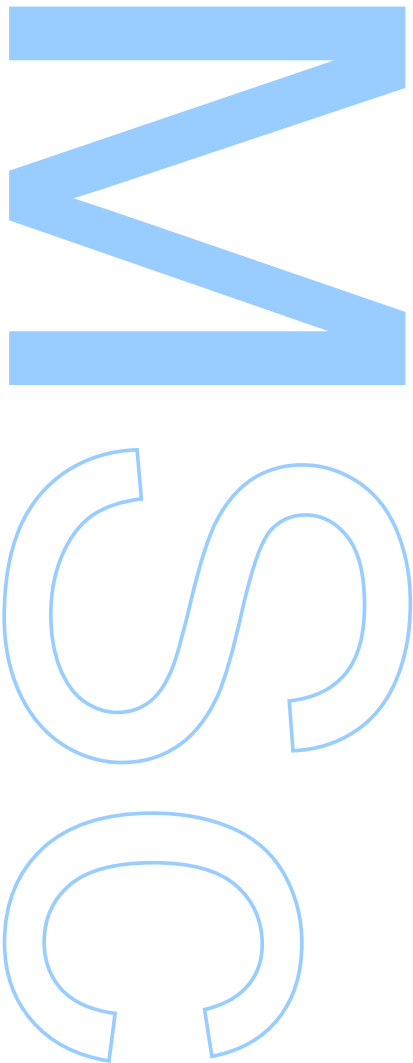
Decoding U.S. Political Discourse: A Natural Language Processing Automatic Approach to Analyzing Major Politicians' Tweets

Margarida Correia Mendonça
MSc. in Data Science
Department of Computer Science
2023

Supervisor
Prof. Dr. Álvaro Figueira, Assistant Professor
Faculty of Science, University of Porto




U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

" To invent your own life's meaning is not easy, but it's still allowed, and I think you'll be happier for the trouble. "

Bill Watterson

Acknowledgements

There are three people without whom this thesis would never have been finished. My mother and father, who were nothing but supportive when I told them my plans all those years ago. I am unbelievably lucky in having them. And Professor Álvaro Figueira, who has been far more patient and understanding than I deserve. Thank you for letting me explore a topic at my will, and thank you for dealing with my erratic motivation.

My family have been my cheerleaders since day one. I would like to particularly thank my sister, who managed to be present although 2733km away, and my Godmother, who made sure I felt her support every week.

My friends have been a constant source of love and happiness. I am blessed to have too many to list, but I hope they all know how important they are. A special thanks to my Cucas & Pitucha, who have been with me even when life gets in the way. And to Luís, who is the best surprise this adventure offered me.

And to Bernardo, my number one fan for the last seven years. I hope someone invents teleportation soon enough. Thank you, truly.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

MSc. Data Science

Decoding U.S. Political Discourse: A Natural Language Processing Automatic Approach to Analyzing Major Politicians' Tweets

by Margarida CORREIA MENDONÇA

As social media becomes increasingly prevalent, its impact on society is expected to grow accordingly. While it has brought about positive transformations, it has also amplified pre-existing issues. A thorough comprehension of this impact, aided by state-of-the-art analytical tools and by an awareness of societal biases and complexities, enables us to anticipate and mitigate potential negative effects.

One such tool is BERTopic. BERTopic is a novel, Deep Learning algorithm developed for Topic Mining. It has been shown to outperform other traditional methods such as LDA, and it is also highly modular, allowing for personalization at each stage.

In this dissertation, we focused on Twitter data from the two years of Congress 117^{th}. We began by conducting a review of the literature on Topic Mining approaches for short-text data. With that knowledge, we explored the potential for optimizing BERTopic, and analysed its success, both from an overarching and a monthly perspective. We found that we could improve BERTopic on coherence and stability, and that defining topics of this Congress include abortion, student debt and Judge Ketanji Brown Jackson. An app for better visualizing Congress topics was developed. We then combined these insights with the Moral Foundations Theory, which argues that there are key aspects that define cultures: Care, Loyalty, Fairness, Authority and Purity. We found that topics that last longer on Twitter have a more pronounced presence of the Care and Loyalty Foundations, and that political ideology is revealed in tweets by different uses of Care, Loyalty, Authority and Purity Foundations.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

Mestrado Integrado em Ciência de Dados

Descodificar o discurso político nos EUA: Uma Abordagem Automática em NLP para Analisar Tweets de Políticos Relevantes

por Margarida CORREIA MENDONÇA

A crescente prevalência das redes sociais é acompanhada pelo aumento do seu impacto na sociedade. Não desvalorizando as transformações positivas que geram, as redes sociais também amplificam problemas pré-existentes. Uma compreensão deste impacto, fomentada por ferramentas analíticas do estado-da-arte e por uma noção dos viéses e complexidades sociais, permitem-nos antecipar e mitigar potenciais efeitos negativos.

Uma dessas ferramentas é o BERTopic. O BERTopic é um algoritmo de Deep Learning recente, desenvolvido para Extração de Tópicos. Mostrou-se já que supera o desempenho de outros métodos tradicionais, como o LDA, e é também altamente modular, permitindo a personalização de cada etapa.

Nesta dissertação focámo-nos em dados do Twitter dos dois anos de trabalho do 117º Congresso. Começámos por conduzir uma revisão da literatura sobre métodos de Extração de Tópicos. Com esse conhecimento, explorámos as alternativas de optimização do BERTopic, e analisámos o seu sucesso, tanto de uma perspectiva global como mensal. Descobrimos que era possível melhorar o desempenho do BERTopic a nível de coerência e de estabilidade, e que os tópicos representativos deste Congresso são Aborto, Crédito Estudantil, e Juíza Ketanji Brown Jackson. Para uma melhor visualização dos tópicos do Congress, desenvolvemos uma *app*. Após isto, explorámos os tópicos extraídos à luz da Teoria de Fundações Morais, que defende a existência de aspetos-chave na identidade de culturas: Cuidado, Lealdade, Equidade, Autoridade e Pureza. Descobrimos que tópicos com maior duração no Twitter têm uma presença mais pronunciada de Cuidado e Lealdade, e que a ideologia política é revelada pelas diferentes utilizações de Cuidado, Lealdade, Autoridade e Pureza nos tweets.

Contents

List of Figures

List of Tables

Glossary

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MFD	Moral Foundations Dictionary
MFT	Moral Foundations Theory
ML	Machine Learning
NLP	Natural Language Processing
pLSA	Probabilistic Latent Semantic Analysis
STTM	Short-Text Topic Mining
SVD	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
VSM	Vector Space Model

This thesis is dedicated to my parents, for their unwavering love and support.

Chapter 1

Introduction

Social media has revolutionized society with impacts across wide-ranging domains. One such domain is politics. The political benefits of social media are plenty - for example, driving civic engagement, increasing both early registration and voting, and mobilizing grassroots movements [1].

An important dimension of this revolution was in challenging the role of traditional media. Before the adoption of social media by the public, the press held control over narrative control, as it served as the primary medium through which politicians communicated with the public. Social media allowed politicians to campaign and communicate directly, informally, and at a lower cost. This context created an opportunity for dialog between elected officials and those they represent [2].

Among social media platforms available today, Twitter* is the leading one adopted by politicians [3]. Created in 2006, as of December 2022 it boasts 368 million monthly active users worldwide [4]. Its current constraint of 280 characters per tweet fosters succinct sharing of opinions and information. This limitation has compelled users to adapt their language, promoting a unique form of brevity in online communication. Beyond this, Twitter's real-time nature allows for rapid dissemination and engagement, driving discussions on a wide variety of topics. Furthermore, hashtags and trending topics function as drivers for tracking popular subjects. Through regular tweeting and engagement with trending topics, politicians can enhance their visibility and cultivate a recognizable personal brand, vital for political success. As a result, Twitter has become a critical catalyst for political success [5].

*or X, as of July 2023

A real-life example of the impact of Twitter in politics is the successful 2016 U.S. presidential campaign of Donald Trump. It employed a variety of data mining and analytics techniques such as A/B testing, voter behavior forecasts, and geo-targeting to identify and engage with specific voter demographics [6]. The use of Natural Language Processing (NLP) allowed the campaign to gain crucial, real-time insights into public opinions and concerns, which in turn helped in crafting messaging and positioning. Some NLP techniques employed included Topic Mining, which extracted the themes under discussion at a given time, and Sentiment Analysis, with which it was possible to explore the feelings associated with those topics.

