Brexit: A Temporal Network Analysis of Sentiment in the Media Coverage

Luís Eduardo de Ávila Ballarati

March 2024

Contents

1 Introduction

1.1 Abstract

This dissertation investigates the evolving sentiment of Brexit coverage in The Guardian over a period spanning from 2007 to 2024. Utilizing a dataset comprising over one million articles, we extracted Subject-Verb-Object (SVO) triplets to construct a series of temporal sentiment networks centered around the term "Brexit". Our analysis focuses on key geopolitical events related to Brexit, examining how sentiment shifted in response to these events and identifying the main words associated with these shifts.

To achieve this, we developed a custom SVO parser and employed advanced natural language processing (NLP) techniques to filter and process the data. The resulting networks were analyzed using various centrality metrics to understand the prominence and influence of specific entities within the Brexit discourse. Before plotting the networks, we explored the lexical diversity and readability of the articles and understood the data, to ensure the reliability and coherence of the content analyzed.

The findings reveal significant in media sentiment around Brexit, particularly during major events such as the Brexit referendum, the invocation of Article 50, and the UK's official departure from the EU. This research contributes to the field of computational social science by demonstrating the utility of network analysis and NLP in understanding the dynamics of media sentiment over time.

1.2 Acknowledgments

Thanks are due to the following: My supervisor, Dr Nello Cristianini for his help and encouragement throughout this project. The Guardian for allowing me an API key to enable me to collect the articles used in this project. My family for their support and encouragement (and often conversations about English Grammar) during the project.

1.3 Introduction

Brexit, the United Kingdom's decision to leave the European Union, has been one of the most contentious and widely discussed political events of the 21st century [13]. Since the 2016 referendum, Brexit has dominated political discourse, significantly impacting the economic, social, and political landscapes in the UK and across the globe [18]. Understanding how media outlets like The Guardian have reported on Brexit provides insights into public sentiment, political bias, and the evolution of narratives surrounding this historic decision [31]. This dissertation investigates how the sentiment toward Brexit, as represented in The Guardian, has changed over time, utilizing a network analysis approach that focuses on Subject-Verb-Object (SVO) triplets extracted from over a million articles [29].

The central problem addressed in this dissertation is the need for a comprehensive understanding of how media sentiment toward Brexit has evolved over time. Traditional content analysis methods often overlook the nuanced shifts in sentiment and the dynamic relationships between key actors and concepts [22]. This research aims to fill this gap by employing an approach that combines natural language processing (NLP) and network analysis to track and visualize these changes. The specific objectives of this study are to extract and analyze SVO triplets from The Guardian articles to understand the relationships between different words related to Brexit; to construct temporal sentiment networks centered around the term "Brexit" and analyze how these networks change in response to key geopolitical events [41]; and to evaluate the readability, lexical diversity, and sentiment distribution of the articles to ensure the reliability of the data and the robustness of the analysis [17].

Media coverage plays a crucial role in shaping public opinion and policy decisions, especially in politically charged contexts like Brexit [19]. By analyzing the sentiment and framing of news articles over time, researchers can uncover biases, trends, and shifts in public discourse that may influence or reflect broader societal attitudes [34]. The Guardian, as one of the leading newspapers in the UK, provides a rich corpus of articles for this type of analysis. Understanding how the sentiment towards Brexit has evolved in The Guardian not only sheds light on the newspaper's editorial stance but also offers a window into the changing public mood and the media's role in framing political debates.

This research builds on a growing body of work in the fields of natural language processing, sentiment analysis, and network analysis. Previous studies, such as [35], have explored the predictive power of sen-

timent analysis in financial news, while [41] demonstrated the utility of network analysis in understanding media narratives during the US presidential elections. Similarly, [22] analyzed the impact of the Fukushima disaster on media coverage, employing big data techniques to uncover shifts in sentiment and framing. This dissertation extends these methodologies to the context of Brexit, combining SVO triplet extraction with temporal network analysis to provide a more nuanced understanding of media sentiment dynamics.

By integrating insights from these previous studies, this research aims to advance the state of the art in sentiment and network analysis. Specifically, it seeks to improve upon existing methods by focusing on the temporal dynamics of sentiment, rather than static snapshots, and by incorporating a richer set of linguistic features through the use of SVO triplets.

This dissertation is structured as follows: Chapter 2 provides a detailed description of the data collection and preparation process, including the use of the Guardian API and the development of the SVO parser. Chapter 3 outlines the methodology for constructing and analyzing the sentiment networks, while Chapter 4 presents the results of the analysis, highlighting key trends and shifts in sentiment over time. Chapter 5 discusses the implications of these findings, situating them within the broader context of media studies and political communication. Finally, Chapter 6 concludes the dissertation with a summary of the main contributions and suggestions for future research.

1.4 Related work

The development of a network tailored to news articles is a multifaceted problem encompassing aspects of NLP and data analysis. One key aspect is the extraction and analysis of SVO triples from news articles to understand the relationships between concepts, like people, companies, and institutions. By identifying subjects, verbs, and nouns triples, one can gain a deeper understanding of the relationships extracted in the text. This approach aligns with the study by [36], which constructed a social network of companies mentioned in financial news articles and built links between co-mentioned organizations. Furthermore, the study by [35] examined a predictive machine learning approach for financial news articles analysis using textual representations such as bag of words, noun phrases, and named entities.

The project's goal to extract meaningful insights from news data is supported by the study of [16], which aims to determine the state of the extant literature in NLP in accounting, auditing, and finance to guide future research efforts. Additionally, the study by [24] investigates how semantic orientations can be better detected in financial and economic news by accommodating the overall phrase-structure information and domain-specific use of language.

The aspect of understanding the dynamics in markets and its implications for investment decision-making is complemented by the research of [8], which explores the relationship between big data, news diversity, and financial market crash. This study is relevant as it delves into the impact of news diversity on financial markets, aligning with the project's aim to contribute to the understanding of market dynamics.

This project approach also builds on the idea presented in [22], which focuses on analyzing the impact of the Fukushima disaster on media coverage of nuclear power using big data analysis of over 5 million English-language science articles between May 2008 and December 2013. The methodology involves extracting narrative networks, sentiment time series, association networks, and SVO triplets to understand how scientific concepts are framed in the media. The study reveals a significant shift in media sentiment towards nuclear power post-Fukushima, moving from a positive portrayal to a negative one, emphasizing risks associated with nuclear power. The analysis includes monitoring attention, sentiment, associations, actors, and actions related to nuclear power before and after the incident, showcasing changes in media discourse.

The project focused on analyzing the framing of science-based issues in mass media by examining how scientific concepts and associated actors are mentioned. The articles were labeled as science articles using Support Vector Machines (SVM) trained on the New York Times and Reuters corpora. They computed attributes like salience and sentiment over time, mined association rules between items to discover their relationships, and extracted SVO triplets to understand actions related to the items. Data processing was done using Apache Hadoop[1], and references to scientific topics, universities, and diseases were extracted from science articles using the GATE architecture, by [15]. Time series data was generated using ElasticSearch[2].

[1] https://hadoop.apache.org
[2] http://www.elasticsearch.org/

The methodology allowed for a detailed analysis of media coverage of science topics, particularly nuclear power before and after the Fukushima disaster.

Another paper this work is related is [40], which focus on automated analysis of the US presidential elections using Big Data and network analysis. They propose a Big Data approach to extract information from news articles related to the elections, forming a semantic graph based on SVO triplets. This is a similar to approach to mine, changing only the content of the information analysed. The methodology involves automated extraction of triplets, creating a network analyzed through graph partitioning, centrality, assortativity, hierarchy, and structural balance. The study reveals insights such as the split between Republican and Democrat camps, central nodes, and media reporting trends favoring Democrats. The methodology includes tools like graph partitioning, centrality measures, and network analysis techniques.

The data for the study on the automated analysis of the 2012 US presidential elections was found in 130,213 news articles related to the elections from 719 news outlets, using state-of-the-art NLP and Artificial Intelligence techniques to extract information about key actors and their relations in the media narrative of the elections. The study involved the automated extraction of SVO triplets from the news corpus, forming a semantic graph that encodes relations among all actors in the corpus, providing a detailed network representation of the campaign coverage.

[41] discuss a methodology to extract narrative networks from large corpora, focusing on SVO triplets to find key entities and actions. It presents experiments analyzing crime and political stories, emphasizing high precision in information extraction. The study uses network centrality measures like Betweenness Centrality, In-Degree, Out-Degree, HITS, and PageRank to identify central entities in crime data for 2002. The methodology involves spectral analysis to rank entities based on political spectrum alignment. Validation includes statistical tests and performance evaluations of modules used in the pipeline. The study aims to automate network creation, validate results, and analyze text properties, such as central entities and relational structures.

Another article that is related to this work is [39]. It presents a system for large-scale Quantitative Narrative Analysis (QNA) of news data, focusing on identifying key actors and actions in news articles. The system characterizes actors by studying their network position, time series properties, subject/object bias, and associated actions. It automates the extraction of Subject-Verb-Object triplets from news data, generating semantic graphs to capture narrative structures. The system ranks actors based on network centrality measures and uses tools like Cytoscape[3] and Gephi[4] for analysis. It aims to automate labor-intensive tasks in social science research, offering scalability and new insights into news narratives.

[38] discuss a methodology for large-scale quantitative narrative analysis in digital humanities and social sciences. It involves automatically transforming a corpus into a network by extracting key actors and objects, linking them, and analyzing the network to extract information about the actors. The methodology includes steps like extracting SVO triplets, weighting entities and actions, filtering triplets, creating networks, and analyzing entities' centrality and subject/object bias. The software pipeline developed for this analysis reuses existing tools and can scale to large corpora, demonstrated on texts from the Gutenberg Project. The study correctly identifies central actors in narratives and reveals insights about their roles and biases.

When it comes to analysing news, [12] proposes a method to measure the state of the economy through textual analysis of business news, using a topic model to summarize news content and quantify attention to different themes over time. By incorporating news attention estimates into statistical models of economic time series, the text-based inputs accurately track various economic activity measures and provide additional forecasting power for macroeconomic outcomes. The model also helps identify news-based narratives underlying shocks in economic data, offering a new perspective on understanding economic conditions. It uses topic modeling to summarize news themes and track attention to topics over time, which can help in understanding relationships and evolving events, akin to my interest in analyzing how relationships evolve over time or in response to events.

In conclusion, the project's focus on developing a network tailored to news articles and extracting meaningful insights from news data aligns with existing research in sentiment analysis and textual analysis. These studies provide a foundation for the project's objectives and methodologies.

[3] https://cytoscape.org
[4] https://gephi.org

1.5 Data description and extraction

The data used for this project was from the Guardian Open Platform[5], a public web service for accessing all the content the Guardian creates, categorised by tags and section. Through an API, 48 datasets were collected divided into 6 tags: Business (11), Culture (9), Science (6), Tech (6), World (8) and Politics (8). The data was scraped using Google Colab Pro[6] and 3 different processing units: CPU (Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13GB of RAM), Nvidia A100 GPU and CPU using High Ram (54GB of RAM). The table 10 found in the appendix provide information of almost all (44/48) data extracted, comparing the performance per processing unit.

The time the A100 took to scrape the information was on average approximately 32 minutes and the dataset had on average 13678 articles. For the CPU, it took on average approximately 81 minutes and the dataset had on average 29249 articles. Lastly, the CPU using Colab's High Ram feature, on average it took approximately 52 minutes and the dataset had on average 23647 articles. Below, the table 1 display the time it took per Section, including all units:

Section	**Average Dataset Size**	**Average Time (m)**
Business	23760.57	63.57
Culture	34900.60	90.27
Politics	25863.08	58.97
Science	7026.50	16.46
Technology	12574.00	29.71
World	30519.88	89.75

Table 1: Average Metrics by Section

When we group by Section all the processing units, it is evident that the "World" and "Culture" tags were the ones with the most articles associated. They were also the ones that took the longest to scrape.

The information collected ranges from 00:00:00 19/07/2007 to 00:00:00 02/07/2024. The articles were collected since 2007 to provide enough range for any analysis in between 2007 and 2024. The date was pick randomly. After collecting the data, the dataset had 1093276 rows divided in 7 columns: "type", "sectionId" and "sectionName", "webPublicationDate", "webTitle", "webUrl" and "content". "Type" indicates the kind of content extracted (article, liveblog, audio and interactive). 98.5% were articles. "SectionId" and "SectionName" were the tags each article was marked, often the same. There were inner tags in the dataset as well: for example, an article extracted as "culture" can have a "music" tag. The most common subtags in the dataset can be found below:

[5]https://open-platform.theguardian.com

[6]When you create a new notebook in Colab, you are given access to a virtual machine with a specific CPU and RAM configuration (https://colab.research.google.com).

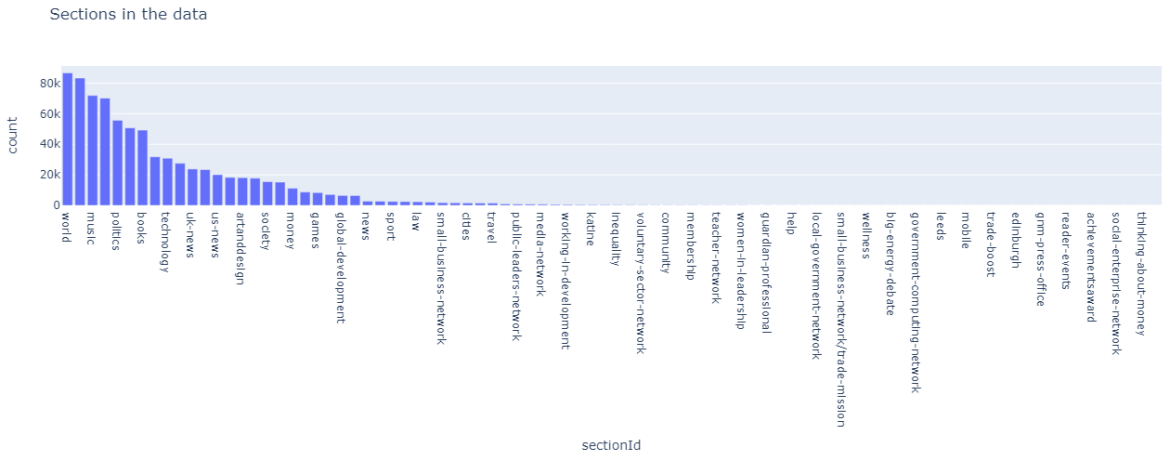

Figure 1: Count of SectionId

The "webTitle" refers to the title of the article and the "webUrl" the article URL. The "webPublicationDate" informs when the article was published. Finally, the "content" is the article body, including only the text. Each article has on average 739.51 words.

1.6 Data preparation

To analyse possible inconsistencies in the data, some techniques were applied. First, empty rows were checked. 47127 rows had at least one value, out of 7, empty. Since the article must have at least the date, content and section not empty, and the rows with at least one empty feature represented only 4.31% of all data collected, they were removed, leaving the dataset with 1046149 articles.

The next step was to check for possible outliers. To do that, some Box Plots were plotted to help identify the distribution of data [33]. This chart is based on 5 metrics: The minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The following is the box plot of the number of words per article:

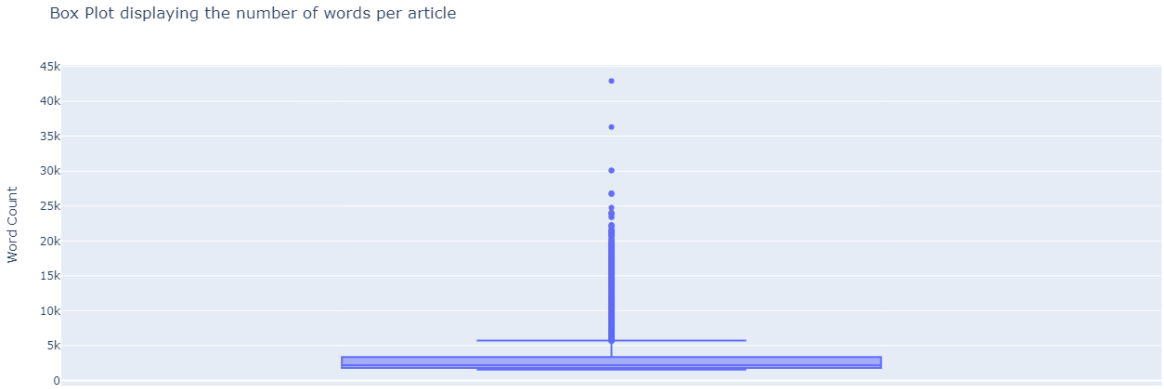

Figure 2: Box Plot of Word Count (including all data)

50% of the articles have between 397 (Q1) and 866 (Q3) words. The median is 615 words. 75% of the

articles have less then 3368.25 words. The inter quartile range (IQR) is 469 (Q3 - Q1).

