Computational Analysis of Trend in House of Commons Debates

Shubham Joshi
Department of Computer Science
University of Exeter
Exeter, UK
sj534@exeter.ac.uk

Supervisor: Dr. Chico Camargo
Department of Computer Science
University of Exeter
Exeter, UK
F.Camargo@exeter.ac.uk

***Abstract*—Text analysis of parliamentary speeches is important for researchers, to understand what is being discussed in the House of Commons (HOC). It's critical to comprehend how the speech semantics vary as the context and the talks change throughout time. This report suggests a novel approach to this task, to analyze Hansard's speeches of HOC by our MP using NLP. The strategy incorporates cutting-edge transformer techniques and improves on earlier attempts to cluster topics from the document. These subjects could help us grasp the issues that MPs are debating. To comprehend the relationship between words in speeches, we also developed a co-occurrence network. We used Orthogonal Procrustes alignment and semantic similarity calculations to understand the semantic shift in the speeches over two separate periods. All of the results were plotted as a prototype in the dashboard.**

***Index Terms*—NLP, Hansard speeches, Orthogonal Procrustes, Dashboard**

I. Introduction

Members of the House of Commons (HOC) in the UK parliament, are chosen by the people, and comprise the government. The government is formed by the party with the most Commons members. MPs discuss current political issues and new legislative initiatives, and "making decisions on financial bills, such as proposed new taxes" [1].

The House of Commons speeches delivered by our MPs are the main focus of this project's research. The parliamentary debates would allow researchers to conduct political analysis, for instance by analyzing the key HOC topics that were under discussion, developing a network for comprehending the speeches and understanding the spread of ideas and points of view over time by analyzing the semantic change in the context, which would help us understand how the government policies will change.

A word's context and meaning change throughout time to the point where a new word differs greatly from an old one. As a result, we must consider how the word has changed semantically as well as its usage context. As an illustration, the semantic meaning of the word "awful" once meant "extremely bad over time," very bad in modern times" [2].

We can get an overview of the main concerns of our MPs by identifying subjects in the speeches. We may better comprehend the topics of discussion in these talks in HOC by plotting a network for them. The relatively untapped field of computational analysis via unsupervised learning based on the linguistic content of documents is therefore becoming a growing area of academic attention.

The strategy described in this study expands on earlier work in computational analysis and makes use of state-of-the-art sentence embedding and document embedding techniques to identify topics and assess the semantic shift. The approach is modular, so researchers interested in other datasets can simply use a different preprocessing regime while keeping other computing processes unaltered. While UK parliamentary speech is the focus of this article, other datasets can also be processed using the same general methodology.

In this project, a more relevant aspect of the parliamentary debates is how the context changes for a specific individual or entity through time. To further understand how the context has changed over a year, we will also attempt to quantify the semantic shift in the context by comparing the cosine similarity between the speeches. To clearly illustrate the semantic similarity, the change detected will be shown over the violin plot. In the study, another Orthogonal Procrustes [28] method is employed to better understand how the semantics of a word varies for a given speech. Finally, we will plot every outcome of the project analysis on the dashboard, which will serve as the prototype. Furthermore, the method is agnostic to the precise definition of semantic change, enabling researchers to use any understanding of the term that is appropriate for a given subject.

II. Background

A. Embeddings: Sentence and Document

In general, word embeddings are a sort of word representation that enables words with similar meanings to have a comparable representation. These dense embeddings are created by the transformer model which was introduced in the 2017 paper "Attention is all you need". [5]. Transformer models are trained on extensive text corpora in the case of "BERT, a corpus of books and all English text passages of Wikipedia" [7], utilising the machine learning method of "attention," boosting significant portions of the training text. The word embeddings produced internally by transformers

take into account the context on either side of each word, in contrast to word2vec [8], which encodes each word as a single, static vector.

The sentence transformers library uses SBERT [6], which utilises the old state-of-the-art (SOTA) models for all common semantic textual similarities, to determine precise sentence similarity (STS). SBERT [6] uses a siamese architecture to focus on sentence pairings. This is analogous to having two parallel, identical BERTs with the same network weights. In addition to a "sentence-level" embedding for the "classification token," the SBERT [6] model generates a 768-dimensional vector for each sentence in the input text.

This was improved via document embedding. Usually, document embedding is calculated using word embeddings. Every word in the text is first given a word embedding, and then the word embeddings are concatenated. To obtain the document embedding of the parliamentary speeches, we employed the Cr5 [9] model with the English language.

B. Co-Occurrence Network

A co-occurrence network [10], also known as a semantic network, is a technique for text analysis that involves a graphic display of possible connections between individuals, groups, ideas, or other things represented in the textual content. With the introduction of text that has been electronically recorded and is text mining-compliant, the development and display of co-occurrence networks [10] have become feasible.

The text used in the parliamentary speeches is displayed using a co-occurrence network [10]. In this case, the word within a speech is represented as a network, and the edges linking it to other words within the same speech indicate their relationship to one another.

C. Clustering: HDBSCAN and KMeans

Clustering low-dimensional data is a crucial step in this research. With non-convex data, the more complex DBSCAN Density-Based Spatial Clustering of Applications with Noise [11] algorithm effectively identifies clusters as regions of high density divided by regions of low density. DBSCAN is noise-resistant as well but finding clusters with different densities presents a challenge for the algorithm.

To discover the clustering that provides the best stability across epsilon, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [12] executes DBSCAN [11] over a range of epsilon values and integrates the result. As a result, HDBSCAN [12] is more resistant to parameter selection and may locate clusters with a range of densities.

The K-means [13] approach for data mining starts with an initial set of centroids that are randomly chosen and are used as the starting points for each cluster. It then does iterative calculations to optimise the placements of the centroids. Since it generalises to clusters of various sizes and shapes, such as elliptical clusters, it is useful for clustering our Hansard speeches.

