Between Brussels and Britain: Uncovering the driving forces of Brexit*

A statistical approach to assessing engagement in a supranational union

Samarth Rajani

April 19, 2024

In a union, all parties must work together to guarantee best outcomes. In the case of the United Kingdom, friction between legislative bodies within the nation and the European Parliament caused the UK to exit the union. This paper looks at a speech based dataset from the UK House of Commons to suggest which political parties were more vocal in the build up to Brexit, giving insight into which political parties were the driving force in this change.

1 Introduction

In the global landscape, where each nation sets to maximize its own gain and self interest, the European Union holds a special place. The only political and economic union of its kind, it spans 27 nations in Europe,

However, there is an inherent concern with being party of a larger authority - there is a risk of a misalignment of potential goals. With 26 other states involved, policy across the union becomes a balancing act between furthering an individual nation's goals without compromising another. This unique situation burdens legislative bodies of the nations involved to consistently discuss and negotiate EU-related matters, potentially diverting money, time and manpower away from domestic issues, all of which are important resources. Therefore, the legislative bodies must balance engaging in EU related discourse and tending to the needs of the local public.

The United Kingdom was an influential part of the European Union. Even though the United Kingdom joined the union later than other European giants like France and Germany (15 years later), it spearheaded efforts in making the union one. Most notably, the UK presented a plan

*Code and data are available at: https://github.com/samarthrajani/-Between-Brussels-and-Britain-Uncovering-the-driving-forces-of-Brexit/tree/main that included 282 legislative actions at the European Council in Milan in June 1985, which was the foundation of the single market within the EU. This enabled complete free movement of goods and services within the EU, and without the involvement of the UK the Union may have taken a more protectionist approach (JACOPO BARIGAZZI 2020). With its economic prowess, UK also commanded areas of foreign policy, free trade and growth among others.

However, the UK's allegiance to the union did not stand the test of time, succumbing to the dilemma described before. The driving force behind the 'Brexit' in 2020 has been a concern for lack of sovereignty, where the public felt the European Commission was taking away the UK's ability to make decisions for itself. David Frost's speech in 2020 best highlighted this position (David Frost 2020) and detailed how the European Union's institutions felt abstract and "disconnected from natural feeling'. After much deliberation since 2016, the UK pulled the trigger in 2020, successfully exiting the Union in dramatic fashion.

Keeping this in mind, it is worth examining how often matter of the European Union dictated discourse in the United Kingdom. Specifically, this paper studies how often each party gave speeches that involved discussions around the European Union right down to the level of individual sentences. Looking at the ParlEE(Sylvester, Christine; Greene, Zachary; Ebing, Benedikt (2022)) dataset, this paper examines how likely is one political party to devote time to discussing EU related matters. It is a reasonable assumption to make that party ideaologies that significantly deviate from and oppose the European Union will cause those parties to be more vocal about EU related matters, thereby increasing the incidence of EU related discourse. This paper utilises this assumption to study major contributors to the Brexit in 2020.

The results reflect support for Brexit reasonably well. The UKIP and GPEW very discussing EU related matters most often, while the Liberal Democrats who oppose Brexit discussed EU matters the least.

The remainder of this paper is structured as follows. Section 2 which discusses the dataset in high detail, Section 3 which outlines the model used for analysis and lastly, Section 4 which explains the results from the model. There is a discussion section Section 5 which developes ideas for future studies, the strengths and the weaknesses of this paper.

2 Data

This paper uses the ParlEE plenary speeches data set to conduct its analysis. This dataset specifically has been obtained from version 1, and uses the UK speech data. The dataset has information on speeches from 2009 to 2019.

For their data collection, the authors chose the government body bearing the most responsibility for legislation. For the United Kingdom they chose the House of Commons. As a starting point, the authors used the The Rauh et al ParlSpeech dataset set since it already contained a large collection of raw speech data for many countries of interest, especially the UK. There was some data missing towards the end of the timeframe of interest (towards 2019), so the authors went through the speeches from UK Parliament (2009-2019) and filled in the missing parts themselves using an API provided to them.