The whiskers extend from the edges of the box to the minimum and maximum values within a certain range. The formula for the whiskers are:

$$\text{Lower Whisker} = Q1 - 1.5 \times \text{IQR} \tag{1}$$
$$\text{Upper Whisker} = Q3 + 1.5 \times \text{IQR} \tag{2}$$

Equation 1: Equations for Lower and Upper Whisker

Any value lower than the lower whisker or higher than the upper whisker is considered an outlier. When we exclude the outliers of the analysis, we end up with 996968 articles. 49181 articles were above or below the limits and were considered outliers. Below is displayed the box plot excluding the outliers.

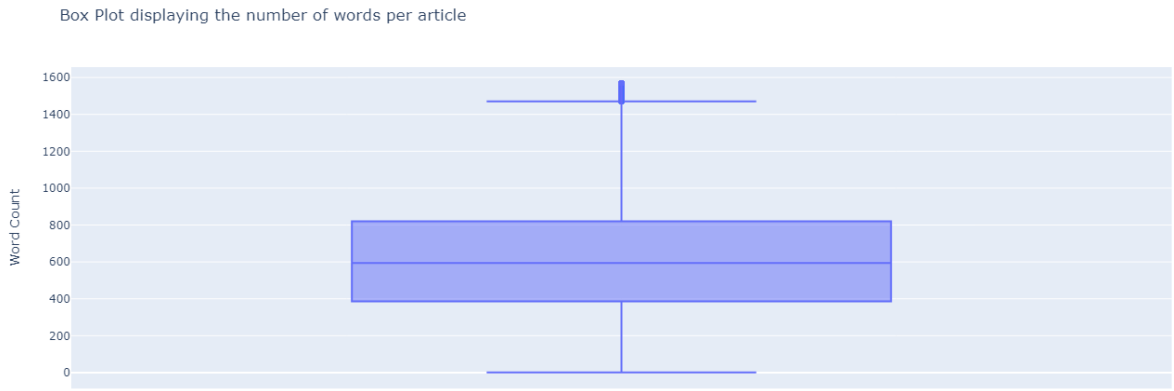

Figure 3: Box Plot of Word Count (excluding outliers)

Out of the 996968 articles, 50% of the articles have between 386 (Q1) and 820 (Q3) words. The median is 594 words. 75% of the articles have less then 820 words. The inter quartile range (IQR) is 484 (Q3 - Q1). Also, in this situation, the lower bound is equal to the minimum value in the dataset (both being 1). This indicates that the lower whisker in the box plot is positioned at 1, meaning there are no outliers below this value according to the box plot criteria (i.e., no values fall below Q1 - 1.5 * IQR). Consequently, values at or above this lower bound should not be excluded, as they are not considered outliers based on the applied definition.

Articles with just 1 word are probably a mistake and, although they should not be excluded from the dataset, they won't be used to provide the subject-verb-object triples.

Excluding the outliers would result in a 4.7% reduction of the data. Since some articles were already deleted due to empty values, excluding more values just because they are large would not make sense.

After dealing with the outliers, the presence was duplicates were checked. Since the collection of the data step was built over key specific dates, the dataset had no duplicates.

1.7 Data analysis and sanity check

This section will use three methods to understand more about the articles and check if they are reliable and make sense. The Flesch Reading Ease and Flesch-Kincaid Grade Level are widely used metrics for evaluating the readability of English text. The Flesch Reading Ease score is calculated based on the average sentence length (in words) and the average number of syllables per word in a given text. The resulting score tipically

ranges from 0 to 100, where higher scores indicate easier-to-read content [17]. For instance, a higher score corresponds to a text that is simpler and more accessible to a broader audience, while a lower score suggests a text that is more complex and challenging to read.

The Flesch-Kincaid Grade Level is closely related but provides a readability score in terms of U.S. school grade levels. This metric estimates the years of education required to understand the text [21]. For example, a Flesch-Kincaid Grade Level of 8.0 suggests that the text is understandable by an average eighth grader. Together, these metrics are valuable tools for assessing the accessibility of written content, whether in academic, professional, or public domains and will be used to assess the accessibility of The Guardian articles.

The Flesch Reading Ease Score is based on the average sentence length (ASL) and the average number of syllables per word (ASW):

$$\text{Flesch Reading Ease} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}) \tag{3}$$

Equation 2: Flesch Reading Ease Formula

Where ASL is the total number of words divided by the total number of sentences and the ASW, the total number of syllables divided by the total number of words.

The score typically ranges from 0 to 100. Higher scores indicate easier readability. 90-100 means very easy to read (easily understood by an average 11-year-old student). 60-70, standard readability (easily understood by 13- to 15-year-old students) and 0-30 means very difficult to read (best understood by university graduates).

The Flesch-Kincaid Grade Level formula also uses ASL and ASW:

$$\text{Flesch-Kincaid Grade Level} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59 \tag{4}$$

Equation 3: Flesch-Kincaid Reading Level

The score represents the grade level needed to comprehend the text. 1.0 means the text is easily understood by a first grader, 8.0 is easily understood by an eighth grader, 12.0 is easily understood by a high school senior and scores above 12 indicate that the text is best understood by college students.

10000 random articles from the corpus were select to display the tests:

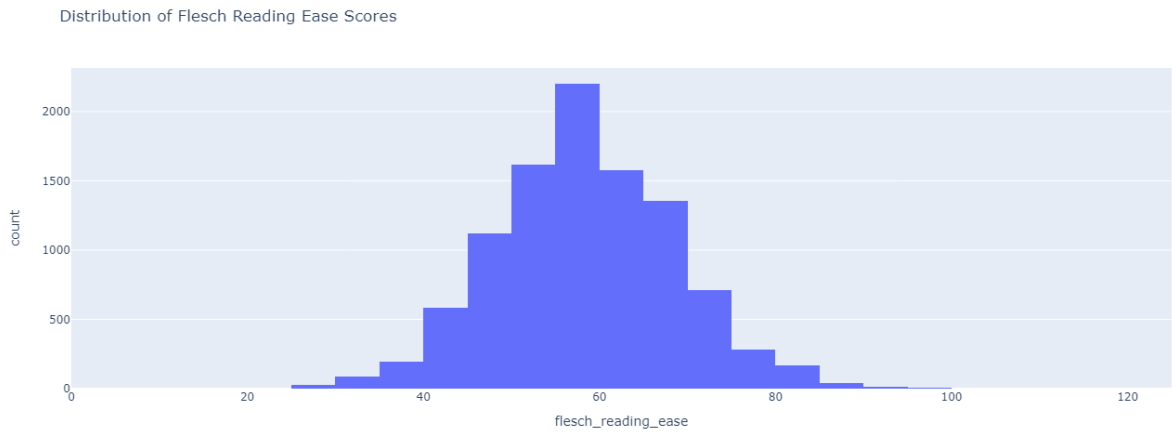

Figure 4: Distribution of Flesch Reading Ease Scores

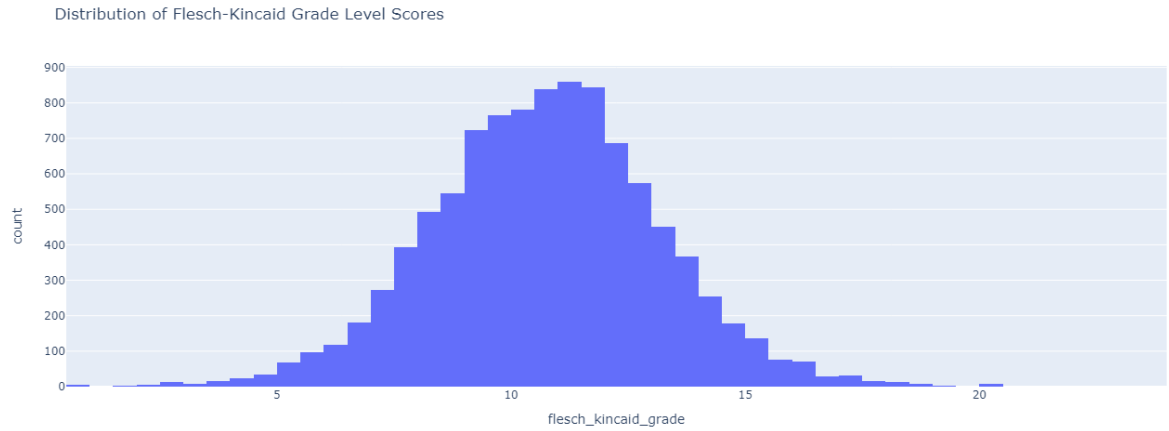

Figure 5: Distribution of Flesch-Kincaid Grade Level Scores

The distribution of the Flesch Ease score ranges between 20-100. This means that the data (at least the sample) ranges between "Very difficult to read" and "Very easy to read", but the main articles (93.92%) were tagged with "Standard readability". The Flesch Kicaid Grade Level result shows that 60.85% of the articles analysed would be easily understood by a high school senior, college students or above.

The next method applied was sentiment analysis. Sentiment analysis is a technique for determining the emotional tone within a body of text. When analysing news, the sentiment over an article might indicate bias over a theme and opinionation. One of the most popular tools for sentiment analysis is the Valence Aware Dictionary and sEntiment Reasoner (VADER) [20]. VADER provides sentiment scores across four categories: Positive, Negative, Neutral, and Compound [20].

Positive, Negative and Neutral scores represent the proportion of the text that corresponds to each sentiment category, with values ranging from 0 to 1. Higher scores indicate a stronger presence of the respective sentiment within the text. The compound score is a normalized, weighted composite score that combines the positive, negative, and neutral scores into a single metric. The compound score ranges from -1 to 1, where a score between -1 and -0.5 indicates a strongly negative sentiment, a score between -0.5 and 0 indicates a negative sentiment, a score between 0 and 0.5 indicates a positive sentiment and a score between 0.5 and 1 indicates a strongly positive sentiment [20].

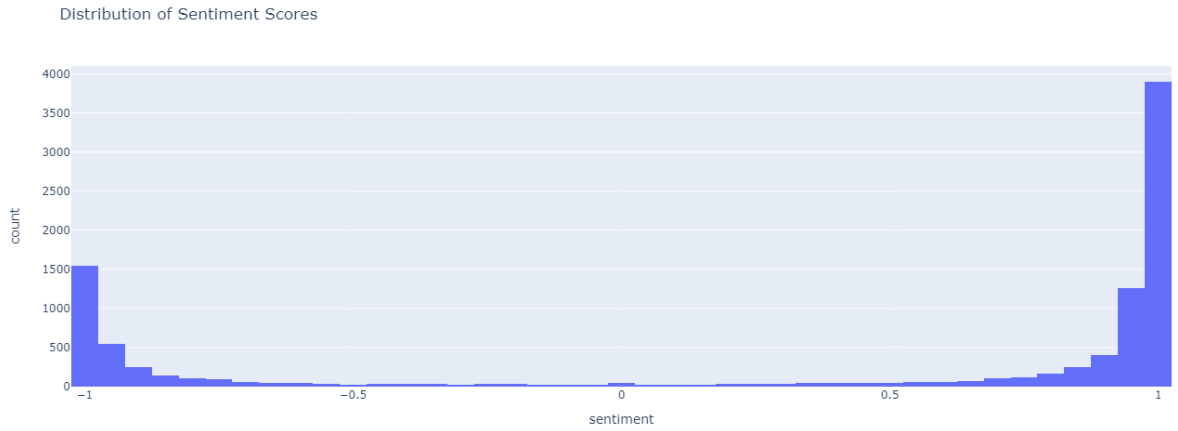

Figure 6: Distribution of Sentiment Scores using VADER

In the analysis 10000 randomly selected articles, the sentiment scores predominantly fall into the categories of strongly positive (compound score close to 1) and strongly negative (compound score close to -1). This distribution suggests that the articles in the sample tend to be highly opinionated, with the majority expressing either a strong positive or strong negative sentiment towards the subjects discussed.

The last metric used was to evaluate the lexical diversity of the articles. Content length and lexical diversity are essential metrics for assessing the quality and comprehensiveness of textual data. Content length refers to the total number of characters or words in a text and is a key indicator of verbosity and depth. Longer texts typically provide more detailed information, while shorter texts might be more concise but potentially less informative [11].

Lexical diversity measures the variety of vocabulary used in a text, calculated as the ratio of unique words (types) to the total number of words (tokens). This metric is crucial for understanding the richness and variety of language within the text. A higher lexical diversity score suggests a broader and more varied vocabulary, which is often associated with well-crafted, engaging, and informative content [25].

Lexical diversity scores can range from 0 to 1. Scores near 0 indicate low diversity, implying repetitive use of words, which might suggest limited vocabulary or a narrow focus [5]. Conversely, scores near 1 indicate high diversity, with a wide range of vocabulary and minimal repetition, often characteristic of well-written literature and academic papers.

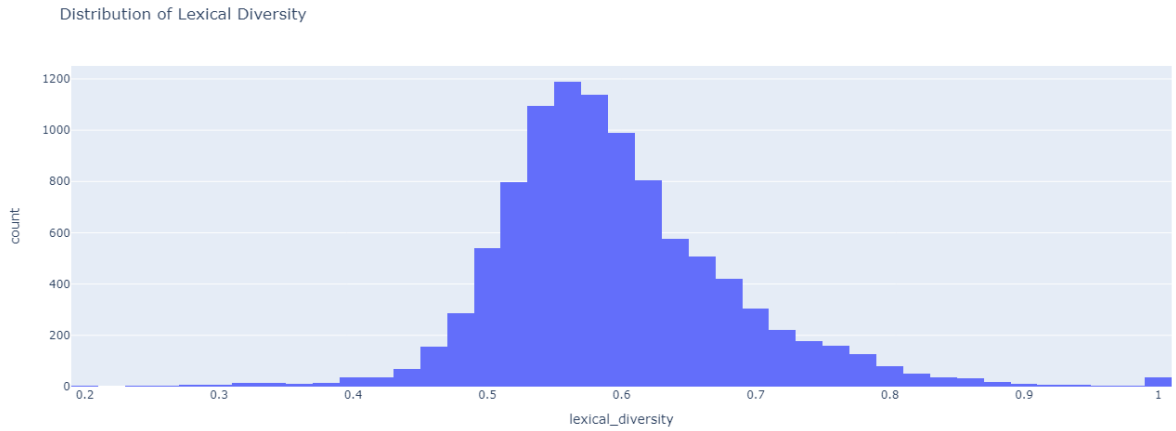

Figure 7: Distribution of Lexical Diversity

In the analysis of the sample, the lexical diversity scores ranged from 0.4 to 0.8, suggesting moderate diversity. The distribution of content length showed that most articles contained approximately 4365 characters, following a right-skewed distribution pattern, which is typical of datasets where longer, more detailed texts are prevalent.

