Framing LREM politicians into a Left-Right matrix
Machine Learning for Natural Language Processing 2021

Guillaume Giacomoni
ENSAE
guillaume.giacomoni@ensae.fr

Manon Verbockhaven
ENSAE
manon.verbockhaven@ensae.fr

Abstract

We used state of the art ml techniques like LSA or MLM in ordrer to classify politicians from the brand new LREM party in order to see if the vocabulary they used was more right or left wing. Results shows that politicias from left or right wing parties are classified well but government's officials tend to be difficult to classify. You can find the code here.

1 Problem Framing

Political communication has evolved considerably over the last ten years. Social networks have become a privileged place for politicians to talk to their possible voters. Today, all prominent politicians have a twitter account that can be administered by one of their collaborators or by themselves. A marked Left-Right divide has shaped the French political landscape for the last forty years. However, five years ago a political party took power in France claiming to be neither left nor right. We then thought it might be interesting to use the tweets of all French politicians and recents NLP techniques to try to show a linguistic difference between left and right that would eventually allow to classify the LREM politicians despite their fierce will to avoid the label.

2 Experiments Protocol

In order to place the LREM party on a right-left fan, we use Word2vec and CountVectorizer to put the tweets into vectors. In a first part we take as training base all the tweets of the different parties except LREM. We first try to do unsupervised learning with a SVM to separate the tweets from the right and the left but resultats were not great. We then used a RandomForest which once trained obtained very good results on the training base but unsatisfactory results on the test base (ie the tweets of the LREM members). Indeed, the scores associated with the classification of the different tweets from LREM members obtained an average close to 0.5 and therefore not very conclusive. How to explain it? Since RandomForest is a linear classifier of the data, this means that the tweets of the right and left parties are characterized by the use of a particular vocabulary and are therefore stigmatized. This is due to the fact that the right and the left are concerned with different issues and have a different electorate ... and therefore have a distinct vocabulary. The LREM party has been very clever since their tweets address a wide range of topics and have a wide and non stygmatic vocabulary. To adress this issue we decided to create 5 categories of tweets in order to treat the problem by topics. We create the topics using the Latent Semantic Analysis (LSA) model. For each new tweet of a member of LREM we label it $\{0, 1, 2, 3, 4\}$, we then search the set of tweets corresponding to this label and train our RandomForest model on this restricted set of tweets in order to classify the starting tweet on the right-left spectrum. Finally we explore de DL approach by training the last layers of the DL CamenBERT model without doing any topic processing because the time needed to run a few epochs was too important ($\gg$ 24H) and the results were conclusive (on a test and train basis) without doing any topic separation.

3 Results/ Conclusion

The results of RandomForest without topic separation do not give conclusive results but that all the different members of LREM are at 0.5 on a scale from 0 to 1 (0: left and 1: right). Separating the training set in order to compare similar tweets seems to be a good idea but our small database does not allow to conclude on the results. CamenBERT's model gets good results on the training

base and on the test base without the need to separate the tweets by category but the training and even the testing is really long.

4 Reference

[1]*CamemBERT: a Tasty French Language Model*, Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, Benoît Sagot, arXiv:1911.03894, 2019.

.