Bachelor's Thesis

On the difficulty of using Contextual Word Embeddings to Measure Political Polarisation in Parliamentary Speech

Author
Giulio Cusenza
giulio.cusenza@student.uni-tuebingen.de
giuliocusenza@gmail.com

Supervisor
Çağrı Çöltekin
cagri.coeltekin@uni-tuebingen.de

A thesis submitted in partial fulfilment
of the requirements for the degree of

Bachelor of Arts
in
International Studies in Computational Linguistics

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

October 2024

Abstract

...

Contents

1 Introduction

Since the field of Natural Language Processing (NLP) was started, there has been significant interest in employing these techniques in various political settings, whether for the analysis of political speech or as tools for politicians themselves. More recently, the exacerbation of public debate has raised the questions of a definition and a measure of political polarization (Lelkes, 2016), with NLP researchers exploring several different methods (Németh, 2023). A particularly intriguing approach was proposed by Ding et al. (2023). Their strategy involves using BERT (Devlin et al., 2019) to obtain contextual word embeddings for topical words from two major broadcasting companies with opposing political alignments, the hypothesis being that the same topical words (e.g., *racism*, *immigration*, *police*) would be employed differently by the two broadcast networks. This difference can thus be quantified through the cosine similarity between the embeddings, providing a measure of polarisation.

As Németh (2023) finds, only 20% of NLP studies on political polarisation use congressional or parliament speeches as their data sources. While broadcast and social media are undeniably important to understand social phenomena, it is likewise important to investigate the extent and influence of polarisation in key decision-making institutions. This study attempts to reproduce Ding et al. (2023)'s study running the analysis on parliamentary data instead of broadcast media. It is worth noting that, despite the authors' claim to include the study's code framework in their GitHub account, there is no mention of the account's name or URL, nor was it possible to contact the authors through the provided e-mail addresses. Therefore, while the technique may be the same, the details of the implementation are necessarily different, hindering the full reproducibility of the study.

The initial plan was to refine the method on Italian data, and then potentially repeat the analysis across multiple countries. This is not as straightforward as it may seem: while in the United States and other countries the two traditional opposing parties still dominate the political scene, many other Western countries have been experiencing a severe party fragmentation (Mair, 2023) and electoral volatility (Blumenstiel and Plischke, 2015). Therefore, identifying the candidate "poles" for such an analysis on polarization is no trivial task, and requires a thorough understanding of each country's political framework. The high electoral volatility also complicates diachronic analysis especially for a study on parliamentary debates, as the amount of data available for each political group shrinks and stretches by each legislative term based on elected representatives.

As the Italian data presented these and other more trivial problems (see Section 2), the study was also conducted on data from the United Kingdom's House of Commons. These data offers advantages for two primary reasons: first, the UK's traditional political parties remain relatively strong; second, the availability of English-language data allows for the application of more sophisticated language models. Despite this,

this study failed to reproduce Ding et al. (2023)'s results in both datasets, finding no evidence for the efficacy of contextual embeddings as measures of political polarisation. The paper concludes with a speculation regarding a few potential reasons for the increased complexity of analyzing parliamentary data compared to broadcast media.

2 Data

ParlaMint[1] is a project by CLARIN[2] providing "comparable corpora of parliamentary debates of 29 European countries and autonomous regions, covering at least the period from 2015 to 2022". This study utilises the corpora from the Senate of the Italian Republic (from March 2013 to September 2022) and United Kingdom's House of Commons (from January 2015 to July 2022) included in ParlaMint's 4.1 release (Erjavec et al., 2024).

The Italian dataset presented significant challenges due to pervasive structural flaws in the annotation and data formatting. These issues primarily arose from difficulties in parsing the plain text files provided by Italian institutions. More specifically, the parser failed to correctly distinguish individual utterances: in some instances, entire dialogues between multiple speakers, along with their names, were included in a single segment. Ideally, these segments should be separated and annotated for each speaker, with speaker names stored as XML attributes. This is a sensitive issue because the other speakers' utterances and even their names can affect the contextual embeddings (e.g., surnames such as senator Alessandra Mussolini's may considerably influence the embedding model for evident reasons). This issue could not be addressed, however, the same analyses were run on the data from the House of Commons as well.

