Are french men and women politicians speaking the same language ?
Machine Learning for Natural Language Processing 2020

Morgane Hoffmann
ENSAE
morgane.hoffmann@hotmail.fr

Melchior Prugniaud
ENSAE
melchior.prugniaud@gmail.com

Abstract

For our project[1], we consider the task of classifying the gender of french politician speakers. In order to capture gender differences in this specific context traditional approaches including an XGBoost which combines homemade features and TF-IDF. Our project then compares this model to a novel approach using neural networks: Bert for Sequence Classification pre-trained on the French language (CamemBERT). Our approaches are tested on a rich corpus of french political speeches from 1980 to nowadays. We report an accuracy of 0.8 on a subsample of 5000 speeches.

1 Problem Framing

If french women and men speak the same language they might speak that language differently. This might be particularly decisive in the political environment. In this project we want to answer two questions :

- Do french political men and women use language differently ? How ?
- Can we train a model that would perform well in classifying those speeches and what would be the decision criterion ?

This question appears interesting to us because we could not find a good answer in the literature and almost none of them was really based on new NLP protocols. Also, few analysis investigated the French language.

In order to answer those questions we first analyze and describe the nature and the existence of gender linguistic differences in the context of political speeches in France by extracting features from the corpus. This approach using words, POS, n-grams and frequency measures is common in the literature. Second we train a classifier using several classical methods recognizing the gender of the person talking. We then compare our baseline to the performance of a BERT like models trained on French. We provide in the last part quantitative and qualitative results of these models.

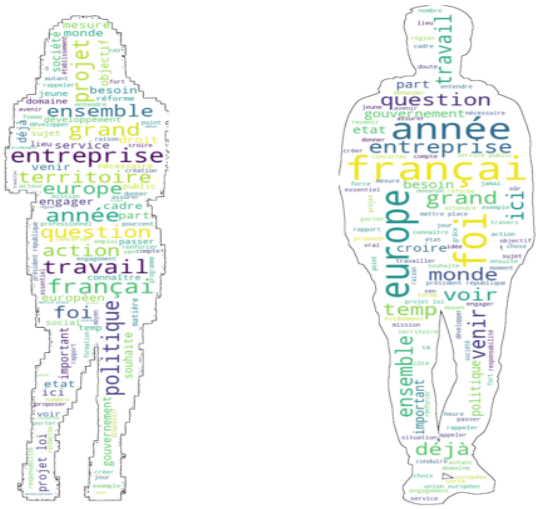

Figure 1: Wordcloud by gender

2 Experiments Protocol

2.1 Data and Preprocessing

We drew this project[2] on data scrapped from the french government database which gathers french political speeches from 1980 to nowadays. The complete database we collected contains around 39 000 speeches in total. However, due to technical constraints we used a subsample of 5000 speeches. Figure 1 displays a wordcloud which gathers most common words by gender.

[1] https://github.com/cerezamo/NLP_project_MHMP

[2] https://colab.research.google.com/github/cerezamo/NLP_project_MHMP/blob/master/NLP_project_MHMP.ipynb

2.2 Features extraction

Before running any kind of model we tried to extract interesting features from the text in order to spot differences between male and female speakers but also highlight potential biases and structure of our corpus. Our work is based on several past studies (Shlomo Argamon and Schler, 2009; Guefrech, 2015) which points out relevant dimensions of gender language differences. Based on these computations, we can for instance observe that women are using more personal pronoun than men. Moreover, we can see with the F measure that women seem more contextual than men.

2.3 Models

We mainly used the scikit learn api (Pedregosa et al., 2011) for building the first bunch of models. The same protocol is applied uniformly for all models. In our sample 75% of the speeches are told by men. Models are applied to both an unbalanced and a balanced dataset. Corpus is split into a 80% training and 20% validation group. We tried several approaches to find a good baseline model which included our homemade features or word embedding method or a combination of both. We then constructed a CamemBERT model using Hugging Face library (Wolf et al., 2019). For this final model we had to overcome the limit input length of 512 tokens and an optimal strategy has not been spotted yet(one method is proposed here (Raghavendra Pappagar and Dehak) and in this tutorial (Olivares). We tried first an unbalanced and balanced sample using only the first 512 tokens. We also tried to split our texts in smaller pieces of 512 tokens maximum and fed the model with the latest. Finally, this last option has been chosen and grouped the class predictions.

3 Results

3.1 Baseline model

The best classifier among our traditional models is the XGBoost classifier (Chen and Guestrin, 2016) including TF-IDF and some of our features. After some fine-tuning, we get an overall F1 score of 0.78 on the test set but also a ROC curve with a good shape (AUC : 0.77). Women are classified slightly better than men. If we dive a bit more into the results, we can observe that some words have more importance in order to classify women versus men and it seems that many words are linked to certain topics. For example, we can see that women tend to speak more about other women and health and on the other hand men are speaking more about economic topics like corporation or development.We observe that the number of pronoun has a big impact on how a text will be classified. Another important feature is the positive sentiment polarity of the text and on the contrary we observe that the XGBOOST has penalized some variables like anger word frequency. Our result are in line with other studies (Lenard, 2016) (use of personal pronoun).

Model	F1 score	Accuracy	AUC
XGBoost	0.778	0.776	0.778
CamemBERT	0.83	0.832	0.831

Table 1: Performance metrics for XGBoost and CamemBERT

3.2 CamemBERT

We observe slightly better performance of BERT over our baseline model. In order to open up the BERT black box we reproduced ex 4 of TP4. From the few sentences we analyzed "assumons" seem to be a female oriented word whereas "mais" is male oriented according to the model. Otherwise we had a hard time understanding and finding clue on the criterion decision of the model through this qualitative analysis.

4 Discussion/Conclusion

Improvements include applying hierarchical transformers for long document which method is presented in (Raghavendra Pappagar and Dehak). Moreover, we could have chosen to balance our sample on all observables to see how the models would behave in this case and avoid theme biases. Besides these technical aspects we had further developments in mind regarding the data we had. In our database we had speeches and interviews. In this work we decided to focus on speeches but interviews might be more relevant as speakers are forced to behave more naturally than in the context of formal declarations in front of an audience. Second, using those interviews, we wanted to construct a "political chatbot" that would answer political related questions. Two models could have been trained on women and men and a single question would have been asked to these two models and answers from both models could have been analysed.

References

Jesus Villalba Yishay Carmiel Raghavendra Pappagar, Piotr Zelasko and Najim Dehak. HIERARCHICAL TRANSFORMERS FOR LONG DOCUMENT CLASSIFICATION.

Armand Olivares. Using bert for classifying documents with long texts.

James W. Pennebaker Shlomo Argamon, Moshe Koppel and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Moadh Guefrech. 2015. Author profiling: Gender and age detection text mining project report.

Dragana Bozic Lenard. 2016. Gender differences in the personal pronouns usage on the corpus of congressional speeches. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 3.2:161–188.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. pages 785–794.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.