The downside of social media in general, and Twitter in particular, has also been well documented. Twitter has been shown to be an efficient vehicle for disinformation and misinformation, as well as being able to expand the power of confirmation bias and polarization of their audiences [7]. While Sentiment Analysis has been used to study these phenomenons [8–10], Moral Foundations Theory is a less-known framework [11]. It argues that moral reasoning is shaped by innate foundations found across cultures. These foundations influence moral judgments and play a substantial role in political ideology [12]. On Twitter, users often align with like-minded individuals, forming echo chambers where moral foundations are reinforced within specific political or ideological groups. This reinforcement further contributes to the deepening of polarization as individuals become increasingly entrenched in their respective moral perspectives [13–15].

1.1 Motivation

Since 2016, there have been rapid advancements in NLP, which can be attributed to the advent of Deep Learning and the availability of large-scale data [16]. One such development was the Transformer architecture [17]. Introduced in 2017, it employs a self-attention mechanism, allowing models to weigh the importance of each word in a sentence, significantly improving the context understanding by capturing long-range dependencies.

The resulting algorithms, built on a Transformer framework, have state-of-the-art performances and capabilities. They include Large Language Models, such as GPT-3, which is an autoregressive model with a large scale of parameters, and BERT, which uses bidirectional context to understand the meaning of words in a sentence. These models are applicable to a wide range of functions. Given the context of Twitter, our focus is on Topic

Mining. More specifically, we aim to understand how the recent, Deep Learning algorithm BERTopic performs on Twitter data, and explore the insights it obtains. BERTopic was introduced in 2019, and its recent nature is responsible for the low number of publications, compared with other, more established algorithms [18].

An additional step in this work is to analyse the topics obtained under the Moral Foundations Theory. Given the rise in political polarization [19], it can help develop new perspectives on the phenomenon, namely, whether more extreme uses of these foundations lead to a longer-lasting topic.

Finally, this work was also motivated by the interest in obtaining a global perspective of Congress topics. Topic Mining is often performed either on an individual, focusing on a concrete person [20], or on a global level, where a theme is under analysis [21, 22]. This work positions itself on an intermediate level, where the subject is the whole of Congress - specifically, the 117^{th}. This provides a comprehensive frame of reference on what were the concerns at this point in time in U.S. politics.

1.2 Goals

This work will focus on studying the capabilities of BERTopic against other traditional Topic Mining algorithms, pursuing an avenue of performance on the U.S. politics Twittersphere. Additionally, it aims to clarify whether the long-lasting nature of the topics is correlated with the language used, under a Moral Foundation framework. The research questions it aims to answer are as follows:

R.Q.1. How does BERTopic differ from traditional Topic Extraction techniques?

Given the novelty of this algorithm, a deep understanding of its functioning will help clarify the main research question, R.Q. 2.

R.Q.2. How can BERTopic be optimized to handle Twitter data?

This is the main research question. Given the novelty of BERTopic, there are few works done with it today. The optimization process will require a strong awareness of the different stages of the algorithm, and a set of evaluation metrics to be compared.

R.Q.3. What topics define Congress 117^{th}?

Is there a variation between parties? Are there overarching topics that will characterize this Congress?

R.Q.4. Is there a correlation between the duration of a Topic and Moral Foundations it displays?

Namely, how can Topic duration be interpreted under Morality Foundations Theory? Is there a significant difference in the Foundations displayed across party lines? How can this knowledge help with fomenting political success on Twitter?

1.3 Structure of this document

This document is structured into seven chapters, including this one. Chapter 2 aims to provide background knowledge on U.S. politics. Chapter 3 reviews the literature on Topic Mining of Short-Text Data. It begins by acknowledging the distinction between traditional and advanced approaches, which are assigned, correspondingly, to Topic Modeling and Topic Classification. It then focuses on the algorithms available for each type of approach: section 3.1.1 focuses on those developed for Topic Modeling, and Section 3.1.2 dives into Topic Classification.

Chapter 4 offers the methodology used at each step. Section 4.1 deals with data selection and extraction, examining the reasons that led to the selection of the Congress members for this study. Section 4.2 presents the preprocessing techniques used on textual data, as well as materializing those applied to our data. Section 4.3 delves into the capabilities of Topic Mining using BERTopic, analysing potential strategies for algorithm optimization. Section 4.4 is concerned with the evaluation of BERTopic optimization.

Chapter 5 begins by providing an exploratory analysis of our data. Section 5.2 presents the results of optimizing BERTopic, comparing its performance across the evaluation metrics selected. Section 5.3 presents the topic results from a global perspective, being extracted on the whole dataset, without a concern for the time interval.

In Chapter 6, we begin by offering the results from a monthly perspective. This section also develops the concept of Topic Longevity (section 6.1), and its relationship with the Moral Foundations framework (section 6.2). A discussion on these results is available on section 6.3.

In Chapter 7 we discuss our main conclusions and answer the research questions. We also present a general summary of the study conducted, the limitations faced, and the possibilities for extending this work.

Finally, the interested reader can find the work developed for this thesis can be found on the GitHub repository bertopic-twitterverse.

Chapter 2

Background

Given the context of this work, this Chapter aims to ensure the reader has the essential knowledge on U.S. politics. We provide some background information on its design, specifically on Congress's structure and functioning. A brief overview of U.S. politicians activity on Twitter is also delivered.

2.1 A Brief Contextualization of U.S. Politics

The United States political system encompasses three branches: Legislative, comprised of Congress, and responsible for law-making; Executive, which implements laws and policies and is led by the U.S. President; Judicial, composed of the Supreme Court and smaller federal courts, who ensure laws are abided.

The U.S. political landscape contains over 400 different parties, but it is dominated by the only two nationally recognized ones. The Democratic Party platform is social liberalism [23] and it believes *"that the economy should work for everyone, health care is a right, our diversity is our strength, and democracy is worth defending* [24]. The Republican Party, also known as GOP ('Grand Old Party), runs on a conservative ideology and stands *"for freedom, prosperity, and opportunity [...]. The principles of the Republican Party recognize the God-given liberties while promoting opportunity for every American"* [25].

2.1.1 *We The People*

The U.S. Constitution has been in operation since 1789 and is the longest-surviving written charter of government in existence [26]. Article I of the Constitution creates a Bicameral Congress consisting of a Senate and a House of Representatives [27]. This solution came as a result of the Great Compromise. At the time of writing the Constitution, Framers - those involved in drafting it - had conflicting positions on defining State representation. Framers from large States defended representation proportional to population, while those from small States argued for equal representation. As a compromise, Congress was separated into two Chambers [28].

The House of Representatives gathers a number of delegates from each State proportional to its population. There are currently 435 Representatives, which are elected by the people every two years. Among other powers, the House can impeach the President and other elected officials, it decides presidential elections if no candidate wins the majority electoral college, and it can initiate tax-raising bills. The House's relevant roles include Speaker of the House, its highest-ranking member and presiding officer, Majority Leader, the second-highest ranking member of the majority party in the House, and Minority Leader, the leader of the minority party in the House.

The Senate is composed of two Senators from each State, totaling 100 legislators. One-third of the Senate is elected every two years, and each Senator runs a six-year mandate. This Chamber holds the impeachment trials initiated by the House, ratifies treaties, and has confirmation power. It has a President of the Senate, a role assigned to the U.S. Vice-President, and similarly to the House of Representatives, it has a Majority Leader and a Minority Leader [29].

2.1.2 The 117^{th} Congress

The 117^{th} United States Congress was convened on January 3, 2021, and resulted from the 2020 elections. The House majority was retained by the Democratic Party, with Speaker Nancy Pelosi, Majority Leader Steny Hoyer, and Minority Leader Kevin McCarthy. After the inauguration of President Joe Biden and Vice-President Kamala Harris on January 20th, Harris' consequent swearing-in to President of the Senate gave Democrats control of this Chamber, with Majority Leader Chuck Schumer and Minority Leader Mitch McConnell.

Some significant moments of the 117^{th} Congress mandate include the January 6^{th} Capitol attack, Donald Trump's impeachment, U.S. sanctions on Russia following the Ukraine invasion, the nomination of Ketanji Brown Jackson to the Supreme Court, and the overturn of Roe v. Wade. Congress also passed relevant bills such as the Inflation Reduction Act, Infrastructure and Jobs Act, CHIPS and Science Act, Honoring our PACT Act, and the Respect for Marriage Act [30].

2.2 American Politics and Twitter

As of 2022, Congress members total 515 Twitter accounts [31], but not all accounts are used the same. Firstly, some members have personal accounts and professional accounts. The former tends to be more informal, with the member managing it directly, while the latter is run by a team. Secondly, Democrats tend to tweet more and have more followers, but engagement is evenly split among parties [32].

