uc3m | Universidad Carlos III de Madrid

Master's degree in Computational Social Sciences

2023-2024

Master Thesis

"NLP Based Analysis and Visualization of the House of Commons Parliamentary Debates (ParlVote)"

Laura Martínez Temiño

Tutor: Carmen Torrijos

Madrid, September 2024

Abstract:

The present work explores the political discourse surrounding Brexit by applying Natural Language Processing (NLP) techniques to observations from 2013 to 2019 of the ParlVote corpus—a dataset of UK House of Commons parliamentary speeches. Through sentiment analysis, topic modelling, and hate speech detection, the study identifies key themes and shifts in discourse that shaped public opinion leading up to and following the Brexit referendum. Although the analysis of speeches reveals a predominant neutrality in sentiment, pronounced negative peaks correlating with critical Brexit milestones can also be found reflecting the polarized nature of parliamentary debates. Topic modelling uncovers recurring issues such as national sovereignty, economic implications, and immigration, while hate speech detection highlights the prevalence of inflammatory rhetoric, especially in pro-Brexit constituencies. These findings offer a methodological framework for future political text analysis and contribute to understanding the influence of political discourse on major national decisions.

INDEX

FIGURE INDEX

1. INTRODUCTION

Brexit has been one of the most consequential political events in recent history and understanding it as an outcome can help us trace back the reasons behind this milestone. Particularly analysing parliamentary debates preceding and following the referendum of June 23, 2016, might allow to grasp the political discourses that shaped public opinions leading up to the Brexit vote.

This idea is supported by Van Dijk (1997) and Lewis (2001), who sustain that parliamentary debates serve as an important platform where political elites articulate their positions and influence societal attitudes. Furthermore, the rhetorical strategies employed in these debates often aim to resonate with public concerns and values, shaping public perception and opinion (Charteris-Black, 2018; Gibbons, 2007; Hauser, 1998). This dynamic has previously been object of study in England where evidence in favour of this impact was found eighteenth-century parliamentary discussions (Innes, 1990).

The UK Parliament comprises two main debating chambers: The House of Commons and The House of Lords. The former serves as the superior legislative chamber, drawing most of the public and media attention and forming the primary focus of this work. Therefore, by examining the discourse within the House of Commons the narratives and arguments that determine voter opinions will be traced, providing a clearer understanding of the motivations behind the referendum results. In this context, the ParlVote corpus emerges as an essential resource for carrying out the analysis, which will revolve around three main Natural Language Processing techniques: sentiment analysis of the parliamentary speeches for profiling the average Brexit voter, topic modelling of the speeches related to Brexit or EU matters to get a glimpse of the main concerns in the House of Commons regarding this topics and lastly hate speech detection to get insights of its use within a formal institutional setting as the UK parliament.

2. BACKGROUND

The computational analysis of political texts has gained significant attention in recent years, particularly the intersection between political science and natural language processing (NLP). Researchers have employed NLP methods to analyse policy positions, perform sentiment analysis, discern topics, and investigate stylistic elements of political discourse. In fact, Abercrombie & Batista-Navarro (2019) conducted a systematic review of 61 studies, highlighting the varied approaches employed to scrutinize opinions and positions in legislative debates and Gurciullo et al. (2015) also applied dynamic topic modelling and topological data analysis to debates within the UK House of Commons, revealing consistent roles for members and parties, as well as patterns of political cohesion. Expanding the scope of analysis, Miok et al. (2022) performed a multi-aspect, multilingual analysis across six European parliaments, examining emotions, sentiment, and speaker attributes and Glavas et al. (2019) studied policy positions, topics, and language style in political texts.

Moreover, the application of advanced NLP techniques, such as transformer models on political discourse analysis has shown promising results. Wankmuller (2021, 2022) demonstrated the superior performance of transformer-based models like BERT, RoBERTa, and Longformer in text analysis tasks compared to traditional machine learning algorithms, even when faced with limited training data. These advancements underline the powerful capabilities of NLP methods in analysing political discourse more accurately and effectively.

Despite these innovations, challenges persist in balancing the interpretability of results with the scalability of automated analyses. Rehbein (2024) and Sawhney et al. (2020) addressed these challenges by exploring semantic and pragmatic representations of political rhetoric, including framing strategies and people-centric messaging in populist discourse.

3. DATA

3.1 Data description

The main data source is the ParlVote corpus (*ParlVote_concat*), which is available in CSV format. This corpus was previously cleaned and prepared for NLP by Abercrombie and Batista-Navarro (2020) and compiles the transcripts of all parliamentary debates of the House of Commons (also known as Hansard), yet for this work only those between January 2013 and November 2019 were used. As the authors explain (ibis.), *"each debate in the House of Commons begins with a motion—a proposal made by an MP. Motions always begin with the words 'I beg to move, ...', which are followed by one or more statements. In response to the motion, MPs may speak, when invited by the Speaker (chief officer of the House), any number of times during a debate"*. Thus, it is structured around the interventions of each speaker regarding each motion that takes place within a given debate

This corpus, due to its design, facilitates the understanding of MPs' positions on critical topics and the emotional tone of their speeches, which is crucial for grasping the broader political climate surrounding Brexit (Abercrombie & Batista-Navarro, 2019). On top of the information contained within the ParlVote corpus, the data was merged to the list of all of the MPs of the UK parliament since 1997, which is available at the TheyWorkForYou (n.d) website, to get their respective constituencies, and later on to data of the Brexit referendum by constituency (House of Commons Library, 2017).

3.2 Data pre-processing.

As mentioned, the data was enriched by adding information about the constituency of each Member of Parliament (MP) from the "TheyWorkForYou" (n.d.) website and processed to ensure consistency across different periods and speakers. Specifically, discrepancies for MPs who changed constituencies following boundary adjustments were addressed to make sure that each speech record was accurately linked to the correct MP and its corresponding constituency. Moreover, each utterance (speech) was cleaned using

regular expressions by removing specific phrases and patterns to standardize the data such as "Orders of the Day — " or "Clause + number —" etc…

Once the utterances text was standardised, topics were assigned to each debate based on the content of the debate titles and motions. This categorization was done using predefined keyword lists associated with various topics like EU/Brexit, Education, Energy/Environment, Employment, Welfare etc... and applied with function that converted debate titles and motion texts to lowercase, checked if any keywords from these lists appeared in the debate titles or motion texts and then assigned the corresponding category all debates. This function was looped through each row of the dataset, and observations without a category were filtered out.

Finally, to run the Hugging Face (HF) models that will later be explained further pre-processing was done using python code. Text entries were converted to lowercase to maintain consistency and two functions were developed to process text based on its length: speeches shorter than 512 tokens[1] were directly tokenized using the respective "*AutoTokenizer"* from the HF library and fed directly into their "*AutoModelForSequenceClassification"* model, while longer speeches were tokenised, split into 512-token chunks, embedded and processed independently, and then the results were aggregated to generate a final label or score. After pre-processing, each model, its tokenizer and configuration were loaded, and the pre-processed text was passed through the models ensuring that the text data conformed to the tokenization limits imposed by them.

This comprehensive pre-processing pipeline not only cleaned and standardized the data but also enriched it by adding relevant metadata, such as constituency information, topic categories and labels, preparing it for in-depth analysis.

[1] Token refers to either words (unigrams), bigrams (sets of two words) or trigrams (sets of three words).

4. METHODOLOGY AND MODELS [2]

4.1 Sentiment analysis on speeches and Brexit vote.

After testing different models for performing sentiment analysis on the parliamentary debates the *twitter-roberta-base-sentiment-latest* provided by CardiffNLP, was chosen. This RoBERTa model is an optimized version of BERT (Bidirectional Encoder Representations from Transformers), designed to handle a wide array of natural language processing tasks, including sentiment analysis. Its architecture benefits from extensive pre-training on large and diverse text corpora, enabling it to capture subtle nuances in language that are often present in parliamentary discourse. In fact, even though there are variants of RoBERTa fine-tuned on political texts, such as political tweets, when tested on interventions from MPs, the general RoBERTa model provided more accurate sentiment classifications compared to other variants. This can be attributed to the broader range of training data, which allows the model to generalize better to the varied subjects of parliamentary speeches (Cardiff University NLP., 2020).

The output of the model comprises a sentiment label—positive, negative, or neutral—along with a corresponding confidence score for each speech of the corpus. It is important to note that while the general RoBERTa model excels in versatility, the sentiment analysis may not be as finely tuned to the specific rhetorical and contextual subtleties inherent in political discourse. Regardless, ensures robust sentiment analysis across diverse topics and styles, making it well-suited for the House of Commons debates.

4.2 Topic modelling on “EU/Brexit” speeches.

To identify and analyse the main themes treated in the parliamentary speeches, topic modelling was conducted using two methods: BERTopic and KMeans clustering.

4.2.1 BERTopic

[2] The complete code and datasets used for the analysis are publicly available on Github and can be accessed at [https://github.com/laura-martinez00/NLP_HouseOfCommons_debates]. This repository contains detailed instructions on how to replicate the experiments and analyses described in this paper.

For the task of topic modelling, the *bert-base-uncased* model developed by Google (available in HF) was employed following the tutorial provided by Plain English (2023) to generate embeddings and clusters for the text data. This model is a pre-trained transformer model based on the BERT architecture that processes text accounting for both the left and right context of each word, capturing contextual relationships in text (Devlin, J et. Al, 2019).

Understanding the contexts of word representations through a bidirectional approach is essential for capturing the complex uses of language in parliamentary debates, as well as for extracting coherent and meaningful topics from the interventions, even when these topics are implicit or span multiple sentences. The model converts text into numerical vectors (embeddings) that capture the semantic meaning of the words, reduces them in dimensionality using UMAP (Uniform Manifold Approximation and Projection), and clusters them using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

The choice of the "Bert-base-uncased" model was driven by its capacity to produce high-quality embeddings that are well-suited for topic modelling tasks across large corpuses such as parliamentary debates. Its uncased variant treats words as case-insensitive, which is appropriate given the formal nature of parliamentary transcripts where capitalization does not typically alter meaning. The resulting topics provided insights on the primary issues discussed and allowed for a deeper understanding of the content and direction of the debates.

4.2.2 KMeans with TF-IDF Vectorization

As a comparative approach, KMeans clustering was used combined with TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This traditional method involves transforming text into numerical vectors using TF-IDF, which captures their importance within the corpus. Vectors are then clustered into pre-specified topics using the KMeans algorithm, which iteratively groups similar texts by minimizing the distance between them and their respective cluster centres. While less sophisticated than BERTopic, this method provided a straightforward and interpretable set of topics, effectively capturing distinct themes within the speeches.

4.3 Hate speech analysis on “EU/Brexit” speeches.

Lastly, in order to identify and analyse hate and offensive language within parliamentary debates, the *unhcr/hatespeech-detection model* (hosted on HF) was used. This model, developed by the United Nations High Commissioner for Refugees (UNHCR), is specifically trained to detect and classify hate speech in text. It is based on the RoBERTa-uncased transformer architecture, standard in natural language processing (NLP) due to its ability to handle complex language patterns and long-range dependencies in text. It is also fine-tuned on a dataset that includes various forms of hate speech, making the model capable of recognizing a broad spectrum of hateful expressions, even subtle or context-dependent ones.