2 Methodology

2.1 Pipeline overview

The methodology applied for the project was the following:

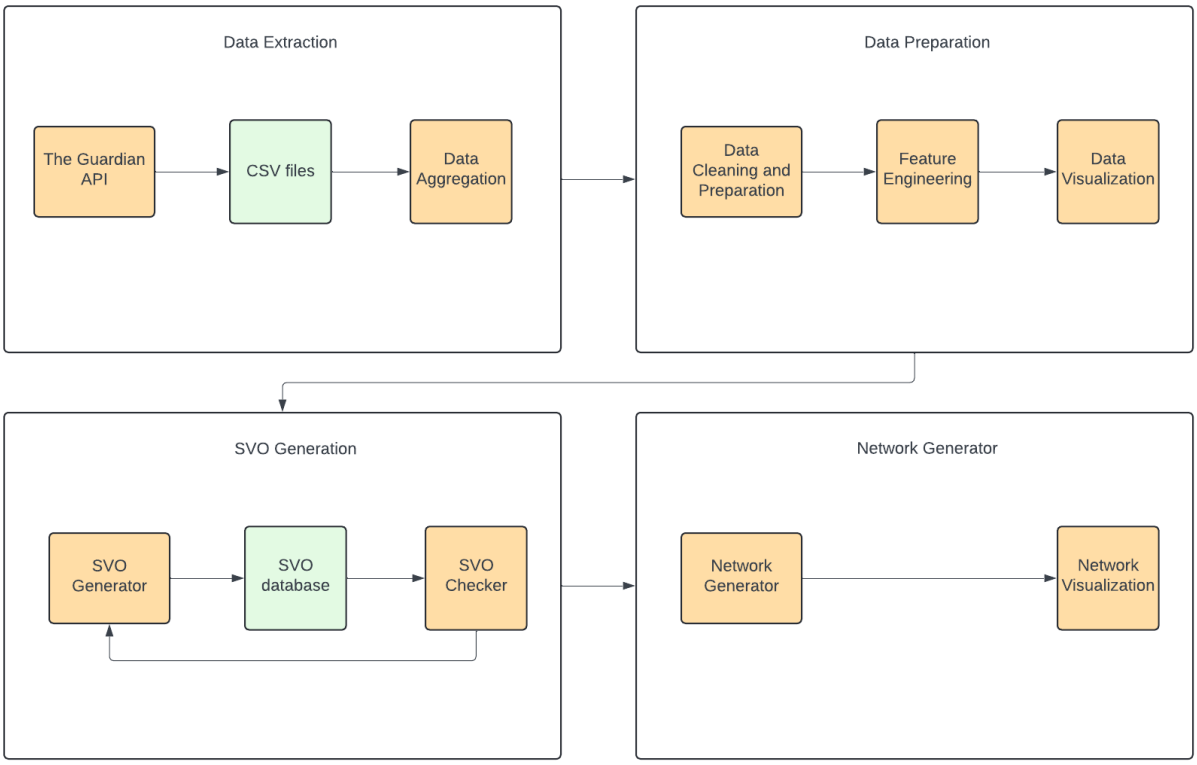

Figure 8: Software/Data Pipeline

Using one API key, the data was extracted and stored in csv files locally. All datasets were merged into one and also stored. After that, the data was cleaned and new columns were introduced to facilitate understanding and manipulating the data. Charts were plotted to help better visualize the information to be dealt with. After that, all the articles were parsed and stored in a new dataset, with only Subject-Verb-Object triplets. A SVO Parser based on Python's Library SpaCy[7] and adapted for this case, was used to collect the triplets from the raw content. With the triplets at hand, the triplets were filtered and then the FP-Growth Algorithm was applied to rank the most common words associated with the central node [6]. "Brexit" was the word used as a central node. With this result, the network was plotted.

2.2 SVO Parser overview

59558455 triplets were extracted from the corpus using the methodology described below. Out of this, 27428837 were unique. There were 76815 different words among the triplets. The number is high because there were a lot of proper nouns. The triplets were stored in pickle datasets to keep formatting. The dataset had three columns, the sectionName, the webPublicationDate and the SVOs. The extra columns were kept to be used as filters when plotting the networks.

2.3 SVO Parser description

2.3.1 Sentence processing

The methodology for parsing the triplets was the following:

[7]https://spacy.io

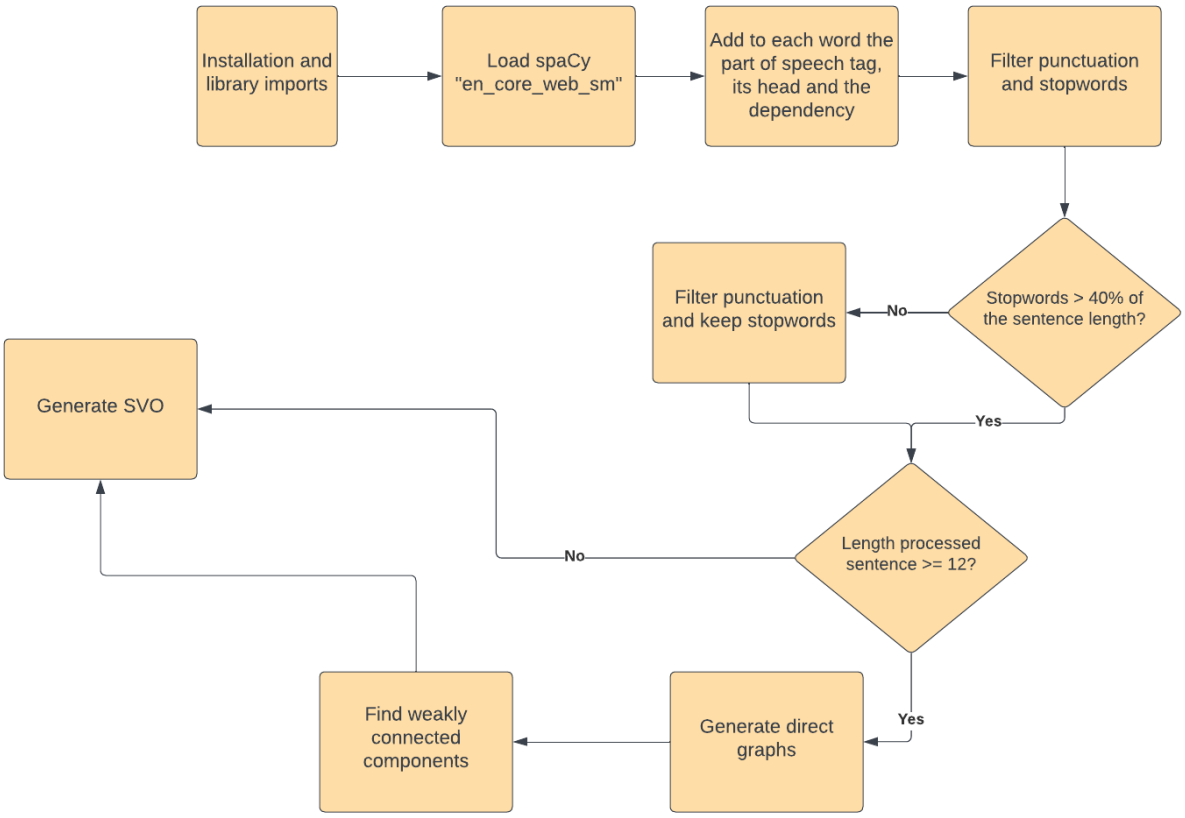

Figure 9: SVO generator Pipeline 1

SpaCy library was chosen to tag the words. After installing and importing it, the next thing to do was load a pre-trained language model in English, "en_core_web_sm". This model is a pre-trained statistical model specifically designed for processing English text.

The model includes components like tokenization and part-of-speech (POS) tagging. Next, we assigned to each word the part-of-speech, its head and its dependency. Part-of-speech tagging is the process of labeling each word in a sentence with its corresponding part of speech, such as noun, verb, adjective, etc [26]. The "head" of a word in dependency parsing refers to the word that governs or controls it syntactically. In other words, the head of a word is the word upon which it depends directly in the sentence's structure. Dependency refers to the relationship between a word (dependent) and its head in the syntactic structure of a sentence. This relationship is described using a dependency label, which specifies the type of grammatical relationship (e.g., subject, object, modifier) [30]. For example, the sentence "The quick brown fox jumps over the lazy dog" would become:

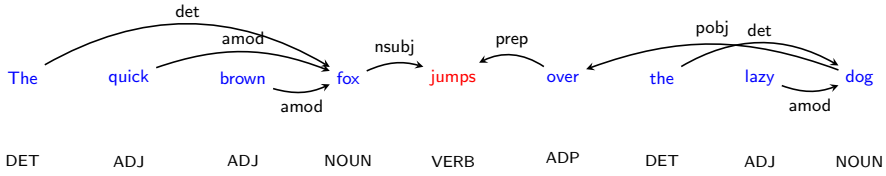

Figure 10: Semantic structure of a sentence

The arrows represent the word linked to its head, with the dependency written in the edge. Below

the sentence, for each word we have the part-of-speech tag. The red word ("jumps") is the ROOT, so its dependency is to itself. For example, the word "quick" belongs syntactically to the word "fox", that syntactically belongs to the ROOT. The dependency connecting "quick" and "fox" is "amod", which means adjectival modifier.

With the sentence processed and tagged with the POS, dependency and head, the punctuation and stopwords are removed. These words include articles, prepositions, conjunctions, and other frequently occurring words that generally do not carry significant meaning or contribute to the core content of the text.

When removing the stopwords, it was found that some sentences lost most of its words and the parser was not able to extract the triplet. To deal with that, a rule was imposed: if, after removing the stopwords of a sentence, less than 40% of the words were kept, only the punctuation would be removed from the sentence. The number 40% was decided after several iterations and analyses contemplating the final SVO triplet. If after removing the stopwords the sentence would keep most of its words, then it would proceed normally.

The next step was to check the processed sentence length. The step after this was to create a direct graph and collect all its weakly connected components. A weakly connected component (WCC) is a subgraph of the original graph where all vertices are connected to each other by some path, ignoring the direction of edges [4].

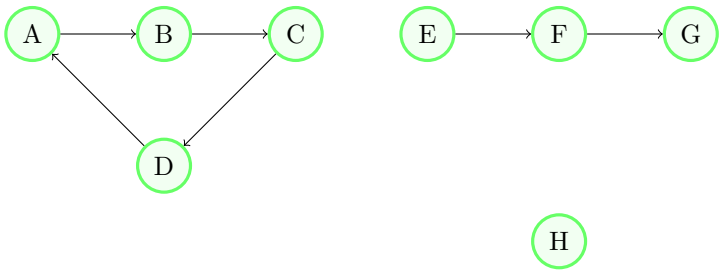

Figure 11: Example of weakly connected components

In the example above, a sentence "A B C D E F G H" would be separated in 3 WCC, and for each one a SVO triplet would be extracted. In large sentences, breaking the information like this enabled an extraction more accurate and that carried more meaning than treating the sentence as one. This approach, however, was not the best when it came to small sentences. Breaking small sentences would destroy the meaning of the sentence and the triplet in the end would carry irrelevant information. To solve that, it was decided that only processed sentences longer than 12 words would be treated as direct graphs and later split in its weakly connected components. The number 12 was found again after several iterations and analyses of the results. If the processed sentence had less than 12 words, it would go directly to the next step, skipping the graph step. That step was designed because, most SVO extractors tested in this study were not collect all the subjects, verbs and objects in a long sentence and information was being left out. To illustrate with a real example, we generated two graphs of the sentence (collected from a business article of the corpus):

Private jet providers are experiencing "unprecedented demand" from wealthy customers seeking to avoid the "mosh pit" of commercial flights on autumn getaways as coronavirus travel restrictions ease.

The first one, no punctuation or stopwords were removed and the result is a connected graph (every pair of vertices in the graph is connected) [4]:

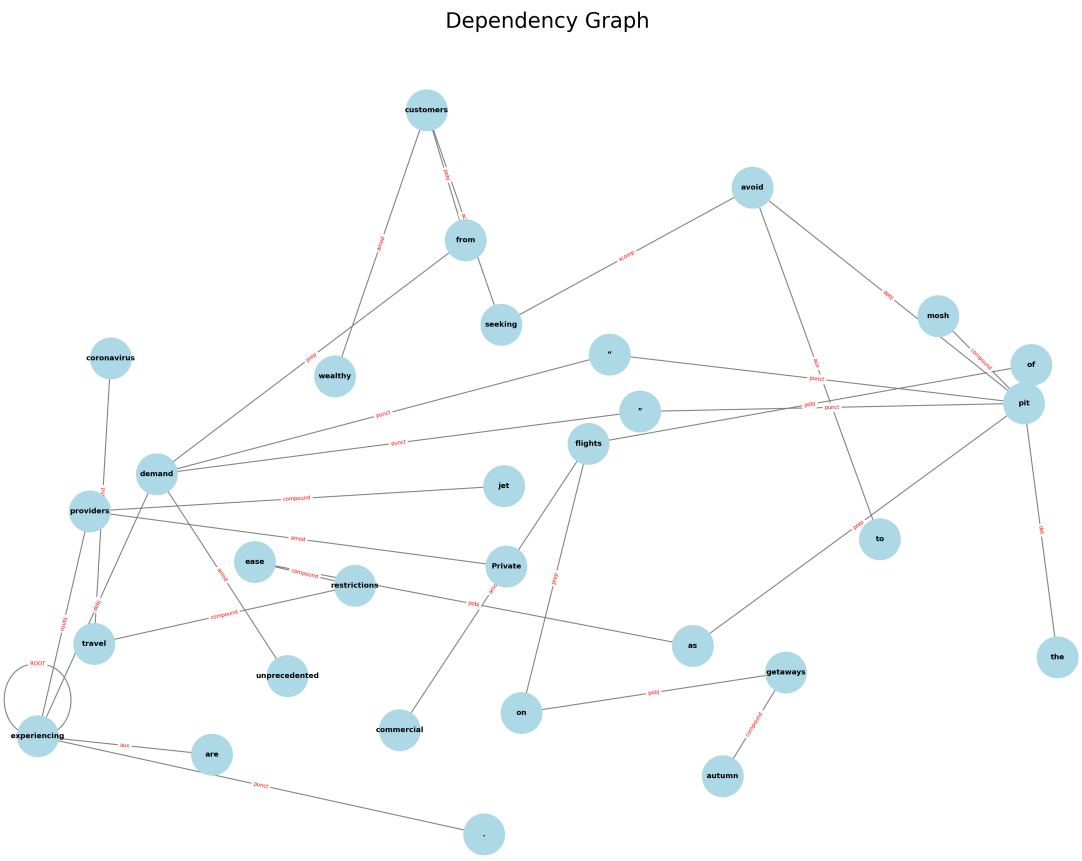

Figure 12: Network of a single corpus' sentence using the dependency between words as label connecting the nodes

The table containing the tagged words used to plot 12 of the sentence can be found in the table 13 in the appendix. The second one, stopwords and puctuation was removed. Important to note that even with stopwords removed from the sentence, some words' head are still stopwords and that is why there are stopwords on the chart and that is the reason the graph is now weakly connected [4].

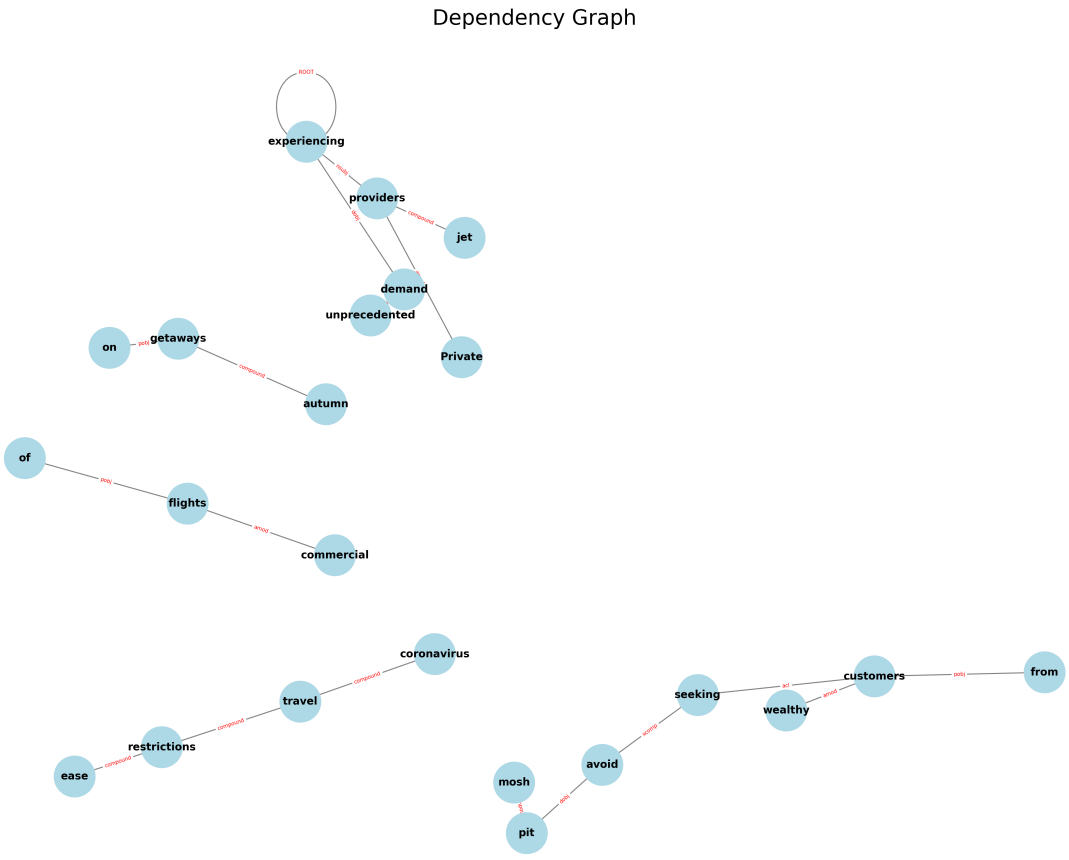

Figure 13: Network of a single corpus' sentence using the dependency between words as label connecting the nodes

The table containing the tagged words used to plot 13 of the sentence can be found in the table 14 in the appendix.