D. Dimensionality Reduction: UMAP

The process of converting highly dimensional data into a representation with fewer dimensions while retaining key facets of the original data structure is known as "dimensionality reduction." While "t-SNE favours the preservation of local structure over global structure" [15] algorithms like PCA "aim to do this by trying to maintain the pairwise distance structure" of the data. UMAP [14] tries to preserve the global structure before performing dimensionality reduction. It is better than t-SNE in the below terms

- "Graph Laplacian is utilised by UMAP" [14] to initialise low-dimensional coordinates, as opposed to "tSNE's usage of random normal initialization." [15]
- Unlike tSNE [15], UMAP [14] employs the stochastic gradient descent (SGD) rather than the conventional gradient descent (GD).
- Instead of perplexity, UMAP uses the number of closest neighbours

E. Orthogonal Procrustes

The Orthogonal Procrustes [28] problem is to identify the orthogonal matrix that maps one set of provided points to another set of supplied points in the shortest amount of time. The one-to-one correspondence of points between the two sets must be known beforehand. It is like finding an orthogonal matrix R that most closely maps A to B in the vector space.

III. Related Work

A. NLP in Politics

Political scientists utilise natural language processing (NLP) to extract useful information from text data and to "gain insight into policy." [16] "Using implied policy positions to select the topics." [17].As far as I can be certain there are no instances where NLP has been used to analyse Hansard data for a short duration of time.

There is some study where the Comparative Legislators Database is being studied in R language by using the legislatoR package. The CLD, which includes political, sociodemographic, career, online presence, public attention, and visual data, has over 45,000 modern and historical politicians from ten nations represented. [18]. It's developed in R and uses the "ParlSpeech V2 dataset, which includes legislative addresses from 10 included democracies (legisltorR)" [19]. Using built-in functions, we may directly retrieve the data from the package.

In 2014 Fromreide et al. [**?**] performed named entity recognition on UK parliamentary data for text extraction to mapped to the entity. According to a literature assessment, legislative data was subjected to sentiment analysis by Kapovciute-Dzikiene and Krupavi- cius [21] where he tried to understand the sentiment of speeches made by our MPs.

Gavin Abercrombie [22] conducted a different study in which he used word embedding to compare parliamentary speeches from the years 1910 to 2010 to do so.

- He used the technology of word embedding to track the change over a decade.

- The comparison of the embedding vector's cosine similarity and its impact on the closest word

In my research, I would examine how MPs' speeches changed over a year. The speeches' cosine similarity will be calculated to track developments between 2020 and 2021.

B. Topic Modeling

Topic modeling is an unsupervised machine learning technique used in natural language processing that may scan a collection of documents, identify word and phrase patterns within them, and then automatically arrange words and phrases that best describe the collection. A popular text-mining technique for locating "latent semantic patterns in a text body is topic modeling." [3]

It has primarily been utilised in the social sciences to locate concepts and topics within a corpus of texts. The papers of "DiMaggio P analyse web content, newspaper articles, books, speeches, and, in one instance, videos." [4]. To determine subjects from unstructured data, topic modeling counts words and groups words with similar word patterns. Since our MPs have given hundreds of speeches, we will attempt to analyse them using a topic modeling technique. A topic model clusters information by identifying patterns like word frequency and word distance. With this knowledge, we can instantly determine what each speech is discussing.

The most popular topic modeling technique, LDA [23], has limitations due to its bag-of-words approach to text representation and the need to define the number of topics in the text.

Top2vec [24] and BERTopic [25] are two contemporary topic modeling methods that use text embedding to try to get over these issues. Similar processes are used in both: dimension reduction with"UMAP [14], text embedding using a sentence transformer, clustering with HDBSCAN [12].

C. Semantic Change

Word definitions change throughout time. As an illustration, the term "gay" recently changed from "carefree" to "homosexual." I can obtain a sense of the word's semantic evolution by observing these kinds of changes. Distributional approaches and "prediction-based word embedding models can be used to capture the temporal change in the word's lexical and semantic properties" [26]. A change in human language is what has caused this transformation.

By comparing the first and last time the change was noticed, training and aligning the embeddings, and then comparing the two reference points, the semantic change may be measured. I'll follow Philippa Shoemark's approach in this project, which he outlined in his paper "A Systematic Comparison of Semantic Change Detection Approaches using Word Embeddings" [27].

- With the corpus of UK parliamentary speech data, train the word embedding via"word2vec [14] at a specified time step of t and t-1.
- Overlapping the two embeddings for all the common words will allow us to align them using Orthogonal Procrustes [28] and monitor how they change over time.
- The cosine similarity and closest neighbour methods are used to measure the change in the embeddings. A change in the closest neighbouring term would indicate a change in the word's semantics.

The semantic changes in the context over a shorter period could be measured by looking at how often the words change. In the article "Short-term Semantic Shifts and their Relation to Frequency Change, Anna Marakasova" [29] discusses how the co-occurrence of phrases, using vector similarity and neighbourhood similarity measures, depicts the semantic shift in the context.

IV. Aims and Objectives

The main research question addressed by this project is whether computer approaches can be used to quickly analyse the semantic shift in a corpus of parliamentary discourse, the UK Hansard Corpus, and automatically identify relevant topics.

A. Utilize the current unsupervised learning technique with the Hansard data.

The most crucial things to do with the Hansard dataset are to build the data pipeline and carry out data pre-processing procedures, like data cleaning and stop word removal. By applying topic modeling to the dataset, unsupervised learning is possible. For topic modeling using Hansard data, Latent Dirichlet Allocation (LDA) [23] is one of the popular approaches I'll be utilising.

In the project, Custom Topic Modeling is used based on the BERTopic [25] approach to enhance the topic modeling using LDA [23]. First, S-BERT was used to get the sentence embedding from the Hansard speeches. After that, UMAP [14] was used to perform the dimensionality reduction, and ultimately, the subjects will be clustered using the HDBSCAN [12] clustering method.

B. Sentence Embedding on Hansard Dataset

Sentence embedding is used after classifying the data set over two years by retraining the SBERT [6] model for the Hansard data.

C. Measure Semantic Change

The change of a word and its nearest neighbour can be determined using the methods given below.

- Use word2vec [14] to extract the vector of text data after dividing the dataset into distinct periods. After that, compute the semantic change in the words by aligning the two vectors using Orthogonal Procrustes [28].
- Create a graph neural network of speeches for two separate years using the co-occurrence network [10]. then locate the closest neighbour for that specific word. If two separate periods have neighbouring words that overlap. By comparing the overlap of the words, we could gauge similarity.

D. Dashboard for Hansard Data

Using Plotly Dash, create a prototype of the dashboard including the findings of my analysis. Install the dashboard on the Heroku platform so it could be independently accessed.