It works by processing input text and converting it into embeddings that capture the semantic content and context of the words. These embeddings are then used to classify the content into categories, such as 'hate speech' or 'non-hate speech,' based on patterns learned during training. This approach enables the identification and understanding of harmful language in parliamentary discourse, as well as the underlying patterns in its usage.

5. RESULTS

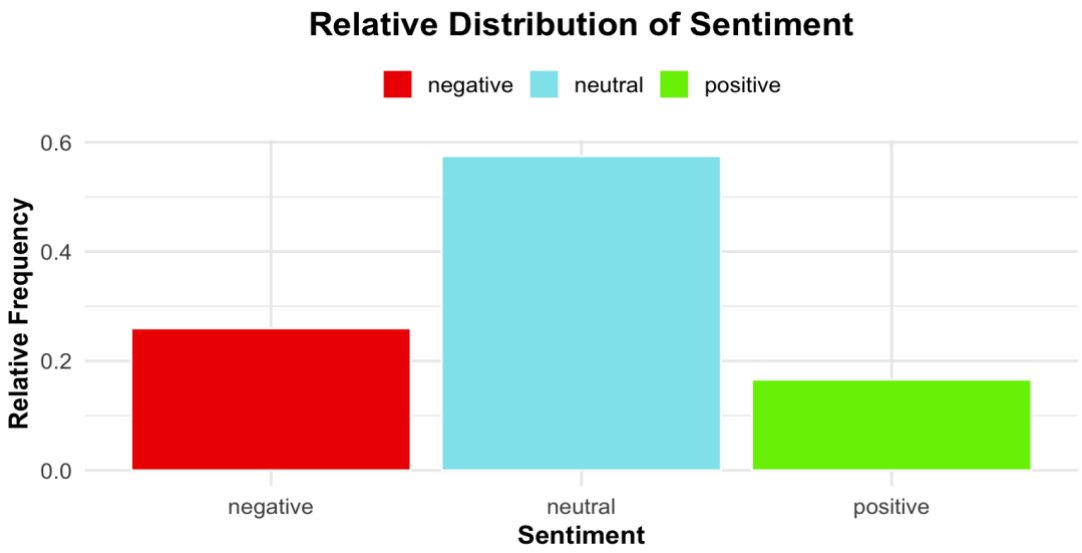
In this section, the results of additional analyses conducted on the databases following the application of NLP methodologies will be presented. These analyses included different approaches to classifying speeches within the corpus or its subsets.

5.1 Sentiment analysis:

Sentiment analysis involves the automatic detection of polarity of opinions expressed in text. In the context of political debates, this technique can reveal how Members of Parliament´s sentiments towards Brexit-related motions are reflected in their speeches, helping to identify patterns among pro-Brexit and anti-Brexit MPs. This section examines the potential correlation between certain sentiments and Brexit voting behaviour. The central research question investigates whether the sentiment conveyed in the speeches of MPs shaped the opinions and voting behaviour of their constituents during the Brexit referendum.

5.1.1 General Sentiment distribution

After running the Hugging Face sentiment classification model on the corpus, the sentiment distribution of the speeches revealed a predominantly neutral distribution with a slight skew towards negativity, reflecting the contentious and procedural nature of Brexit debates leading up to the referendum.



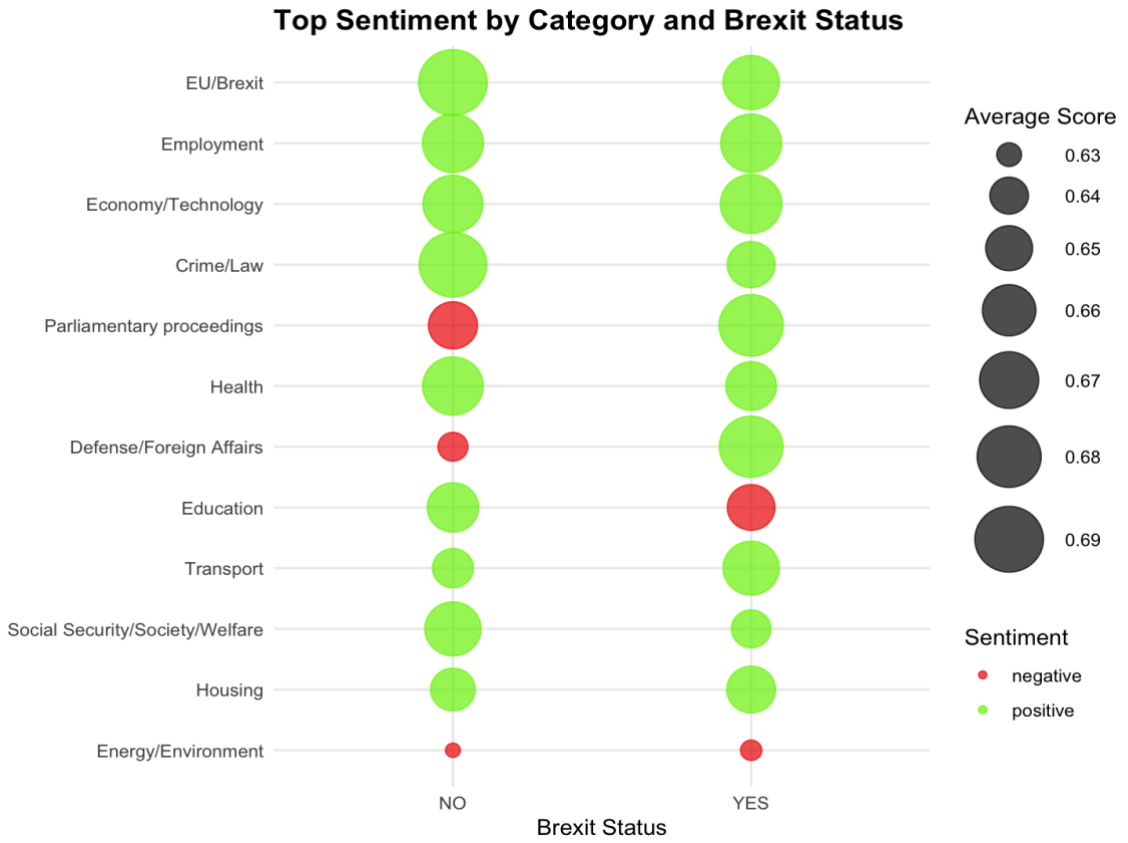
Plot 1: *Relative distribution of sentiment of the House of Commons speeches.*

Also, the general distribution of sentiment through time was plotted to visualize times in which speeches were more negative or positive (Annex).

The share of both positive and negative speeches is quite similar with constant peaks through time but the period of mid to late 2015 stands out, where the speeches became more polarized with less neutral sentiment. This corresponds with the intense political climate leading up to the Brexit referendum, suggesting that MPs were more emotionally charged as they debated the future of the UK's relationship with the EU.

5.1.2 Pro-Brexit and Anti-Brexit sentiment profiles.

To get a better picture of the general sentiment associated to the different speeches the average sentiment for each of the topics previously assigned was analysed using pro-Brexit and anti-Brexit sentiment profiles. In order to do so the sentiment of the speeches was grouped by topic, sentiment and Brexit outcome, calculated the average score, chose the sentiment with the highest score and then plotted the results.

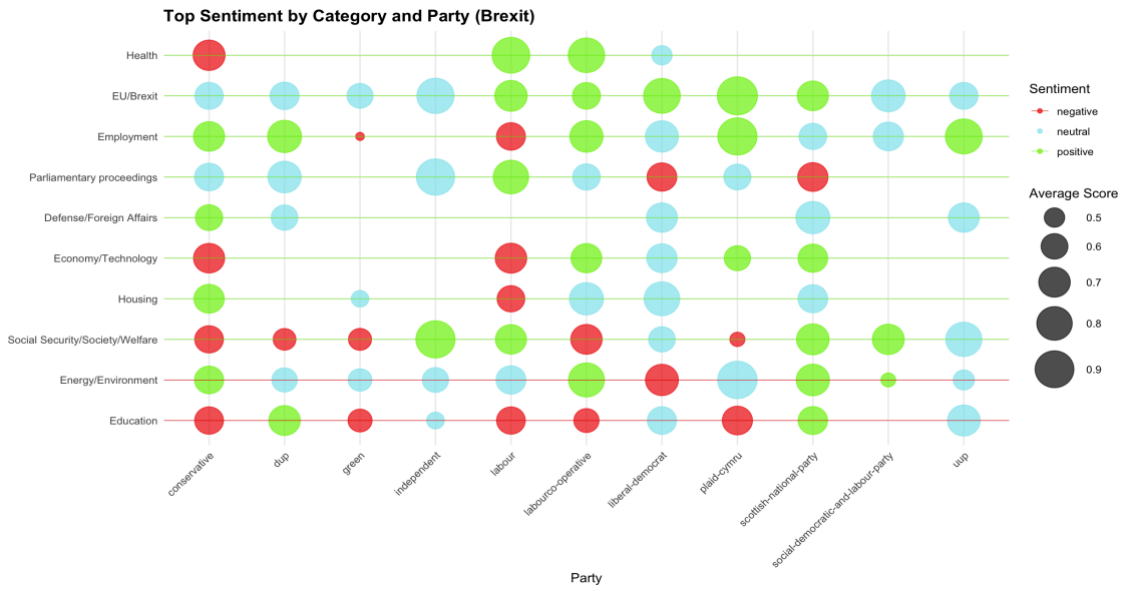


Plot 2: *Predominant sentiment for each topic by Brexit profile.*

As observed, the topics where sentiment differs most between constituencies that supported Brexit and those that did not are parliamentary proceedings, defence/foreign affairs, and education. Parliamentary proceedings may indirectly relate to debates or laws concerning EU matters. Defence and foreign affairs are closely tied to the UK's relationship with the rest of the world and the European Union, reflecting the focus on sovereignty and national pride. Education topics tended to be more negatively perceived in anti-Brexit areas; it is a sector in which UK universities are notably recognized for attracting top students globally and might reflect concerns over the impact of Brexit on academic collaboration and student mobility.

To delve deeper, the stance of UK Parliament parties on key topics in constituencies that voted either for or against Brexit was examined. The data was filtered to include only records from 2016, the year of the Brexit referendum to keep consistency between the parties' positions and the Referendum outcomes. This analysis presents an additional grouping by party and consistency in ideas and positions within them was assumed. The general pro-Brexit and anti-Brexit sentiment profiles were used as a baseline for the Y-axis lines corresponding to each topic. As a result, two plots were generated: one for

constituencies that voted **'NO'** to Brexit and another for those that voted **'YES**.' These plots provide a detailed understanding of how different parties and topics were perceived in constituencies, depending on the Brexit outcome (anti-Brexit profile in the Annex).



Plot 3: *Most frequent sentiment by party and debate category (Pro-Brexit MPs).*

The plots reveal that in pro-Brexit constituencies, the prevailing sentiment among parties was generally positive, especially regarding employment, defence, foreign affairs, housing, and EU/Brexit discussions. Particularly, the Conservative Party's positive sentiment on defence and foreign affairs aligns with the pro-Brexit narrative of enhancing UK sovereignty and global standing. This suggests confidence in Brexit's potential benefits, such as job creation and stronger national security.