There are some members who have taken to Twitter better than others. Alexandra Ocasio-Cortez (D*) is the most-followed House Representative, having 3 million followers on her personal account. The second-most followed is Nancy Pelosi (D), with 2.2 million followers, and Adam Schiff (D), with 1 million. Other active accounts included Jim Jordan (R†), Bernie Sanders (I‡), and Mitt Romney (R) [33].

*Democrat
†Republican
‡Independent

Chapter 3

Literature Review

A well-known fact in the data science community is that the proliferation of the Internet, and subsequently of social media, has generated massive amounts of data [34]. It is also well understood that it requires proper manipulation and analysis to extract useful knowledge. Textual data represents approximately 75% of all data on the web [35], which presents its own processing challenges. Textual data is unstructured, in the sense that it does not have an underlying schema. It is more ambiguous - making use of synonyms, slang, and abbreviations - and often requires an understanding of its context. Spelling errors and lack of grammar further complicate its analysis. When studying textual data extracted from social media, its informal and short-text nature deepens these challenges.

NLP is a branch of Computer Science that aims to overcome these obstacles, enabling computers to understand, interpret, and generate natural, human language. In NLP, text data organized into datasets is called a corpus, while documents are the sources of the data, such as books, news articles, and emails [36]. NLP has a wide range of applications: Document Summarization consists of identifying and extracting relevant from lengthy documents and providing overviews; Information Retrieval is the task of categorizing documents according to their main subjects, which is crucial for search engines or recommendation systems; Sentiment Analysis involves analyzing the emotions expressed in a document, to understand how the approached topics are perceived.

To successfully perform these tasks, there are several NLP techniques available. In Part-of-Speech Tagging, each word in a sentence of the corpus is assigned a part-of-speech (e.g., noun, verb, adjective). Named-Entity-Recognition identifies and categorizes entities into names of persons, organizations, locations, among others. Another NLP technique

is Topic Mining*. It consists of using algorithms to automatically identify and categorize meaningful topics in a document. The objective is to uncover underlying themes within the corpus without prior knowledge of those topics. Since we wish to explore the subjects of politicians' tweets, Short-Text Topic Mining (STTM) is the focus of this Chapter. In terms of literature, we build upon the review of Murshed et al. [37]. We focus on publications from 2022 and 2023 that look into Twitter Topic Mining. In the cases where this was too restrictive, we look into publications on other social media platforms.

3.1 Topic Mining of Short-Text Data

In a broad sense, Topic Mining can be divided into two categories: Topic Modeling and Topic Classification. The former is the task of finding topics from unlabeled documents, while the latter refers to the automatic labeling of those topics. This separation is a result of Machine Learning (ML) progress over the past ten years, which has brought forth new algorithms using novel approaches and achieving higher performance. This chapter presents the most relevant models for Topic Mining in a chronological fashion. Fig. 3.1 offers a different perspective, with models organized according to techniques used: advanced techniques encompass ML models, while traditional ones use classical, statistical approaches. This taxonomy is a simplified adaption of Murshed et al. [37].

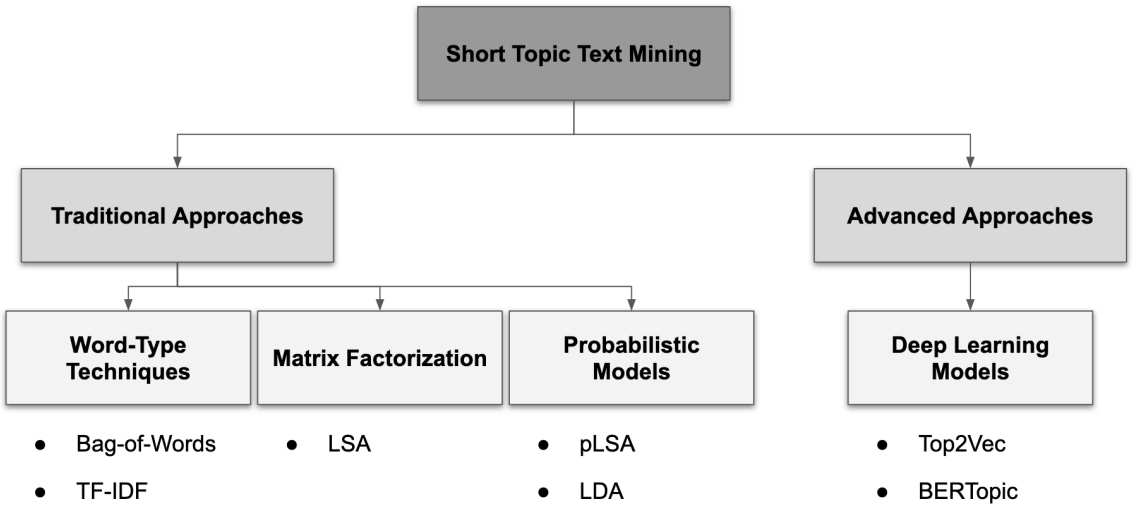


FIGURE 3.1: Taxonomy of Topic Mining models

*Also known as Topic Extraction and Topic Discovery.

3.1.1 Topic Modeling

Topic Modeling consists of extracting latent topics from a corpus of unlabeled documents. Latent topics are those that are not immediately perceivable in the document, instead being suggested by a collection of words. Topic Modeling first originated as text representation metrics. The **Bag-of-Words** (BoW) was an initial attempt, where word frequency is used as a classifier feature [38], as was **Term Frequency-Inverse Document Frequency** (TF-IDF), which favors words that are common across a single document, and uncommon across all documents [39]. Topic Modeling was performed with these metrics through clustering techniques such as Hierarchical Clustering [40] and K-Means clustering [41]. More elaborate approaches were developed with the aim of not only producing document clusters but also generating topic interpretations of such clusters. A first major progress was due to the **Vector Space Model** (VSM), where a document is represented as a vector of words*. This model, however, lacked the ability to capture polysemy - a word with multiple meanings - and synonymy - multiple words with the same meaning - and provided a small reduction in dimensionality [42].

The **Latent Semantic Analysis** (LSA) model by Deerwester et al. [43] has a main foundation in the distributional hypothesis. In essence, the idea is that texts with similar meanings are expected to have similar representations, which translates to being closer to each other in the vector space.

To achieve this, a term-document matrix X is constructed, where rows represent terms (words) and columns represent documents. Each cell represents the frequency of a term in a particular document. Since documents can vary widely in length and in terms, this matrix tends to be large and highly sparse - a challenge for efficient analysis. This obstacle is overcome through a matrix factorization technique called Singular-Value Decomposition (SVD). Its goal is to find the lower-dimensional representations of terms and documents that preserve the semantic relationships in X. This way, LSA reduces the dimensionality of the vector space while capturing the patterns in the data.

Specifically, X is decomposed into its three constituent parts: the term-topic matrix U (size $m.r$), the diagonal matrix of singular values Σ, and the document-topic matrix V (size $r.n$). U represents the relationship between terms and the r selected topics, where its elements indicate the strength of this relationship. V represents the relationship between documents and topics, where the elementsand its elements indicate the strength

*JR Firth summarized this idea with the saying *"You shall know a word by the company it keeps"*

of the association. Σ's values correspond to how much each of the r latent topics explain variability in the data. Formally:

$$X = U\Sigma V^T \tag{3.1}$$

LSA was widely adopted at the time of its creation and is still used today, despite the high computational cost of calculating the SVD and the often hard-to-interpret generated feature space. Valdez et al. [44] used LSA to analyze tweets on the 2016 U.S. Election, showing topics paralleled the most frequent policy-related Internet searches at the time. More currently, Sai et al. [45] used LSA to guide the identification of fake news on Twitter, while Chang et al. [46] aimed to gain insights into public perception of the Russia-Ukrain conflict on Twitter through LSA. While works have shown that alternative models outperform LSA [47], others have tried to overcome LSA's limitations: Karami et al. [48] proposed Fuzzy LSA Topic Mining in health news tweets; Kim et al. [49] presented Word2Vec-based LSA for blockchain trend analysis.

In 1999, Hofmann introduced **probabilistic LSA** (pLSA) as an alternative to LSA under a probabilistic framework [50]. It assumes topics are distributions of words and its aim is to find a model of the latent topics that can generate the data in the document-term matrix. Formally, this distribution is defined as:

$$P(d, w_{di}) = P(d) \sum_{z=1}^{K} P(w_{di}|z) \cdot P(z|d) \tag{3.2}$$

where d are documents, $d \in D = \{d_1, d_2, ..., d_m\}$, w are words, $w \in W = \{w_1, w_2, ..., w_n\}$ and z are the latent topics, $z \in Z = \{z_1, z_2, ..., z_K\}$.

