DEMA^{2}IN at Touché: Salient Events Extraction for Ideology and Power Identification in Parliamentary Debates

Notebook for the Touché Lab at CLEF 2025

Benjamin Callac[1,*,†], Anne-Gwenn Bosser[1], Florence Dupin de Saint-Cyr[2] and Eric Maisel[1]

[1] *Ecole Nationale d'Ingénieurs de Brest, Lab-STICC*

[2] *Université Paul Sabatier, IRIT*

Abstract

Parliamentary debates are a rich source of information for analyzing political ideologies and power structures. In this work, we explore an approach to train a model with minimal text input, by focusing on the extraction of salient events from transcribed speeches from the ParlaMint corpora. Rather than relying on full-text inputs, we investigate whether a distilled representation of meaningful events is sufficient to train classifiers for downstream political tasks. Although the reduced input leads to performance slightly below baseline this aligns with expectations given the minimal training data.

Keywords

Event Extraction, NLP, Political Debates, CLEF, Touché

1. Introduction

Parliamentary debates play a critical role in the life of a democracy. They require knowledge of the political landscape to accurately interpret the ideologies and power dynamics expressed by political actors. As such, analyses of political discourse often rely on manually reviewing the data, which can be both time-consuming and limited in scope.

With the advent of machine learning and large language models, the field of natural language processing has had tremendous advancements in recent years. Notably, in the fields of sentiment analysis and text classification, comprising the tasks of extracting the meaning of a document, these technologies have achieved impressive performance levels by learning complex patterns and contextual nuances from vast corpora.

In this context, Touché at CLEF suggests a task aiming at the identification of power and orientation based on parliamentary speeches [1].

Our approach to this task is to verify whether or not summarizing a speech and extracting the events related within keeps this type of classification task robust. Rather than relying on explicit indicators of stance, manner of speech, or action proposals, our method aims to capture the underlying description of a situation and events from political actors. This approach offers a more abstract way to analyze political discourse, emphasizing the events that are reported rather than how they are framed rhetorically. As such our evaluation aims to assess whether this event-centered summarization retains meaningful information for a model to distinguish between ideological orientations and between the party in power or the opposition.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

benjamin.callac@enib.fr (B. Callac); bosser@enib.fr (A. Bosser); dupin@irit.fr (F. D. d. Saint-Cyr); maisel@enib.fr (E. Maisel)

2. Dataset

Touché [2] provides a dataset to work with on this task, it is composed of a selection of speeches from the ParlaMint corpora [3]. The dataset is comprised of parliamentary speeches in 29 European languages and has been modified to exclude any information about parties and speakers susceptible to aid in the identification of the labels we are trying to predict. Each different language is contained in its own file, all the relevant labels, however are combined in a single file. The dataset file is composed of the following fields :

id is a unique (arbitrary) ID for each text.

speaker is a unique (arbitrary) ID for each speaker There may be multiple speeches from the same speaker.

sex is the (binary/biological) sex of the speaker. This information is collected from varying sources (typically data published by the respective parliament), and in some cases it may be unspecified or unknown.

text is the transcribed text of the parliamentary speech. Real examples may include line breaks, and other special sequences escaped or quoted.

text_en is an automatic English translation of the corresponding text. This field may be empty (obviously) for speeches in English, but the translations may be missing for a small number of non-English speeches as well.

orientation is the binary/numeric label (0 is left and 1 is right). Orientation labels are based on Wikipedia.

power is the binary label for power role (0 is opposition, 1 is coalition), this information is based on the information provided by the ParlaMint contributors. This value is not always present, either due to parliamentary systems with no defined coalition/opposition, or unknown orientation information for some speakers (e.g., PMs with no party affiliation). Missing values are indicated as 'NA'.

populism is a populism index based on multiple expert surveys (to increase the coverage). We focus on a particular dimension of populism in this task: the position of the party of the speaker in populist - pluralist spectrum. This is measured on a 4-point ordinal scale (1: Strongly Pluralist, 2: Moderately Pluralist 3: Moderately Populist, 4: Strongly Populist). Not all values are present in all parliaments. Many parties/speakers are not covered by the data, and some values are missing due to failure to match the survey identifies/names and ParlaMint identifiers. Missing values are indicated as 'NA'.

We focus solely on speeches from the British Parliament. This allowed us to obtain a baseline of our system's performance without having to factor in automatic English translations while still working on enough data for our investigation.