V. Methods And Experiment Design

We used the procedure shown in Figure1 to analyse the HOC parliamentary data. The Hansard data was first gathered through the Zenodo website. After performing a data cleaning operation on it, we conducted data exploration in search of important HOC subjects. Next, we used LDA to do topic modeling and plot the co-occurrence network for the legislative discussions. As a result of our inability to obtain quality topics, we carried out custom topic modeling. We searched the speeches within a year for semantic parallels. Additionally, Orthogonal Procrustes [28] was utilised to determine word similarity. After compiling all of our analyses, we plotted the dashboard prototype using Plotly Dash.

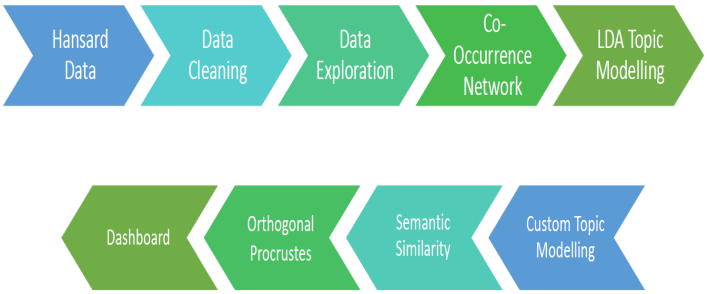


Fig. 1: Hansard Data Flow Diagram

A. Software Tools

This project's code is primarily written in Python 3 and uses the machine learning tools pandas, NumPy, and matplotlib for visualisation. I utilised the hugging face transformer approach for SBERT [6] models. Spacy and nltk for text processing will be utilised as the other NLP libraries. The Windows 10 OS is used to process code. The Heroku platform is used for the deployment of the dashboard. The Github repository could be found at https://github.com/shubh802/hoc-hansard

B. Data

This research's main goal is to examine the themes and semantic shifts in a corpus of House of Commons speeches from the UK Parliament. The House of Commons speeches in the United Kingdom are contained in Hansard [30], which is the project's main data source. The data is not immediately accessible in Hansard, but it is accessible as a zip file from several places, including zendo. It includes Hansard speeches from 1979 to 2021 under a Creative Commons Attribution 4.0 International License.

- Speeches from Hansard, 1979–2021 (Version 3.1.0): The zendo website has processed the "Hansard" [15] dataset in a CSV format as zip files, making it much simpler to use Python. It contains discussions from 1979 through 2021 and can be downloaded from URL https://zenodo.org/record/4843485/

Each column records the speech's transcript, the debate in which it was delivered, the speaker's name and unique ID, as well as other pertinent information, with each row representing a single "contribution" made by an MP in Parliament. Appendix TableIII reproduces a sample row of this raw data.

C. Data Prepossessing

Python was used to load the data into a pandas dataframe, and pandas were used for all preprocessing operations. 2,694,375 entries totalling 42 years' worth of legislative debates can be found in the raw dataset. We would examine the talks within the periods of 2020-04-29 to 2021-04-29 and 2019-04-29 to 2020-04-29 because the total number of years is too large. By picking a recent historical era, it is easier to evaluate outcomes using current political domain knowledge. Additionally, it will make it easier for us to see the semantic shift in the speech environment over a shorter period. This might eventually enable us to draw some conclusions.

After the data was filtered by date, non-spoken entries were eliminated by preserving speeches with the speech_class value of "Speech"; this eliminated, for instance, descriptions of procedural events and vote records, both of which are kept in Hansard. Id, display_as, party, constituency, mnis_id, time, column, oral_heading, year, hansard_membership_id, speakerrid, person_id, speakername, and url were also deleted because they were unnecessary. Then, entries that were not made by a specific MP were eliminated. Following cleaning, columns providing information characterising the debates such as "speech_date," "speech_class," "major_heading," and "minor_heading" are saved. The MP-specific data was then further processed using these special columns.

We first lowered the text of each MP's speech before removing the punctuation. Regex expression was also utilised to filter out simply the speeches' text. The phrases "thing," "give," "try," "look," "therefore," "go," "hon", "use," and "health" were also eliminated from the speeches since they were just being repeated too often and made it difficult to comprehend the context. After that, each contribution's text was lemmatized using the language model "en_core_web_sm" from spaCy. Each contribution's lemmas were inserted as a new column into the dataframe. Rows with no lemmas, or contributions without any non-stop word material, were eliminated. In total it reduced 2% of the raw data. In Appendix TableIV, a sample of the data at this preprocessing stage is displayed with "speech_processed" as the column with processed speeches. These processed data for the two years was saved as a pickle file to be used further in the project.

D. Hansard Data Exploration

To understand what topics are the most frequently discussed in HOC, all of the speeches for 2021 have been grouped based

on the major_heading daywise. Figure 1 displays the typical HOC discussion themes for 2021.