Given that pLSA has been built upon, its usage in NLP tasks is very infrequent. Kumar and Vardhan [51] used pLSA for topic-based sentiment analysis of tweets, and Shen and Guo [52] used it to add context to a translation task. While an advancement to LSA, pLSA is prone to overfitting, and new models would attempt to overcome this limitation.

Latent Dirichlet Allocation (LDA) was first proposed in 2003 by Blei [53] and is a widely used text classification algorithm. Its fundamental idea is that *"documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words"*. The number of latent topics is user-defined. Given parameters α and β, the joint distribution of a topic mixture θ, a set of K topics $\mathbf{Z}$ and a set of N words $\mathbf{W}$ is given by:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{Z_{j,t}}) \tag{3.3}$$

Fig. 3.2 (from Anastasiu et al. [54]) illustrates the LDA algorithm and how each parameter interacts with all others.

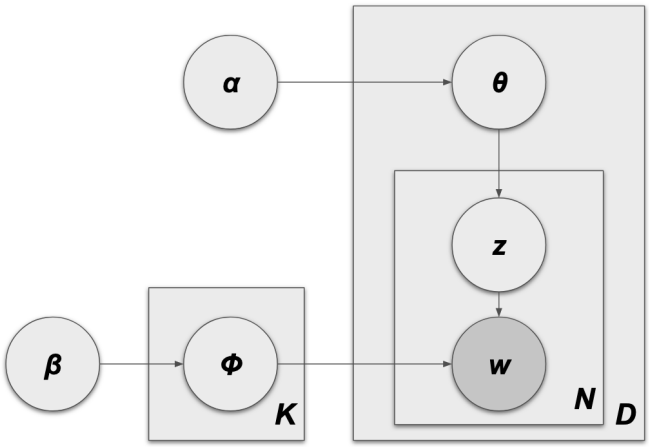


FIGURE 3.2: LDA algorithm

The following example helps to better understand the mechanisms:

Let us have a corpus of ten documents, five words, and three topics: M = 10, N = 5, K = 3. The documents in the corpus follow a Dirichlet distribution, Dir(α), which defines how they relate to the topics. In this example, consider a triangle, where each vertex corresponds to a topic. The triangle shading is darker in the corners and lighter in the center, suggesting the probability of the documents is higher near the individual topics/vertices than their combinations. Dir(β) defines how the topics relate to the words. In this example, consider a tetrahedron (which has all four vertices equidistant to the center) where each vertex corresponds to a corpus word. Similarly to the triangle, the shading in this figure is lighter in the center and darker in the corners, suggesting the probability of the topics is higher closer to the individual words than to their combinations.

From the Dirichlet distributions, a new document, D, can be generated as set in 3.3. D will have a mixture, θ, of representations of each topic - for instance, 70% of Topic 1, 20% of Topic 2, and 10% of Topic 3. These percentages create the multinomial distribution of **Z** ($P(Z|\theta)$), giving the topics of the words in D. Having the topics, it is necessary to find the words, **W**. Similarly, the multinomial distribution of **W** is subject to β. For each topic in D, a word is selected with probability $P(W|\phi)$.

Having D, it is possible to compare it with the original documents and select the parameters that generate the closer ones, i.e., that maximize 3.3.

To account for different length documents, a Poisson distribution is attached to the formulation in 3.3. Gibbs sampling was later suggested as an efficient method for estimating θ and ϕ [55].

LDA has been a popular and widely adopted model for several tasks. In the case of Topic Mining of Twitter data, it has been applied to market research ([56–58]), health ([59–61]), and specifically COVID-19 ([62–64]), as well as other subjects such as climate change [65], layoffs [65] and hate speech [66].

As the limitations of LDA seem to impact performance in STTM, research has shifted to modifying the original LDA model. Table 3.1 lists some LDA iterations, with a simple description of each one.

TABLE 3.1: LDA iterations for STTM

Model	Objective
TM-LDA	Temporal feature-based TM
RO-LDA	Solve lack of local word co-occurrence
TSVB-LDA	Increase accuracy
BR-LDA	Remove background words
Logistic LDA	Handle arbitrary input (e.g. images)
TIN-LDA	Explore the interest of microblog users
TH-LDA	Hierarchical dimensions for high semantics
FB-LDA	Handle the binary weighing of words
MBA-LDA	Select best topic representation words
SS-LDA	Does not require annotated training data
Twitter-LDA	Achieve high accuracy with Twitter data

Of the listed models, **Twitter-LDA** has achieved particularly interesting results in performing STTM [67]. As its name suggests, it has been tailored to perform LDA on Twitter data by including two main differences:

1. It has expanded the preprocessing steps. Instead of removing all non-alphanumeric characters, it saves those that are relevant in tweets - hashtags, mentions, and URLs - and uses them for further information (e.g. as potential topic labels).

2. Words in a tweet are either topic words or background words, and each has underlying word distributions. A given user has a topic distribution, and follows a Bernoulli distribution when writing a tweet: firstly, the user picks a topic based on their own topic distribution; at the time of choosing each next word, it either chooses a topic word or a background one.

3.1.2 Topic Classification

The development of ML and neural networks has brought forth new, better-performing Topic Mining models. On top of extracting latent topics, they are capable of performing unsupervised classification of those topics.

A key concept in Topic Classification is word embeddings. Word embeddings consist of transforming individual words into a numerical representation, i.e., vectors. This vector aims to describe the word's characteristics, such as its definition and context. It is then possible to identify similarities (or dissimilarities) between words - for instance, understanding synonyms of different terms, or ironic uses of the same one. Once the embeddings have been generated, it is possible to position the words in the vector space, with similar words closer to each other and dissimilar words further apart.

Although word embeddings and document-term matrices (such as those generated by TF-IDF) have the same goal of representing words in a numerical fashion, they do so differently. In a matrix, all words in a corpus are represented and separated by document. It is, therefore, a large, sparse object where context and meaning are lost. Additionally, vectorization is corpus-dependent, which complicates training in different datasets.

Word2Vec by Mikolov et al. [68] consists of a group of two-layer neural networks used to generate word embeddings. It makes use of two architectures with opposite goals:

- **Continuous Bag-of-Words**: predicts a word based on its neighbors. The architecture consists of an input layer (the neighbor words), a hidden layer, and an output layer (the predicted word). The hidden layer learns the vector representation of the words.

- **Skip-Gram**: predicts neighbors based on a word. The architecture consists of an input layer - the word -, a hidden layer, and an output layer (the neighbors). The hidden layer outputs the vector representation of the individual word.

Doc2Vec was proposed in 2014 by Le and Mikolov [69]: It expanded on the ideas of Word2Vec: instead of words, the model is able to learn vector representations of documents. To do so, it includes an additional vector to represent the document as a whole. Doc2Vec presents two possible approaches:

- **Distributed Memory**: predicts the next word given an input layer of document vector and a 'context' vector of current words (those surrounding the target word). The latter consists of a sliding window, which traverses the document as the target word changes. These two vectors are concatenated and passed through a hidden layer, where their relationships are learned. The outputs consists of a softmax layer, predicting the probability distribution of the target word. During training, the hidden and output layer's weights are adjusted to minimize prediction error. The document vector is also updated at this stage.
- **Distributed Bag-of-Words**: predicts a set of words given only the document vector. The difference of this approach to Distributed Memory is that the input layer only consists of the document vector. There is no consideration of word order or context within the document, which simplifies the prediction task.

While neither Word2Vec nor Doc2Vec are used for Topic Mining, introducing these models is necessary to explain Top2Vec. **Top2Vec** is an unsupervised model for automatic Topic Mining that does not require an *a priori* user-defined number of topics [70]. It does so in five steps:

1. **Create embeddings**: generates embedded document and word vectors, using Word2Vec and Doc2Vec
2. **Reduce dimensionality**: reduces the number of dimensions in the documents using UMAP (Uniform Manifold Approximation and Projection). UMAP is a dimensionality reduction algorithm useful for complex datasets. It also has the advantage of clustering similar data points close to each, helping to identify dense areas. UMAP calculates similarity scores across pairs of data points, depending on their distances and the number of neighbors in the cluster (which is user-defined). It then projects them into a low-dimensional graph, and adjusts distances between data points according to their cluster [71]. By using this algorithm, this step maintains embedding variability while decreasing the high dimensional space, and it also manages to identify clusters in the data.