Moreover, speakers are categorized as either `#guest` or `#regular`, with the latter tag assigned to elected representatives of the Senate. However, many guest speakers – typically government officials – are marked as `#guest` only at their first intervention in a session, while subsequent utterances are incorrectly labeled as `#regular`. This issue was easily resolved through a simple script.

It is worth to note the peculiarities of parliamentary data compared to the broadcast media language used in (Ding et al., 2023)'s study. A striking difference is quantitative: parliaments only have a few sessions per week, and part of them are spent in rituals and votes. On the other hand, broadcasting companies often have multiple television or radio channels with uninterrupted streams of news and talks, making for a significantly larger amount of data over the same time span. Figures 1 and 2 show the number of occurrences of the topical words selected for this study over the analysed periods in the House of Commons and Italian Senate respectively. The lower

[1] clarin.eu/parlamint
[2] clarin.eu

occurrences in the Italian date may be due to the parliament's structure (the other house, the *Camera*, is not included), to the dataset problems mentioned previously, or to a possible tendency to periphrases. It is interesting to observe the similarities between some of these graphs. For example, in both cases the topic of immigration peaks in 2018, while energy and Russia sharply increase after Russia's full-scale invasion of Ukraine in 2022.

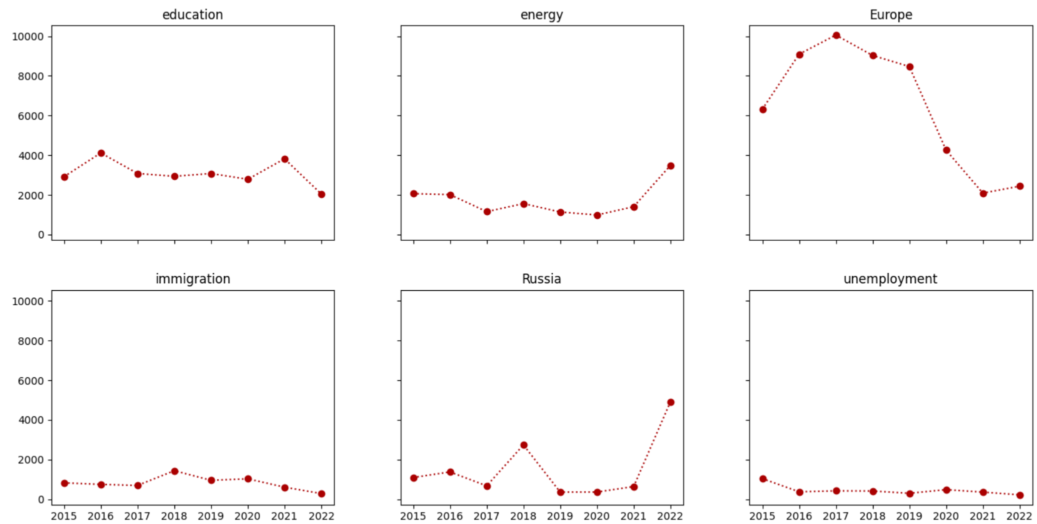

Figure 1: Number of times some selected topical words were pronounced by major parties in the UK's House of Commons from 2015 to 2022.