3. System overview

In this section, we describe how we processed the data before feeding it to a model for fine-tuning. For the model we opted to use Mistral-7b v0.2 Instruct. Approaches in past editions [4] of this task tried to preprocess and augment the dataset as much as possible to give models more context to make predictions out of, with great results. In our system however we aim to verify whether or not subtracting elements to only keep the key events described in the speeches is enough to make accurate predictions.

We largely base this work on CALLMSAE by [5] in which, via cascading LLM prompting, they are able to build causal graphs describing the events and their relations — hierarchical, temporal or causal. We found however that the generation of causal graphs did not lead good results on a large part of the data and ended up focusing on salient events generation.

3.1. Data processing

As described in [5] we first prompt a LLM to summarize a speech, because LLMs, much as humans, tend to only relate the most relevant events when asked to summarize a document.

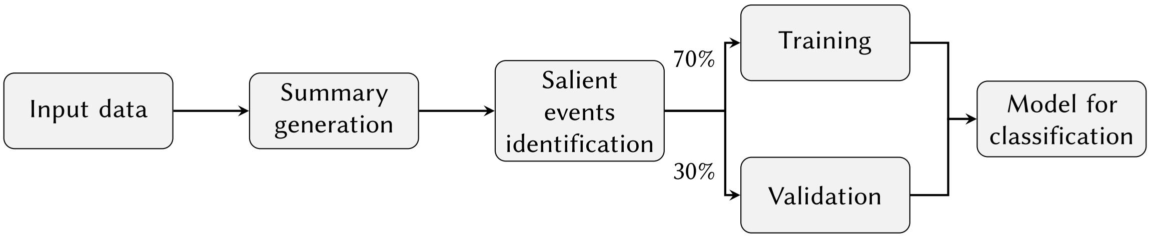

Figure 1: Method illustration, based on [5]. The input data is comprised of the English texts of the training dataset.

Prompt 1 - Summarization

```
You are a helpful assistant. Write a detailed summary of the document below.
Document: """ text """ Summary:
```

Then we instruct another LLM to extract the events from this summary: the event generation is often lacking events. To prevent this issue and to get the most from the summarized text we elected to run this process multiple times. The results of these multiple processing are then aggregated with another prompt asking to combine all the different lists while eliminating duplicates and entries with semantically close meanings. During our different attempts we observed the final event list would not change substantially when combining the results of more than 4 different promptings, so that is the number of total passthroughs we finally opted to use.

Prompt 2 - Salient Event extraction

```
A structured event is something that happened as described in the text. A
structured event is represented as a tuple, which consists of actors, a
trigger, and objects. Could you list all the structured events in the
following article? Example: 1. (John; married; Alice). 2. (Alice; was
hired; by Google). Format the results in between parenthesis as in the
example. Article: """ summary """
```

Prompt 3 - Event list aggregation

```
You are a helpful assistant. Given a list of events if there are semantically
similar event descriptions, choose the most accurate and typo-free version
of the event, do not lump details together. Then, return the events list
with no duplicate using the following format: Example: 1. (John; married;
Alice) 2. (Alice; was hired; by Google) 3. (John; was hired; by Google)
Events : """ events_string ""“
```

For the most part of the data, following this processing gives us a relevant event list of what transpires from the speech. However with the sheer amount of data, there is no way to know for sure if there are important events missing from the lists. Also something that we do not account for in our approach is the possibility of hallucinations from the LLM.

3.2. Classification process

Classification was made by fine-tuning the pre-trained BERT base for all the different labels to predict. To account for the two different tasks that are the identification of power and identification (a multilabel classification task with binary output) and the populism identification (a single label classification task with output ranging 4 values) we trained two separate classifiers. The first classifier performs a multi-label classification on both power and orientation, while the second one performs a single label classification task on populism only. Both classifiers are based on the same architecture and use identical hyperparameters except for the problem type. We use the salient events data obtained previously for this fine-tuning step. Each input corresponds to a list of events represented as a single string and is tokenized using the Standard BERT tokenizer with truncation and padding up to 128 tokens. For the multi-label setup, the corresponding binary labels were encoded into a multi-hot vector of length two (orientation, power). We then trained a classification head on top of the BERT base encoder using the Hugging Face AutoModelForSequenceClassification. We do the fine-tuning using the Hugging Face Trainer API. Ultimately, training was conducted on GPU with a batch size of 8, a learning rate of 2×10^{-5} for 5 epochs. Training and validation were done using an 80/20 split. For inference, we applied our fine-tuned multi-label BERT model to the separate test set consisting of unseen speeches represented by extracted salient events. Each data point was passed through a tokenization pipeline and fed into the model in batches. The multi-label classifier produced independent sigmoid-activated probability scores for each label while the single label classifier produced logits for the four populism classes, to which we applied an argmax operation to assign the most likely class label to each event list.