3. **Segment clusters**: the model identifies dense areas of documents in the space: if a document belongs to one of the areas, it is given a label, otherwise, it is considered noise. This is performed with HDBSCAN, an hierarchical clustering algorithm that classifies data points as either Core Points or Non-Core Points. Core Points are those with more than ϵ neighbors. After classifying all observations, a Core Point is assigned to cluster 1. All other points in its neighboring region are added to the cluster, and so are the neighboring points of these. When no other Core Points are able to be assigned to this cluster, a new Core Point is assigned to cluster 2. This process is repeated until all data points have been assigned to a cluster [72].

4. **Compute centroids**: for each of these dense areas, the centroid of the document vectors is calculated which is the topic vector. the centroid of each dense area is calculated - this is the topic vector

5. **Find topic representations**: the topic representations consist of the nearest word vectors to the topic vector

In the context of short-text data, Top2Vec has been used to analyze social media of patients with rare diseases, being shown that it performs better than LDA [73]. Zengul et al. [74] showed that Top2Vec results are more closely correlated to LDA's than to LSA's, and than LDA's to LSA's. Vianna and Silva De Moura [75] achieved successful results in extracting topics from the abstracts of legal cases - which are much shorter than the complete document, and Crijns et al. [76] managed to scrap web texts that identify innovative economic sectors or topics. Bretsko et al. [77] applied Top2Vec to abstracts of academic papers.

Although Top2Vec was only introduced in 2020, the approach is now considered dated. It has been overcome by Transformers as the state-of-the-art models in NLP [78].

BERTopic belongs to BERT, a family of language models based on Transformers. Transformers are a type of Deep Learning architecture that has earned significant attention due to their ability to handle sequential data efficiently and capture complex patterns in text.

The fundamental innovation brought forth by Transformers is the self-attention mechanism, which allows the model to weigh the importance of different words in a sentence while considering their relationships. This enables Transformers to capture both local

and global dependencies in the data, making them particularly effective for tasks involving long-range dependencies, such as language translation, text generation, and language understanding.

The transformer architecture consists of two main components:

- **Encoder-Decoder Structure**: the encoder takes in the input sequence and processes it through multiple layers of self-attention and feed-forward neural networks, capturing contextual information. The decoder then generates the output sequence based on the encoded representation of the input and its own self-attention mechanism.

- **Self-Attention Mechanism**: self-attention allows each word in a sequence to consider the relationships with all other words in the sequence. This is achieved by calculating weighted representations of the input words, where the weights are determined by the relevance of each word to the current word being processed. The self-attention mechanism consists of three main components: query, key, and value. These components are used to compute attention scores that determine how much each word contributes to the representation of the current word. Multiple attention heads are used in parallel to capture different aspects of the relationships.

Transformers have become the foundation for many state-of-the-art NLP models, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), T5 (Text-to-Text Transfer Transformer), and more.

In 2020, BERTopic was presented as a *"a topic model that leverages clustering techniques and a class-based variation of TF-IDF to generate coherent topic representations"*. It aims to overcome other models' limitations, such as not being able to take into account the semantic relationships between words.

BERTopic mines topics in five stages (Fig. 3.3, from [79]). Although the default algorithms have been selected for the reasons presented below, BERTopic is highly modular, and users can customize it at each step. It also allows the fine-tuning of each default algorithm through its corresponding hyperparameters.

BERTopic begins by converting documents into embedding representations. The default algorithm for this is Sentence-BERT, which achieves state-of-the-art performance on such tasks [80]. The dimensionality of these embeddings tends to be quite high, with some models achieving ten thousand dimensions [81]. UMAP is used to reduce this to

2D or 3D since it preserves local and global features of high-dimensional data better than alternatives such as PCA or t-SNE [71]. With data in a more feasible vector space, it can now be clustered. HDBSCAN allows noise to be considered outliers, does not assume a centroid-based cluster, and therefore does not assume a cluster shape - an advantage to other Topic Modeling techniques [72]. The next step is to perform c-TF-IDF. This is a variation of the classical TF-IDF: firstly, generate a bag-of-words at the cluster level, concatenating all documents in the same class. From this, apply TF-IDF to each cluster bag-of-words, resulting in a measure for each cluster, instead of a corpus-wide one.

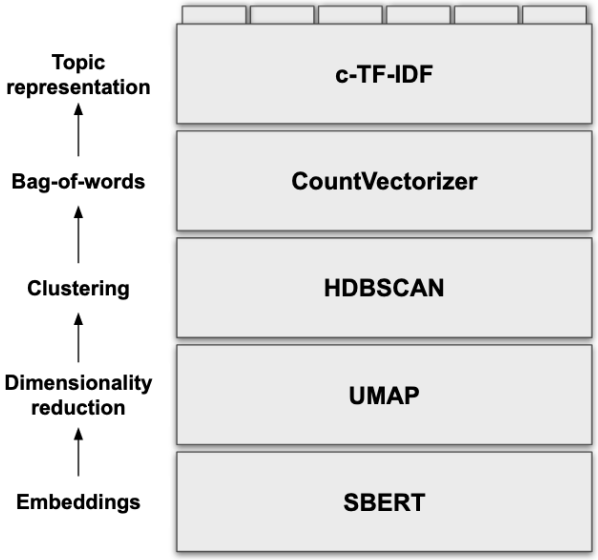


FIGURE 3.3: BERTopic algorithm

BERTopic has been increasingly adopted in STTM. Hägglund et al. [82] aim to identify the topics discussed on Twitter by autistic users, LI [83] apply it to examine how tweets can predict tourist arrivals to a given destination. Strydom and Grobler [84] analyse COVID-19 misinformation on Twitter, Turner et al. [85] look for concept drift on historical cannabis tweets and Koonchanok et al. [86] attempt to track public attitudes towards ChatGPT on the same platform. Grigore and Pintilie [87] train BERTopic to help identify potential patients of eating disorders based on their social media. Finally, concerning hate speech on social media, Mekacher et al. [88] analyse the risk of banning malicious accounts, showing that users that were banned from Gettr and not Twitter were more toxic on the latter. Schneider et al. [89] present a pipeline for detecting hate speech and its targets on Parler.

While BERTopic has been shown to be competitive against other models, it presents some disadvantages. It assumes that each document only has one topic, which does not necessarily correspond to the truth. It also uses bag-of-words to generate topic representations, meaning it does not consider the relations between those words. As a result, words in a topic might be redundant for interpreting the topic.

Chapter 4

Methodological Approach

Given the novelty and performance of BERTopic described in Chapter 3, and given the research questions in Chapter 1, this Chapter starts by offering a description of the methods used to select and extract data (Section 4.1), followed by the methodology followed to preprocess it (Section 4.2). Section 4.3 elaborates on BERTopic capabilities, and Section 4.4 offers a discussion on the techniques used for model evaluation.

4.1 Data Selection and Extraction

This work focuses on Twitter activity of the members of the 117^{th} U.S. Congress. This encompasses 102 senators and 413 Representatives Twitter handles, totaling 515 accounts [31]. To reduce the number of accounts in the analysis, a first look was taken into how active they were. Tweet Congress is a grassroots project that informed users about Congress Twitter activity. While no longer available at the time of this writing, it was possible to use it for data selection. Among other information, it listed the top 10 most active users.

Besides these names, those who played an important role in the Congress structure - namely, the majority and minority leaders - were also selected. Additionally, an effort was made to balance the number of Democrats and Republicans, and both personal and professional accounts of a given selected name were included.

Table 4.1 lists the final selection. It consists of 27 members with 40 accounts. Thirteen are Democrats (D), 13 are Republicans (R) and one is Independent (I). Eleven are Representatives and 14 are Senators, with the remaining two corresponding to the U.S. President and Vice-President. Table A.1 provides further information on each member.

Tweets of the above-mentioned accounts were extracted using the Twitter API (v.1.1). For each account, the tweets corresponding to the duration the 117^{th} Congress - from January 2021 to December 2022 - were obtained, amounting to 27782. The retrieved data included the tweet text, its publishing date and time, and the account handle.

