The political speeches of Eric Zemmour

Thibaud Cazanave
Affiliation
thibaud.cazanave@ensae.fr

Anatole Sadri
Affiliation
anatole.sadri@ensae.fr

1 Problem Framing

We are studying Eric Zemmour speaches during the 2022 French presidential election thanks to NLP tools in order to get a better insight into his political strategy. Eric Zemmour has delivered his speeches in several French cities from December 2021 to the end of March 2022. These cities differ in their demographics, economic conditions and political identities.

1. For instance Cannes and Toulon are associated with conservative positions on cultural issues, security and immigration. Those were promoted by right and far right elected politicians over the last decades : Eric Ciotti, Thierry Mariani, Marion Maréchal Lepen, Stéphane Ravier.

2. On the other side Saulieu, Saint-Quentin, Chaumont-sur-Tharonne are more rural and modest cities with probably more social expectations ans less political legacy from inhabitants with a most social background.

We could expect that the main ideas of his speeches are driven by these sociological features. As a politician, Eric Zemmour may try to gather different categories of French voters and could have an incentive to fit the content of his speeches to his audience expectations. However it is also possible that his speeches reflect more the issues on the agenda at a given time of the campaign, with some topics being imposed such as the Ukrainian war.

We are thus wondering whether the geography or the chronology is a better tool to understand what are the main topics of Eric Zemmour speeches.

2 Experiments Protocol

We have tried to address the following issues.

- **Data cleaning and tokenization**: which is the best way to tokenize the data in our context ? We have also tried to remove the verbs with tagging methods to only focus on nouns or adjectives.

- **Summary statistics LDA** on the whole corpus: what are the main topics expressed in Eric Zemmour speeches with simple figures and plots ?

- **Word embeddings**: we want to measure words similarity within the corpus with Word2Vec. There are two goals:
  - What are the words associated with Zemmour main topics during the campaign: ”France”, ”Reconquête”, ”Immigration”, ”puissance” for instance ?
  - Is the word similarity robust to a corpus split into two or more clusters (rural areas against the others for instance) ? We are expecting that ”France” will be more related to agriculture and countryside in the first case for instance.

- **Clustering**: we are trying to make clusters from Zemmour speeches and match them on a geographic or a temporal scale. We wanted to use both TD-IDF and Word2Vec, where TD-IDF works as a baseline.

- **Sentiment analysis**: we are trying to compute a positive or negative score for each speach, which may help to rank them and match their score to a geographic or temporal scale.

- **Text generation** : we are using a LSTM and a n-gram model to make a text generation from Zemmour corpus. We want to use the clusters to train the model on different dataset and compare their output. Does the text generation reproduce the main topics associated with the underlying demographics and sociological features of the clusters ?

3 Results

We have found three main clusters by using Kmeans on the TD-IDF vector representation of speeches. We have tried to make a relevant intepretation of them

1. A "national" cluster gathers the first (Villepinte) and the last (Trocadero) speech of Eric Zemmour with the speech that he delivered in Cannes in the middle of the campaign. We have called it "national" because we expect these speeches to target a very large audience (youtube, real time information TV channels). These speeches do not seem to be really influenced by the place: Trocadéro and Villepinte were probably chosen because Paris is very well connected and in the middle of the 12 billion inhabitants Ile-de-France area.

2. An "international" cluster gathers speeches related to immigration (Calais), geopolitics and the Ukranian crisis. Both the location (Calais) and the date of the speech matter here. Eric Zemmour talked a lot about peace in Savoie because of the Russian invasion, not because of the Chamberry demographics. These speeches deal with the international relationships and the French position and power in the long run. That is why we also find Mont-St-Michel in that cluster.

3. The last cluster gathers the vast majority of the speeches. We have named it "rural" or "social" because these speeches took place mostly in rural areas and their main topics were more about voters concerns of the daily file such as purchasing power or education.

All in all it seems that the geography is still a better dimension to understand the content of his speech than their date.

From a more technical point of view, we have succeeded and failed in the following fields.

- **Data cleaning and tokenization**: the regexp tokenizer proved to be slightly better but its added-value was not striking. The tagging methods failed to make a clear difference between nouns and verbs so we have decided not to use it to clean the text data.

- **Summary statistics LDA** on the whole corpus: the LDA was quite successful on

- **Word embeddings**:We have not formally optimized the best parameter values of the word2vec model but we have looked at the variance of the similarity scores output to check whether they could differentiate the words or not.

- **Clustering**: we have only used the TD IdF method. The Word2Vec model was too complex for a limited number of speeches in this specific task.

- **Sentiment analysis**: we have computed a positive score for each speech. We could try to make an interpretation for the most positive and negative speeches. This method brought some added-value to our clusters by adding a new dimension

- **Text generation** : we have succeed in generating text with the LSTM model but we could not train it only on some clusters because the data would have been too limited. This last step is more a bonus one because it cannot really help us to measure the quality and relevance of our clusters.

4 Discussion/Conclusion

We have managed to make a relevant clustering of Eric Zemmour speaches. We have tried to check if these clusters were robust to alternative measures such as sentiment analysis, word similarity and even text generation (we wanted to see the impact of the speeches on which the LSTM model was trained in the output, be it more focused on social or security issues). The more quantitative analysis was probably too ambitious for our amount of data but the more qualitative parts give an interesting insight into Eric Zemmour speeches and their connection with the place they took place.

Instructions (to remove):

1. write and present clearly and synthetically your project

2. Add all the references in Appendix (use .bib (**?**))

3. 1 page maximum

4. You can add any non-textual content (plots, table, images, schemes...)

5. Should point to your notebook (colab or github/gitlab (e.g. with footnote [1])) .

Abstract

[1] https://nlp-ensae.github.io/