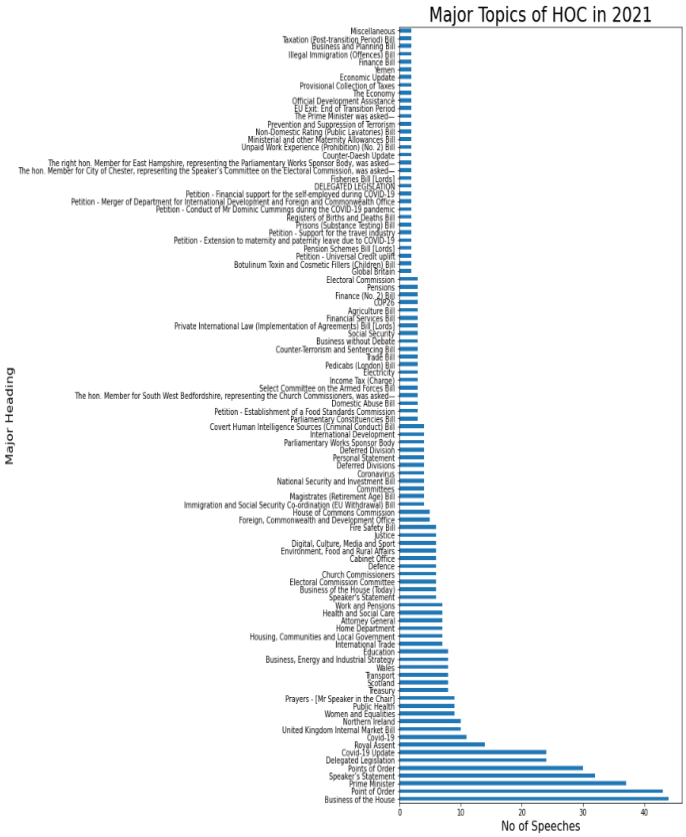


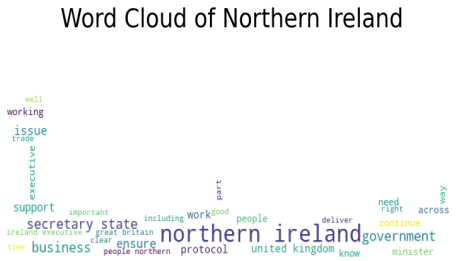
Fig. 2: Major Discussion in HOC

Table I shows the number of speeches for each major heading covered in HOC.

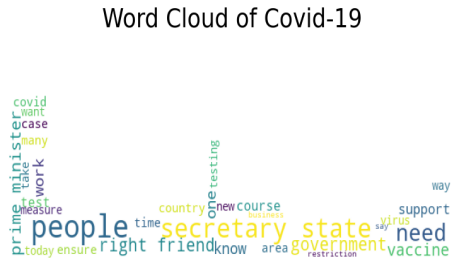
Major Heading	No of Speeches
Business of the House	44
Point of Order	43
Prime Minister	37
Speaker's Statement	32
Points of Order	30
Delegated Legislation	24
Covid-19 Update	24
Royal Assent	14
Covid-19	11
United Kingdom Internal Market Bill	10

TABLE I: Speeches in HOC

We looked more closely at the talks on Northern Ireland and COVID 19 and produced a word cloud of them in Figure 2. Since Covid-19 prevention and the Northern Ireland Protocol became the government's key priorities, important actions have been implemented. Like the HOC parliament image, the shape of the word cloud is masked.



(a) Northern Ireland



(b) Covid-19

Fig. 3: Word Cloud

E. Co-Occurrence Network

We grouped the major heading with Covid-19 debates to make the discussion on Covid-19 clear. For the Covid-19 speeches in 2020 and 2021, we plotted a co-occurrence network to better understand the parliamentary debate. We filtered the network by removing words that appeared less than 2 times in a speech. The Covid-19 pandemic was the main topic of research because it was so damaging and we wanted to know what our government was doing to deal with it.

We could observe the dense network for Covid-19 in 2021 in Figure 4, which shows a lot of discussions have happened for Covid-19 in parliament. The government is more concerned about safety, making the proper decisions, and taking steps to manage Covid-19, as evidenced by looking at the neighbouring words of "government" in Figure6. Priorities for the "virus"7

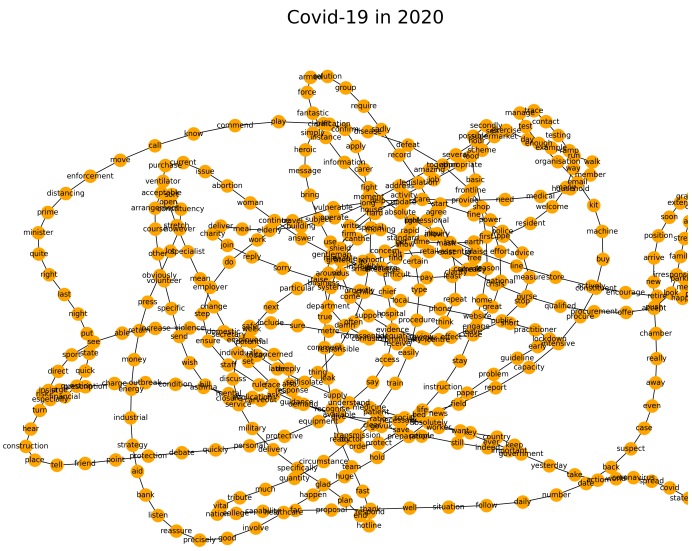


Fig. 4: Co-Occurrence Covid-19 Network 2020

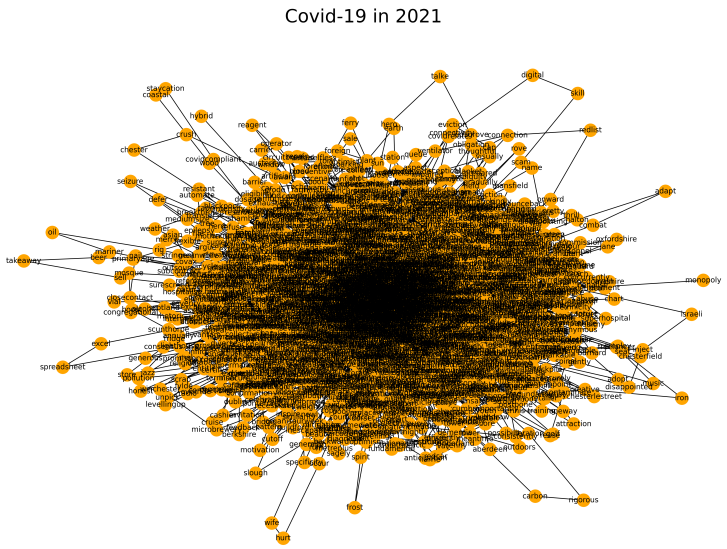


Fig. 5: Co-Occurrence Covid-19 Network 2021

included stopping its spread, providing aid, and learning how to stop it.

F. Topic Modeling: LDA

To identify abstract "topics" in a text as a collection of words that can be manually labelled, we used the LDA [23] approach. We first chose ten subjects for the Hansard talks. Next, we examined coherence in Figure 6 to see how semantically connected the topics are, and perplexity to assess how well our LDA model [23] has been trained.