TABLE 4.1: Final selection of Congress Twitter accounts

Member name	Twitter handle(s)	Chamber	Party
Adam Schiff	@RepAdamSchiff @AdamSchiff	Senator	D
Alexandria Ocasio-Cortez	@AOC @RepAOC	Representative	D
Andy Biggs	@RepAndyBiggsAZ	Representative	R
Bernie Sanders	@BernieSanders @SenSanders	Senator	I
Charles Schumer	@SenSchumer @chuckschumer	Senator	D
Cory Booker	@SenBooker @CoryBooker	Senator	D
Elizabeth Warren	@ewarren @SenWarren	Senator	D
Jim Jordan	@Jim_Jordan	Representative	R
Joaquin Castro	@JoaquinCastrotx	Representative	D
Joe Biden	@JoeBiden @POTUS	President	D
John Cornyn	@JohnCornyn	Senator	R
John Kennedy	@SenJohnKennedy	Senator	R
Kamala Harris	@KamalaHarris @VP	Vice-President	D
Kevin McCarthy	@GOPLeader	Representative	R
Lee Zeldin	@RepLeeZeldin	Representative	R
Marco Rubio	@SenRubioPress @marcorubio	Senator	R
Marjorie Taylor Greene	@RepMTG	Representative	R
Marsha Blackburn	@MarshaBlackburn	Senator	R
Matt Gaetz	@RepMattGaetz	Representative	R
Mitt Romney	@SenatorRomney @MittRomney	Senator	R
Nancy Pelosi	@TeamPelosi @SpeakerPelosi	Representative	D
Patty Murray	@PattyMurray	Senator	D
Pramila Jayapal	@RepJayapal @PramilaJayapal	Representative	D
Rand Paul	@RandPaul	Senator	D
Rick Scott	@SenRickScott	Senator	R
Steny Hoyer	@LeaderHoyer @StenyHoyer	Representative	D
Ted Cruz	@SenTedCruz	Senator	R

4.2 Data Preprocessing

4.2.1 Preprocessing in Theory

Data preprocessing is the first step when building a model. It reduces data dimensionality and its proper application can have a large impact on the final results [90]. Although the process often has to be fine-tuned to the data at hand, it can be structured into five main stages. Firstly, tokenization consists of converting sentences and paragraphs into smaller units, such as words. This allows the machine to both understand the larger text, and its smaller parts [91]. While the simplest tokenization breaks a sentence down to words according to spaces, challenges arise when there are contractions, symbols or lack of punctuation.

The second step is normalization. This includes lowercasing all letters and removing punctuation and odd characters. Then, stopwords are removed, which consist of common words in a given language, such as prepositions or pronouns, that add no significant meaning to a text. In English, such words can be 'and', 'the', 'or', for example, [92]. By normalizing data and removing stopwords, dimensionality is significantly reduced [93].

This step is followed by stemming and/or lemmatization. Stemming is the process of reducing words to their base form. For example, 'run', 'running', and 'ran' would all be reduced to 'run'. This further normalizes the text. Stemming can have poor results: over-stemming occurs when two words of different roots are reduced to the same form, and under-stemming occurs when words of the same root fail to be reduced to the same form. Currently, there are three commonly used stemming algorithms:

- **Porter Stemmer**: developed in 1980, it simplified previous options (such as Lovins') by implementing a group of five heuristics. It is fast and easy to use, but it is only available for English text [94].

- **Snowball Stemmer**: developed as an improvement over Porter's, it can be used for languages other than English. It is also faster and considered to be more aggressive, in the sense that words are further reduced in this algorithm [95].

- **Lancaster Stemmer**: considered the most aggressive algorithm, because its stricter rules tend to lead to over-stemming. Some implementations of Lancaster allow setting user-defined rules [96].

A stemmer operates on a single word, disregarding its context, and it can create a non-word. Lemmatization, on the other hand, is the process of reducing words to their base form (lemma, i.e., dictionary form), while accounting for context. For example, the stemming of 'better' would be 'better'. Its lemma can be 'good', if in the context of the text it is used as an adjective*, or it can be 'better' if it is a verb†. It follows that lemmatization tends to be slower than stemming. Lemmatization is usually performed by using a database of words and their relationships. **WordNet** is one such resource for the English language, which is used by WordNet Lemmatizer.

The success of preprocessing is contingent on choosing the adequate steps and algorithms for a given dataset.

4.2.2 Preprocessing in Practice

Fig. 4.1 illustrates the steps taken to preprocess the extracted Twitter data. There are several Python packages available for NLP preprocessing: NLTK, SpaCy, NLP Stanford, among others. Given the simplicity of the tasks, either of these could have been chosen. NLTK was selected due to its ease of use, a rich library of resources, and previous experience using it.

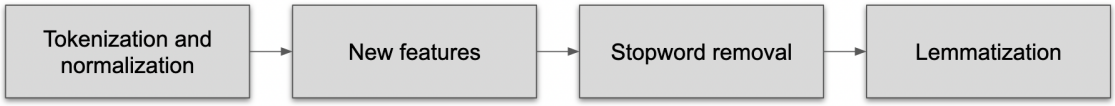


FIGURE 4.1: Preprocessing steps

The first step was to tokenize and clean the tweets. This included:

- Replacing web characters, such as *
*, *%quot;*, *'* and *&*, with their intended meaning
- Removing hashtags, hyperlinks, and mentions, creating new tweet features with each one
- Remove 'RT' from tweets that added no further information
- Remove punctuation
- Remove tweets of length 0, which could result from the previous steps

*"The green apple is better."
†"There are ways to better your score."

Besides the hashtags, mentions, and hyperlink new features, the original data also included a feature of *color*, corresponding to the tweet author party - Democrat (blue), Republican (red), Independent (white), and of *account type*, either professional or personal. Having cleaned the tweets, stopwords, as defined by the NLTK corpus, were removed, and lemmatization was performed. Lemmatization, instead of stemming, was used, because short-text data already has small windows of context, and further removal would be ill-advised. Since removing stopwords can reduce tweets to length 0, those were once more dropped.

4.3 Topic Extraction with BERTopic

As mentioned in Chapter 3, BERTopic is a high-performance Topic Mining algorithm. Given the fact that it is a recent algorithm and its state-of-the-art performance [97, 98], it was selected for this work.

BERTopic is also highly modular. To maximize the benefits of this, a search for the best algorithms at each stage of the process was performed, using a fraction (20%) of the total data. Table 4.2 shows the options considered for each step of BERTopic. While it would be interesting to test all possible combinations, it would result in over 100 models. Instead, each option of a given stage is evaluated separately, while keeping all other stages with default algorithms/parameters [99].

TABLE 4.2: BERTopic personalization options considered

Stage	Algorithms / Parameters
Embeddings	SBERT, SpaCy, Word2Vec
Dimensionality reduction	UMAP, PCA, Base dimensionality model
Clustering	HDBSCAN, k-Means
Vectorizer	*ngram_range*, *min_df*, *max_features*
Topic representation	*reduce_frequent_words*, *bm25_weighting*

4.3.1 Embeddings

BERTopic's default algorithm for embedding is SBERT, which is transformer-based and has been shown to outperform other methods [80]. SBERT also has a wide variety of pre-trained models, including multi-lingual and multi-modal. In the context of the chosen data, "all-MiniLM-L6-v2" is used.

SpaCy offers the option to use both transformer and non-transformer models. Since the former have consistently outperformed the latter, they are used. A challenge with using SpaCy transformers is the intensive use of memory, which could spoil results in the complete dataset.

Word2Vec, which was described in Chapter 3, is also included in this analysis. While it is still a neural network, it is not transformer-based, offering the chance to add an additional comparison in this stage. Additionally, it is implemented through Gensim, which is considered "the fastest library for training of vector embeddings" [100] due to its efficient implementation [101].

4.3.2 Dimensionality Reduction

UMAP is the default dimensionality reduction technique used by BERTopic. PCA (Principal Component Analysis) can also be selected, and it is often found to be faster than UMAP. PCA is a linear dimensionality reduction technique that assumes correlation between the dataset features. It aims to transform them into a new coordinate system based on a set of uncorrelated variables (principal components) through their eigenvalues. This is expected to capture the most significant variability in the data [102]. It is also possible to skip the dimensionality reduction stage, setting it to the base model. This might be used for simpler datasets, where high dimensionality is not an issue, or for cases where the previous algorithms do not perform adequately.

4.3.3 Clustering

While HDBSCAN is the default algorithm, k-Means gives the option of not allowing outliers, and therefore they are compared against each other. K-Means is an unsupervised clustering algorithm that partitions a dataset into k clusters. It iteratively assigns data points to the nearest cluster centroid based on Euclidean distance and then recalculates the centroids as the mean of all points assigned to each cluster, until convergence. The original centroid is chosen randomly from the data points, and the *k-1* remaining centroids are determined as the most distant from the first. In doing so, k-Means attempts to minimize the within-cluster variance [103]. K-Means is particularly useful when the number of clusters is known *a-priori*, while HDBSCAN is more versatile and more adequate to high-dimensional data.

4.3.4 Vectorizer

The Vectorizer stage uses CountVectorizer to create its vectorization and, instead of algorithms, it can be personalized by its parameters [104]. These parameters are tightly related to the specific characteristics of our data, such as how relevant n_grams are. *ngram_range* allows us to decide how many tokens are in each entity, in a topic. It uses the range of n_grams in the data, selecting up to is maximum as a topic representation. Its default is (1,1).