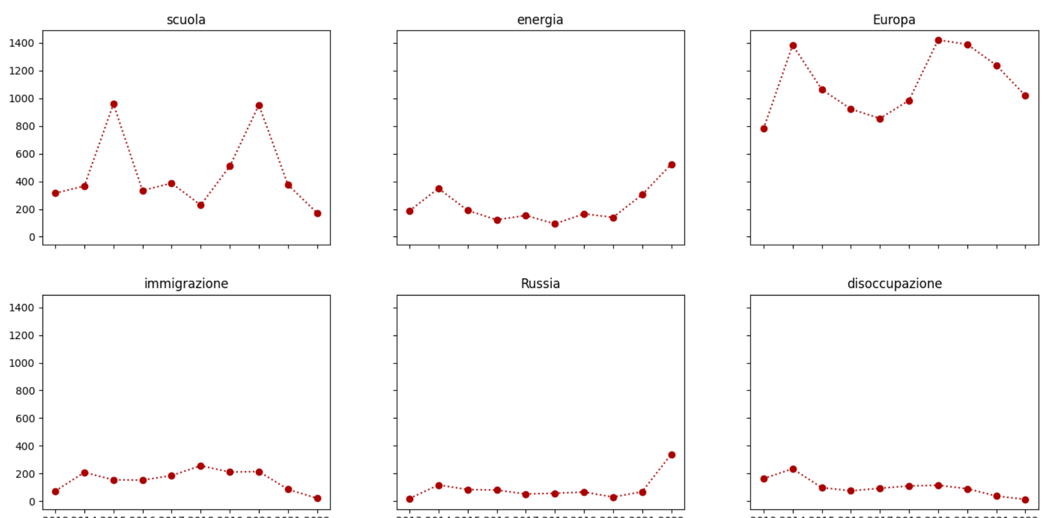

Figure 2: Number of times some selected topical words were pronounced by major parties in the Italian Senate from 2013 to 2022.

Resuming the comparison with broadcast media, it is the qualitative differences that are particularly incisive: the speech from broadcast news reports differs extremely from that of elected politicians discussing major issues. First, in the case of aligned broadcasting companies we can observe the phenomenon of *agenda setting* (McCombs and Shaw, 1972): the coverage of a news determines how much importance the public audience attaches to that event. For example, right-wing broadcasting companies may talk about immigration in the context of a crime committed by an immigrant, while the left-wing may typically focus on human rights in detention facilities. This does not happen in parliamentary debates: all parties contribute to the selection of the issues to be discussed, and all parties take part to the same discussion with one another. This significantly limits the variation of context in the different groups.

Another qualitative difference is in the register: parliamentary speech is often full of sophisticated rituals and grandiloquent expressions, which are definitely a minority in the training data of common language models. Critique, mockery, sarcasm, and other forms of reporting the adversary's speech can also prove tricky for a contextual embedding model, as they add words to the context that are actually not seriously meant by the current speaker.

3 Method

As mentioned previously, this study aims to reproduce Ding et al. (2023)'s results on parliamentary speech. While in their study there are two selected groups for Democrats and Republicans – the US-American broadcasting companies CNN and Fox News –, most countries do not have a two-party system, prompting a country-by-country analysis of each political landscape. In the case of the House of Commons, the two traditional parties – Conservative (CON) and Labour (LAB) – were selected for the analysis. As for the Italian Senate, a classification by the five major parties was adopted: Forza Italia (FI), Fratelli d'Italia (FdI), Lega (LN), Movimento 5 Stelle (M5S), Partito Democratico (PD). Each term, in Italy elected MPs form parliamentary groups according to their coalitions and shared political views. Generally, this results in major parties attracting "satellite" MPs from much smaller parties. For the sake of simplicity and continuity, in this study each parliamentary group was renamed after its major party. The party Lega was considered the successor of Lega Nord.

Next, a set of topical words was arbitrarily chosen, drawing inspiration from (Ding et al., 2023)'s original set. For the House of Commons these are *education*, *energy*, *Europe*, *immigration*, *Russia*, *unemployment*. The same ones, in Italian, were chosen for Italian Senate: *scuola*, *energia*, *Europa*, *immigrazione*, *Russia*, *disoccupazione*. Note that this approach can be problematic in a low-data setting, as these topics are often mentioned through synonyms (e.g., "migration"), periphrases or euphemisms (e.g., "migratory crisis"), or simply through morphologically related words (e.g., "immigrants"). The last point can prove especially challenging in morphologically rich languages such as Italian, and even lemmatisation does not provide a solution as the embeddings will be affected by the different morphosyntactic features (or even semantic in the case of words such as *energy* - *energies*). This results in a heavy filtering of the available data, including only those utterances that exactly match (except for capitalization) a word in the set.

After extracting the relevant utterances, contextual embeddings were computed for the topical words, and each embedding stored together with the year and the name of the group (i.e., political party of the speaker). To achieve this, the models `bert-base-uncased`[3] and `bert-base-italian-uncased`[4] were used as tokenizers and embedders for the English and Italian data respectively. A significant amount of words appeared to not be included in the vocabulary of the Italian tokenizer, possibly due to the high register of the discussions. For example, the phrase "l'apertura, per quanto cauta - diciamo pure timida - del Consiglio Europeo ad un riorientamento" was split into the following tokens: `l`, `'`, `apertura`, `,`, `per`, `quanto`, `cau`, `##ta`, `-`, `diciamo`, `pure`, `timi`, `##da`, `-`, `del`, `consiglio`, `europeo`, `ad`, `un`, `rio`, `##rie`, `##nta`, `##mento`; not recognising the words meaning *cautious* (feminine), *shy* (feminine), and *reorientation*. The embeddings used in this study are the sum of the last four layers obtained after feeding the sentences to the model.