2.3.2 SVO generator

With the sentences processed and prepared to be transformed in SVOs, the algorithm applied to extract it was the following:

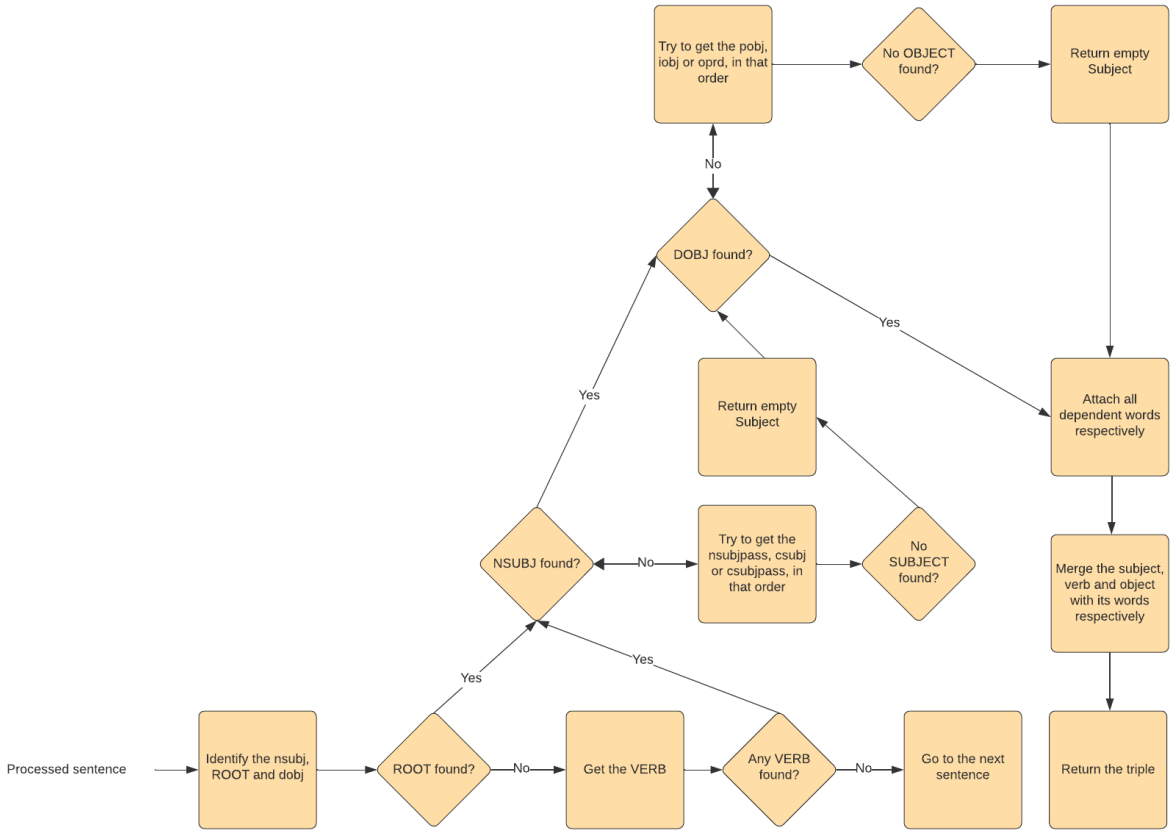

Figure 14: SVO generator Pipeline 2

The objective is to identify not only the main subject, verb and object of the processed sentence but also its dependents (like adjectives and adverbs, for example). First, the algorithm tries to identify the main subject, verb and object. That means, check if there is a nominal subject (nsubj), a root verb (ROOT) and a direct object (dobj). If it succeeds, then it attaches all words that are its dependents, like noun, adverbial, adjective and numeric modifiers, respectively and returns the triplet in the tuple format:

$$\text{(subject + subject dependents, verb + verb dependents, object + object dependents)} \tag{5}$$

If no ROOT verb is found, then triplet selects the verb(s) available. If no verb is found in the processed sentence, then the algorithm moves to the next processed sentence. If no nominal subject is found, the algorithm looks for a passive nominal subject, clausal subject and clausal passive subject, in that order. If nothing is found, it returns empty. Lastly, if no direct object is found, the algorithm searches for an indirect object, object of preposition and object predicate. If nothing is found, it returns empty. The algorithm then repeats the process of attaching every dependent to it respective head and returns the triplet.

2.3.3 Performance of the parser

To evaluate the accuracy of the parser, a score function was implemented. To serve as input, 15 semantically and syntactically different sentences were used. Also, the corrected SVO triplet was assigning to each one to be used as a comparison between the parsers. These SVOs triplets were manually assigned. The results can be found in 8 and 9. For example, the sentence:

The quick brown fox jumps over the lazy dog.

Would have to generate the triplet:

[["quick brown fox", "jumps over", "lazy dog"]]

This sentence follow the structure of active voice in the simple present tense. The parser had to able to deal with different semantic structures, such as passive voice, for example. Detailed information about the sentences used to test the parser is provided in the tables 11 and 12 in the appendix.

To compare the performance of the parser, python's library textacy[8] was used. That library is used to perform a variety of NLP tasks, built on the spaCy library. The function "extract.subject_verb_object_triples" extracted SVO triplets automatically and was used as a second parser. To compare the performance of them, the function "rank_svo_solution" was designed. It compared its output with a correct reference for Subject-Verb-Object (SVO) triplets, passed as an argument as well. This comparison was essential to ensure that the parser correctly identified the grammatical components of sentences.

The function starts by initializing a score variable to 0. It immediately checks if the parser's output is empty. If it is, the function returns empty, indicating that the parser was not able to produce any SVO triplets. After the first check, the function decomposes both the correct SVO and the parser's SVO into individual components: subject, verb, and object. This separation allows the function to compare each component of the triplets independently.

The function then splits each subject into individual words and converts them into sets to handle potential duplicates and unordered words.

For each unique word in the parser's subject that also appears in the correct subject, the score is incremented by 1. For words in the parser's subject that do not appear in the correct subject, the score is decremented by 1. For each word in the correct subject that is not in the parser's subject, the score is decremented by 1 again. If the parser's subject has more words than the correct subject, an additional penalty is applied based on the difference in the number of words.

In other words, if there is extra words or missing words, there is a penalty. That was implemented because often an adjective that belongs to the subject was extracted in the object tuple, leading to a misinterpretation of the sentence. The function repeats the matching process for the verbs and objects, following the same logic as for the subjects.

After comparing all three components (subject, verb, object), the function returns the final score. This score reflects how well the parser's output aligns with the correct SVO triplets.

The "rank_svo_solution" function provides a quantifiable method to evaluate the accuracy of SVO parsers. By systematically scoring the parser's output against a reference, this function helped in assessing the effectiveness of different parsing algorithms and identifying areas for improvement.

The scoring function effectively highlighted the strengths and weaknesses of both parsers.

In the example sentence "The quick brown fox jumps over the lazy dog," both parsers demonstrated varying degrees of success. Our parser correctly identified the subject and verb but missed the object entirely, resulting in a score of 3. The Textacy parser failed to extract any SVO triplet. This outcome suggests that while our parser was able to capture some essential components, it still struggled with the complete extraction of SVOs in straightforward sentences.

For sentences with more complex structures, such as passive voice or sentences with multiple clauses, both parsers generally performed poorly. For instance, in the sentence "She was given a bouquet of flowers by her friends," our parser scored -7, failing to accurately identify the subject and verb phrases. The Textacy parser, on the other hand, scored 1, indicating a slightly better performance by capturing the main subject-verb-object triplet more effectively.

Sentences like "Running through the park, John saw a beautiful sunrise" and "The project that we started last month is almost finished" were challenging for both parsers. Our parser often missed the correct syntactic structure or introduced extraneous elements, resulting in negative scores. For example, our parser scored -11 for the sentence involving a relative clause, highlighting its difficulty in handling embedded clauses. The Textacy parser also struggled but showed some ability to extract partial triplets, resulting in a score of -9.

One notable issue for our parser was the misinterpretation of adjectives or adverbial modifiers as part of the object or other components of the SVO triplet. For instance, in the sentence "The book, which was

[8] https://textacy.readthedocs.io/en/latest/

written by a renowned historian, provides an in-depth analysis," our parser misinterpreted the modifying clause and extracted incorrect components, resulting in a score of -7. This demonstrates that our parser needs improvement in distinguishing between core sentence elements and modifiers.

In cases where the sentence structure was particularly complex or convoluted, such as "Despite the fact that the weather forecast predicted heavy rain and thunderstorms throughout the weekend, the outdoor music festival proceeded as planned," our parser frequently failed to extract any meaningful triplet or produced outputs with misplaced words, leading to scores as low as -13. The Textacy parser, while also challenged, managed to extract some correct elements, resulting in slightly better scores.

Generally, our parser displayed an ability to capture parts of the SVO triplet, especially in simpler sentences. However, it struggled significantly with more complex structures, passive voice, and sentences with embedded clauses or modifiers. This indicates a need for further refinement, particularly in handling non-standard syntactic arrangements and distinguishing between different types of sentence components.

The Textacy parser showed a more consistent performance across different sentence structures. While it also faced difficulties with complex sentences and failed to capture complete SVO triplets in some cases, it often performed better than our parser. This suggests that the underlying NLP models and training data used by Textacy/spaCy have a more robust handling of natural language nuances, which our parser could benefit from.

The analysis indicates that while both parsers have limitations, the Textacy parser generally provides more reliable results, particularly for sentences with complex structures. Our parser, however, demonstrates potential in simpler cases but requires significant improvements to handle more sophisticated language features effectively. However, since our parser can always retrieve something, it was used to collect the triplets. Not only that, but after analysing the sentences presented in the corpus, they tend to be simple, as seen in 4, 5 and 7, which leads us to believe that our edited parser will perform well is most scenarios.

2.4 Network Generator

With the triplets extracted, the next process was to aggregate and understand the relationships in the data. The dataset head:

sectionName	**webPublicationDate**	**svo)**
Business	2022-01-31 22:25:24+00:00	[(stock market, closed hit, shares), (turbule...
Business	2022-01-31 19:35:49+00:00	[(vodafone, came ago remarkably,), (daring ac...
Business	2022-01-31 19:24:56+00:00	[(tesco, closing launched, jack shoppers chain...
Technology	2022-01-31 19:18:10+00:00	[(sony, agreed buy, bungie maker latest), (wa...
Politics	2022-01-31 18:53:47+00:00	[(demands, accept blamed,), (long queues, cau...

Table 2: Dataset containing the svos head

The algorithm chose to aggregate the data was the FP-Growth [32]. To do that, the Python library *mlxtend*[9] was used. In the library, *fpgrowth* function expects data in a *one-hot encoded pandas DataFrame*. So, the first step was to convert the data to the right format:

Index	**Row Number**	**Triplets**
0	1235	[brexit, deepened, crisis]
1	8410	[brexit, getting, worse]
2	20887	[brexit, added, caused, unnecessary, difficult...]
3	26138	[brexit, based]
4	26479	[brexit, created, expect, issue, return]

Table 3: Sample Triplets Dataset (yet to be one-hot encoded)

A filter was applied selecting triplets that contained the word that would be the central node of the network. That saved time and memory, making the process faster. In the example above, only triplets

[9]https://rasbt.github.io/mlxtend/

containing the word "brexit" were selected.

The data was then converted to a python list containing only the triplets column, called transaction data. Then, the function *TransactionEncoder* transformed transaction data into a one-hot encoded format, suitable for frequent pattern mining. The *TransactionEncoder* first identifies all unique items across transactions during the fitting process and then transforms each transaction into a binary vector, indicating the presence or absence of these items. This encoded format is then converted into a pandas[10] DataFrame, providing a structured and interpretable dataset for subsequent analysis.

The next step was to identify frequent itemsets from the one-hot encoded dataset with a minimum support threshold, here 0.01. The minimum support threshold is a parameter that determines the minimum frequency at which an itemset (a collection of items) must appear in the transaction dataset to be considered "frequent" [1]. In other words, a collection of items must appear in at least 1% of all transactions in the dataset to be considered frequent. This filters out infrequent itemsets that appear less often and focuses the algorithm on itemsets that are more common within the dataset.

$$\text{Support(Itemset)} = \frac{\text{Number of Transactions Containing the Itemset}}{\text{Total Number of Transactions}}, range : [0, 1] \tag{6}$$

It sets a baseline for how often an itemset should occur in the dataset to be included in the analysis. This helps in focusing on more relevant and significant patterns, making the algorithm more efficient. If an itemset appeared less frequently than 1%, it was ignored by the algorithm. Since we have a lot of triplets and most of them were different, the number of words that appeared in more than one triplet is low. For that reason, the minimum support had to be low, there are few triplets that are similar.

This was followed by generating association rules from these itemsets using a specified confidence metric to measure the reliability of the rules [1]. The confidence metric measured the reliability of an association rule. It is the proportion of transactions containing the antecedent (the "if" part of the rule) that also contained the consequent (the "then" part of the rule). Confidence is an indicator of how often the rule has been found to be true.

$$\text{Confidence(Antecedent} \rightarrow \text{Consequent)} = \frac{\text{Support(Antecedent} \cup \text{Consequent)}}{\text{Support(Antecedent)}}, range : [0, 1] \tag{7}$$

Confidence is used to evaluate the strength of association rules derived from frequent itemsets. Higher confidence values indicate stronger associations between the antecedent and the consequent, suggesting a more reliable rule.

The rules dataset head:

Antecedents	**Consequents**	**Antecedent Support**	**Consequent Support**	**Support**	**Confidence**
(added)	(brexit)	0.012607	0.999325	0.012607	1.000000
(having)	(brexit)	0.013620	0.999325	0.013620	1.000000
(means)	(brexit)	0.024651	0.999325	0.024651	1.000000
(economy)	(brexit)	0.019811	0.999325	0.019811	1.000000

Table 4: Association Rule Metrics

The last step is to plot the network where the central node is the word filtered, in the example "brexit" and the other nodes are the antecedents and consequents, weighted by the support.