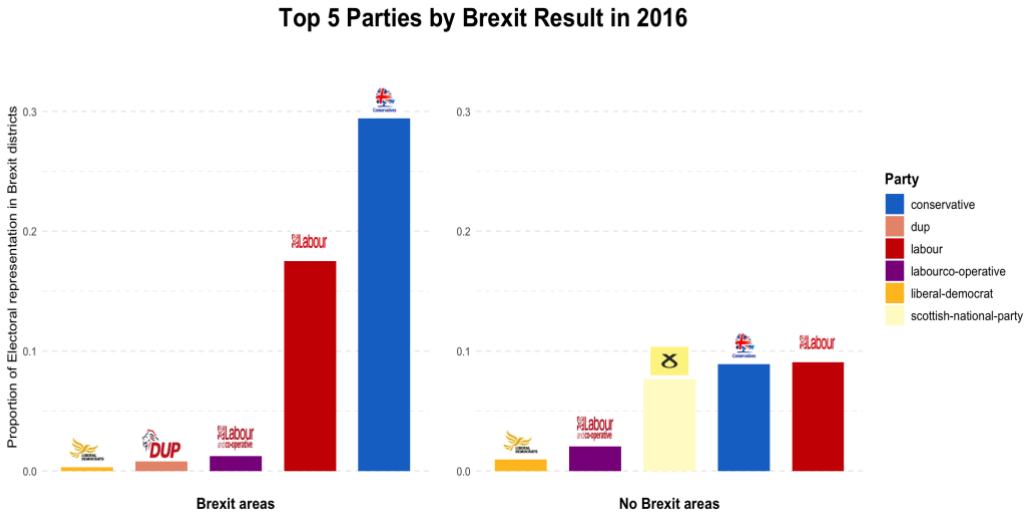
However, concerns were also expressed, particularly in education, technology, and economic policies, where both Labour and Conservative parties showed negative sentiments. These findings indicate that while pro-Brexit constituencies were generally hopeful, significant apprehension about potential sector-specific disruptions persisted.

Conversely, in constituencies that opposed Brexit, the sentiment was more cautious or negative, especially concerning social welfare, environmental/energy issues and defence/foreign affairs. Parties like Labour and the Liberal Democrats expressed ongoing concerns about Brexit's impact on social and economic stability and the UK's global standing. Despite these worries, there was pragmatic acceptance in some areas, with positive sentiments towards employment and economic technology even in anti-Brexit constituencies. This complex sentiment landscape highlights the deep divisions in the UK

regarding Brexit's effects, reflecting the challenges faced by UK parliamentary representatives in balancing national interests with regional and party-specific priorities.

5.1.3 Parties in power in 2016 by Brexit result

To enhance the previous analysis, the dominant party in areas that voted for Brexit versus those that did not was identified. The data was grouped by Brexit results, calculating the relative frequency of the top five parties in the House of Commons based on their Brexit stance. The results were then plotted using relative frequency.



Plot 4: *Top five parties governing in Pro-Brexit and Anti-Brexit areas.*

Results show how the Conservative Party, which strongly supported Brexit, holds a dominant position in areas where the referendum resulted in a majority vote to leave. Conversely, the Labour Party, which exhibited internal divisions regarding Brexit, shows significant representation in both Brexit and non-Brexit areas, although it is more prevalent in the latter. This distribution underscores Labour's attempt to balance its appeal across constituencies with differing views on Brexit.

A particularly notable finding is the strong representation of the Scottish National Party (SNP) in areas that opposed Brexit, where they account for 7.67% of constituencies. This is consistent with Scotland's overall preference to remain in the EU, where 62% of voters supported remaining (Toszek, B.H., 2020). The SNP's unequivocal anti-Brexit stance resonated with the Scottish electorate, solidifying its position as the predominant political force in Scotland. This reflects broader regional distinctions within the UK, where

national identity and the prospect of Scottish independence played a significant role in the referendum (Henderson et al., 2017).

5.1.4 Quantitative analyses

To investigate the relationship between political party affiliation, the sentiment expressed in parliamentary speeches, and the subsequent impact on Brexit voting outcomes an analysis of several stages will be performed:

Single-Factor ANOVA:

ANOVA Summary Table

Term	Degrees of Freedom	Sum of Squares	Mean Square	F Value	p-value	Signif. Codes[1]
party	10	1.135	0.11348	19.96	<0.001	***
Residuals	196	1.114	0.00569	NA	NA	

[1] Significance codes: *** p < 0.001

Table 1: *Single-Factor ANOVA (party).*

An initial one-way ANOVA was conducted to determine whether the party of the MP had a significant effect on the average leave percentage within constituencies. The results *(F(10, 196) = 19.96, p < 0.001)* confirmed that the political party of the MP significantly influenced the Brexit outcome in their respective constituencies. This suggests that party alignment played a critical role in shaping the electorate's stance on Brexit.

Multi-Way ANOVA:

Multi-Way ANOVA Summary Table

Term	Degrees of Freedom	Sum of Squares	Mean Square	F Value	p-value	Signif. Codes[1]
party	10	1.135	0.11348	8.35	<0.001	***
sentiment	2	0.000	0.00000	0.00	1	
category	11	0.000	0.00000	0.00	1	
party:sentiment	19	0.000	0.00000	0.00	1	
party:category	82	0.000	0.00000	0.00	1	
Residuals	82	1.114	0.01359	NA	NA	

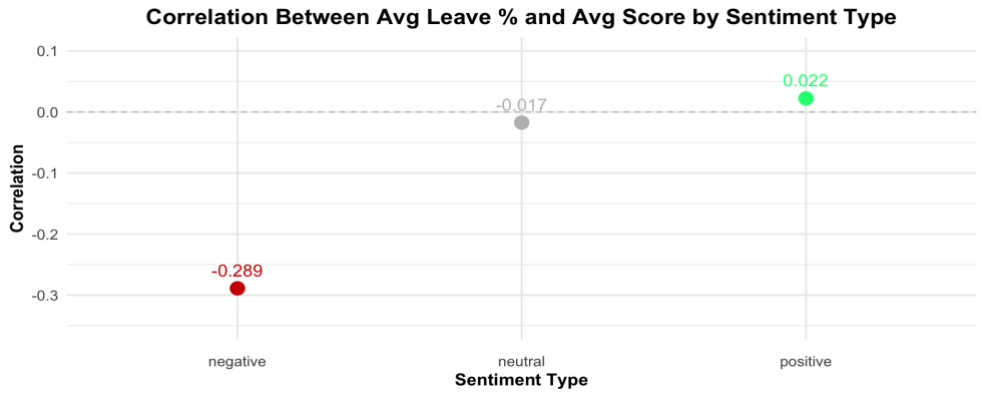
[1] Significance codes: *** p < 0.001

Table 2: *Multi-Way ANOVA (party*category*sentiment).*

Subsequently, a multi-way ANOVA was performed to assess the interaction effects of party affiliation, sentiment type, and category on the average leave percentage. While the

party remained a significant factor *(F(10, 82) = 8.35, p < 0.001)*, neither the sentiment nor the category or their interactions, showed significant effects. This indicates that while the political party of the MP is a decisive factor, the specific sentiments or topics addressed in their speeches do not independently alter the average leave percentage.

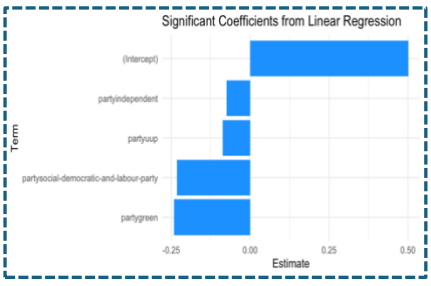
Correlation Analysis:

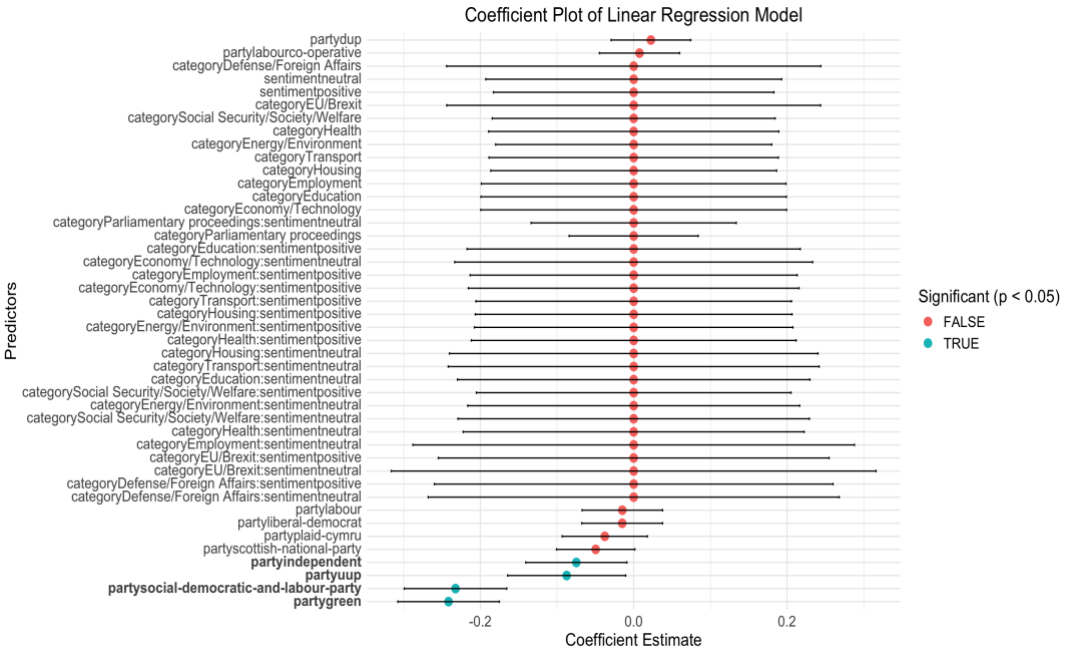


Plot 5: *Correlation between the average share of "leave" votes and sentiment type.*

Further analysis tested involved calculating the correlation between different sentiment types and the average leave percentage even when the previous analysis showed this relationship to be insignificant. Yet although no causation was found, the correlation analysis revealed a negative correlation for negative sentiment, suggesting that constituencies where MPs expressed more negative sentiments had lower leave percentages.