The authors ironed out typical wrinkles with government related datasets - removing boilerplate information, page numbers, eliminating corrupted social characters and introducing consistent naming conventions for party and speaker alike. This made the data uniform across countries, improving readability.

2.1 Variables

The dataset has variables explained below :

- **Date** : The date on which a specific sentence from the speech has been said is recorded in the 'date' variable. This is recorded in the 'DD/MM/YYYY' format, with duplicated values for multiple sentences spoken in the same speech as well as if there were multiple speeches on the same day.
- **Agenda** : The variable 'agenda' is the title given to a set of speeches given pertaining to the issue at hand in the House of Commons at the time. This is provided by official sources.
- **Speechnumber** : The variable 'speechnumber' gives a unique identifier to all the sentences spoken in the same speech. So if the first speech in the data set has 5 recorded sentences, they will all be allotted the number 1.
- **Sentencenumber** : The variable 'sentencenumber' assigns a unique identifier to every sentence spoken within the same speech. Continuing the example above, every sentence in speech number 1 is allotted a sentence number 1 through 5. This variable resets from speech to speech, such that the first sentence in every speech will always have number 1.
- **Speaker** : The variable 'speaker' indicates the name of the member making the speech. Measures have been taken to eliminate inconsistencies, such as removing any nicknames, variations in spelling or honorary titles and using only one standard naming convention. The only exception is the usage of ministerial titles in the speaker column, since it is worth indicating when the same speaker is speaking as 'the Minister' versus just an individual. This is because that being in that position may lead to influence in the opinion of the speaker as demanded by the given title.
- **Party** : The 'party' variable records the alignment of every speaker. The parties recorded are APNI, Conservative Party, DUP, GPEW, Labour Party, LibDem, other, Plaid Cymru, SDLP, SNP, UKIP and UUP.
- **Text** : The 'text' variable records the raw text of sentences part of a speech given in parliament. This has already been cleaned as stated previously, and therefore is ready to be studied across the other variables like party, agenda, etc.

- **Parliament**The 'parliament' variable stores the name of the legislative body. For the UK and for the purposes of this paper, this variable is 'UK- HouseofCommons' everywhere.
- **iso3country** : Similarly, the variable 'iso3country' records the country in question, which for this paper is the United Kingdom throughout, recorded as 'GBR'.

The authors augmented the collected data so as to make it enable analysis beyond what was already possible. These variables are described below:

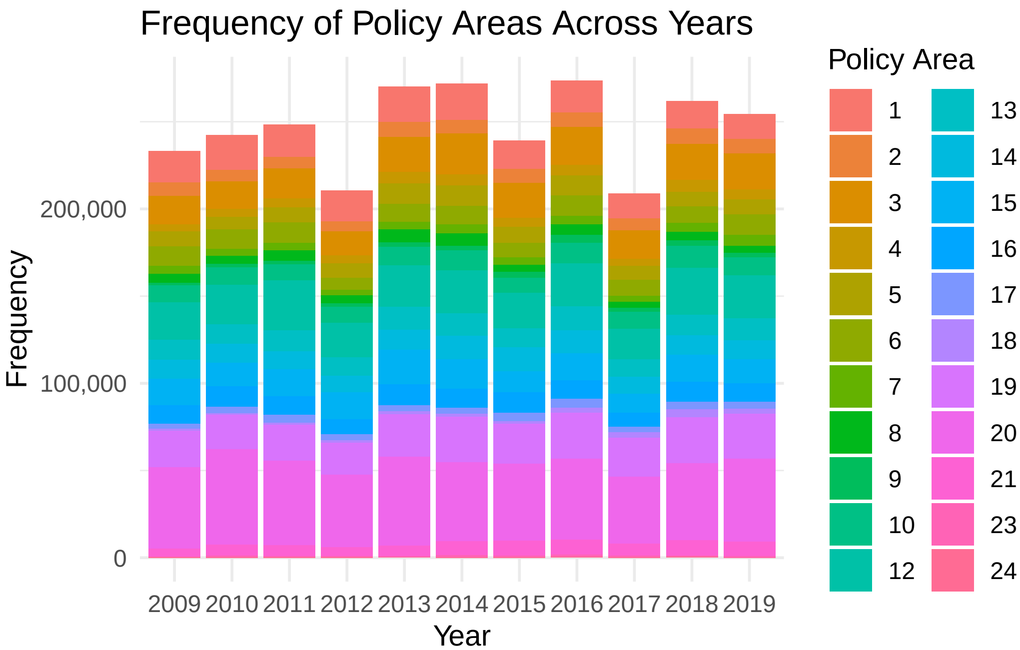
- **EU** : The variable 'eu' is a dummy variable introduced by the authors that determines whether the given text or sentence discusses anything related to the European Union. The dummy is '1' if it does discuss the European Union and '0' if it doesn't.
- **Policyarea** The variable 'policyarea' assigns a number in accordance with the specific policy topic of a text. The numbers used are from the CAP major categories coding scheme.