We ultimately reduced our topics to six. We plotted the subjects to see what they looked like. We could still see some overlaps in the subjects using pyLDAvis, which produced

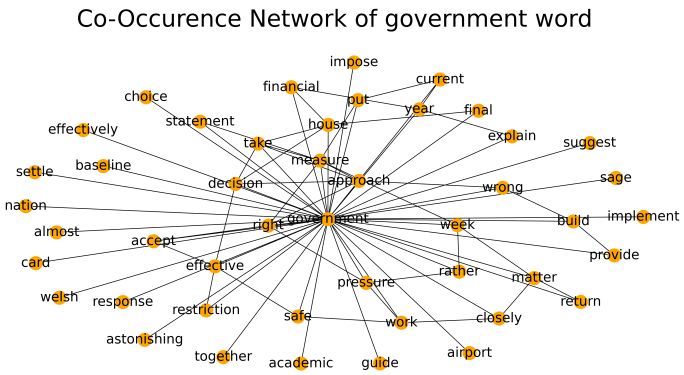


Fig. 6: Government Word

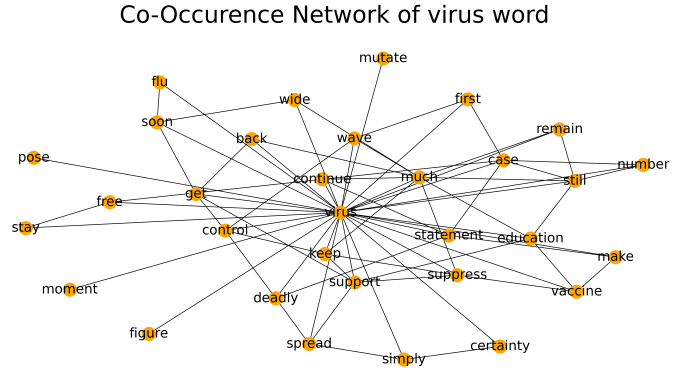


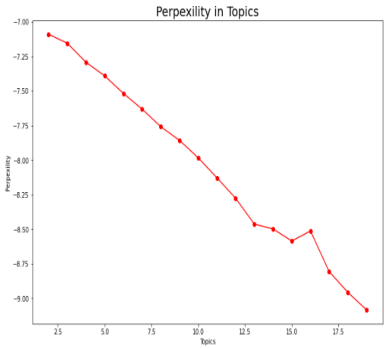
Fig. 7: Virus word

the visual representation, indicating that our model was not accurately predicting separate topics. To improve this we modelled it with Custom Topic Modeling

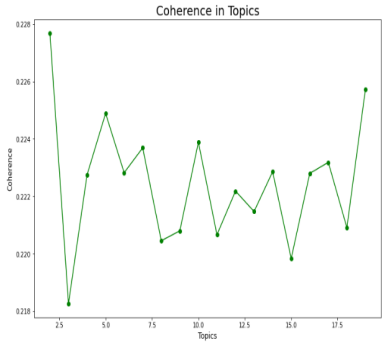
G. Custom Topic Modeling

At first, I used the BERTopic [25] library, a topic modeling method that makes use of transformers and c-TF-IDF to build dense clusters that enable understandable topics while preserving key terms in the topic descriptions. Guided, (semi-)supervised, and dynamic topic modeling is supported by BERTopic [25]. Even after adjusting the parameters, just 4 topics for legislative speeches in 2021 were found. We started separating the BERTopic [25] library's procedure and experimented by swapping out its modules to comprehend the topics in HOC debates. The processed text is transformed into an embedding to obtain the topics, and the vector embedding is then subjected to dimensionality reduction. Then, using unsupervised clustering, the reduced dimensions are grouped. Based on categorical tf-idf [31], which compares words in one cluster with terms in another cluster and assigns a score, the subjects are selected from the clusters. The topic of the generated cluster is determined by the words with the highest score.

First, we utilised SBERT [6] to obtain a sentence's embedding in the speech, then UMAP [14] to reduce the number



(a) Perplexity



(b) Coherence

Fig. 8: LDA Model Test

of dimensions. To obtain the topics following c-tf-idf [31], utilise HDBSCAN [12] for unsupervised clustering. Since we were able to identify fewer topics with multiple overlaps. We experimented by changing the HDBSCAN [12] clustering with KMeans [13] clustering and changing the sentence embedding with a document embedding. Finally, we were able to identify the 9 topics for the speeches in 2021, which demonstrate the difference in topics as shown in Appendix TableV by applying the document embedding from Cr5 Model [9] with UMAP [14] as the dimensionality reduction technique and KMeans [13] for clustering. We drew the elbow graph using the KMeans [13] clustering and discovered that the speech covers nine topics in Figure**??**. We chose document embedding because it allowed us to efficiently compare and locate subjects within other speeches by giving us an embedding for the entire speech as a document.

H. Semantic Similarity

a) Overlapping Words: We tried to measure the semantic similarity in the co-occurrence network by overlapping the neighbouring words for the years 2020 and 2021. We initially did with the government word in 2020 and 2021. Only 3 of the words overlapped in the two-time frames, making the semantic score 0.33 which is the lowest.

b) Semantic Speech Comparison: The comparison of the many speeches is grouped by major heading and is done for the 2020 and 2021 timeframes. By first utilising SBERT [6] to determine the speech's sentence embedding, we next attempted

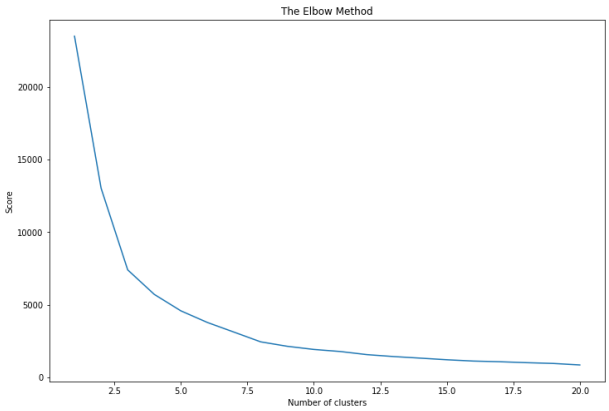


Fig. 9: Elbow Method for KMeans Clustering

to compare it to other talks under the same major heading by determining the cosine similarity i.e. the angle between the two speeches that make them comparable. It didn't yield very excellent outcomes. To determine the semantic change in the context of speeches within a year time frame, we finally used the document embedding via the Cr5 Model [9] to obtain the embedding of each speech within the major heading in 2021 and compared it with the document embedding from speeches within the same major heading in 2020 by using the parwise_cosine similarity from sklearn metrics.