Linear Regression Analysis:





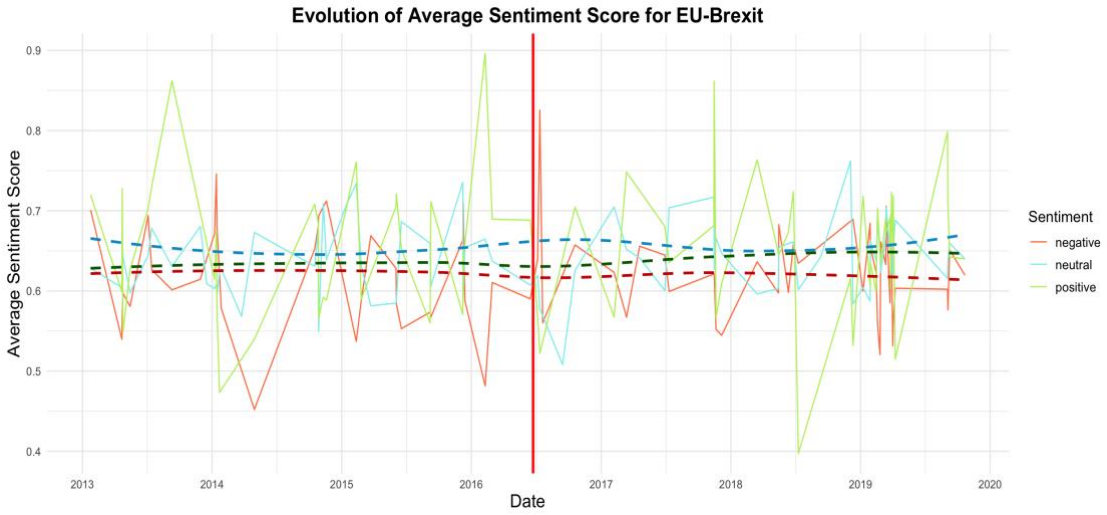
Plot 6: *Linear regression analysis of the effect of party, sentiment and category on voted "leave" share.*

Finally, a linear regression model was employed to confirm the significance of the effects of party affiliation and the interaction between sentiment and category on the average leave percentage. The model identified significant negative coefficients only for certain parties: for the Green Party, Independent Party, Social Democratic and Labour Party, and UUP, associating them with lower leave percentages. Results also suggest that while sentiment and topic interactions were not independently significant, the overall messaging and party identity were crucial in shaping the electorate's stance on Brexit. On top of this, the **GitHub** [3] repository contains residual plots (Q-Q and Cook's Distance) that confirm that the assumptions of linear regression are met.

5.1.5. EU/Brexit Speeches' sentiment evolution

Lastly, to visualize sentiment trends regarding Brexit-related discourse from 2013 to 2019, the **sentiment evolution of EU/Brexit** debates through time was plotted:

[3] *https://github.com/laura-martinez00/NLP_HouseOfCommons_debates*



Plot 8: *Sentiment evolution of EU/Brexit debates.*

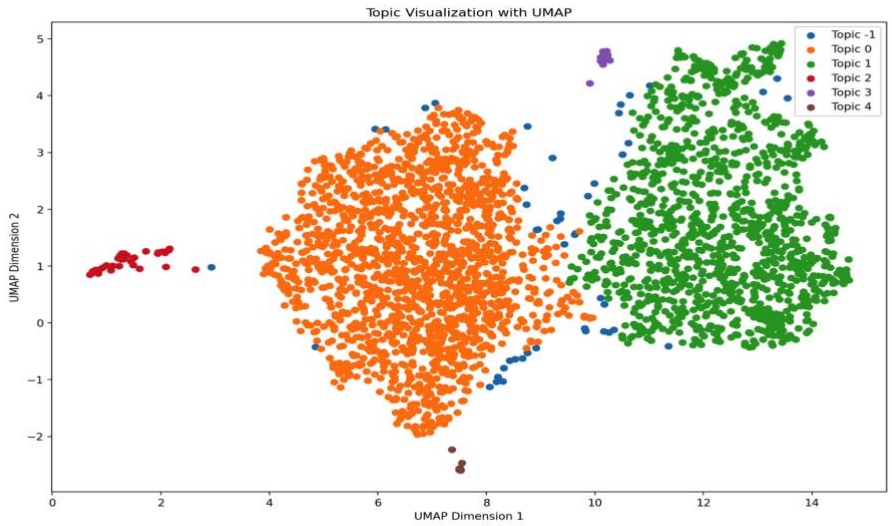
The graph shows an increase in negative sentiment leading up to and following the Brexit referendum, with a notable cluster of negatively classified speeches around the time of the June 2016 referendum. This reflects the intense and often uncertain and contentious nature of the debates as MPs grappled with the implications of the referendum outcome (Gavin, N.T., 2018). Post-referendum, the stabilization of sentiment suggests that as the UK government moved towards implementing Brexit, the discourse may have become more focused on pragmatic concerns, such as negotiating the terms of exit and managing the economic and social fallout. This aligns with the literature indicating a shift in public and political focus from ideological debates to practical considerations following the referendum (Hobolt, 2016).

5.2 Topic modelling on Brexit/EU topics.

Topic modelling is a method for identifying themes and subjects within corpus of text corresponding to “EU/Brexit” topics. In the context of Brexit, it can help categorize the various arguments and concerns raised by MPs during parliamentary debates. Thus, by implementing topic modelling on the ParlVote corpus, major themes discussed in the debates can be mapped and how these themes evolved over time tracked, offering insights into the shifting focus of parliamentary discussions on Brexit. As mentioned, two different methodologies were implemented and both outcomes will be shown:

5.2.1 BERTopic

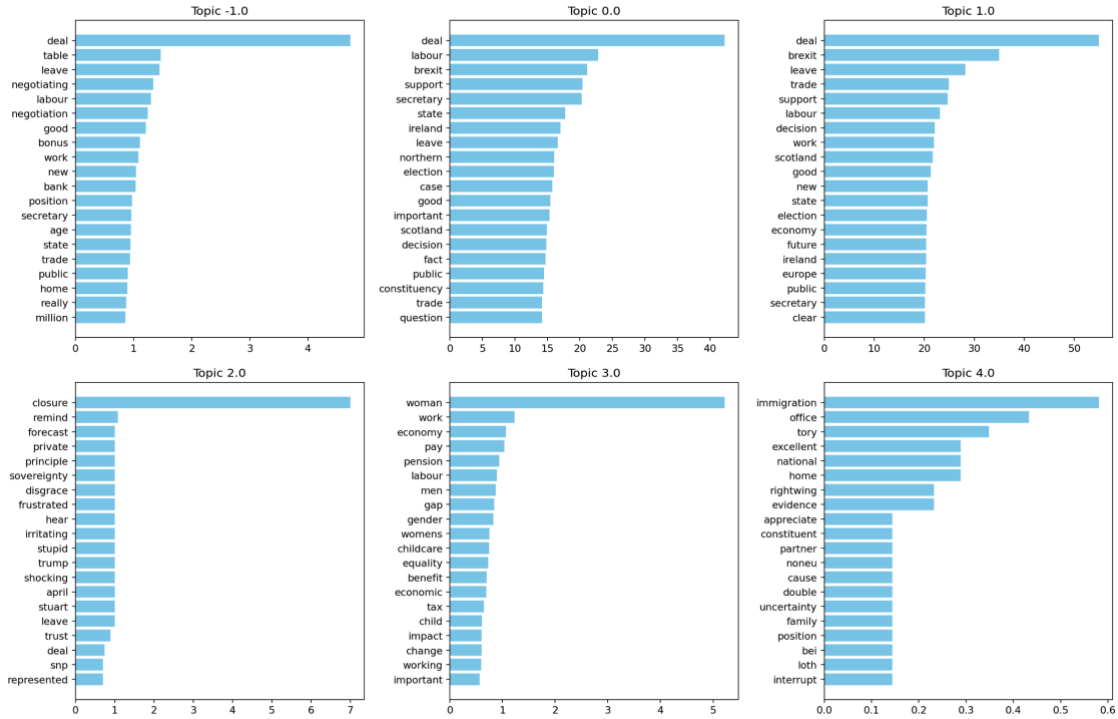
The following UMAP plot displays the topic clusters identified by BERTopic, where each point represents a speech:



Plot 8: *BERTopic topic clusters*

At first glance two well-defined yet quite similar clusters can be recognised (Topics 0 and 1) and then five other smaller ones, from which Topic -1 represents outliers. BERTopic, by default, uses the HDBSCAN algorithm, which labels some points as noise if they don't fit well into any cluster. The topics overlap may reflect the intertwined nature of Brexit political discussions, where similar themes are present across different debates, while topics 2 (red) and 4 (brown) are more isolated, showing discussions that are less connected to the broader debates.

Also, the top 20 words of each identified topic were displayed on a plot. Bear in mind that the topics identified by BERTopic do not necessarily match the topics in the original database, especially if there is a semantic or structural difference in how topics are represented in the embeddings. The words assigned to each topic were the following:



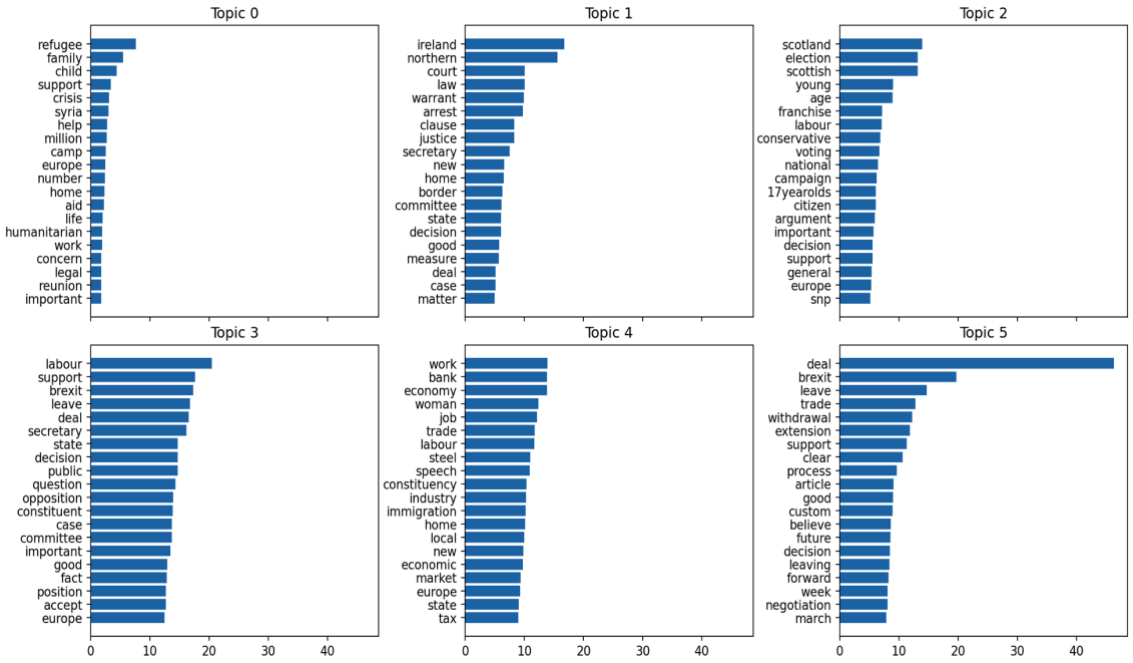
Plot 9: *BERTopic top 20 words per topic.*

Topics 0 and 1 correspond to the two big clusters, which share a significant thematic overlap. Topic 0, characterized by terms such as "deal," "labour," "Brexit," and "support," likely captures political discussions, party positions and parliamentary strategies while Topic 1, captures terms like "Brexit," "trade," "leave," and "economy" and seems to encapsulate broader discussions on the Brexit process, negotiations, and the economic implications thereof.

In contrast, Topics 2 and 4, represented by more isolated clusters on the UMAP plot, highlight specialized discussions within the Brexit debate. Topic 2, with keywords such as "closure," "remind," and "forecast," appears to land into specific economic or political matters, possibly reflecting debates on the outcomes and strategies related to Brexit. Topic 4, defined by terms like "immigration," "office," and "tory," suggests focused discussions on immigration policies and their alignment with Conservative Party (Tory) stances, which were critical components of the Brexit narrative. Lastly, Topic 3 appears to deal with gender and economic equality issues, with terms like "woman," "work," "pay," "pension," and "labour" being prominent. This could reflect discussions on how Brexit impacts women and work-related economic factors.