Another parameter allows us to set the minimum frequency a word must have to be included in a topic representation (*min_df*). While BERTopic is designed to remove low-frequency words with the c-TF-IDF calculation (see next stage), removing them beforehand helps reduce the topic-term matrix, potentially speeding up the algorithm.

Conversely, CountVectorizer also offers a parameter to select the top n features of a topic representation (*max_features*). This limits the size of the topic-term matrix, making it less sparse. This parameter and the previous have similar impacts - reduce the size of the topic-term matrix. It is expected that, when optimizing BERTopic, the two parameters will balance each other out to achieve the 'best-sized' matrix. In cases where the matrix is very small (for example, there are few documents), these parameters often have to be set to default values, or else BERTopic is not able to run.

4.3.5 Topic Representation

c-TF-IDF is calculated for a term **x** in a class **c**:

$$W_{x,c} = \|tf_{x,c}\| \cdot \log(1 + \frac{A}{f_x}) \tag{4.1}$$

where $\mathbf{tf}_{x,c}$ is the frequency of word **x** in class **c**, $\mathbf{f}_c$ is the frequency of word **x** across all classes, and **A** is the average number of words per class. There are two parameters that can be tuned in this definition of c-TF-IDF [105]. *bm25_weighting* is a boolean parameter that indicates what is the weighing scheme in use. The bm25 weighing scheme is defined by $\log(1 + \frac{A - f_x + 0.5}{f_x + 0.5})$. For smaller datasets, it can be more robust to any stopwords that might exist.

The other parameter that can be tuned is *reduce_frequent_words*, which is also a boolean term and adds a square root to the term frequency after normalizing the frequency matrix: $\sqrt{\|tf_{x,c}\|}$. It is useful in situations where very common words exist, but were not included as stopwords.

Combining both parameters results in:

$$W_{x,c} = \sqrt{\|tf_{x,c}\|} \cdot \log(1 + \frac{A - f_x + 0.5}{f_x + 0.5}) \tag{4.2}$$

4.4 Topic Evaluation with BERTopic

In order to select the best combination of BERTopic options at each stage, evaluation metrics must be defined. The task of evaluating topic modeling is challenging. On the one hand, it often does not exist a groundtruth model to which new models can be compared. On the other hand, their performance can be seen as subjective - clusters of the same dataset can have distinct relevance for distinct objectives.

Following Abdelrazek et al. [106], the metrics considered were:

- **Perplexity** measures the model's capability of generating documents based on the learned topics. It shows how well the model explains the data by analysing its predictive power, and therefore it can be seen as a measure of model quality.

- **Topic coherence** is a measure closely related to interpretability. Since a topic is a discrete distribution over words, it is important that these words are coherent inside each topic. This means they are similar to each other, making it an interpretable topic, instead of being only a result of statistical inference. Word similarity can be measured with cosine similarity, which ranges from 0 to 1. Higher values of cosine similarity suggest closely related words.

- **Topic diversity** looks at how different the topics are to each other. Low diversity is a result of redundant topics, implying difficulty in distinguishing them from the remaining. A suggested technique for measuring diversity is counting the unique words present in the top words of all topics [107].

- **Stability** in topic modeling refers to whether the results obtained are consistent across runs. Since topic modeling applies concepts from statistical inference, is expected some variation when modeling the same data for several runs. It also follows that more relevant topics will be consistent throughout iterations. One way to measure the stability of a model is to use the Jaccard similarity coefficient of the top words across topics, across iterations. Jaccard similarity ranges from 0 to 1, where identical topics share a coefficient of 1.

Chapter 5

Results and Discussion

The current section begins by offering an exploratory analysis of the data. It follows with the results of the optimization process of BERTopic (Section 5.2), as described in 4, and the results of applying the optimized algorithm to the data at hand (Section 5.3).

5.1 Exploratory Data Analysis

The aim of performing an exploratory data analysis is to obtain a broader comprehension of the dataset. We recall that our dataset consists of tweets from selected members of the 117^{th} U.S. Congress. This Congress was in session from January 3^{rd}, 2021, to January 3^{rd}, 2023.

In fig. 5.1 we show the tweeting frequency throughout this interval. There is an increase in tweeting activity from January 2022, reaching an all-time high on September 28^{th}. Hurricane Ian was formed on September 23^{rd}, and reached Florida on September 28^{th} as Category-4. It was one of the most impactful storms in U.S. history, causing 155 deaths and an estimated $113 billion losses. As suggested by Fig. B.1, the surge in tweets is directly correlated with this event, with terms such as 'Hurricane Ian', 'Ian', 'Florida', 'FL' and 'storm' being atypically frequent.

As mentioned previously, tweeting activity varies with the user. Fig. 5.2 shows the total number of tweets published by each account in the given timeframe. @MittRomney, @SenRubioPress and @StenyHoyer were among the least frequent tweeters, achieving between 100 and 200 tweets. Nonetheless, Mitt Romney's professional account, @Senator-Romney produced 802 tweets, the third highest value. Similarly, Marco Rubio's personal

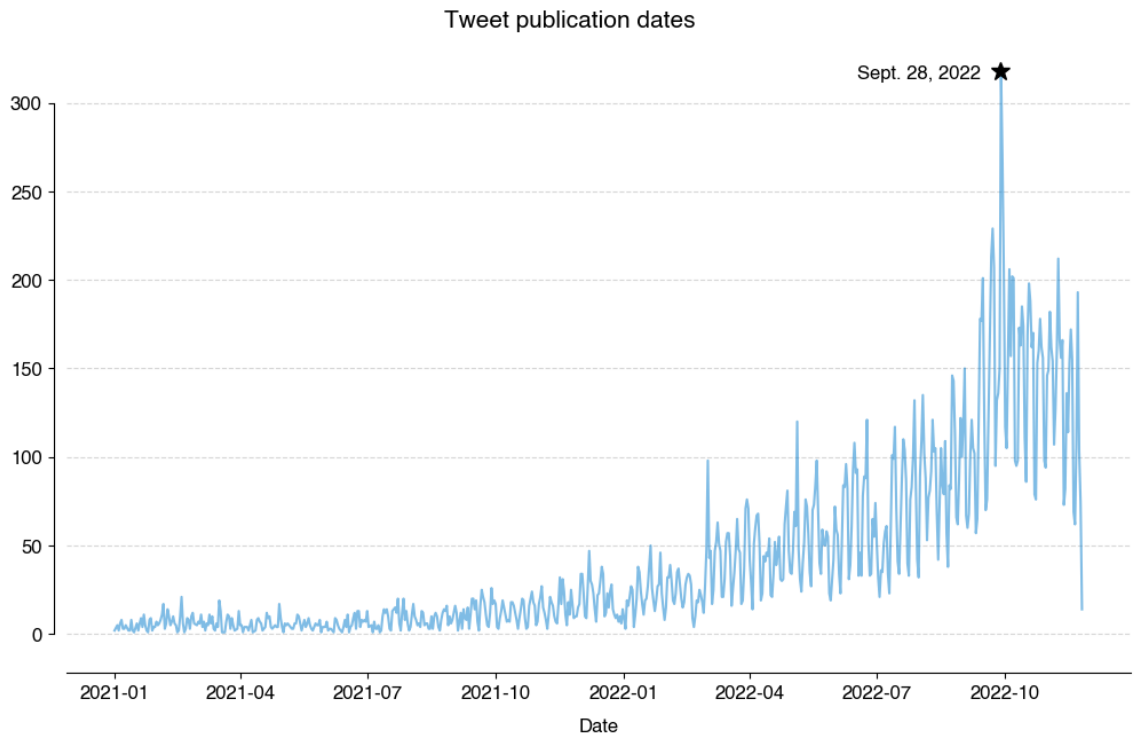


FIGURE 5.1: Frequency of tweets

account, @marcorubio, tweeted 796 times. Although @LeaderHoyer does not amount to this, it also compensated its user's low activity, with a total of 583 tweets.

Most accounts tweeted over 700 times. Pramila Jayapal was the most active user (1616 tweets), with her personal account @PramilaJayapal as the most active overall (817 tweets) and her professional account @RepJayapal tweeting 799 times. Bernie Sanders was responsible for 1561 tweets: 782 on his personal account @BernieSanders and 779 on his professional account @SenSanders. Alexandria Ocasio-Cortez tweeted a total of 1225 times: @AOC, her personal account, with 644 tweets, and @RepAOC, her professional account, with 581 tweets.

Both Pramila Jayapal and Alexandria Ocasio-Cortez are Democrats, while Bernie Sanders is the only Independent individual. The trend of more frequent tweeters being *blue* is further illustrated in Fig. 5.3: *blue* users represent 54.65% of total tweets, *red* users 39.73% and *white* 5.62%.