Looking at the images, it was easy to see which speeches were most similar to those in 2021 and 2020 in Figure10, as well as which ones were least similar, according to Figure11.

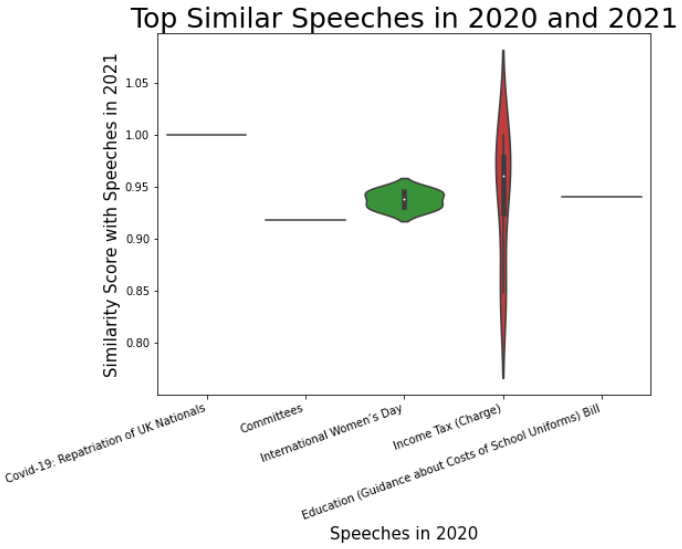


Fig. 10: Top 5 Semantic Similar Speeches

I. Orthogonal Procrustes

In 2020 and 2021, we investigated the Orthogonal Procrustes [28] of the Covid-19 speeches. To see how the words' semantics changed across these periods. The speeches in both time frames were subjected to word2vec [14] embedding. The

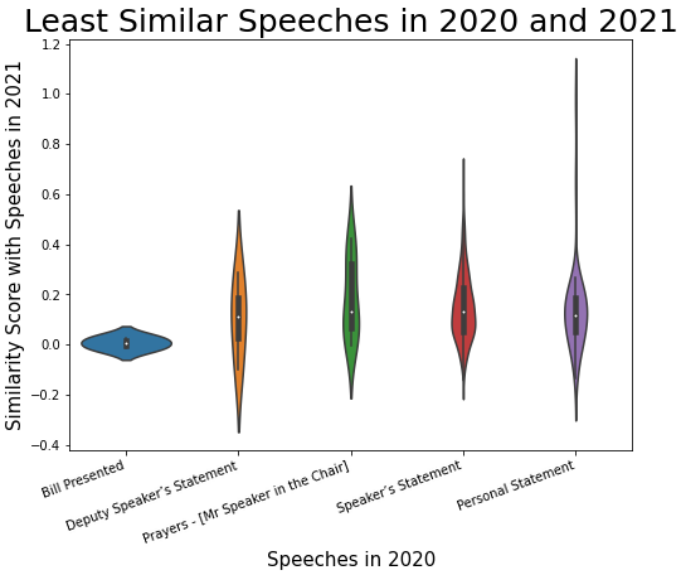


Fig. 11: Least 5 Semantic Similar Speeches

common words from the time frame vectors were stacked one on top of the other in an orthogonal alignment. The cosine distance was then obtained using the Scipy spatial module and used to determine the cosine similarity for the orthogonally aligned word.

In tableII we could observe the semantic change in the word's for Covid-19 speeches in the year 2020 and 2021 with the frequency of the words in both years.

Word	Semantic Change	Frequency 20	Frequency 21
not	0.735125	75	2001
people	0.621964	45	1691
right	0.751094	40	1200
get	0.709065	37	1028
friend	0.737107	36	1026
secretary	0.873057	32	952
state	0.885768	32	884
test	0.610378	31	864
testing	0.516114	27	843
need	0.659850	27	824

TABLE II: Semantic Change via Orthogonal Procrustes

J. Dashboard

All of the studies conducted for this research study resulted in the creation of a dashboard prototype, which includes all of the findings as static images. Plotly Dash is used to build the dashboard, and the Heroku server is used to deploy the application. The URL https://hoc-hansard.herokuapp.com/ provides access to the dashboard.

VI. Results

The Business of the House was the primary topic of discussion in HOC legislative debates in 2021, with a total of 44 speeches. Prime Minister, Covid-19 Update, and Covid-19 were some of the topics of conversation. The 35 total speeches for topics related to COVID reveal that our MPs are talking more about the Covid-19 pandemic crisis.

We could infer from the word cloud3 that the government is more concerned with its citizens, its economy, and its efforts to take vaccine- and test-based covid prevention measures. This hypothesis was further supported by the co-occurrence network, where it was clear that the government was worried about the citizens and their safety and was keeping a careful eye on the issue and taking precautions to stop the virus from spreading while keeping an eye out for its mutation. Figures4 and 5 made it evident that Covid-19 had a denser network in 2021 than in 2020, indicating that there had been an increase in the debate about it in 2021 in HOC.To better comprehend the Covid speeches, we also attempted to look at the word frequency count for 2021 in Figure12. We were able to identify the speakers with the highest frequency. Vaccine and test support are two of the HOC's most hotly debated subjects.

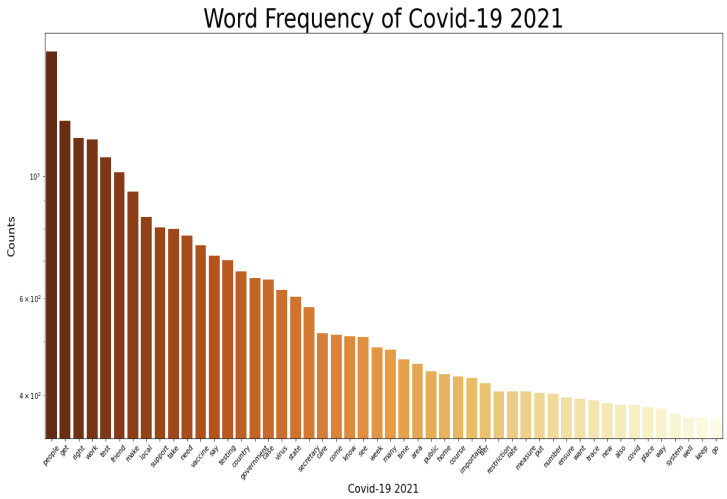


Fig. 12: Word Frequency in 2021

The LDA results for topic modeling weren't great because we could notice topic overlap in the pyLDAvis display in Appendix Figure 1. So we started putting the Custom Topic Modeling into practice. By first employing the Cr5 model to embed documents, followed by UMAP to reduce dimensionality and Kmeans to cluster the data. The conversation in HOC has been for individuals in terms of probability and forensics, and we were able to identify 9 subjects that characterise this, some of which revolved around the prince and the queen. We might spend days observing other topics, such as postmasters, proxies, Yemen, and Palestine. Other topics of tracing and flu in parliamentary speeches.