[10] https://pandas.pydata.org

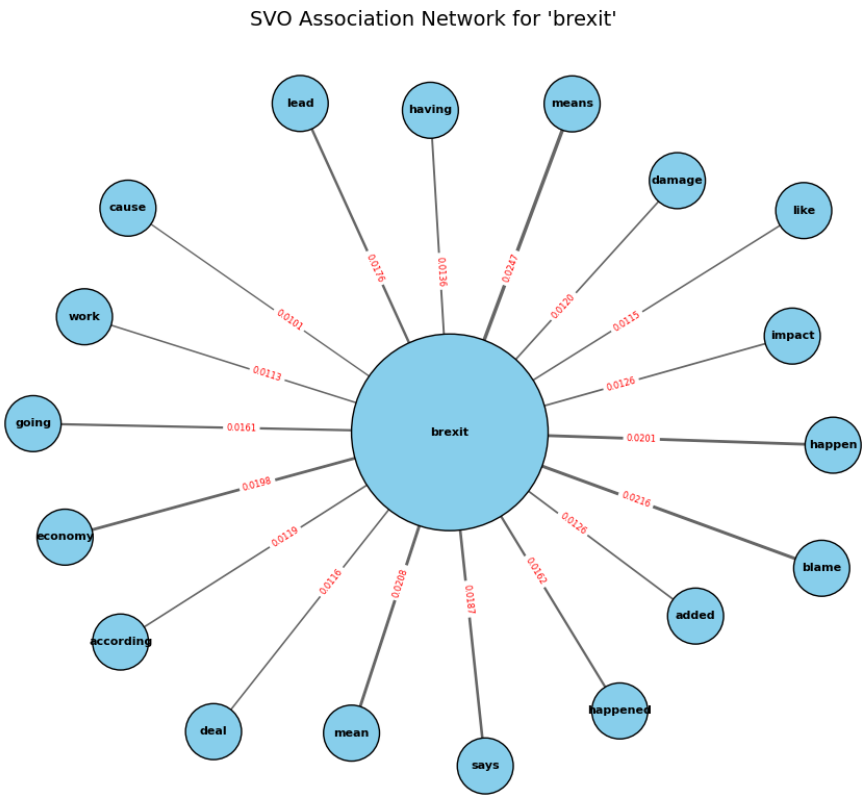

Figure 15: Network with "brexit" as central node

In the network above, all the dataset was used. It used 8884 triplets. We can see some interesting words associated with the brexit, like "impact", "damage", "cause" and "blame" that are usually associated with negativity. Also, words like "economy", "deal" and "work" might indicate the practical issues considered during the brexit. The minimum support used was 0.01 and the node sizes represent the degree centrality 8, that measures the number of edges connected to a node. The more connections, the higher the centrality, the bigger the node. The number in the edges are the support.

3 Results

The next step was to plot some metrics relating to the network. All the metrics were used analysing the network plotted that covered all articles available that contained the word "brexit", with no date filters applied 15. Degree centrality was the first metric used [43]. In a network, the degree centrality of a node is the number of edges connected to it. For directed networks, it is distinguish between in-degree, the number of incoming edges to a node and out-degree, the number of outgoing edges from a node.

$$C_D(v) = \frac{\deg(v)}{N-1} \tag{8}$$

Where $C(u, v)$ is the adjacency matrix of the graph, with a value of 1 if there is an edge from u to v and 0 otherwise. In the network, degree centrality can help identify the most active verbs (high out-degree) or the most referenced objects (high in-degree). As a result, it is possible to see that the top 4 nodes by in-degree and out-degree centrality are:

Node	**Degree Centrality**
uk	0.014722975590856257
deal	0.01394808213870593
economy	0.011235955056179775
opportunities	0.008911274699728787

Table 5: Top 4 nodes by in-degree centrality

Node	**Out-Degree Centrality**
brexit	0.6416117783804727
says	0.017047655947307245
means	0.01162340178225494
affect	0.01162340178225494

Table 6: Top 4 nodes by out-degree centrality

Next, it was evaluated the betweenness centrality [43]. Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes in the network. It quantifies a node's importance in facilitating communication or acting as a bridge within the network. Its formula is:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{9}$$

where:

σ_{st} is the total number of shortest paths from node s to node t, and σ_{st} (v) is the number of those paths that pass through v. The top 4 nodes by betweenness centrality recovered were:

Node	**Betweenness Centrality**
brexit	0.0015
damage	0.0001
deal	0.0001
impact	0.0001

Table 7: Top nodes by Betweenness Centrality

The next metric used was eigenvector centrality [2]. This considers not just the number of connections a node has but also the quality (i.e., centrality) of those connections [23]. A node with high eigenvector centrality is connected to other well-connected nodes. The formula is represented below:

$$x_i = \frac{1}{\lambda} \sum_j A_{ij} x_j \tag{10}$$

The formula indicates that the eigenvector centrality x_i of a node i is proportional to the sum of the centrality scores x_j of its neighboring nodes j, weighted by the adjacency matrix A_{ij}. The top 10 eigenvector found in the network were:

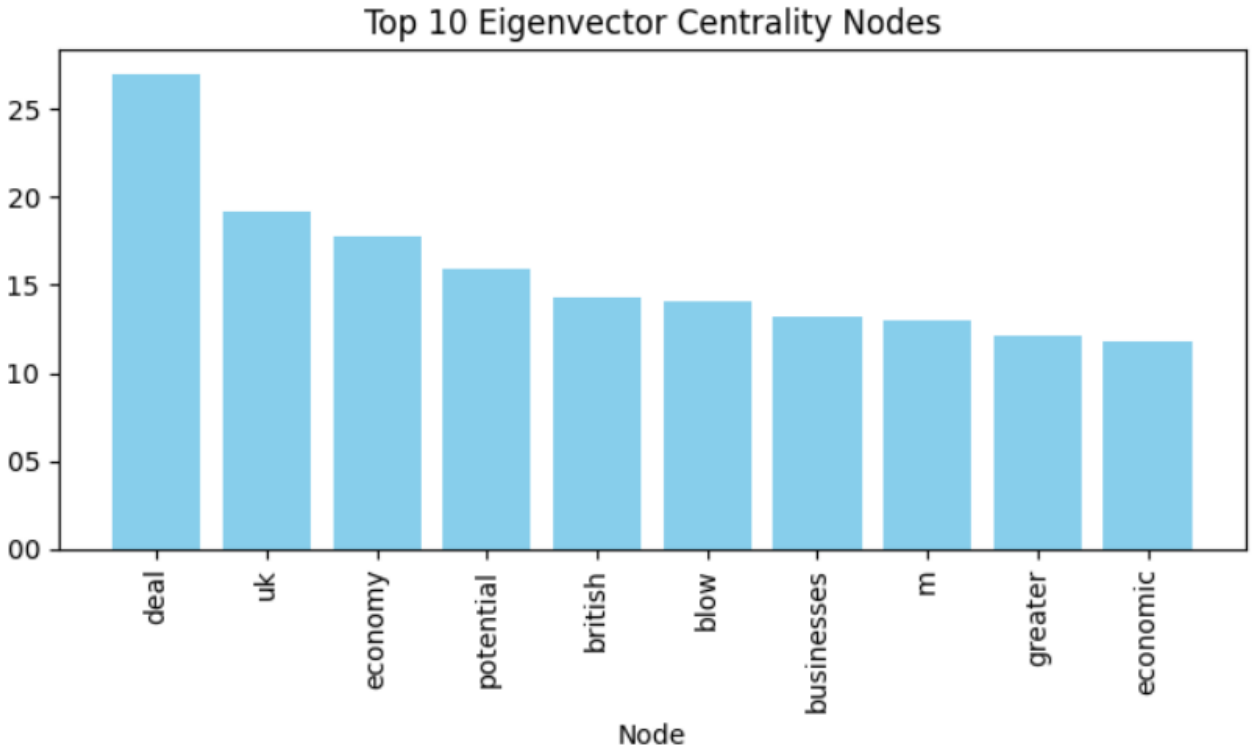

Figure 16: Top 10 nodes considering eigenvector centrality

After that, the metric density was applied [3]. Density measures how many edges are in a network compared to the maximum number of possible edges. Its formula is:

$$D = \frac{2E}{N(N-1)} \tag{11}$$

Where e is the number of edges and N is the number of nodes. For a SVO network, density can indicate how interconnected the subjects, verbs, and objects are. In other words, a graph's density measures how close the graph is to being a complete graph, where every pair of distinct vertices is connected by a unique edge. The graph density was 0.0007. This indicates that the graph is very sparse, out of all the possible edges that could exist between nodes in a complete graph, only 0.07% of those edges actually exist.

In practical terms, a density close to 0, like 0.0007, signifies that the graph has very few connections relative to the number of nodes. This is common in real networks where connections are rare. A low density can imply that the graph is less connected or that the network has a high number of nodes with very few edges between them. In this example, it can mean that the network might consist of several isolated or weakly connected groups. A density of 0.0007 suggests a large graph with relatively few edges, representing a real-world scenario where connections are limited or represent specific, sparse relationships. Since the density was very low, this might indicate that the graph might be disconnected. To confirm that hypothesis, the average length of the shortest path between all pairs of nodes was plotted [27]. Its formula is described below:

$$L = \frac{1}{\frac{N(N-1)}{2}} \sum_{i \neq j} d(v_i, v_j) \tag{12}$$

Where $d(vi, vj)$ is the shortest path distance between nodes vi and vj. The result was that the graph was, in fact, disconnected, meaning that there are at least two nodes in the graph that cannot be reached

from each other through any path of edges. Finally, the clustering coefficient was found [37]. The clustering coefficient measures the likelihood that a node's neighbors are also connected. It provides insights into the local density of the network. For a node v,

$$C(v) = \frac{2T(v)}{\deg(v)(\deg(v) - 1)} \tag{13}$$

Where ev is the number of edges between the neighbors of v and kv is the number of neighbors of v. A high clustering coefficient indicates that the verbs and objects linked to a particular subject (in our example "brexit") are also closely related to each other, forming tightly-knit clusters.

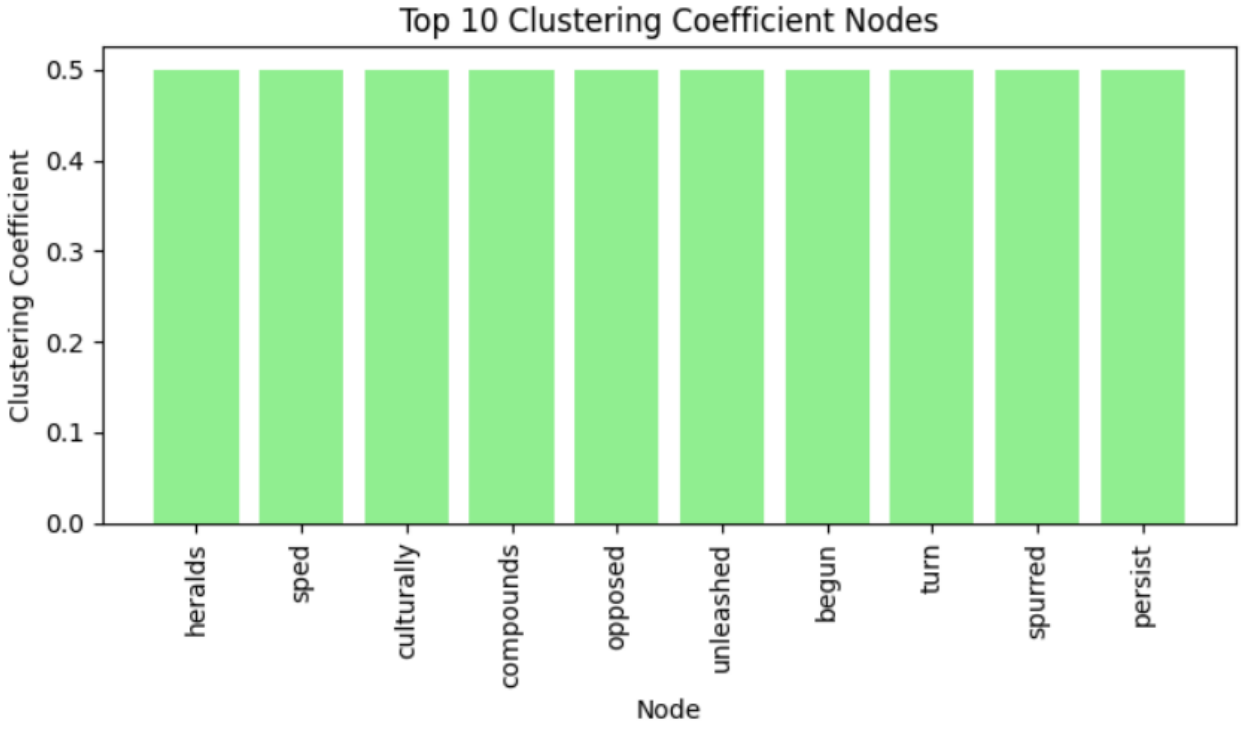

Figure 17: Top 10 nodes clustering coefficient nodes

Finally, there were selected some important dates regarding the Brexit and filtered networks were plotted. For each important date, a network was plotted using the confidence metric as weight.

3.1 23/06/2016: The Brexit Referendum

On 23/06/2016, the UK held a referendum to decide whether to remain in or leave the European Union [28]. The result was a narrow victory for the "Leave" campaign, with 51.89% voting to leave and 48.11% voting to remain. This decision set in motion the process of the UK's withdrawal from the EU [42]. The network selected covered 1 year, dating from 23/06/2015 to 23/06/2016. The filtered dataset had 103461 articles, 4812703 svo tripets and, out of these, 710 containing the word "brexit". The minimum support applied was 0.02.

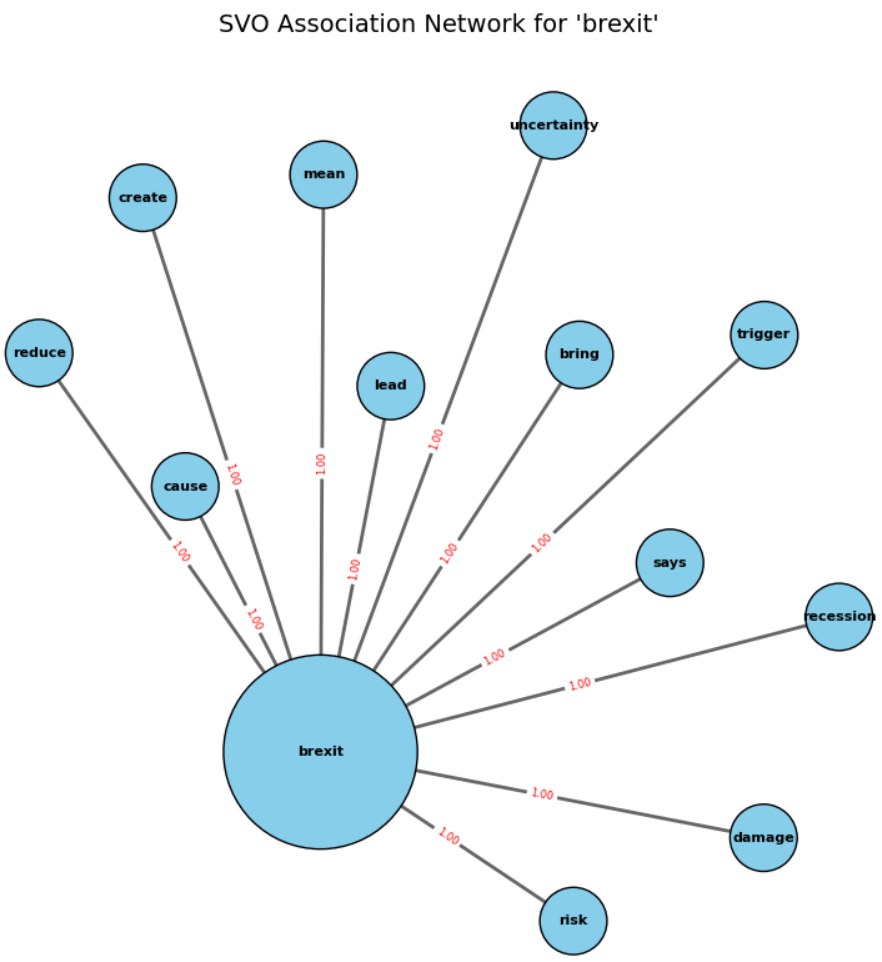

Figure 18: SVO association network from 23/06/2015 to 23/06/2016

During this period, the discourse around Brexit was characterized by uncertainty and concern about the potential consequences of leaving the European Union.

The most prominent nodes in the network, such as "reduce," "uncertainty," "trigger," "recession," "risk," and "damage," indicate a predominantly negative sentiment. These terms suggest that much of the media coverage was focused on the potential economic and political risks associated with Brexit. The frequent appearance of terms like "uncertainty" and "recession" reflects widespread concerns about the stability and future of the UK economy outside the EU.

The overall framing in this period appears to be negative, with the narrative centered around fear and potential negative outcomes. This aligns with the broader public debate during this time, which was heavily polarized between those fearing economic downturns and loss of global influence and those advocating for sovereignty and control over immigration.