3.3. Parameters

Because of the large amount of data to process through multiple LLMs in a downstream task, we put a certain amount of effort in making the processing faster. We used VLLM to load the models and make inferences with, prompting the model in batches of 5 prompts at a time. The default sampling parameters were set at .3 for the temperature and 512 for the maximum tokens. Although, this last value changes depending on the prompt, we used a maximum of 300 tokens for the summary generation. We do believe we could have gone lower for the amount of tokens but that would have meant taking the risk of impacting the quality of the output without being able to observe it directly.

4. Results

In this section we shall review and be critical of the results we obtained for the submission on the Touché test data.

During the validation phase, we obtained the metrics shown in table 1. We observe right at the validation that while the F1-score is not terribly low, the accuracy however is barely better than a coin-flip.

Metric	Value
Loss	0.554
Accuracy	0.592
F1-score	0.796

Table 1
Validation metrics after the training phase.

As we only used speeches in the English language, we can only match our results with the baseline on the very same data. The baseline is a simple system, vectorizing the documents with TF-IDF and making predictions through linear regression. We compare the results of our system with the baseline results in the tables 2, 3, 4.

Team	Precision Orientation	Recall Orientation	F1 Orientation
baseline	0.77	0.771	0.77
dema2in (ours)	0.727	0.724	0.719

Table 2
Orientation metrics of the baseline and our approach.

Team	Precision Populism	Recall Populism	F1 Populism
dema2in (ours)	0.561	0.556	0.558
baseline	0.718	0.517	0.501

Table 3
Populism metrics of the baseline and our approach.

Team	Precision Power	Recall Power	F1 Power
baseline	0.784	0.762	0.766
dema2in (ours)	0.737	0.727	0.729

Table 4
Power metrics of the baseline and our approach.

In these results we observe our model is barely below the baseline in terms of F1-Score except for populism. This outcome, while slightly underwhelming considering the heavy preprocessing applied to the data was anticipated due to the nature of the data we fine-tune the model with. A substantial amount of potentially useful data for the model to train with was excluded, thus limiting to the broader context that could have contributed to better performances. Although the subset of data used to train the model was easier for humans to understand, it lacked information that could have been present in the entire texts.

5. Discussion

There are limitations in this experiment. It would have been interesting to use the same approach using different models to see how they compare, especially considering the numerous powerful models available. In the future we would like to compare the results using Chat-GPT, BERT, RoBERTa, to see if they yield different results. Furthermore, the evaluation setup did not allow for a thorough analysis of the impact of each stage in the pipeline — summarization quality, the precision of event extraction, and the influence of lost contextual information. These components likely have a significant cumulative effect on the final classification performance, but their individual contributions were not quantified. Future developments could benefit from a more detailed evaluation of each component in the pipeline, as well as maybe more advanced summarization and event extraction methods.

Acknowledgments

Many thanks to my supervisors Anne-Gwenn Bosser, Florence Dupin de Saint-Cyr and Eric Maisel who have trusted me in trying out this approach. This work was supported by the French Defence Innovation Agency (AID) and the Region Bretagne.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4o in order to: Grammar and spelling check. Abstract drafting. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes

in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[2] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[3] Ç. Çöltekin, M. Kopp, K. Meden, V. Morkevicius, N. Ljubešić, T. Erjavec, Multilingual Power and Ideology Identification in the Parliament: A Reference Dataset and Simple Baselines, 2024. doi:10.48550/arXiv.2405.07363. arXiv:2405.07363.

[4] O. Palmqvist, J. Jiremalm, P. Picazo-Sanchez, Notebook for the Touch{é} Lab at CLEF 2024 (????).

[5] X. Tan, Y. Zhou, G. Pergola, Y. He, Cascading large language models for salient event graph generation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 2223–2245. URL: https://aclanthology.org/2025.naacl-long.112/.