Ding et al. (2023) define a semantic polarity measure SP as follows:

$$SP(A_t, B_t) = 1 - \frac{\sum_{w_i^a \in A_t, w_j^b \in B_t} \cos(w_i^a, w_j^b)}{|A_t||B_t|}$$

where:

t: a specific year (for diachronic analysis)
A_t: set of word embeddings from group A in year t
B_t: set of word embeddings from group B in year t
w_i^a: word embedding in A_t
w_j^b: word embedding in B_t

This is essentially one minus the average cosine distance between all pairs (w_i^a, w_j^b), the idea being that the farthest the embeddings in A are from the embeddings in B, the more polarised the two groups. This score ranges between 0 (not polarised) and 2 (completely polarised). However, an $SP > 1$ is only possible with a significant amount of negative cosines, but these are rare and often not easily interpretable in the case of word embeddings. As the embeddings compared here are from the same topical words – just used by different political groups – it is extremely unlikely to get even a single negative cosine. Therefore, we will observe only values of SP between 0

[3]huggingface.co/google-bert/bert-base-uncased
[4]huggingface.co/dbmdz/bert-base-italian-uncased

and 1. While this measure is useful for comparison (e.g., comparing SP in different years), it is not much meaningful to read it as an absolute value.

4 Results

If an effect of the political group is captured by the contextual embeddings, then a three-dimensional projection obtained through principal component analysis (PCA) should reveal some regions where the embeddings of each political group cluster together. These may be blurred and more or less overlapping, but if the effect is strong they should be clearly visible. However, no such clusters could be found in the PCA projections. As an example, Figure 3 shows the results of the PCA over the embeddings for the word *immigration* in the English and Italian data in year 2016. As you can see, the projections are labeled differently for each political group and there seem to be no distinct regions. The results were analogous across different years and topical words.

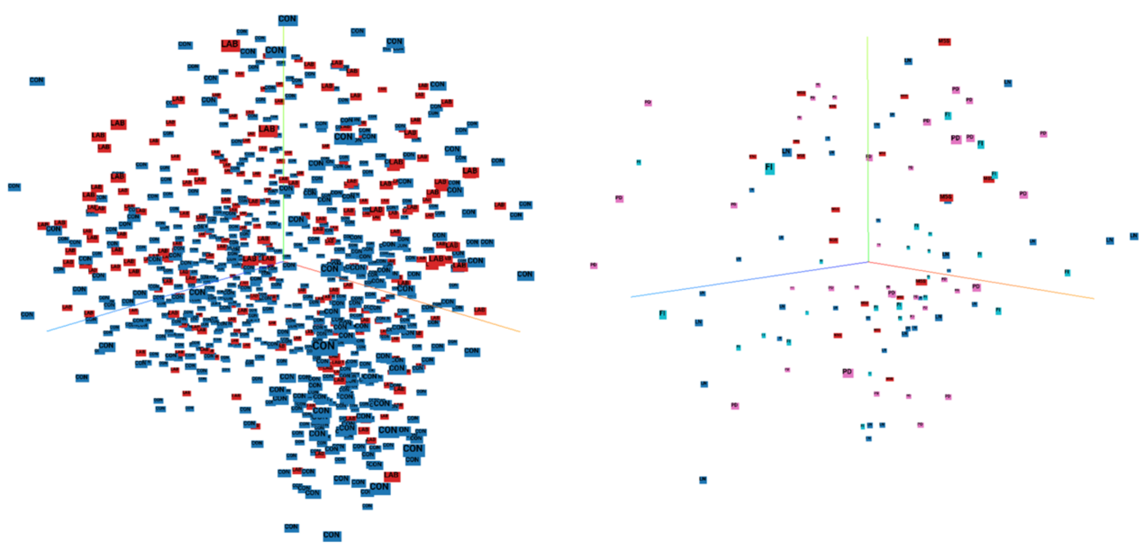

Figure 3: PCA projections of the contextual word embeddings for the words *immigration* and *immigrazione* pronounced respectively in UK's House of Commons (on the left) and Italy's Senate (on the right) in 2016. Each projection is labeled with the political group of the MP who uttered word. The scarcity of Italian data is likely due to the dataset problems mentioned in Section 2.