Nodes like "economy" and "deal" showing high betweenness centrality suggest that discussions were not just focused on the concept of Brexit itself, but also on the practical implications and negotiations that would follow a decision to leave the EU.

The other network that was plotted was from 23/06/2016 until 29/03/2017, to understand the feeling towards the brexit from the media perspective since the "Brexit Referendum" to Article 50, triggered by Prime Minister Theresa May. The second network had 59613 articles and 3019833 svo triplets. Out of those, 1675 contained the word brexit. The minimum support applied was 0.012.

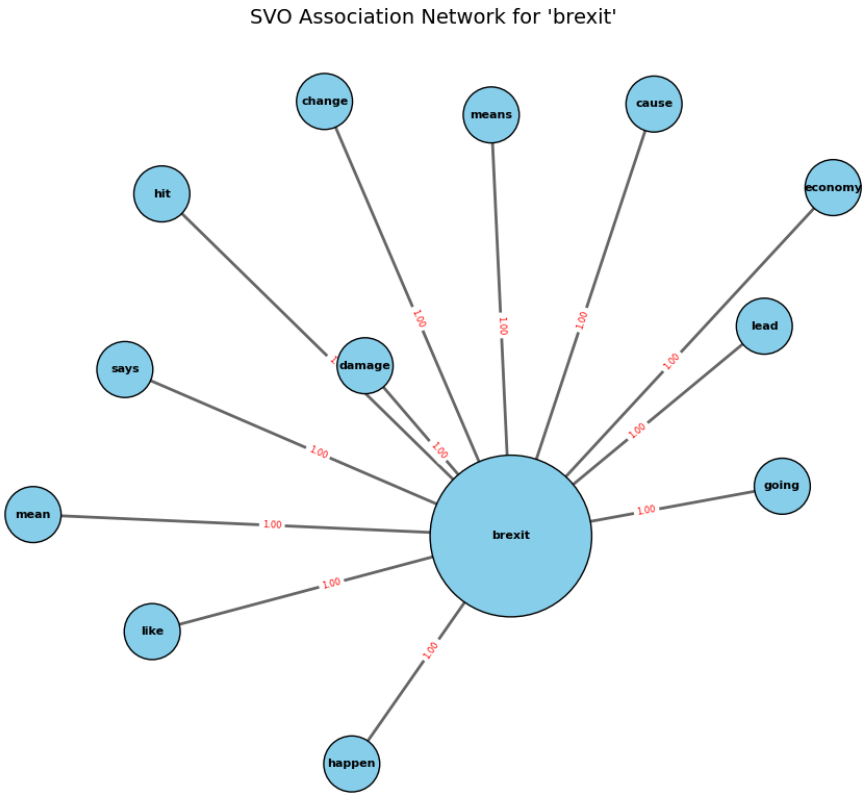

Figure 19: SVO association network from 24/06/2016 to 29/03/2017

The network covering the period from 24/06/2016 to 29/03/2017 captures the immediate aftermath of the Brexit referendum and the early stages of the UK's formal departure process from the EU.

After the referendum, the network still reflects a predominantly negative sentiment, but with an increased focus on terms like "like" and "change,"." This shift indicates a transition in the media narrative from debating the pros and cons of Brexit to dealing with its practical implications and the steps needed to formalize the UK's exit from the EU. This change in focus suggests that the media narrative began to center more on the mechanics of leaving the EU rather than just the potential impacts.

This network highlights a period of transition in the Brexit discourse, where the immediate shock and uncertainty post-referendum gave way to a more structured debate on how Brexit would be implemented

and what the future relationship with the EU might look like.

3.2 29/03/2017: Article 50 Invocation

Prime Minister Theresa May triggered Article 50 of the Lisbon Treaty, officially starting the two-year countdown to the UK's exit from the EU. This marked the beginning of formal negotiations between the UK and the EU on the terms of the withdrawal [9]. The third network had 83251 articles and 4211129 svo triplets. Out of these, 2194 contained the word "brexit" and were used. The minimum support applied was 0.012.

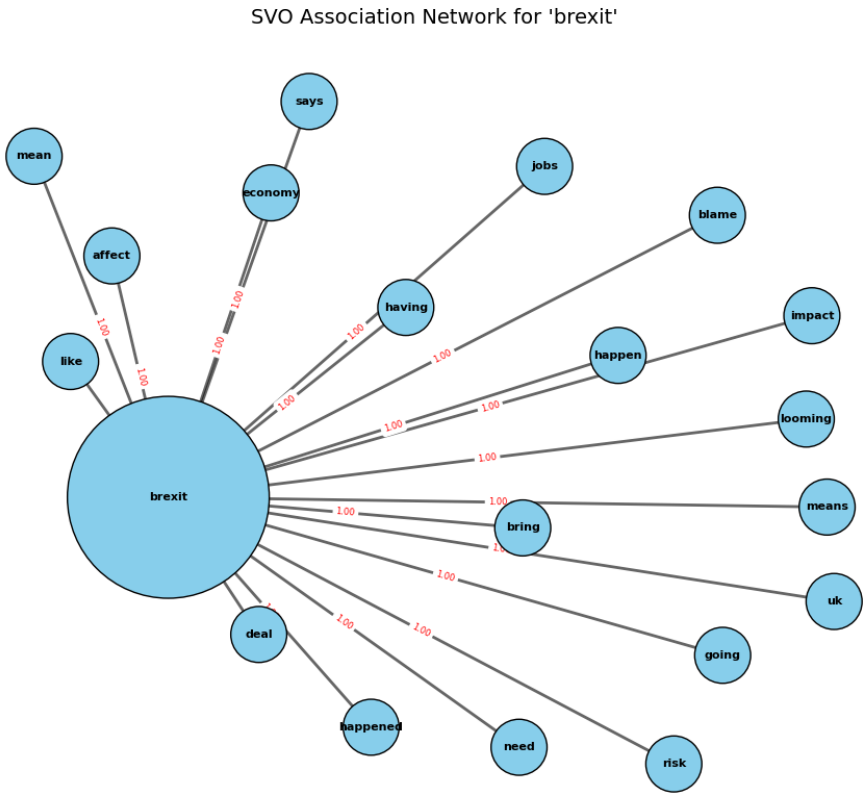

Figure 20: SVO association network from 30/03/2017 to 15/01/2019

As expected, terms related to negotiations and the impact of the brexit ("jobs," "deal," "economy", "impact") became more prominent. This shift reflects the media's focus on the back-and-forth between the UK and EU as both parties began to outline their positions and the potential outcomes of the Brexit process.

These terms being central in the network suggests they were key themes in the media narrative, reflecting public concern and media emphasis on the outcome of the Brexit negotiations. The high betweenness centrality of "deal" and "negotiation" indicates their role as connecting points in the discussion, bridging various aspects of the Brexit debate.

The discourse remains highly polarized, with negative nodes ("risk", "looming", "affect") and positive

possibilities ("like"). This fragmentation reflects the divided public opinion and the complex, often contradictory expectations surrounding Brexit.

Compared to the networks before and immediately after the referendum, this network shows a more structured narrative around specific negotiation points. While economic concerns remain, the discussion has shifted more towards the specifics of what Brexit will look like, rather than the abstract concept of leaving the EU.

This network analysis reveals a media narrative that evolved from initial shock and uncertainty to a more detailed examination of the practicalities and negotiations involved in leaving the EU. The persistent polarity and emerging focus on negotiations underscore the complexity and contentiousness of the Brexit process.

3.3 15/01/2019: The UK House of Commons reject the agreement

The UK House of Commons overwhelmingly rejected the withdrawal agreement that had been negotiated between the United Kingdom and the European Union. The rejection of the deal intensified the Brexit crisis, creating significant uncertainty about the future relationship between the UK and the EU. Following this vote, the UK government faced a no-confidence motion and subsequent discussions on possible alternatives, including a no-deal Brexit, renegotiation of the agreement, or even a second referendum. The decision marked a critical moment in the UK's journey to leave the European Union, reflecting deep divisions within the country and among its political leaders [14]. The forth network had 19824 articles and 987070 svo triplets. Out of these, 826 contained the word "brexit" and were used. The minimum support applied was 0.014.

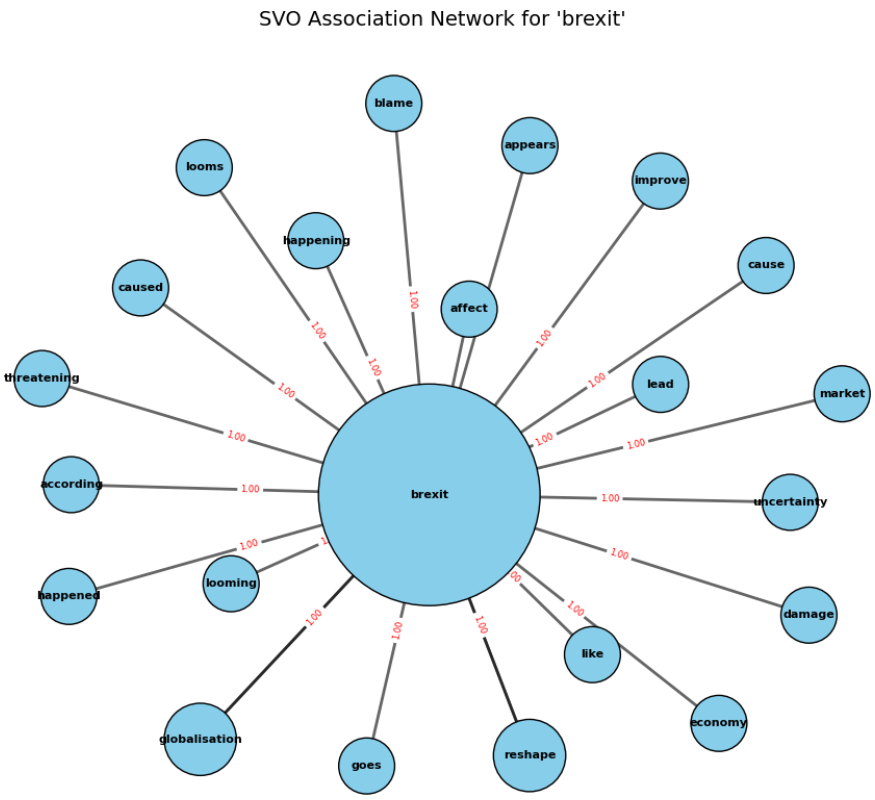

Figure 21: SVO association network from 15/01/2019 to 24/07/2019

The network reflects a critical moment in the Brexit process, characterized by heightened uncertainty and political turmoil.

It shows an intensification of negative sentiment compared to previous networks. Central nodes such as "threatening," "looms," "blame,", "damage" and "uncertainty" indicate a pervasive sense of instability and disappointment following the Commons' decision. This sentiment shift reflects the immediate political crisis and the uncertainty regarding the UK's future relationship with the EU. Some interesting new nodes appeared as well, like "globalisation", "reshape", "market" and "improve", indicating the polarization towards the theme and the more in-depth discussions regarding it.

The network during this period shows a marked increase in references to "no-deal Brexit" and alternative Brexit strategies, reflecting the urgent need for a new direction. This change highlights a shift from discussions about the mechanics of leaving the EU to the reality of navigating political deadlock and public frustration.

3.4 24/07/2019: Boris Johnson becomes the prime minister of the United Kingdom

Boris Johnson became the Prime Minister of the United Kingdom, following his victory in the Conservative Party leadership election. Johnson, who was a prominent figure in the Brexit campaign, succeeded Theresa May after she resigned due to her inability to secure parliamentary approval for her Brexit deal. Upon taking office, Johnson pledged to deliver Brexit by the October 31 deadline, with or without a deal, signaling a shift towards a more hardline approach to the UK's departure from the European Union. His appointment marked a significant moment in British politics, with a focus on completing the Brexit process and addressing other domestic challenges [7]. The fifth network had 29489 articles and 1509170 svo triplets. Out of these, 669 contained the word "brexit" and were used. The minimum support applied was 0.013.

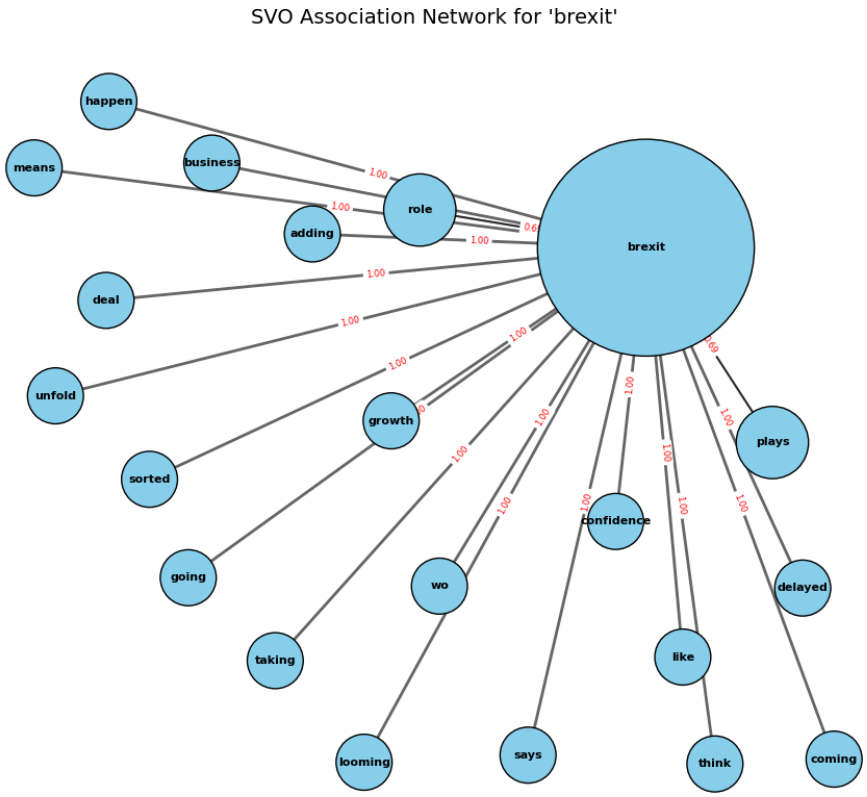

Figure 22: SVO association network from 25/07/2019 to 31/01/2020

While earlier networks exhibited significant polarity with both negative and positive clusters, this period's network shows a consolidation around more definitive, albeit polarized, positions. The discussions became less about whether Brexit should happen and more about how and under what terms, reflecting a narrowing of the debate as the final deadline approached.New words appeared in the network like "business", "adding", "role", "unfold", "confidence", "growth", "sorted" and "delayed", reflecting Johnson's firm stance on delivering Brexit by the set deadline, regardless of whether a deal was secured. This network indicates

a shift in sentiment from negotiation and compromise to decisiveness and action, aligning with Johnson's public promises to "get Brexit done."

Comparing this network with those before Johnson's premiership, there is a noticeable reduction in references to extended negotiations or second referendums, suggesting a diminished focus on reversing or delaying Brexit. The narrative became more about preparing for an inevitable exit, with increased references to potential consequences and strategic outcomes.

3.5 31/01/2020: The UK Leaves the EU

The UK officially left the European Union at 23:00 GMT on January 31, 2020. This marked the end of 47 years of membership and initiated an 11-month transition period during which the UK and the EU negotiated their future relationship [10]. The sixth network had 57254 articles and 3578605 svo triplets. Out of these, 371 contained the word "brexit" and were used. The minimum support applied was 0.015.

