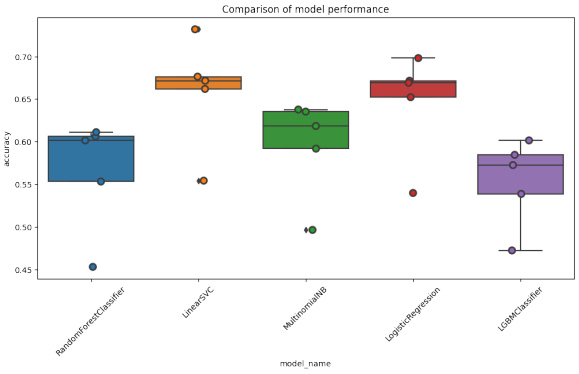

Figure 1: Selection of the challenger model

- Implementation

We use BERT in its distilBERT form to save memory and calculation power. We fine-tune its pre-trained embedding, then use distilBERTforSequenceClassification to perform a classification on our speeches, using the huggingface pytorch, transformers, and tensorflow libraries. It is trained using backpropagation and attention masks. We chose to train it in 4 epochs, althoug this induced slight overfitting (see figure below), because 4th epoch still significantly improved accuracy.

Figure 2: Loss for BERT training on train and validation set

3 Results

Model	Accuracy over training set	Accuracy over test set
BERT	0.44	0.48
SVM + TFIDF	0.76	0.75

Table 1: Evaluation metrics for two classification methods over our dataset

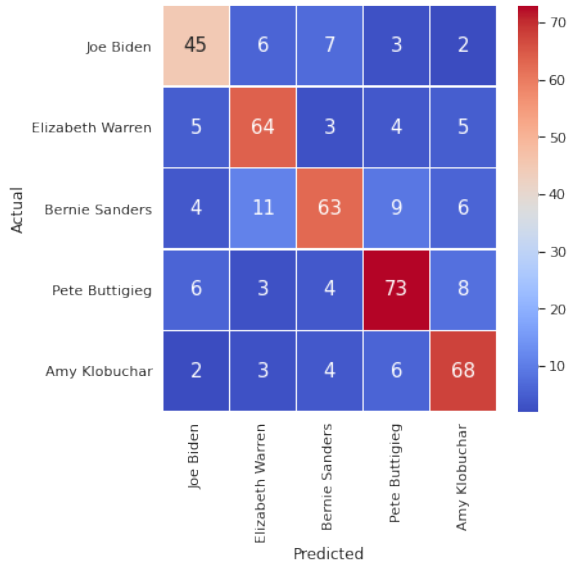

Figure 3: Confusion matrix of SVM predictions

The accuracy our implementation of BERT reaches is fairly low: only 48% of its predictions are accurate. Accuracy and F1 scores are particularly low for candidate Elizabeth Warren. This could indicate Warren has a less distinctive style than other candidates. A look into which speeches are wrongly predicted tells us that most short interventions are afflicted, as well as sentences where the speaker is interrupted.

We wanted to see if we could get better results using a different, possibly simpler model such as SVM over TF-IDF features. Table 1 shows this is the case, since by optimizing some of its parameters we were able to achieve 75% accuracy.
The confusion matrix for SVM shows that errors are distributed among candidates. It is noted, however, that on 11 occasions "Sanders" was predicted "Warren". These errors can be explained by the ideological proximity of the two candidates. By looking in detail at the prediction errors, it includes very generic phrases that are difficult to attribute to a particular candidate (e.g., "please let me answer now, it is important"). In addition, during a debate candidates interrupt each other. Yet, cut sentences where the key element is missing are difficult to classify: (ex: "I'm running because so many people...").

4 Discussion/Conclusion

We conclude from our analysis that "simpler" Machine Learning methods seem to result in better accuracy over our dataset. It is possible that a complex neural network method is not the most suited for small datasets like the one we used.

Finally, support-vector machine models predict relatively accurately which candidate is speaking despite a relatively small database.Despite being on the same political "side", Democratic candidates seem to have characteristics that distinguish them. It would be interesting to explore these differences in oratory/electoral strategies with topic modelling or regular expression methods. Moreover, in potential future debates between Trump and Biden it would be interesting to see if the latter changes his strategy and adapts his speech to his opponent.

References

https://huggingface.co/transformers/model_doc/distilbert.html
https://mccormickml.com/2019/07/22/BERT-fine-tuning/