The CAP coding scheme is shown in the table below :

Table 1: Policy Areas and CAP Numbers

Policy.Area	CAP.number
Macroeconomics	1
Civil Rights	2
Health	3
Agriculture	4
Labour	5
Education	6
Environment	7
Energy	8
Immigration	9
Transportation	10
Law & Crime	12
Social Welfare	13
Housing	14
Domestic Commerce	15
Defence	16
Technology	17
Foreign Trade	18
International Affairs	19
Government Operations	20
Public Lands	21
Culture	23
Fisheries	New

The following graph shows the occurrence of the given policy types over the years:



As can be seen, there is very little variation in discussion of specific policy areas over the course of 10 years. This could be due to the dataset being extremely large, as it would iron out any major differences due to sheer sample size. There is also a case to be made that the CAP classifier used by the authors to assign the policy area to sentences may not be completely accurate, although that discussion is not relevant to this paper.

3 Model

We use a linear logistic regression model in order to estimate how likely each individual party is to talk about European Union related affairs. Using the 'eu' binary values, the regression is run using the Alliance Party of Northern Ireland (APNI) as the predictor category. We will also be using Bayesian Estimation in our model, using the Goodrich et al. (2022) package.

3.1 Model set-up

Define y_i as discussion of the EU in a specific sentence of a speech. Then x_i is the party affiliation.

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i + \gamma_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect there to be some differences between different partys' and their affinity to discuss EU related matters if their ideologies demand that kind of discussion.

For example, if the Conservative party leans towards a smaller role of governments and institutions in daily lives, they would advocate for a small government political model. However, if the policies of the European Union dictate increased involvement, it is more likely to be discussed by that specific party. This enables us to draw conclusions as to which parties are more likely to disagree with the involvement of the European Union.

It is worth noting that our variable 'eu' works as a proxy for the expression of disagreement for every party. Its accuracy as a proxy will be discussed further into the paper.

4 Results

Our results are summarized in Table 2.

The summary table provides the coefficients for our regression.

The reference category for this regression is the The Alliance Party of Northern Ireland (APNI). All the coefficients show how more or less likely is each party to mention the European Union in their speeches in log odds. For example, the conservative party is 1.2 log odds more likely to discuss the EU in their speech as compared to APNI.

Our model seems to be a good predictor for support for Brexit. The UK Independence Party (UKIP) has most strongly favoured Brexit according to Matthew Smith (2019), which as can be seen has a 1.7 log odds coefficient which is the second largest. The highest coefficient is held by the Green Party of England and Wales (GPEW), which also held pro-Brexit views until 2019, which is in line with our model timeframe.

Table 2: Logistic model of likelihood of a party affiliation speaking about the EU

	Log likelihood of EU discussion
(Intercept)	-4.454
partyCon	1.176
partyDUP	1.068
partyGPEW	2.038
partyIndependent	1.221
partyLab	0.996
partyLibDem	0.784
partyPlaidCymru	1.022
partyRespect	0.685
partySDLP	1.273
partySNP	1.450
partyUKIP	1.681
partyUUP	1.116
Num.Obs.	784 555
ELPD	$-117\,744.1$
ELPD s.e.	537.8
LOOIC	235 488.2
LOOIC s.e.	1075.7
WAIC	235 486.7
RMSE	0.18

Furthermore, the Liberal Democrat party (LibDem) have been against Brexit, beieving that being a part of the European Union is in the best interests of the UK government. This can be seen as a coefficient of 0.8 log odds is one of the lowest in the summary table.

