Predicting the political trend of deputies from their interventions at the Parliament

Gabriel Watkinson and Jérémie Stym-Popper

Master MVA - ENS Paris-Saclay, Gif-sur-Yvettes, France

gabriel.watkinson@ensae.fr
jeremie.stym-popper@ensae.fr

MATHÉMATIQUES VISION APPRENTISSAGE

1. Introduction

With the political debate being more and more open to the public, having tools to analyse the discussions taking place is really valuable. This evolution in the Parliament translates into many more data of all type being openly available, from the live debates broadcasted on television and online, to the reruns, annotated with timestamps, theme and speakers, to the full man-made retranscription of all seances, available on the website of the French Assemblée Nationale.

Therefore, NLP models can be used to conduct analysis on this textual data. The most basic task we wanted implement, is to predict the political party of a deputy from his addresses at the Parliament. This can then be used to regroup speakers with similar ideas, associate new speakers to a political idea, to analyze the opinion of a deputy on a number of subjects, etc.

2. The dataset

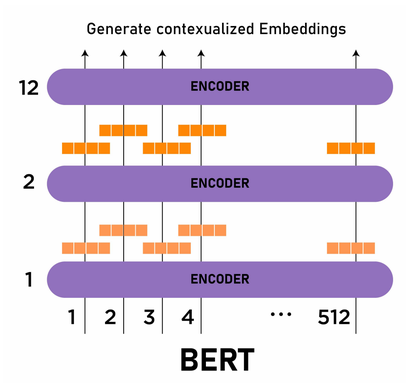
The transcriptions of the session at the national assembly are made publicly available on the website `assemblee-nationale.fr/`. There we can find all the interactions, the statements of the deputies, the questions, the reactions and the organisation of the session. We gathered the data through the api made available by `nosdeputes.fr`

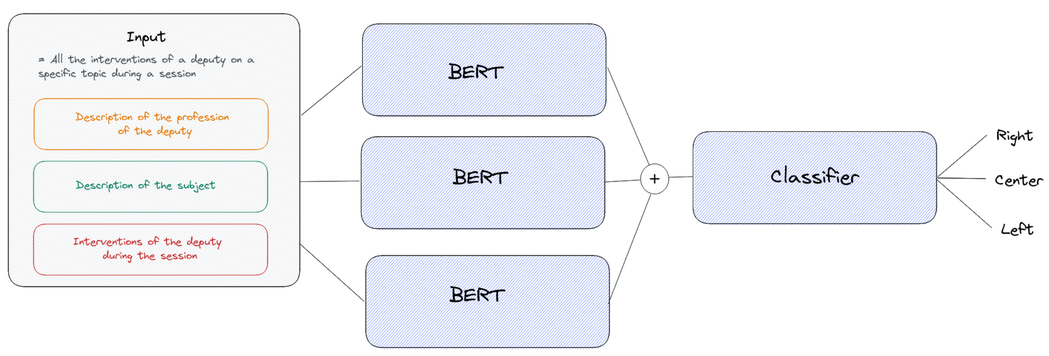
The intuition we have is that the words spoken by the deputies , the tone, the prefered subjects, etc., vary from party to party. The previous figure shows an extreme example were we can easily identify the parties.

In the end, we focused on the interventions, the titles of the session that gives the themes of the interventions, and we also added the profession of the deputy. Other data is available and could improve the model.

3. The methodology

We took inspiration from Ahmed, Nabib [1] and Terechshenko, Zhanna et al. [4] to build a model for politics. We relied on the well known BERT [2] architecture, and added an extra layer to classify the data. We tried both the BERT multilingual and the CamemBERT [3] weights.





In practice, we had three textual inputs, all the interventions of a deputy on a specific theme during a given session. We also added the profession. We encoded them each with a BERT model, then added or concatenated the embeddings before building a basic multi-layer perceptron to learn the classification. We also restrained the labels to the right, to the left or to the center in order to have a classifier that works across multiple legislations.

4. The experimental setup

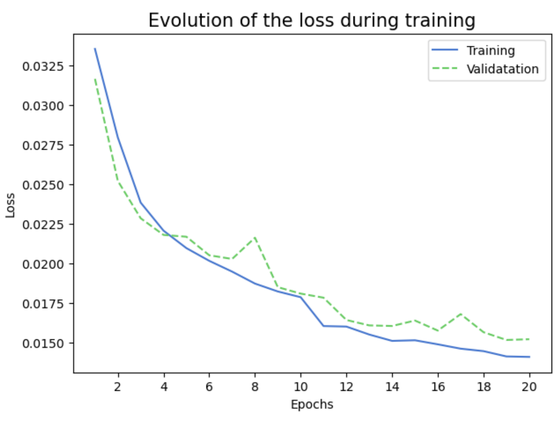
We build a dataset containing one observation for each session, theme and deputy. The theme and profession were used as extra inputs to improve the model, but the main data are the speeches of the deputy. We restrained ourselves to the 2017-2022 legislature.

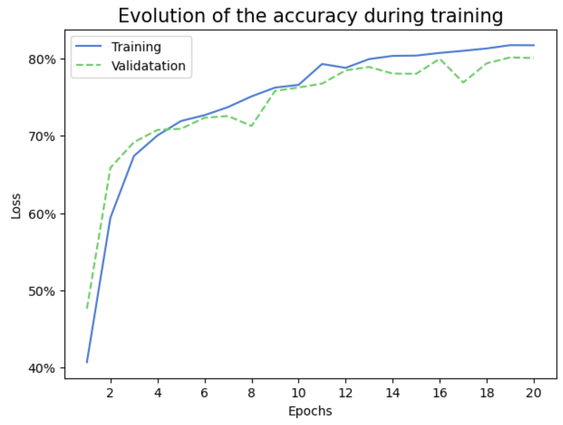
We first cleaned the texts from HTML residuals. Then, we tokenized it with BertTokenizer, we padded the inputs to 512 for the interventions, 64 for the theme and 16 for the profession and batched the data with size 32.

The classifier consists of a simple three layers perceptron, whose input is either the sum or the concatenation of the embeddings of the three inputs. The intermediate dimension we used is 256. We implemented the training loop using Pytorch Lightning, using the ADAM optimizer.

5. First results

We trained our model for 20 epochs and obtained decent results, around 80% weighted accuracy. We tried many different variations of the model, but the one presented above performed the best with a smaller memory footprint, and faster training.





6. Discussion

As of now, we considered the deputies independently from each other, however, the data corresponds to a discussion between them. Therefore, a more complex architecture (sequence to sequence), taking into account all the information in a session would be much more interesting. It would also be interesting to look at the evolution between legislatures.

7. Conclusion

The fact that our classifier is way better than a random one shows that the lexical field of the deputies is linked to their political party, thus their ideas. Thanks to all this information, we can build a model that helps to see more clearly inside the semantic and the prevailing opinion of each speech.

8. References

[1] N. Ahmed. *Natural Language Processing Techniques for Political Opinion and Sentiment*. PhD thesis, Harvard University, 2022.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] L. e. a. Martin. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[4] Z. e. a. Terechshenko. A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN 3724644*, 2020.