5.2.2 KMeans

As an alternative, **KMeans clustering** with TF-IDF vectorization was also used. This approach, while less sophisticated than BERTopic, provided a straightforward interpretable method for identifying six topics within the Brexit speeches:



Plot 10: *KMeans top 20 words per topic.*

Topic 0, with terms such as "refugee", "family", "child", "support", "Syria", "camp", and "crisis" signals a focus on humanitarian concerns, particularly related to refugees and asylum seekers. This also touches on international crises and the UK's role and legal responsibility of providing aid and support during Brexit.

Topic 1 is characterized by terms such as "Ireland", "northern", "court", "law", and "justice", pointing to discussions about the enforcement of laws, the justice system, and the unique challenges posed by maintaining legal integrity while addressing the concerns specific to the Northern Ireland border in the context of Brexit.

In **Topic 2,** the key terms "Scotland", "election", "Scottish", "young" and "voting" suggest a focus on Scottish independence and and the debate over voting rights for younger citizens, particularly those aged 17. This topic likely reflects the tensions between Scottish nationalism and the broader UK political structure, as well as the potential for another referendum on Scottish independence in the wake of Brexit.

Topic 3 presents terms such as "labour", "support", "brexit", "leave", and "secretary", indicating discussions related to parliamentary procedures, political strategy, the positions of key political parties, especially the Labour Party and how different political factions and committees navigated the Brexit process.

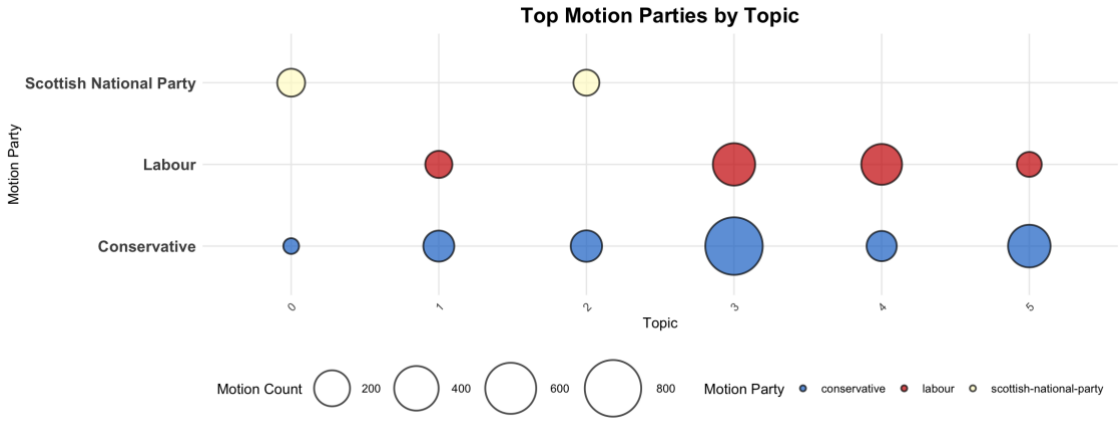
The terms "work", "bank", "economy", "tax", "woman", and "back" in **Topic 4** indicate a focus on the economic implications of Brexit, particularly related to labour markets, trade relations, banking, economic growth, and the broader economy. This topic likely captures debates about the role of women in the workforce, and the strategies needed to mitigate economic risks.

In **Topic 5** terms such as "deal", "Brexit", "leave", "trade", and "withdrawal", clearly point to discussions about the specifics of the Brexit deal and the withdrawal process including negotiations, legal and procedural steps, the withdrawal agreement, and the future relationship between the UK and the EU.

The comparison between these two methods highlighted several key differences. BERTopic, with its use of transformer-based embeddings effectively captured the text's semantic structure and dynamically determined the number of topics based on the data but produced overlapping topics due to speech homogeneity (Brexit category). In contrast, the KMeans with TF-IDF approach simpler and faster, offering clearer results, though it required predefined topic numbers.

5.2.3 Top Party starting motions related to Brexit.

Using the topics detected with the KMeans method, an analysis of the initiation of parliamentary motions reveals a clear dominance of the Conservative Party and the Labour Party in shaping the discussions in the House of Commons related to Brexit matters and especially parliamentary procedures (Topic 3). Notably, the Scottish National Party (SNP) emerges as a significant actor in the humanitarian and refugee issues (Topic 0), and in the debates surrounding Scottish independence and voting rights (Topic 2), as it probably led many motions concerning Scotland's political future post-Brexit.



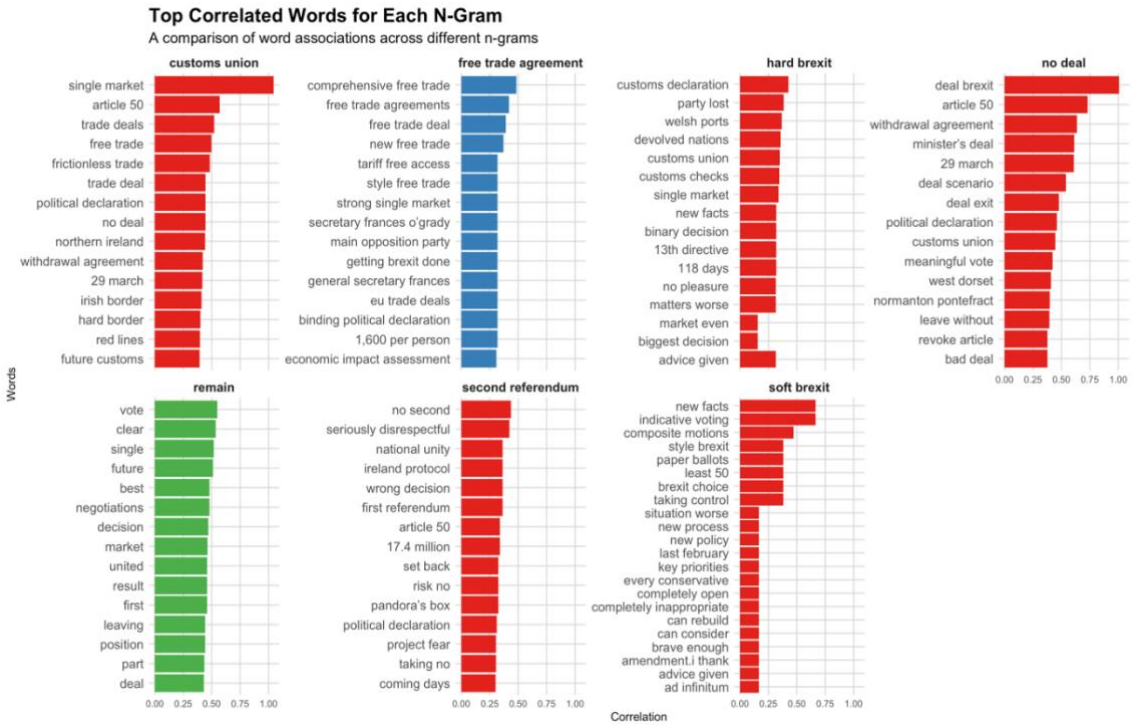
Plot 12: *Party that initiated most motions (debates) by topic.*

5.2.4 Word correlations

Referendum possible outcomes

In the wake of the 2016 referendum the UK could take in its exit from the EU. To better understand the discourse surrounding the possible scenarios—such as "Hard Brexit," "Soft Brexit," "No-Deal Brexit", “Customs Union”, “No deal”, “Free Trade Agreement”, “Second Referendum” and "Remain"—an analysis was conducted to identify the most correlated words and phrases associated with each of the different outcomes to the referendum.

The data was first pre-processed as follows: first filtered with a custom list of stop words and n-grams, then tokenized into unigrams, bigrams and trigrams. The n-grams with lowest frequency were filtered out and pairwise correlations between words within the same speeches were calculated to identify strongly associated terms before plotting the results:



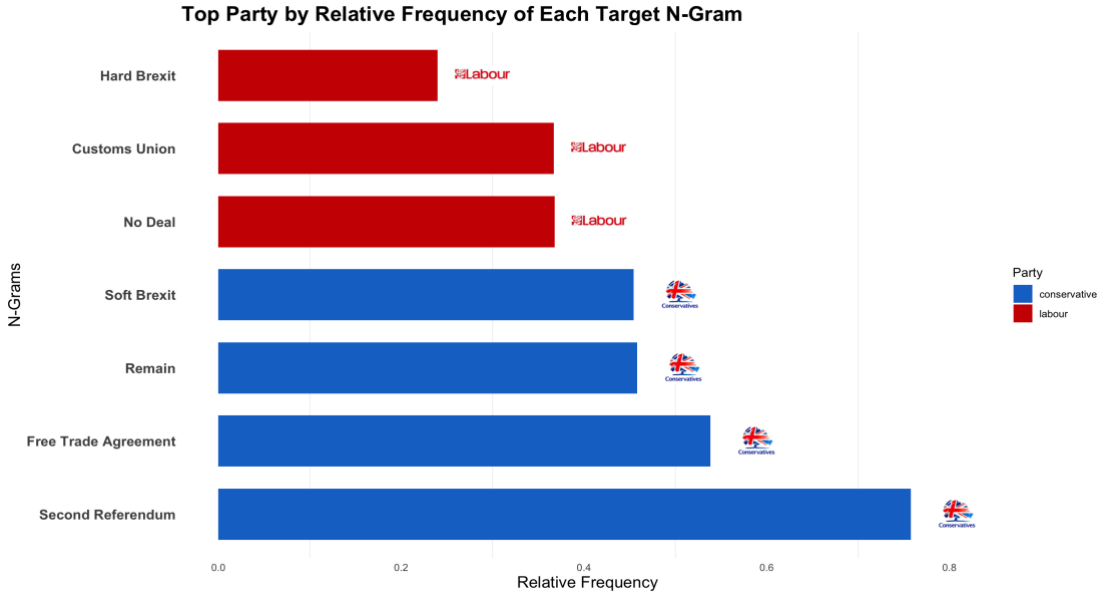
Plot 12: *Top n-grams correlated with each type of referendum outcome.*

The "customs union" scenario is associated with terms like "single market," "Northern Ireland," and "frictionless trade," emphasizing a desire to maintain close economic ties with the EU while avoiding a hard border in Ireland. However, it involves a break, as outlined in Article 50 of the Treaty on European Union (TEU), which details the legal process for a member state to withdraw. In contrast, the "free trade agreement" scenario focuses on "comprehensive free trade" and "tariff free access," aiming to secure favourable trade terms without staying in the customs union or single market.

The "hard Brexit" scenario, with words like "customs declaration" and "binary decision," reveals the logistical challenges and political divisions that imply leaving the EU without agreements in place. The "no deal" scenario is closely tied to legal and procedural terms such as "withdrawal agreement" and "article 50", stressing the uncertainty and high stakes of leaving the EU without a formal agreement. The "remain" scenario, associated with words like "vote," "future", and "negotiations", captures the arguments for staying within the EU, highlighting the benefits of market access and political stability. Meanwhile, the "second referendum" scenario reflects the divisive debate over revisiting the Brexit decision, with terms like "national unity" and "wrong decision" indicating concerns about potential risks and repercussions. Finally, the "soft Brexit" scenario, with words like

"indicative voting" and "composite motions", suggests a focus on procedural compromises to minimize disruption while fulfilling the Brexit mandate.