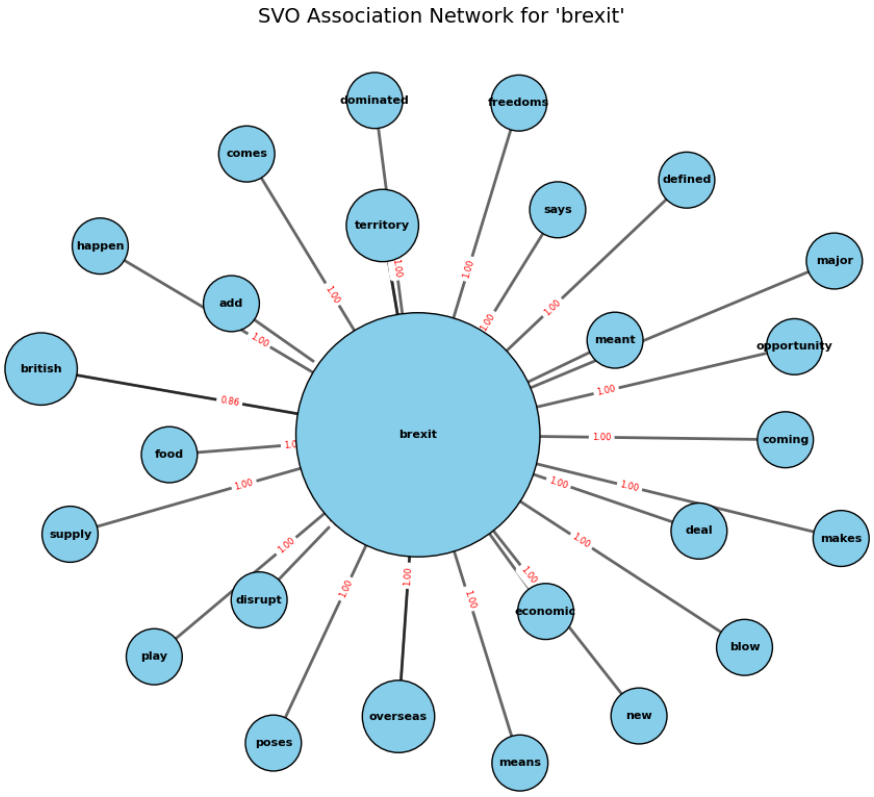

Figure 23: SVO association network from 31/01/2020 to 31/12/2020

The network for the period following the UK's official departure from the European Union reflects the evolving discourse around Brexit as it moved from theoretical discussions and negotiations to practical implementation and impact assessment.

"Freedom," "territory," "opportunity": these new nodes suggest a shift towards discussing Brexit as a means of gaining autonomy and defining a new national identity. The appearance of "freedom" and "opportunity" indicates that some narratives were framing Brexit as a positive move, focusing on the potential benefits and new possibilities it might bring to the UK. Terms like "Economic," "deal," "supply," and "disrupt" reflect concerns about the practical and economic implications of Brexit. Discussions around "deal" and "economic" likely relate to the ongoing negotiations and their expected impact on trade and the economy. The presence of "supply" and "disrupt" suggests that there were significant concerns about potential disruptions to supply chains and economic stability during the transition period. New terms like "New," "overseas," and "means" indicate a focus on the UK's new status and relationships outside the EU. "New" likely refers to the changes and adaptations required in various sectors, while "overseas" points to the UK's attempts to establish new trade relationships globally. The term "means" suggests discussions around the strategies and methods needed to navigate this new reality.

In comparison to earlier networks, certain terms related to negotiation and uncertainty (e.g., "risk," "uncertainty") seem to have diminished. This change reflects the shift in focus from whether Brexit would happen and the associated negotiations to the realities of post-Brexit life and its immediate impacts.

The presence of words like "deal," "blow," and "coming" indicates that while there was some optimism, there was also acknowledgment of the challenges and adjustments required in this new phase.

The network analysis for the period following January 31, 2020, indicates a significant shift in the Brexit discourse. While earlier networks were dominated by uncertainty and negotiation, this network reflects a balance between optimism about new opportunities and concerns about the economic and practical implications of leaving the EU.

3.6 31/12/2020: The date the UK is set to leave the EU. The transition period was not extended

The United Kingdom officially left the European Union, marking the end of the transition period that followed its formal exit from the EU on January 31, 2020. This date marked the conclusion of nearly a year of negotiations between the UK and the EU on their future relationship, particularly focusing on trade, security, and other forms of cooperation. The transition period was not extended, despite calls from some quarters for more time to finalize arrangements. As a result, the UK fully exited the EU's single market and customs union, ushering in a new era of UK-EU relations defined by a newly agreed trade and cooperation agreement. The end of the transition period represented a significant milestone in the Brexit process, affecting various aspects of life in the UK, from trade and travel to immigration and regulatory standards [10]. The last network had 57250 articles and 3336604 svo triplets. Out of these, 649 contained the word "brexit" and were used. The minimum support applied was 0.013 and the date was since the UK left the UE until a year later.

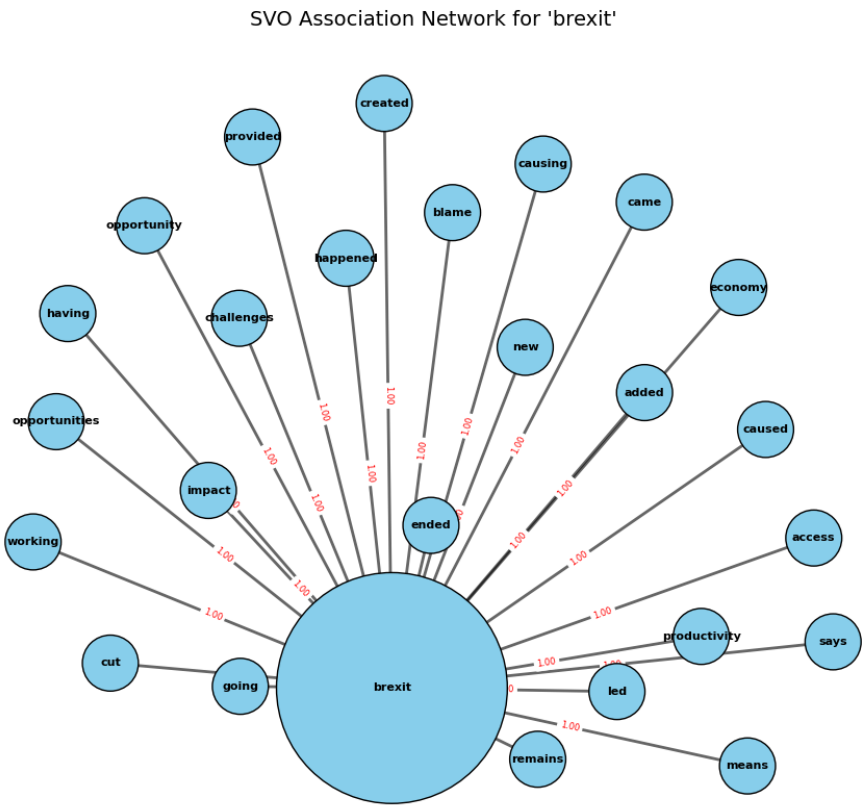

Figure 24: SVO association network from 2021-01-01 to 2022-01-01

The network illustrates the media and public discourse during a crucial transition as the UK navigated its new relationship with the EU under the recently agreed trade and cooperation agreement.

"Challenges," "impact," "opportunities": These nodes indicate a narrative focused on the dual aspects of Brexit as both a challenge and an opportunity. The presence of "challenges" and "impact" suggests ongoing concerns about the negative consequences of Brexit, particularly regarding economic and logistical adjustments. Simultaneously, terms like "opportunities" and "opportunity" reflect a discourse that also acknowledges potential benefits, likely in terms of new trade deals, regulatory freedoms, and other national strategies.

"Economy," "access," "productivity": These nodes suggest a continued focus on economic themes. "Economy" remains a key topic, reflecting concerns about the financial implications of Brexit. "Access" likely refers to discussions around market access, particularly the UK's new terms of trade with the EU and other global partners. "Productivity" might relate to debates on how Brexit could affect economic efficiency and competitiveness.

"Blame," "causing," "led": These terms reflect a discourse around accountability and causality, suggesting that there were narratives focused on attributing responsibility for the outcomes of Brexit, whether positive or negative. The presence of these nodes indicates a continued politicization of Brexit, where different actors may be blamed or credited for its consequences.

The network analysis for the period following December 31, 2020, reveals a nuanced and evolving Brexit discourse. While the formal process of leaving the EU has ended, the narrative remains highly dynamic, reflecting ongoing debates about the benefits and challenges of Brexit. The focus has shifted from the logistics of negotiation to the practical realities of life outside the EU, with discussions centered around economic impacts, new opportunities, and the complex adjustments required. This network captures a period of adaptation and reevaluation, highlighting that Brexit's story is far from over as the UK continues to navigate its new path on the global stage.

4 Discussion

This study aimed to analyze the temporal network of sentiment around Brexit as reported in The Guardian. By extracting Subject-Verb-Object (SVO) triplets and applying the FP-Growth algorithm, we were able to identify significant patterns in how Brexit-related events were framed and discussed in the media.

The networks showed distinct shifts in sentiment corresponding to major Brexit-related events. For instance, the network generated around the time of the Brexit Referendum on 23/06/2016 showed a overall negativity. Words like "reduce", "uncertainty", "trigger", "recession", "risk" and "damage" were found in the first network, from 23/06/2015 to 23/06/2016, suggesting a negative sentiment towards the brexit. The second network, overall negative words remained, but new words were introduced. Those might be interpreted either positively or negatively, showing a high degree of sentiment polarity, reflecting the divisive nature of the event. The network for the period following the invocation of Article 50 on 29/03/2017 displayed increased negativity, likely due to the uncertainty and contentious negotiations that ensued.

Nodes such as "Brexit," "deal," "economy," and "UK" consistently appeared as central in the networks across different periods. This suggests that these topics were at the core of the Brexit discourse. The high betweenness centrality of nodes like "deal" and "economy" indicates their role as bridges in the narrative, linking various aspects of the Brexit discussion together.

The analysis of lexical diversity and sentiment scores revealed that articles tended to have moderate lexical diversity, indicating a varied vocabulary but within a certain thematic constraint. The sentiment analysis using VADER showed a bimodal distribution with peaks at strongly positive and strongly negative sentiments. This suggests that the discourse around Brexit was highly polarized, with few articles adopting a neutral stance, which, by analysing the networks, makes sense.

The use of network analysis provided a nuanced view of the media's portrayal of Brexit over time. By focusing on SVO triplets, we were able to capture not just the sentiment but also the relationships between different entities involved in the discourse. This approach allowed us to move beyond simple sentiment scores and explore the underlying structure of the narrative.

The application of the FP-Growth algorithm was particularly effective in identifying frequent itemsets, which provided insights into the most commonly associated terms with "Brexit" and other key entities. This method allowed for a more granular understanding of the discourse, highlighting how certain terms and sentiments clustered together around specific events.

One of the unexpected results was the very low density of the generated networks. This suggests that the discourse around Brexit, while highly interconnected around key events, was not uniformly dense, indicating periods of focused discussion interspersed with less connected narratives. This finding aligns with the observation of distinct shifts in sentiment, where the narrative focus could have been tightly centered around specific events, leading to sparse connectivity in other areas of the network.

Another unexpected outcome was the high clustering coefficient of certain nodes, which indicated that while the overall network was sparse, certain groups of words were closely related, forming tight-knit clusters. This could suggest a segmented discourse where certain topics were discussed in isolation from others, reflecting the fragmented nature of media coverage.

5 Conclusion

This dissertation set out to explore the evolving sentiment of Brexit coverage in The Guardian from 2015 to 2021 through the use of advanced NLP techniques and network analysis. The primary objectives were

to extract SVO triplets from a large corpus of news articles, construct temporal sentiment networks around Brexit, and analyze the resulting patterns in sentiment and relationships among key entities over time.

The project successfully developed and applied a custom SVO parser to extract triplets from over one million articles, leading to the creation of temporal sentiment networks. These networks provided insights into how media sentiment around Brexit evolved, especially in response to major events.

The analysis highlighted significant shifts in sentiment and identified central entities within the Brexit discourse, such as "Brexit," "deal," "economy," and "UK." These entities consistently appeared as key nodes in the networks, indicating their importance in the narrative.

The use of the FP-Growth algorithm effectively identified frequent itemsets and allowed for a more detailed understanding of the associations between terms related to Brexit. This provided a granular view of how sentiments clustered around specific events and entities.

While the primary objectives of the study were met, there were certain limitations and unexpected outcomes. One of the significant findings was the very low density and disconnected nature of the networks. This suggests that while key events were well-covered, the overall discourse was not uniformly dense, indicating that some aspects of Brexit were discussed in isolation. And, although sentiment analysis was conducted using VADER, the scope was somewhat limited to basic sentiment categories. A more nuanced sentiment analysis, potentially incorporating irony, sarcasm, or more complex emotional tones, could have provided deeper insights.

The results of this project have largely met the initial expectations, with some surprising findings. The identification of distinct shifts in sentiment around major Brexit events aligns with the hypothesis that media coverage would reflect the contentious and evolving nature of Brexit negotiations. The significant factors in these findings include the ability to capture temporal changes in sentiment and the identification of central entities within the Brexit narrative. These results are significant because they provide a clearer understanding of how media sentiment evolves in response to political events and how key terms are framed in public discourse.

However, the sparsity and segmented nature of the networks were unexpected. This finding indicates that the Brexit narrative was more fragmented than anticipated, with clusters of tightly related terms suggesting isolated discussions. This contributes new insights into the nature of media coverage, highlighting the need for a more connected and comprehensive approach to understanding complex political events.

If given the opportunity to redo the project, a different approach might involve a more comprehensive sentiment analysis incorporating more sophisticated models to capture a wider range of emotions and tones. Additionally, expanding the network analysis to include different types of relationships (beyond SVO triplets) could provide a more holistic view of the discourse.

The overall methodology of using NLP and network analysis was appropriate for the research aims. The custom SVO parser and the FP-Growth algorithm provided valuable insights and met the project's objectives effectively. However, incorporating additional data sources, such as social media or other news outlets, could have further enriched the analysis. If the project were to be repeated, incorporating these elements and applying more advanced machine learning models for sentiment and network analysis would likely yield even more comprehensive results. In addition, the parser could be improved as well, adding more edge cases and being able to handle even more different semantic structures. Our parser needs enhanced algorithms to better identify and separate core components from modifiers, and handle passive constructions and clauses more accurately. In addition, leveraging pre-trained models and embeddings, similar to those used in spaCy, could improve the parser's accuracy and generalization to various syntactic forms.

By addressing these areas, our parser could achieve more accurate SVO extraction, aligning more closely with the correct grammatical analysis of sentences.

6 Next steps

Future research could extend the analysis to other major newspapers or media outlets to compare the evolution of sentiment across different platforms. This would provide a more diverse perspective on media bias and sentiment. Also, including social media data, such as Twitter or Facebook posts, could offer insights into public sentiment and how it interacts with media coverage. Also, introducing a more diverse set of training data, including a wider range of sentence types, could help improve the parser's ability to

handle different syntactic and semantic structures. For the network, delete common words that appear often in every network and don't change the frequency would also make the analysis more interesting.

Implementing more sophisticated sentiment analysis techniques, such as deep learning models that can detect sarcasm, irony, and more complex emotional tones, would enhance the understanding of media sentiment and developing dynamic network analysis techniques that can track changes in real-time would allow for more responsive and up-to-date insights into evolving media narratives.

Incorporating additional linguistic features beyond SVO triplets, such as modal verbs, negations, and rhetorical devices, could provide a more nuanced understanding of the discourse, and improving the parser itself to capture more sentences and be more accurate could benefit the dataset used as input. Also, implementing more edge-cases in the parser to make it more reliable and complete would help.

By building on the foundations laid in this dissertation, future research can continue to unravel the complexities of media sentiment and its impact on public perception, especially in politically charged contexts like Brexit.