The word overlapping method was used to attempt to measure the semantic shift in parliamentary speeches, but the

results were less than encouraging. If we use the pairwise cosine similarity for the speeches of 2020 and 2021. We could easily observe in the violin plot Figure10 the highest similarity of speeches in "Covid-19: Repartition of UK Nationals", and" Committees" as 1, with the median semantic change in "International Women's Day" of 0.95. Since income tax changes every year, there has been a varied semantic similarity in the resemblance of the speeches on the "Income Tax (Charge)" topic. These were also verified by comparing the text in these speeches and compared manually with the text comparator to confirm these results from the algorithm. In figure11, you can see the talks that are the least comparable. The speeches from "Bill Presented," "Deputy Speaker's Statement," "Speaker's Statement," and "Personal Statement" were the least similar. The bills that are presented in the parliament, and the deputy speaker's speeches change every year, so these speeches are bound to show less similarity to the previous year.

Another technique to assess the semantic shift in a term over the two distinct periods of 2020 and 2021 is Orthogonal Procrustes [28]. The Orthogonal Procrustes [28] for Covid-19 speeches were examined, and we produced a thorough analysis of the term in TableII. It is clear from the chart that certain words, such as "not," exhibit the greatest semantic change in terms of frequency in 2020 and 2021; this could also be due to the context. If we pay close attention to word changes in 2021, the meaning shift will be significant as well as the frequency of words.

A few additional words, such as test and testing, also exhibit semantic changes of 0.61 and 0.51, respectively, indicating that the government's emphasis on testing has increased in 2021 with a frequency of 843 and a significant difference of 833 from the previous year. Although the word "get" showed a significant meaning change of 0.70. This may occur as a result of how certain terms are employed in certain contexts.

VII. Discussion

In conclusion, the clusters created by this method of topic analysis can identify significant linguistic variance in speeches that are situated within a larger topic. The descriptions (especially LDA) comprise thematically consistent and easily understandable sets of phrases, and the topic clusters are very well-coordinated. The most prevalent term in both the LDA and document embedding descriptions was typically the manual topical label that was most suited for each cluster. These findings support the validity of using document embeddings, the excellence of dimensionality reduction and clustering offered by UMAP [14] and KMeans [13], and the quality of the text embedding (semantically comparable utterances have been embedded next to one another).

The theoretical interpretation of what the LDA [23] reflect is ambiguous, as the data demonstrate. The difference between these concepts is not well defined, which may be a result of problems with consistent conceptualization that affect the accuracy with which the topics are produced. It's also feasible that specific texts in a corpus are only distinguishable because of variation, which calls for knowledge of the intertextual context. As shown by their superior performance on natural language understanding benchmarks, transformer models can represent or identify such high-level constructs. However, their performance has not yet reached human-level levels.

Given that we could compare the speeches within two years using pairwise cosine similarity, the semantic shift we were able to detect in the Hansard speeches over a brief period produced substantial results. The violin plot, whose change depends on the similarity measurements calculated, could be used to measure changes. This is also significant since we can compare the speeches' real text to identify changes and confirm whether the conclusions we have drawn are accurate or not.

The semantic change in a word for the two-time frames can be determined by computing the word2vec for the text and then aligning the two vectors orthogonally on top of one another. However, to orthogonally align the word2vec, we must have identical words in both time frames, which may not always be the case in legislative debates. We removed words that were unique to either period to make them similar; occasionally, this may have resulted in the elimination of words that were key to the MPs' statements. Plotly Dash was used to develop the dashboard, which was then deployed on the Heroku platform. The dashboard just displays static photos of the parliamentary speeches.

A. Limitation and Further Work

Along with the broad range of subjects and events discussed in parliament, the format and style may have harmed the outcomes. Due to the dialogical character of parliamentary debate. It can turn the speeches into a procedural.

The text embedding technique used by the transformer may have other drawbacks. Transformers are intended to input text sequences, and what sets them apart from earlier state-of-the-art embedding algorithms is the context-dependent word embedding that results from their bidirectional architecture. Therefore, it is uncertain if creating static vocabulary embeddings from single words as inputs is reliable and useful. This may have caused some cluster descriptions to include topically incorrect but conceptually related phrases that would have been incorporated in a cluster differently if their context had been taken into account.

The custom topic modeling reveals several topics where the phrases "divorce" and "forensic" are merged for Topic 1, demonstrating that there is room for improvement in the methodology. Since the algorithmic pipeline is modular, it would be simple to swap out the existing Cr5 model for document embedding with a better one that, for example, performs better on tasks involving natural language understanding and has been modified to avoid ingrained biases. We could even completely swap out the clustering technique in it.

In speeches' major headings' semantic change reveals information about the themes' changes for the upcoming year. We won't be able to capture a topic if it is mentioned elsewhere but not under the same major heading next. Future research could also investigate how particular changes in speech's semantics.

Although the Orthogonal Procrustes [28] is a useful tool for determining how the words have changed in meaning. However, if the words are not present in the same context over the following year, we won't be able to gauge the actual shift and some of the contexts in our MPs' speeches will be lost. The dashboard prototype is static, however, it might be enhanced to obtain dynamic data.

VIII. Conclusion

In the report, a computational method for analysing the discussions in the UK HOC parliament is described. The number of remarks made by our MPs based on the major heading has been highlighted, reflecting the facts surrounding the important themes of discussion in the parliament. Co-occurrence networks with the term as the node and its connections to other words within the same speech were plotted on the graph of speeches to better comprehend the context of the arguments. The topic modeling approach helps us understand the discussions by giving us a sense of the issues being discussed in HOC.