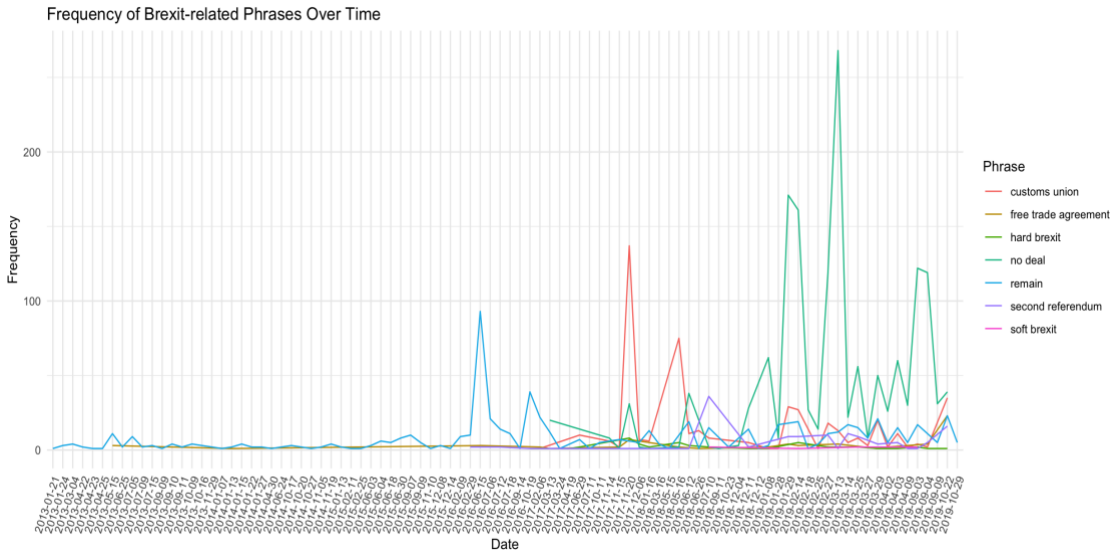
On a further analysis, the **frequency and usage of unigrams, bigrams, and trigrams related to Brexit** was analysed across different political parties, aiming to uncover how each party framed and discussed key issues during parliamentary debates:



Plot 13: *Party that most frequently used each referendum outcome n-gram.*

Results clearly show the divergent strategies of the Labour and Conservative parties in the Brexit discourse. The Labour Party predominantly drove discussions around "Hard Brexit," "Customs Union," and "No Deal," reflecting their concern with the more severe implications of Brexit and the potential (economic) risks associated with leaving the EU without robust agreements in place. On the other hand, the Conservative Party led the narrative on "Soft Brexit", "Remain", "Free Trade Agreement", and "Second Referendum." This suggests that Conservatives were more focused on negotiating favourable trade terms, exploring softer exit options that maintained closer ties with the EU, and countering calls for a second referendum to fulfil the Brexit mandate with a focus on trade and sovereignty.

Finally, analysing these n-grams frequency over time might also shed some light on the different stages of the Brexit process:

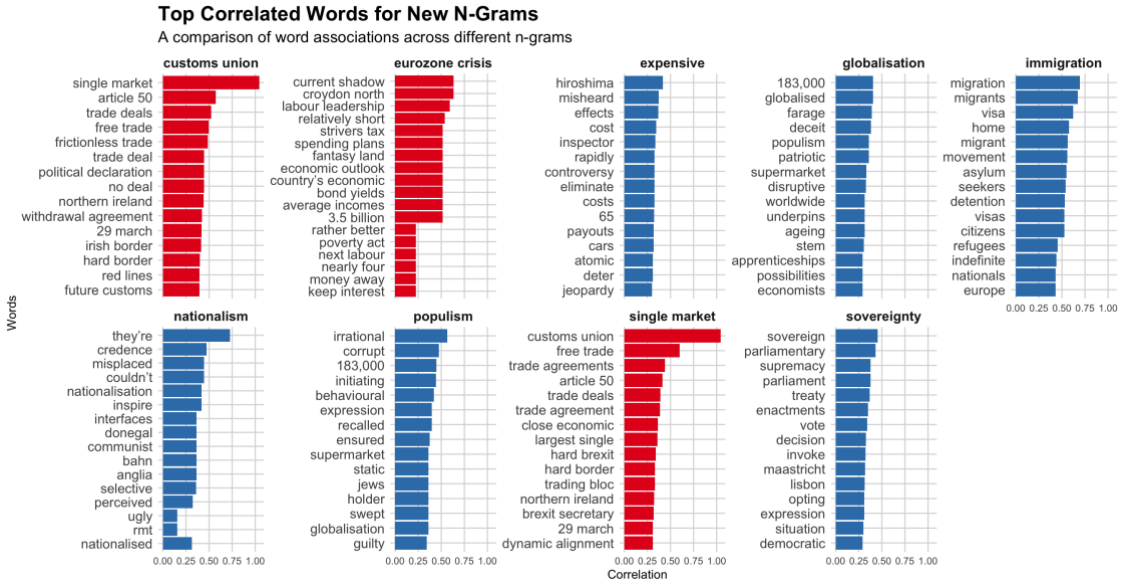


Plot 14: *Frequency of usage of Brexit-related n-grams over time.*

The data shows clear spikes in the discussion of terms like "no deal" and "hard Brexit" during critical moments near the deadline, reflecting a concern over the possibility of the UK exiting the EU without an agreement or pursuing a more extreme form of separation. In contrast, terms like "customs union" and "free trade agreement" appear less frequently but show increases during times when softer Brexit options were considered. Discussions around "remain" and "second referendum" also experienced bursts of activity, particularly when there was time for reconsidering the Brexit decision.

Hot Topics:

The same word correlation analysis was performed for words considered to be “hot topics” in the Brexit realm. The terms analysed were "single market", "customs union", "immigration", "sovereignty”, “expensive", "migrant crisis", "populism", "nationalism", “globalisation" and “eurozone crisis":

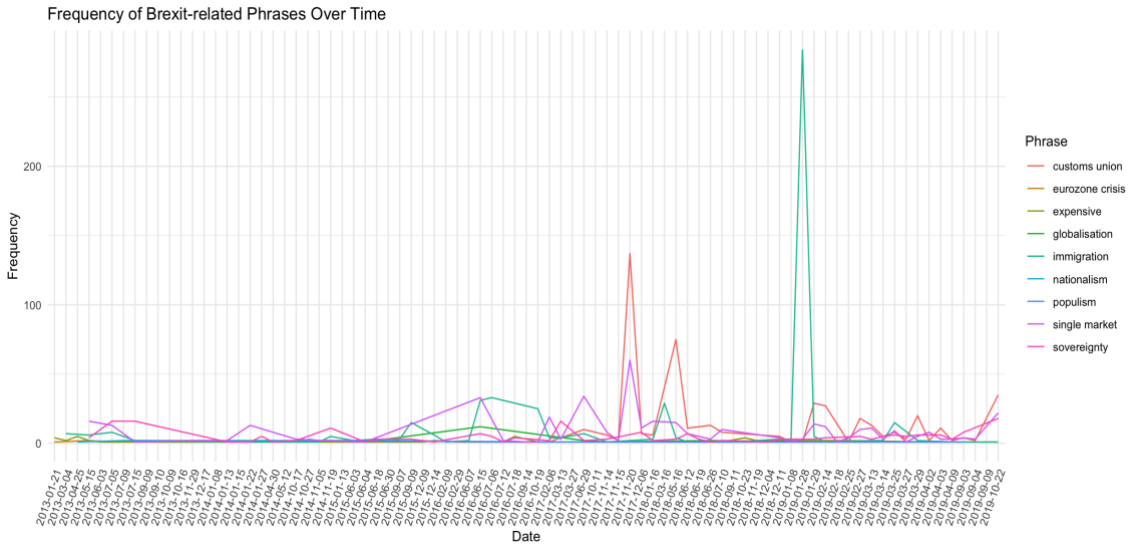


Plot 15: *Top n-grams correlated with each hot topic.*

The "customs union" n-gram is closely associated with terms like "single market", "article 50", and "trade deals," emphasizing the UK's efforts to maintain close economic ties with the EU while negotiating its withdrawal. It also reflects concerns about trade continuity, particularly in the context of avoiding a hard border in Ireland, as evidenced by the correlation with "Irish border" and "Northern Ireland." The "eurozone crisis" n-gram, on the other hand, is linked to phrases such as "economic outlook", "current shadow", and "economic downturn," pointing to the economic challenges faced by the EU, which influenced the UK's approach to Brexit.

The n-grams "nationalism" and "populism" reveal the underlying ideological shifts that have influenced Brexit debates. Words related to "Nationalism" suggest a narrative focused on the legitimacy and consequences of nationalist sentiment. Meanwhile, "populism" is correlated with words like "irrational", "inappropriate", and "behavioural", indicating concerns about the rise of populist movements and their impact on political stability and governance. The "single market" n-gram is strongly connected to words that remark the importance of economic considerations in the Brexit negotiations, and of maintaining market access and minimizing trade disruptions. Finally, the "sovereignty" n-gram is associated with words that capture the core Brexit argument of reclaiming national sovereignty, highlighting the tension between domestic governance and European integration.

Lastly, this section will investigate the frequency of these hot topics through time:



Plot 16: *Frequency of Brexit hot topic n-gram usage over time.*

Terms like "single market," "immigration," and "customs union" exhibit significant spikes. For instance, the substantial increase in discussions around " immigration" in the beginning of 2019 indicate concerns on the implications and ways to deal with immigration when Brexit became effective. Similarly, the rise in mentions of "customs union" and "single market" align with debates over the economic implications of Brexit and the potential impacts on trade and border arrangements, particularly in relation to Northern Ireland. Other phrases, such as "globalisation" and "nationalism," show more consistent but lower-level activity over time.

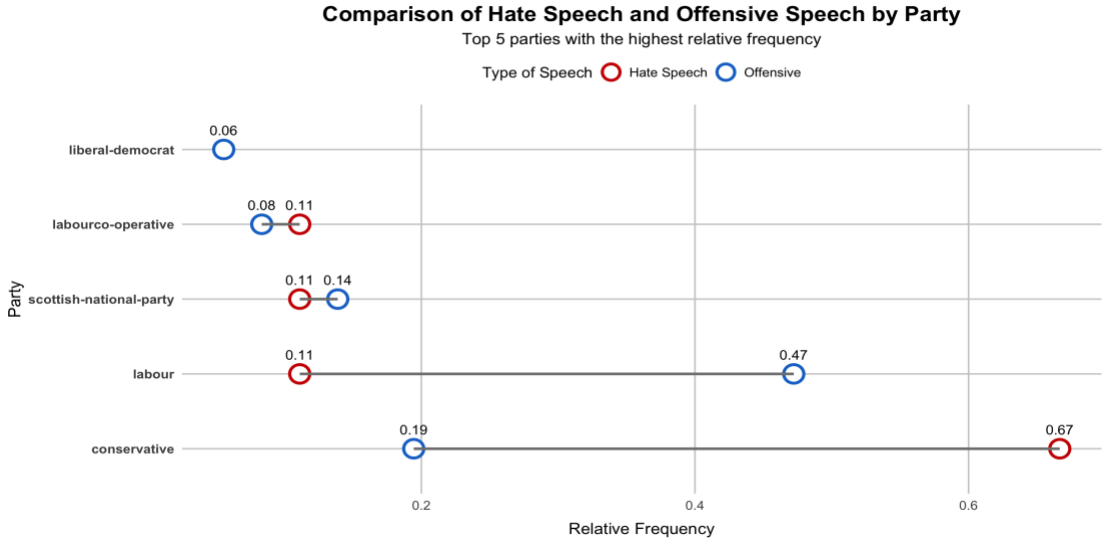
5.3 Hate speech analysis on debates

Hate speech detection is important for understanding political discourse. Research has shown that emotive and extreme rhetoric is more likely to appear in high-stakes political debates (Osnabrügge et al., 2021). Thus, identifying and analysing instances of hate speech in parliamentary debates with the *unhcr/hatespeech-detection model* can shed light on the impact of hate speech in parliamentary debates.