7 Appendix

Sentence	**Sentence SVO**	**Structure**
The quick brown fox jumps over the lazy dog.	[*quick brown fox*, *jumps over*, *lazy dog*]	Active Voice, Simple Present Tense.
She was given a bouquet of flowers by her friends.	[*She*, *was given*, *a bouquet of flowers*]	Passive Voice, Simple Past Tense.
Running through the park, John saw a beautiful sunrise.	[*John*, *saw*, *beautiful sunrise*]	Present Participle Clause.
If it rains tomorrow, we will cancel the picnic.	[*we*, *will cancel*, *picnic*]	Conditional Sentence, First Conditional.
Although he was tired, he continued to work on the project.	[*he*, *continued to work*, *project*]	Complex Sentence, Subordinate Clause.
Do you know where the nearest bookstore is?	[*you*, *know*, *nearest bookstore*]	Indirect Question.
The project that we started last month is almost finished.	[*project that we started last month*, *is*, *almost finished*]	Relative Clause.
She not only sings but also plays the guitar.	[*She*, *plays*, *the guitar*]	Correlative Conjunction.
Having finished his homework, he went out to play.	[*he*, *went out*, *play*]	Perfect Participle Clause.
In the morning, we went for a long walk.	[*we*, *went for*, *a long walk*]	Prepositional Phrase.
Despite the fact that the weather forecast predicted heavy rain and thunderstorms throughout the weekend, the outdoor music festival proceeded as planned, with thousands of attendees enjoying the live performances, food stalls, and various other attractions without any significant disruptions.	[*the weather forecast*, *predicted*, *heavy rain and thunderstorms*], [*the outdoor music festival*, *proceeded*,], [*thousands of attendees*, *enjoying*, *live performances and various other attractions*]	Complex Sentence with Multiple Clauses.

Table 8: Parser Comparator 1

Sentence	**Sentence SVO**	**Structure**
The book, which was written by a renowned historian and meticulously researched over a span of ten years, provides an in-depth analysis of the socio-economic factors that contributed to the rise and fall of ancient civilizations, drawing parallels with modern-day societal trends and challenges.	[*The book*, *provides*, *an in-depth analysis of the socio-economic factors*]	Complex Sentence with a Relative Clause.
While the initial reaction to the proposed changes in the company's policy was overwhelmingly negative, with numerous employees expressing concerns about potential job cuts and reduced benefits, the management team held a series of town hall meetings to address these concerns and clarify the long-term vision and benefits of the new policy.	[*the initial reaction to the proposed changes in the company's policy*, *was*, *overwhelmingly negative*], [*the management team*, *held*, *a series of town hall meetings*]	Compound-Complex Sentence.
The technological advancements in artificial intelligence and machine learning, which have been accelerated by significant investments from both the private sector and government agencies, are expected to revolutionize various industries, including healthcare, finance, and transportation, by automating complex processes, improving decision-making, and enhancing overall efficiency.	[*The technological advancements*, *are expected to revolutionize*, *various industries*]	Complex Sentence with a Relative Clause.
After months of extensive negotiations and despite numerous setbacks, the international trade agreement was finally signed by all participating countries, paving the way for reduced tariffs, increased market access, and stronger economic cooperation, which are anticipated to stimulate global economic growth and foster closer diplomatic ties.	[*all participating countries*, *signed*, *the international trade agreement*], [*the international trade agreement*, *was signed*,]	Complex Sentence with Multiple Clauses.

Table 9: Parser Comparator 2

Processing Unit	Dataset Size	Section	Time (s)
A100	10628	Science	2176
	8468	Science	1099
	8339	Science	1051
	7362	Science	747
	3370	Science	318
	3992	Science	535
	10114	Technology	1342
	14855	Technology	2444
	1540	Technology	236
	36657	Technology	4999
	7005	Technology	705
	5273	Technology	968
	37903	Culture	5635
	35992	Culture	5306
CPU	37939	World	8393
	31127	World	5650
	33445	World	5676
	31881	World	5114
	23087	World	3549
	33974	World	4728
	21691	World	2714
	31015	World	7258
	37662	Culture	5498
	37762	Culture	6657
	35719	Culture	4203
	28458	Culture	4423
	34902	Culture	6300
	14788	Business	3019
	32553	Business	4741
	36406	Business	4532
	16671	Business	1934
	17686	Business	3664
	37956	Business	6927
	10264	Business	1883
CPU High Ram	10791	Culture	1519
	21411	Culture	2882
	32851	Politics	3782
	32851	Politics	3942
	32851	Politics	4126
	21411	Politics	3020
	32851	Politics	3977
	9344	Politics	1114
	19691	Politics	3499
	22410	Politics	3658

Table 10: Processing Unit and Data Collection Information

Correct SVO	**Edited Parser SVO**	**Other Parser SVO**	**Edited Parser Score**	**Other Parser Score**
[['quick brown fox', 'jumps over', 'lazy dog']]	[('quick brown fox', 'jumps', 'lazy ')]	[]	3	[]
[['She', 'was given', 'bouquet flowers']]	[('bouquet', 'given', 'friends')]	[('[She]', '[was, given]', '[bouquet]')]	-5	4
[['John', 'saw', 'beautiful sunrise']]	[('John', 'saw Running', 'beautiful sunrise ')]	[('[John]', '[saw]', '[sunrise]')]	3	2
[['we', 'will cancel', 'picnic']]	[('picnic', 'cancel rains', 'tomorrow')]	[('[we]', '[will, cancel]', '[picnic]')]	-4	4
[['he', 'continued to work', 'project']]	[('he', 'continued was work', 'the ')]	[('[he]', '[continued]', '[to, work, on, the, project]')]	-1	-7
[['you', 'know', 'nearest bookstore']]	[('you the nearest bookstore', 'know Do is', '')]	[]	-10	[]
[['project we started last month', 'is', 'almost finished']]	[('', 'finished', 'month')]	[('[we]', '[started]', '[that]')]	-10	-10
[['She', 'plays', 'the guitar']]	[('She', 'plays also', 'the guitar ')]	[]	2	[]
[['he', 'went out', 'play']]	[('homework', 'play', '')]	[]	-6	[]
[['we', 'went', 'long walk']]	[('morning', 'went', 'long ')]	[]	-3	[]

Table 11: Comparison of SVO Extraction and Scoring 1

Correct SVO	Edited Parser SVO	Other Parser SVO	Edited Parser Score	Other Parser Score
[['the weather forecast', 'predicted', 'heavy rain and thunderstorms'], ['the outdoor music festival', 'proceeded', ''], ['thousands of attendees', 'enjoying', 'live performances and various other attractions']]	[('weather forecast', 'proceeded Despite planned', 'thunderstorms heavy rain '), ('thousands', 'enjoying', 'live stalls performances ')]	[('[weather, forecast]', '[predicted]', '[rain, thunderstorms]'), ('[thousands]', '[enjoying]', '[performances, food, stalls, attractions]')]	-3	2
[['The book', 'provides', 'an in-depth analysis of the socio-economic factors']]	[('book', 'provides drawing', 'parallels analysis '), ('economic factors', 'contributed fall', 'ancient ')]	[('[book]', '[provides]', '[analysis]'), ('[which]', '[was, written]', '[historian]')]	-7	-4
[['the initial reaction to the proposed changes in the company's policy', 'was', 'overwhelmingly negative'], ['the management team', 'held', 'a series of town hall meetings']]	[('changes', 'proposed', ''), ('numerous employees', 'held address', 'series vision concerns ')]	[('[employees]', '[expressing]', '[concerns]'), ('[management, team]', '[held]', '[series]')]	-13	-17
[['The technological advancements', 'are expected to revolutionize', 'various industries']]	[('finance healthcare', 'enhancing', 'overall '), ('complex processes', 'automating', 'significant ')]	[('[advancements]', '[are, expected]', '[to, revolutionize, various, industries, ,, including, healthcare, ,, finance, ,, and, transportation, ,, by, automating, complex, processes, ,, improving, decision, -, making, ,, and, enhancing, overall, efficiency]'), ('[which]', '[have, been, accelerated]', '[investments]')]	-13	-36
[['all participating countries', 'signed', 'the international trade agreement'], ['the international trade agreement', 'was signed', '']]	[('numerous setbacks', 'signed despite paving finally', 'trade'), ('countries', 'participating', ''), ('closer diplomatic ties', 'stimulate foster', 'stronger ')]	[('[trade, agreement]', '[was, signed]', '[countries]'), ('[which]', '[are, anticipated]', '[to, stimulate, global, economic, growth, and, foster, closer, diplomatic, ties]')]	-12	-11

Table 12: Comparison of SVO Extraction and Scoring 2

Word/Token	**Dependency**	**Head**
Private	amod	providers
jet	compound	providers
providers	nsubj	experiencing
are	aux	experiencing
experiencing	ROOT	experiencing
“	punct	demand
unprecedented	amod	demand
demand	dobj	experiencing
”	punct	demand
from	prep	demand
wealthy	amod	customers
customers	pobj	from
seeking	acl	customers
to	aux	avoid
avoid	xcomp	seeking
the	det	pit
“	punct	pit
mosh	compound	pit
pit	dobj	avoid
”	punct	pit
of	prep	pit
commercial	amod	flights
flights	pobj	of
on	prep	flights
autumn	compound	getaways
getaways	pobj	on
as	prep	pit
coronavirus	compound	travel
travel	compound	restrictions
restrictions	compound	ease
ease	pobj	as
.	punct	experiencing

Table 13: Dependency Relations in the Example Sentence

Word	Dependency	Head
Private	amod	providers
jet	compound	providers
providers	nsubj	experiencing
experiencing	ROOT	experiencing
unprecedented	amod	demand
demand	dobj	experiencing
wealthy	amod	customers
customers	pobj	from
seeking	acl	customers
avoid	xcomp	seeking
mosh	compound	pit
pit	dobj	avoid
commercial	amod	flights
flights	pobj	of
autumn	compound	getaways
getaways	pobj	on
coronavirus	compound	travel
travel	compound	restrictions
restrictions	compound	ease

Table 14: Dependency Relations in the Example Sentence withou stopwods and punctuation

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. "Mining associations between sets of items in large databases". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2019, pp. 207–216.

[2] Alan L Andrew. "Eigenvectors of certain matrices". In: *Linear Algebra and its Applications* 7.2 (1973), pp. 151–162.

[3] Hayat Dino Bedru et al. "Big networks: A survey". In: *Computer Science Review* 37 (2020), p. 100247.

[4] Claude Berge. *The theory of graphs*. Courier Corporation, 2001.

[5] Yves Bestgen. "Measuring lexical diversity in texts: The twofold length problem". In: *Language Learning* (2023).

[6] Christian Borgelt. "An Implementation of the FP-growth Algorithm". In: *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*. 2005, pp. 1–5.

[7] *Boris Johnson becomes PM with promise of Brexit by 31 October*. Accessed: 2024-08-30. July 2019. URL: `https://www.theguardian.com/politics/2019/jul/24/boris-johnson-becomes-pm-with-promise-of-brexit-by-31-october`.

[8] Sabri Boubaker, Zhenya Liu, and Ling Zhai. "Big data, news diversity and financial market crash". In: *Technological Forecasting and Social Change* 168 (2021), p. 120755.

[9] *Brexit: Article 50 has been triggered - what now?* Accessed: 2024-08-30. Mar. 2017. URL: `https://www.bbc.com/news/uk-politics-39143978`.

[10] *Brexit: What you need to know about the UK leaving the EU*. Accessed: 2024-08-30. Jan. 2020. URL: `https://www.bbc.com/news/uk-politics-32810887`.

[11] Bram Bulté and Alex Housen. "Defining and operationalising L2 complexity". In: *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* 32 (2012), p. 21.

[12] Leland Bybee et al. *The structure of economic news*. Tech. rep. National Bureau of Economic Research, 2020.

[13] Harold D Clarke, Matthew J Goodwin, and Paul Whiteley. *Brexit*. Cambridge University Press, 2017.

[14] *Commons votes to reject Government's EU Withdrawal Agreement*. Accessed: 2024-08-30. Mar. 2019. URL: https://www.parliament.uk/business/news/2019/march/mps-debate-and-vote-on-the-withdrawal-agreement-with-the-european-union/.

[15] Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. *Text processing with gate*. Gateway Press CA, 2011.

[16] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. "Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research". In: *Intelligent Systems in Accounting, Finance and Management* 23.3 (2016), pp. 157–214.

[17] Rudolph Flesch. "A new readability yardstick." In: *Journal of applied psychology* 32.3 (1948), p. 221.

[18] Diana Gutiérrez-Posada, María Plotnikova, and Fernando Rubiera-Morollón. ""The grass is greener on the other side": The relationship between the Brexit referendum results and spatial inequalities at the local level". In: *Papers in regional science* 100.6 (2021), pp. 1481–1501.

[19] Catherine Happer and Greg Philo. "The role of the media in the construction of public belief and social change". In: *Journal of social and political psychology* 1.1 (2013), pp. 321–336.

[20] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[21] J Peter Kincaid et al. "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel". In: (1975).

[22] Thomas Lansdall-Welfare et al. "On the coverage of science in the media: A big data study on the impact of the Fukushima disaster". In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE. 2014, pp. 60–66.

[23] Warih Maharani, Alfian Akbar Gozali, et al. "Degree centrality and eigenvector centrality in twitter". In: *2014 8th international conference on telecommunication systems services and applications (TSSA)*. IEEE. 2014, pp. 1–5.

[24] Pekka Malo et al. "Good debt or bad debt: Detecting semantic orientations in economic texts". In: *Journal of the Association for Information Science and Technology* 65.4 (2014), pp. 782–796.

[25] David Malvern et al. *Lexical diversity and language development*. Springer, 2004.

[26] Christopher D Manning. "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" In: *International conference on intelligent text processing and computational linguistics*. Springer. 2011, pp. 171–189.

[27] Guoyong Mao and Ning Zhang. "Analysis of average shortest-path length of scale-free network". In: *Journal of Applied Mathematics* 2013.1 (2013), p. 865643.

[28] Hannah Marshall and Alena Drieschova. "Post-truth politics in the UK's Brexit referendum". In: *New Perspectives* 26.3 (2018), pp. 89–105.

[29] Dinesh Nagumothu et al. "Linked data triples enhance document relevance classification". In: *Applied Sciences* 11.14 (2021), p. 6636.

[30] Joakim Nivre. "Dependency grammar and dependency parsing". In: *MSI report* 5133.1959 (2005), pp. 1–32.

[31] Kevin O'Rourke. *A short history of Brexit: From Brentry to backstop*. Penguin UK, 2019.

[32] Mining Frequent Patterns. "Mining Frequent Patterns Without Candidate Generation". In: ().

[33] Kristin Potter et al. "Methods for presenting statistical information: The box plot." In: *VLUDS*. 2006, pp. 97–106.

[34] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de-Albornoz, and Laura Plaza. "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it". In: *Expert Systems with Applications* (2023), p. 121641.

[35] Robert P Schumaker and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system". In: *ACM Transactions on Information Systems (TOIS)* 27.2 (2009), pp. 1–19.

[36] Robert P Schumaker et al. "Evaluating sentiment in financial news articles". In: *Decision Support Systems* 53.3 (2012), pp. 458–464.

[37] Sara Nadiv Soffer and Alexei Vazquez. "Network clustering coefficient without degree-correlation biases". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 71.5 (2005), p. 057101.

[38] Saatviga Sudhahar and Nello Cristianini. "Automated analysis of narrative content for digital humanities". In: *International Journal of Advanced Computer Science* 3.9 (2013), pp. 440–447.

[39] Saatviga Sudhahar, Roberto Franzosi, and Nello Cristianini. "Automating quantitative narrative analysis of news data". In: *Proceedings of the Second Workshop on Applications of Pattern Analysis*. JMLR Workshop and Conference Proceedings. 2011, pp. 63–71.

[40] Saatviga Sudhahar, Giuseppe A Veltri, and Nello Cristianini. "Automated analysis of the US presidential elections using Big Data and network analysis". In: *Big Data & Society* 2.1 (2015), p. 2053951715572916.

[41] Saatviga Sudhahar et al. "Network analysis of narrative content in large corpora". In: *Natural Language Engineering* 21.1 (2015), pp. 81–112.

[42] *The Brexit referendum and the British constitution*. Accessed: 2024-08-30. May 2019. URL: https://www.economist.com/briefing/2019/05/30/the-brexit-referendum-and-the-british-constitution?ppccampaignID=&ppcadID=&ppcgclID=&ppccampaignID=&ppcadID=&ppcgclID=&utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=18151738051&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gad_source=1&gclid=CjwKCAjwlbu2BhA3EiwA3yXyu_xdpalLQUGJYPAKbwrplA4oB3yKGcfDfjWW_llb7syFUAW3PVaixhoC0RsQAvD_BwE&gclsrc=aw.ds.

[43] Junlong Zhang and Yu Luo. "Degree centrality, betweenness centrality, and closeness centrality in social network". In: *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*. Atlantis press. 2017, pp. 300–303.