Calculating the cosine similarity of the speech, which is then compared with the identical speeches in different time frames, is a novel method for identifying a semantic shift in the speeches of parliamentary debates. The violin plot, which clearly outlines the discussion points that are exhibiting the notable changes and the ones that remain the same throughout time, allows us to measure and explain the change in detail. Orthogonal Procrustes alignment is another method, stated in the report with the word frequency giving us an indication of stressed words over time, of quantifying the semantic change in words over time.

The prototype of the dashboard was created to show all the results and analysis done on the Hansard speeches. The dashboard is deployed on the Heroku server and can be accessed by the URL https://hoc-hansard.herokuapp.com/

IX. Declarations

Declaration of Originality: I acknowledge that I am familiar with and understand the University of Exeter's policy on plagiarism, and I certify that this assignment is entirely original with the exception of the citations that show otherwise.

Declaration of Ethical Concerns: There are no moral concerns raised by this work. There are no human or animal subjects involved, and no personal information about human subjects has been handled. Additionally, no acts that could compromise security or safety have been done.

References

[1] The two-house system - UK parliament. (n.d.). Retrieved March 20, 2022, from https://www.parliament.uk/about/how/role/system/

[2] Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web (DETECT '11). Association for Computing Machinery, New York, NY, USA, 35–40. DOI:https://doi.org/10.1145/2064448.2064475

[3] Asmussen, Claus Boye, and Charles Møller. "Smart literature review: a practical topic modeling approach to exploratory literature review." Journal of Big Data 6, no. 1 (2019): 1-18.

[4] DiMaggio P, Nag M, Blei D. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics. 2013;41(6):570–606. https://doi. org/10.1016/j.poetic.2013.08.004.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[6] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] Church, K.W., 2017. Word2Vec. Natural Language Engineering, 23(1), pp.155-162.

[9] Josifoski, Martin, Ivan S. Paskov, Hristo S. Paskov, Martin Jaggi, and Robert West. "Crosslingual document embedding as reduced-rank ridge regression." In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 744-752. 2019.

[10] Özgür, Arzucan, Burak Cetin, and Haluk Bingol. "Co-occurrence network of reuters news." International Journal of Modern Physics C 19, no. 05 (2008): 689-702.

[11] Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." ACM Transactions on Database Systems (TODS) 42, no. 3 (2017): 1-21.

[12] McInnes, L., Healy, J. and Astels, S., 2017. hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), p.205.

[13] Kodinariya, T.M. and Makwana, P.R., 2013. Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), pp.90-95.

[14] McInnes, L., Healy, J. and Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[15] Chan, D.M., Rao, R., Huang, F. and Canny, J.F., 2018, September. t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data. In 2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (pp. 330-338). IEEE.

[16] Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review, 97(02): 311–331.

[17] Gary King and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. International Organization 57(3).

[18] Sascha Göbel and Simon Munzert. 2021. The Comparative Legislators Database. British Journal of Political Science.

[19] Rauh, Christian; Schwalbach, Jan, 2020, "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies", https://doi.org/10.7910/DVN/L4OAKN, Harvard Dataverse, V1

[20] Hege Fromreide, Dirk Hovy, and Anders Søgaard.2014. Crowdsourcing and annotating NER for Twitter drift. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

[21] Jurgita Kapovciute-Dzikiene and Algis Krupavicius. 2014. Predicting party group from the Lithuanian parliamentary speeches. Information Technology And Control, 43(3):321–332.

[22] Gavin Abercrombie and Riza Batista-Navarro. 2019. Semantic Change in the Language of UK Parliamentary Debates. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 210–215, Florence, Italy. Association for Computational Linguistics

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, 'Latent Dirichlet Allocation',J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[24] D. Angelov, "Top2Vec: Distributed Representations of Topics," arXiv:2008.09470 [cs, stat], Aug. 2020, arXiv: 2008.09470. [Online]. Available: http://arxiv.org/abs/2008.09470

[25] M. Grootendorst, "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics." 2020, version Number: v0.7.0.[Online]. Available: https://doi.org/10.5281/zenodo.4381785

[26] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In

Proceedings of the 27th International Conference on Computational Linguistics, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
[27] Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 66–76, Hong Kong, China. Association for Computational Linguistics.
[28] Wedin, P.Å. and Viklands, T., 2006. Algorithms for 3-dimensional Weighted Orthogonal Procrustes Problems.
[29] Anna Marakasova and Julia Neidhardt. 2020. Short-term Semantic Shifts and their Relation to Frequency Change. In Proceedings of the Probability and Meaning Conference (PaM 2020), pages 146–153, Gothenburg. Association for Computational Linguistics.
[30] HC Deb (20 January 2009). vol. 500, col. 1990. Available at:http://www.publications.parliament.uk/pa/hcdeb1990.
[31] Simeon, Mondelle, and Robert Hilderman. "Categorical proportional difference: A feature selection method for text categorization." In Proceedings of the 7th Australasian Data Mining Conference-Volume 87, pp. 201-208. 2008.

APPENDIX

speech	The House being met; and, it being the
display_as	Unknown
party	NaN
constituency	NaN
mnis_id	NaN
date	28984
time	NaN
colnum	1
speech_class	Procedural
major_heading	Preamble
minor_heading	NaN
oral_heading	NaN
year	1979
hansard_membership_id	NaN
speakerid	NaN
person_id	NaN
speakername	Unknown
url	NaN

TABLE III: Hansard Raw Speeches

speech	What assessment her Department has
date	4/29/2020
speech_class	Speech
major_heading	International Development
minor_heading	Covid-19: Developing Countries
speech_processed	assessment department made effect covid pandem...

TABLE IV: Hansard Processed Speeches

	Words
Topic 1	divorce, forensic, probation, holocaust, grenfell
Topic 2	duke, prince, philip, highness, queen
Topic 3	object, friday, november, resumed, october
Topic 4	proxy, certified, nominated, ayes, divided
Topic 5	aria, cooperative, selfemployment, kickstart, freelancer
Topic 6	postmaster,subpostmasters, greensill, lobbying, nazanin
Topic 7	sri, xinjiang, yemen, palestinian, saudi
Topic 8	fishery, fisherman, sovereignty, fish, unfettered
Topic 9	dental, bame, variant, tracing, flu

TABLE V: Topics from Custom Topic Modeling