Projecting a multidimensional space to three dimensions comes at the cost of information loss. Several tests were conducted with three-component PCA. In all of these, PCA never described more than 30% of the variance. The semantic polarity measure defined in Section 3 provides us with a tool to truly check whether there is an effect of the political group on the distribution. The strategy adopted here is to compare the semantic polarity within a group A – $SP(A,A)$ – with the semantic polarities between A and any other group: $SP(A,B)$, $SP(A,C)$, $SP(A,D)$, etc. If group A is using the word differently from the other groups, then $SP(A,A)$ should be lower than the other semantic polarities. Geometrically, this is equivalent to what we were doing earlier searching visually for clusters in three dimensions: now we are measuring whether the embeddings in group A are closer to each other than to the embeddings from other groups. The difference is that we get a precise measure and we work on the multidimensional space, thus not suffering from any information loss.

While mathematically equivalent, it might not be very intuitive to talk about "polarisation" within a party. It is more meaningful to consider $SP(A,A)$ as a measure of how differently a certain word is used within group A. If $SP(A,A)$ is low, this means that the multidimensional "cloud" is more compact and thus the group's members are using the word in a similar fashion to each other. If $SP(A,A)$ is high, then we will observe a less concentrated, wider cloud, suggesting that the group's members are using the word in dissimilar ways. It is important to mention that the formula for SP within the same group was modified to exclude the distances between each embedding and itself, as in a scenario with few data points this distance – which is equal to 0 – can really drag down the average.

Figure 4 illustrates the semantic polarity results for the word *immigration* used by the main British political parties from 2015 to 2022. Figure 5 shows the same, mutatis mutandis, for the Italian case.

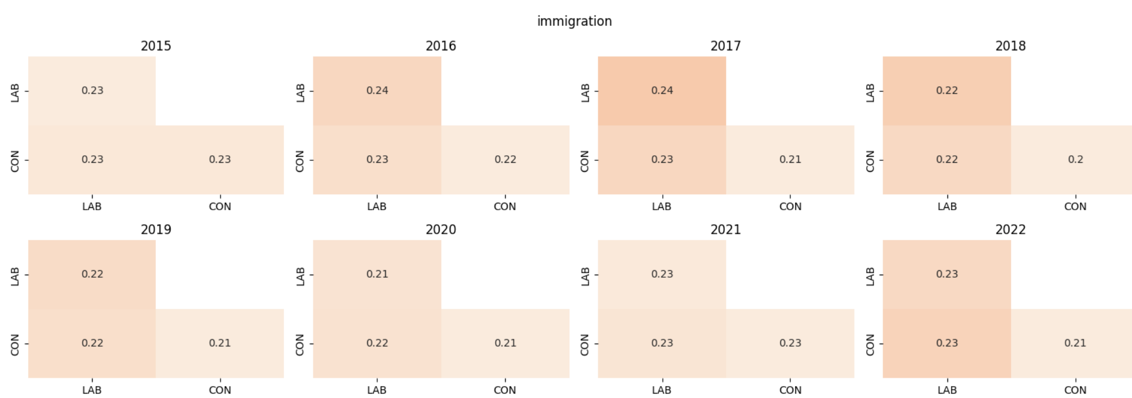

Figure 4: Semantic polarity values within and across major political parties for the use of the word *immigration* in the House of Commons from 2015 to 2022.