5.3.1 Identification of Parties Using Hate Speech:

The first part of the analysis seeks to identify the political parties that use hate and offensive speech the most. To achieve this, speeches labelled as "Hate speech" and "Offensive" were filtered and the relative frequency of these labels for each party

calculated. The following dumbbell plot compares the of top five parties based on their use of hate and offensive speech highlighting the political groups that most frequently use inflammatory language.

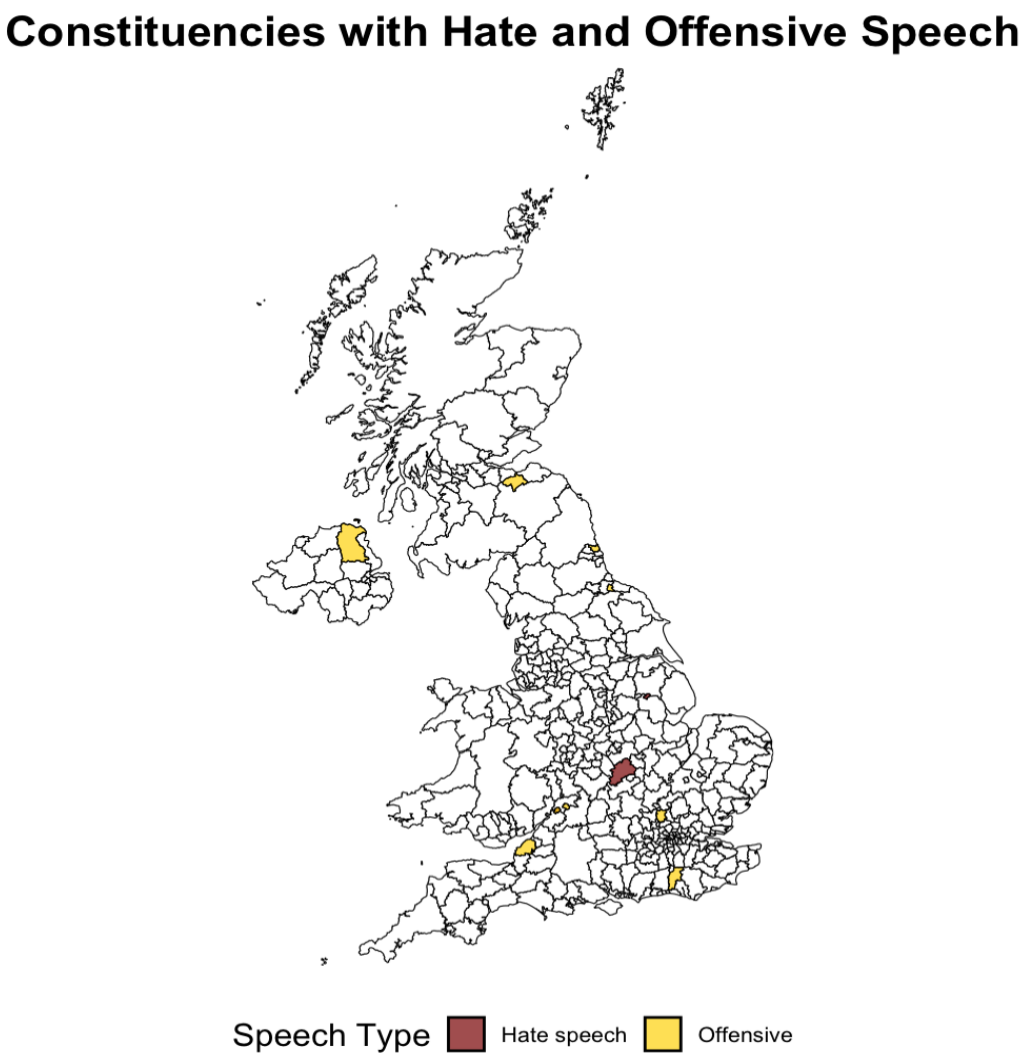


Plot 17: *Relative use of Offensive and Hate speech by party.*

The Conservative Party is notably associated with the highest relative frequency of both offensive speech and hate speech, with offensive speech being particularly prominent. This suggests a significant inclination towards using more inflammatory language within this party compared to others. On the contrary Labour, Liberal Democrats and the Scottish National Party are more frequently associated with offensive speech rather than hate speech. This distribution of speech types highlights the varying rhetorical strategies and communication styles adopted by different political factions, as well as their tendency to employ inflammatory language.

5.3.2 Geographic Distribution of Hate Speech:

Following the line, it was explored whether there is a geographic pattern to the use of hate and offensive speech by mapping these instances on parliamentary constituencies. The code created joins the speech data with geographic shapefiles representing constituencies and plots a map showing areas where hate or offensive speech is prevalent:



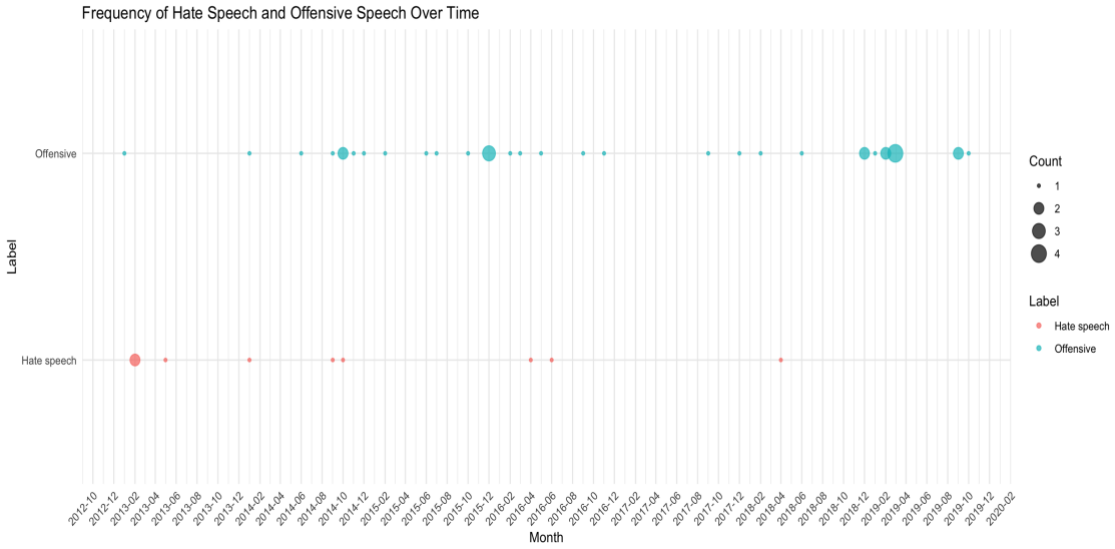
Plot 18: *Geographic distribution of Offensive and Hate speech.*

The map highlights specific areas across the United Kingdom, suggesting potential regional trends. For example, constituencies like Gloucester and Lincoln are marked for both hate speech and offensive speech, while others such as Cheltenham, Midlothian, and North Somerset are predominantly linked with offensive speech, yet most of them appear to be Brexit supporters (BBC News, n.d.).

The clustering of these constituencies in certain parts of the country, particularly in the Midlands, Southwest, and Northern England, may indicate underlying regional socio-political dynamics influencing the use of inflammatory language. However, given the limited number of observations involved, these findings should be interpreted cautiously, as they may not represent widespread regional trends but rather isolated occurrences.

5.3.3 Temporal Trends in Hate Speech:

The analysis also examines the frequency of hate and offensive speech over time by summarizing and plotting the number of instances by month.

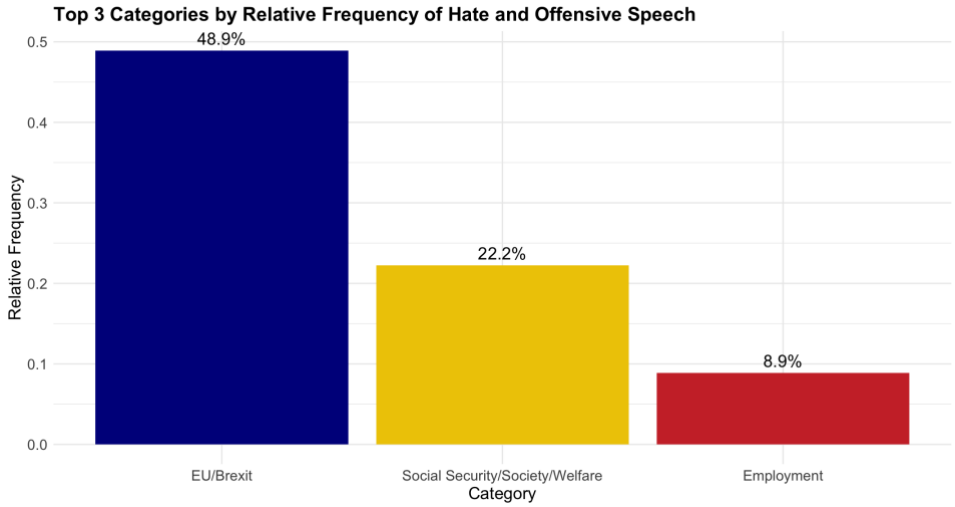


Plot 19: *Offensive and Hate speech usage in the House of Commons Brexit-related debates over time.*

Results show that offensive speech is more frequent and consistently present across the timeline, with a noticeable clustering of instances (offensive language) in late 2018 and early 2019. This period corresponds with the finalization of political debates over the Withdrawal Agreement, which set the terms for the UK's departure from the EU and the European Atomic Energy Community (EURATOM). In contrast, instances of hate speech are relatively rare, with only a few occurrences scattered over the timeline.

5.3.4 Most Conflictive Categories:

Finally, the categories of discussions that are most frequently associated with hate and offensive speech were identified by calculating their relative frequency and plotting the top three categories where inflammatory language is most common.



Plot 20: *Top 3 debate categories with highest concurrence of Offensive of Hate speech.*

Results show that Brexit has been a trigger for extreme language, as it touches on issues of national identity, sovereignty, and the future of the UK in the international community (Clarke, Goodwin, & Whiteley, 2017). Social Security/Society/Welfare and Employment follow, indicating that discussions around sensitive issues like immigration, fairness, entitlement, national identity, welfare, and job security also provoke significant inflammatory language.