It is worth stating that the reference ctaegory variable here is Alliance Party of Northern Ireland (APNI), which has holds an anti-Brexit stance.

5 Discussion

5.1 Disagreement and discourse - a relationship

While this paper studies a very specific question, the strategy used to model has implications that may be more far reaching than one would anticipate at face value. The strategy is simple - the more strong the disagreement, the more it is voiced in an environment conducive to that sort of discussion. This means it could have implications not only for the EU, but for studies that have nothing to do with socioeconomic outcomes.

For example, there exist many small institutions that are not usually at the center of statistical models, like a high school council. However, if there is a key issue up for discussion, it seems safe conjecture that students who disagree with actions of the school are more likely to voice their concerns, even repeatedly. This easy-to-miss behavioural strategy can be extrapolated to other important areas to enable or bolster a study.

5.2 U

5.3 Limitations

5.3.1 Small timeframe

The one obvious limitation of this paper is that it only uses data from 2014 to 2016 in its analysis to judge which parties are being vocal. This was done due to resource limitations in the logistic analysis - the runtime for merely 2 years was close to 5 hours, severely limiting the ability to measure the relationship between party affiliation and talking about the EU in the House of Commons. Since the Brexit referendum took place in 2016, it is possible that the propensity to talk about the EU was in fact talking just about the referendum, and therefore might not be the vocalisation of disagreement. This causes our estimates to overestimate the the friction between the political parties and European Parliament, making our conclusions inaccurate. This is a fundamental reverse causality problem.

5.3.2 Construction of the 'eu' column

It is worth noting that the 'eu' column was not obtained from any existing dataset from the UK House of Commons. Instead, the authors utilised a classifier that had been trained on existing EU topical coded speech data from the Comparative Agendas Project (CAP) (Baumgartner et al. 2019).

In the release note, the authors discuss how a binary classification of EU related issues may have been too strict, and thereby generated more false negatives (classifier incorrectly classifies a sentence discussing EU Governance as not discussing EU governance) in the pursuit of less false positives (classifier incorrectly classifies a sentence as discussing EU Governance when it does not). This causes a potentially downward bias in our logistic regression estimates, underestimating the extent to which EU related matters have dictated discourse.

5.4 Next steps - An Instrumental Variable Framework

This paper lays the groundwork for a more detailed study into which parties played the strongest role in Brexit. With the dataset of speeches already available, devising an Intrumental Variable Framework to determine party influence seems like a suitable way forward.

When studying the relationship between two variables, there is a concern of reverse causality. For example, if we were to study the impact on institutions on the GDP of a nation, our estimates would be biased since GDP could also affect and improve an institution. To tackle this, an Instrumental Variable approach can be used.

In this setting, to help predict influence on brexit, we can use frequency of EU in speeches as the instrument. While this will require further testing and OLS regression to prove that frequency of EU speeches are correlated with supporting Brexit, as well as discussions about the Exclusion Restriction. However, those are beyond the scope of this course.

References

David Frost. 2020. *David Frost Lecture: Reflections on the Revolutions in Europe*. Prime Minister's Office. https://no10media.blog.gov.uk/2020/02/17/david-frost-lecture-reflections-on-the-revolutions-in-europe/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

JACOPO BARIGAZZI. 2020. *Britain Leaves Its Mark on Europe*. Politico. https://www.politico.eu/article/britain-leaves-mark-eu-brexit/.

Matthew Smith. 2019. *Which Parties Are Pro and Anti Brexit?* YouGov. https://www.politico.eu/article/britain-leaves-mark-eu-brexit/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sylvester, Christine; Greene, Zachary; Ebing, Benedikt. 2022. *ParlEE Plenary Speeches Data Set: Annotated Full-Text of 21.6 Million Sentence-Level Plenary Speeches of Eight EU States*. Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZY3RV7.

UK Parliament. 2009-2019. *Commons Hansard*. UK Parliament. https://www.parliament.uk/.