If the hypothesis that such contextual word embeddings are able to capture political polarisation were correct, we would observe paler colours along the diagonal (i.e., low "polarisation" values within the parties) and darker colours at the intersections of politically distant parties (i.e., high polarisation values). However, as you can see,

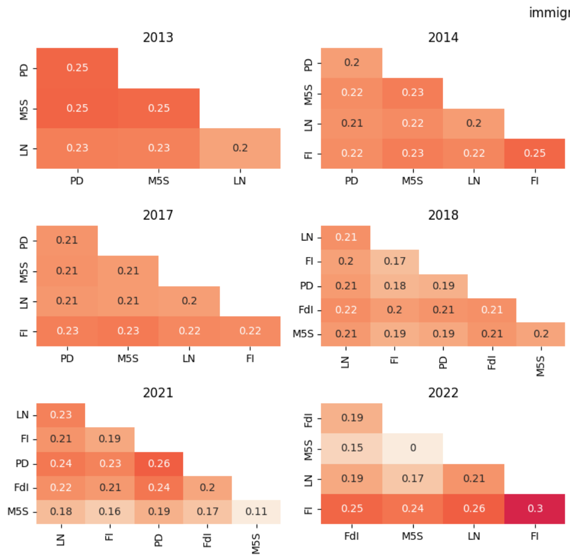

Figure 5: Semantic polarity values within and across major political parties for the use of the word *immigrazione* in the Italian Senate from 2013 to 2022.

no such patterns are recognisable in the matrices, with SP sometimes even peaking within-party. This suggests that there is not evident effect of political party captured by these contextual embeddings. The results for other topical words are analogous.

The slight fluctuations of SP are due to the fact that the contextual embeddings, albeit from different groups, are derived on the same word form (here, *immigration*). Stronger fluctuations in 2021 and 2022 in the Italian graph are due to fewer available data (see *immigrazione* in Figure 2), and thus less averaging-out of the cosine distances.

5 Discussion

6 Conclusion

7 Materials

References

Blumenstiel, J. E. and T. Plischke (2015). Changing motivations, time of the voting decision, and short-term volatility – the dynamics of voter heterogeneity. *Electoral Studies 37*, 28–40.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Ding, X., M. Horning, and E. H. Rho (2023, June). Same words, different meanings: Semantic polarization in broadcast media language forecasts polarity in online public discourse. *Proceedings of the International AAAI Conference on Web and Social Media 17*, 161–172.

Erjavec, T., M. Kopp, M. Ogrodniczuk, P. Osenova, M. Agirrezabal, T. Agnoloni, J. Aires, M. Albini, J. Alkorta, I. Antiba-Cartazo, E. Arrieta, M. Barcala, D. Bardanca, S. Barkarson, R. Bartolini, R. Battistoni, N. Bel, M. d. M. Bonet Ramos, M. Calzada Pérez, A. Cardoso, Ç. Çöltekin, M. Coole, R. Darģis, R. de Libano, G. Depoorter, S. Diwersy, R. Dodé, K. Fernandez, E. Fernández Rei, F. Frontini, M. Garcia, N. García Díaz, P. García Louzao, M. Gavriilidou, D. Gkoumas, I. Grigorov, V. Grigorova, D. Haltrup Hansen, M. Iruskieta, J. Jarlbrink, K. Jelencsik-Mátyus, B. Jongejan, N. Kahusk, M. Kirnbauer, A. Kryvenko, N. Ligeti-Nagy, N. Ljubešić, G. Luxardo, C. Magariños, M. Magnusson, C. Marchetti, M. Marx, K. Meden, A. Mendes, M. Mochtak, M. Mölder, S. Montemagni, C. Navarretta, B. Nitoń, F. M. Norén, A. Nwadukwe, M. Ojsteršek, A. Pančur, V. Papavassiliou, R. Pereira, M. Pérez Lago, S. Piperidis, H. Pirker, M. Pisani, H. v. d. Pol, P. Prokopidis, V. Quochi, P. Rayson, X. L. Regueira, A. Rii, M. Rudolf, M. Ruisi, P. Rupnik, D. Schopper, K. Simov, L. Sinikallio, J. Skubic, L. M. Tungland, J. Tuominen, R. van Heusden, Z. Varga, M. Vázquez Abuín, G. Venturi, A. Vidal Miguéns, K. Vider, A. Vivel Couso, A. I. Vladu, T. Wissik, V. Yrjänäinen, R. Zevallos, and D. Fišer (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1.

Lelkes, Y. (2016, 03). Mass Polarization: Manifestations and Measurements. *Public Opinion Quarterly 80*(S1), 392–410.

Mair, P. (2023). *Ruling the void: The hollowing of Western democracy*. Verso books.

McCombs, M. E. and D. L. Shaw (1972, 01). The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly 36*(2), 176–187.

Németh, R. (2023, 04). A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*.