6. LIMITATIONS OF THE STUDY

The study's limitations include potential biases in the pre-trained models and the computational intensity required for analysis, which led to the need of performing analyses on a reduced subset of the data (“EU/BREXIT” debates). Also, the hate speech detection model, while generally effective, struggled to detect more subtle forms of hate speech or offensive language, especially when embedded in complex parliamentary rhetoric. This challenge is stressed by the fact that only 36 speeches were classified as Offensive and just 9 as Hate speech out of 1,377 speeches, reflecting potential gaps in the model's sensitivity. Additionally, the models, such as RoBERTa, might not have fully captured the nuances of parliamentary language, particularly in contexts involving sarcasm or indirect speech, which could impact classification accuracy. As mentioned, the computational demands of processing such large dataset required adjustments, such as chunking longer speeches, potentially affecting the continuity and coherence of the analysis. These limitations nontheless set a foundation for future research and improvement in parliamentary discourse analysis.

7. FINAL CONCLUSIONS

The application of Natural Language Processing (NLP) techniques to the ParlVote corpus has produced numerous insights on the political discourse leading up to and following the Brexit referendum. This study offers a detailed analysis of the rhetorical strategies, sentiments, and thematic concerns that shaped the Brexit debate within the UK House of Commons from 2013 to 2019 - a period marked by intense political and public scrutiny due to Brexit negotiations.

The sentiment analysis revealed that parliamentary discourse was predominantly neutral with a notable rise in negative sentiments, particularly during critical Brexit-related landmarks, which reflected the polarized and controversial nature of these discussions. It further revealed significant differences in sentiment between pro-Brexit and anti-Brexit constituencies, especially on topics related to defence, foreign affairs and education.

Topic modelling, testing both BERTopic and KMeans with TF-IDF vectorization methods, uncovered the central themes that dominated the Brexit discourse. Issues related to trade, the economy, national sovereignty, and immigration were recurrently discussed, highlighting the complex and evolving nature of the Brexit debate as well as the nation's main concerns. Findings also showed how political parties shaped the narrative with their rhetorical strategies, reflecting the shifting priorities and concerns of MPs as the Brexit process unfolded.

The hate speech analysis, while detecting relatively few cases of explicit hate speech, identified a concerning prevalence of offensive language, especially in constituencies that supported Brexit. This suggests that the political tensions of the Brexit debate may have contributed to a more hostile verbal environment. The geographic distribution of offensive language, with clusters in certain regions, points to underlying dynamics and patterns that require further investigation (especially under the urban-rural cleavage).

By integrating sentiment analysis, topic modelling, and hate speech detection, this study offers practical tools for dissecting and understanding complex political phenomena. However, it is essential to acknowledge the limitations encountered, including the

potential biases inherent in pre-trained NLP models, the computational intensity required for processing large datasets, and the challenges in detecting subtle forms of hate speech.

BIBLIOGRAPHICAL REFERENCES

Abercrombie, G., & Batista-Navarro, R. (2020). ParlVote: A Corpus for Sentiment Analysis of Political Debates. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).

Abercrombie, G., & Batista-Navarro, R. (2019). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, *3*(1), 245-270.

BBC News. (n.d.). EU referendum results. BBC News. Retrieved from https://www.bbc.co.uk/news/politics/eu_referendum/results

Bhatia, S., & P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

Cardiff University NLP. (2020). cardiffnlp/twitter-roberta-base-sentiment-latest. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

Charteris-Black, J. (2018). *Analysing political speeches: Rhetoric, discourse and metaphor*. Retrieved from Google Books

Clarke, H. D., Goodwin, M., & Whiteley, P. (2017). Brexit: Why Britain Voted to Leave the European Union. *Cambridge University Press*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*.

Gavin, N.T. (2018). Media definitely do matter: Brexit, immigration, climate change and beyond. *The British Journal of Politics and International Relations, 20*, 827 - 845.

Gibbons, V. (2007). *Lights, camera, inaction? The media reporting of parliament*. Parliamentary Affairs. Retrieved from Oxford Academic

Glavas, G., Nanni, F., & Ponzetto, S.P. (2019). Computational Analysis of Political Texts: Bridging Research Efforts Across Communities. *Annual Meeting of the Association for Computational Linguistics*.

Gurciullo et al., (2015). Complex politics: A quantitative semantic and topological analysis of uk house of commons debates. arXiv preprint arXiv:1510.03797.

Hauser, G.A. (1998). *Vernacular dialogue and the rhetoricality of public opinion*. Communications Monographs. Retrieved from Taylor & Francis

Henderson, A., Jeffery, C., Liñeira, R., Scully, R., Wyn Jones, R., & Lodge, G. (2017). How Brexit Was Made in England. The British Journal of Politics and International Relations, 19(4), 631-646.

Hobolt, S. B. (2016). The Brexit Vote: A Divided Nation, a Divided Continent. *Journal of European Public Policy*, 23(9), 1259-1277.

House of Commons Library. (2017, February 6). *Brexit: votes by constituency*. UK Parliament. Estimates of constituency-level EU Referendum result [Data set].. https://commonslibrary.parliament.uk/brexit-votes-by-constituency/

Innes, J. (1990). *Parliament and the shaping of eighteenth-century English social policy*. Transactions of the Royal Historical Society. Retrieved from Cambridge Core

Lewis, J. (2001). *Constructing public opinion: How political elites do what they like and why we seem to go along with it*. Retrieved from Google Books

Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *Cognitive Computation*, *14*(1), 353-371.

Osnabrügge, M., Hobolt, S. B., & Rodon, T. (2021). Emotive rhetoric in legislative debates. Journal of Political Science, 57(1), 45-62.

Plain English. (2023). Exploring Hugging Face Topic Modeling. *Medium*. Retrieved from https://python.plainenglish.io/exploring-hugging-face-topic-modeling-7590efe7c3d3.

Rehbein, I. (2024, May). Resources and Methods for Analysing Political Rhetoric and Framing in Parliamentary Debates. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024* (pp. 36-37).

Sawhney, R., Wadhwa, A., Agarwal, S., & Shah, R. (2020, December). GPolS: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 4847-4859).

TheyWorkForYou. (n.d.). *MPs by name*. TheyWorkForYou. https://www.theyworkforyou.com/mps/

Toszek, B.H. (2020). The Battle of Brexit. Analysis of the 2019 United Kingdom General Election Results. *Polish Political Science Yearbook*.

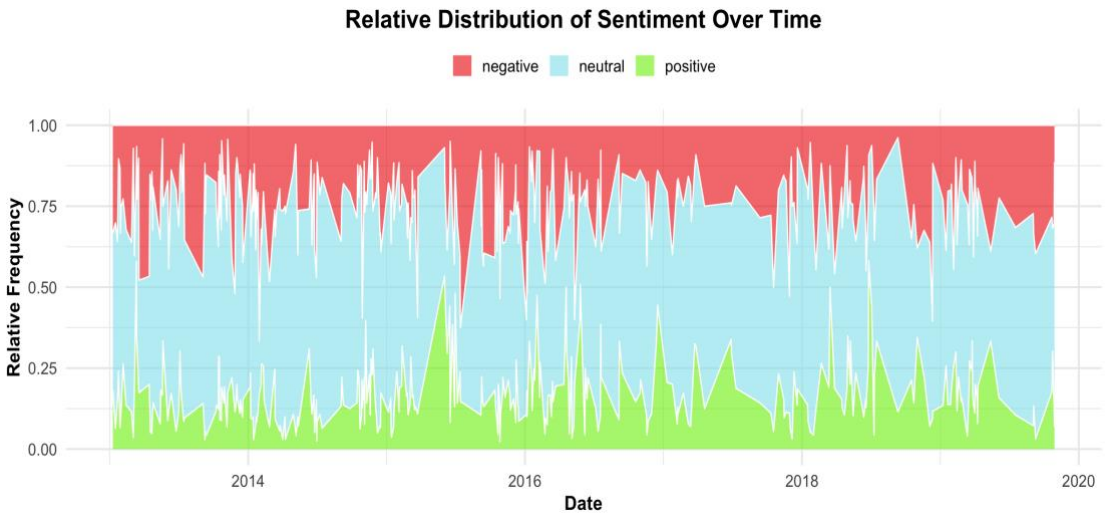
UNHCR. (n.d.). *Model Name*: Hate Speech Detection Model. Retrieved from https://huggingface.co/unhcr/hatespeech-detection.

Van Dijk, T.A. (1997). *Discourse as interaction in society*. Retrieved from Google Books

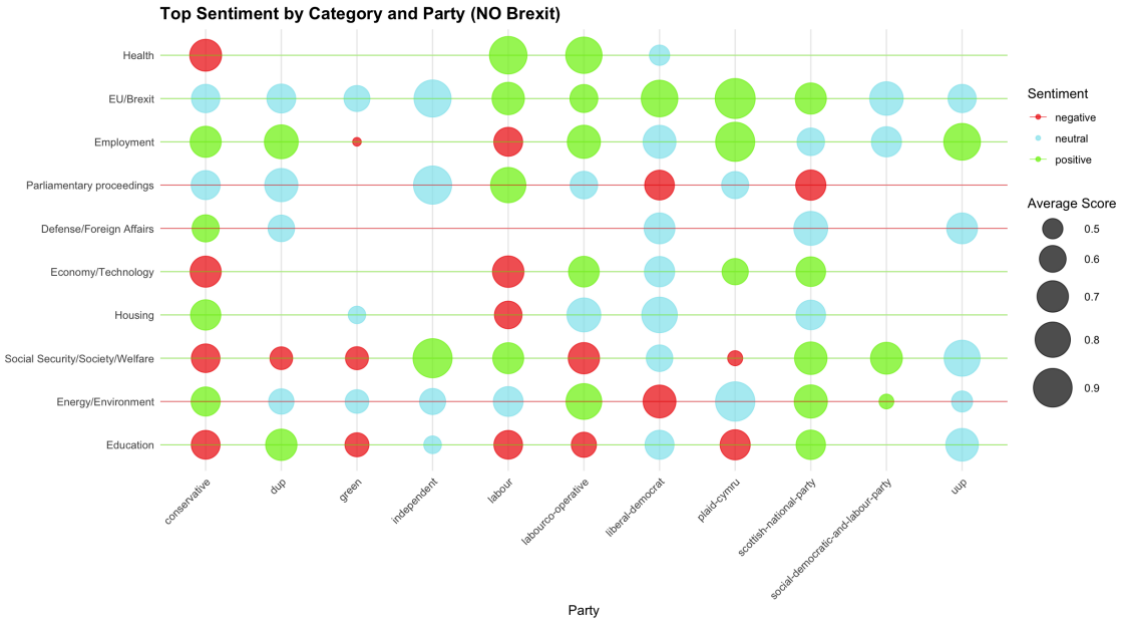
Wankmüller, S. (2021). Neural transfer learning with Transformers for social science text analysis. *arXiv preprint arXiv:2102.02111*.

Wankmüller, S. (2022). Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research*, 00491241221134527.

ANNEX



Plot: *Relative distribution of sentiment of the House of Commons speeches over time.*



Plot: *Most frequent sentiment by party and debate category (Anti-Brexit MPs).*

uc3m | Universidad Carlos III de Madrid

"NLP Based Analysis and Visualization of the House of Commons Parliamentary Debates (ParlVote)"

Laura Martínez Temiño

Madrid, September 2024