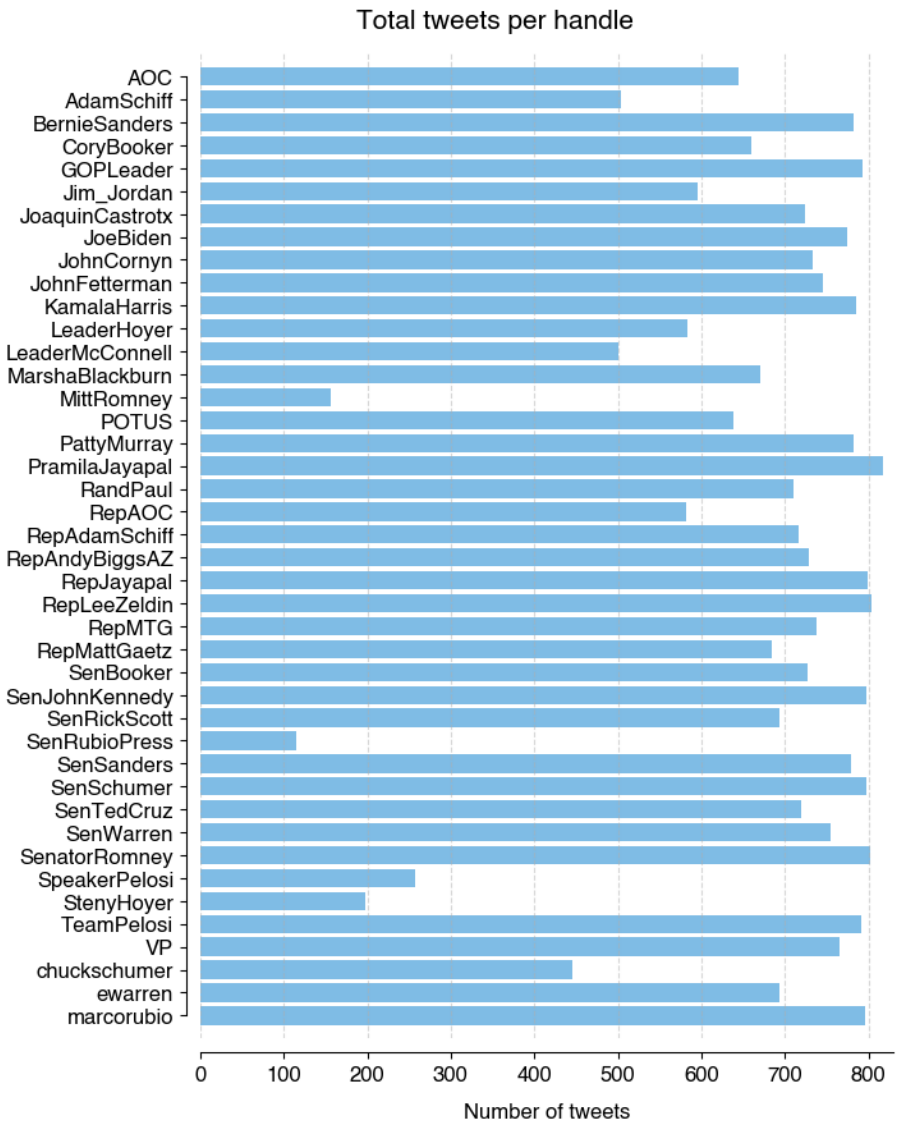


FIGURE 5.2: Total tweets per user

Given that Bernie Sanders is the only user of color *white*, his activity is over twice as high as what would be expected in a uniform distribution*. Given that there is an equal amount of Democrats and Republicans in the dataset, this suggests Democrats are more frequent tweeters.

Fig. 5.3 also shows how total tweets are distributed across account types. Of the 42 accounts, 18 (42.86%) are personal and 24 (57.14%) are professional. Personal accounts tweeted 40.14% of all tweets; professional accounts tweeted 59.96%. Adjusting for the imbalance of both types suggests no relevant difference between the two.

*If all users tweeted the same, each would represent 2.38% of the dataset

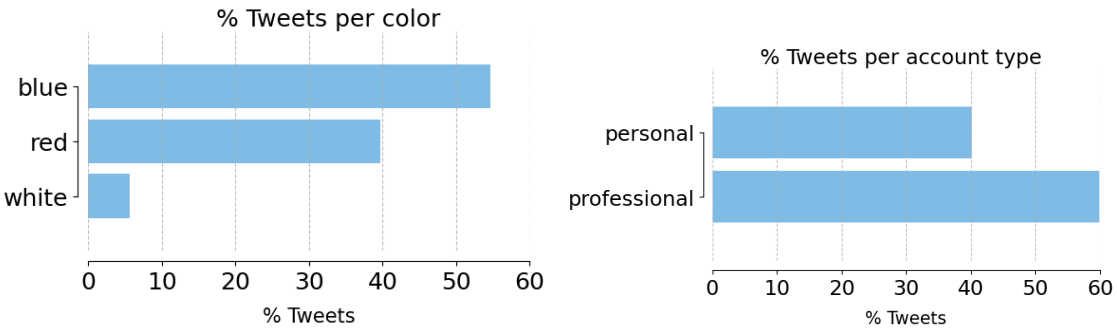


FIGURE 5.3: Distribution of tweets according to color (left) and account type (right)

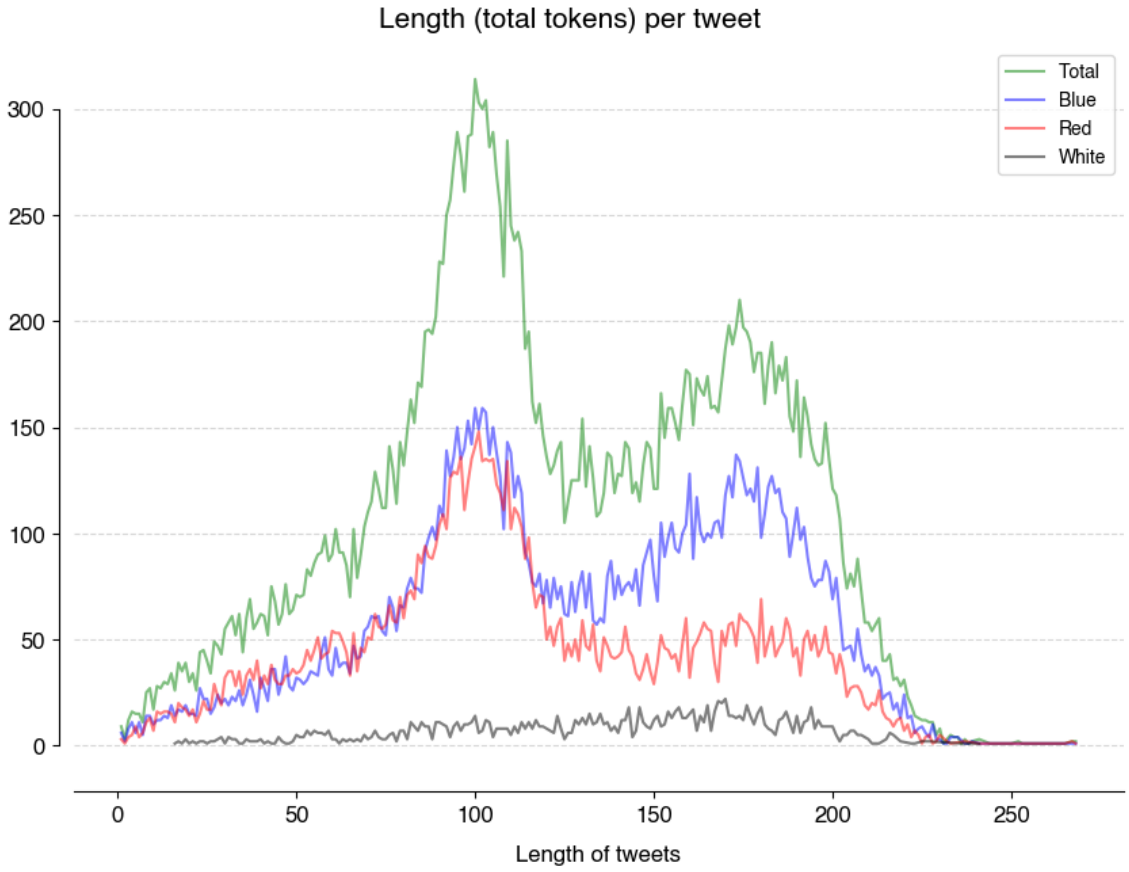
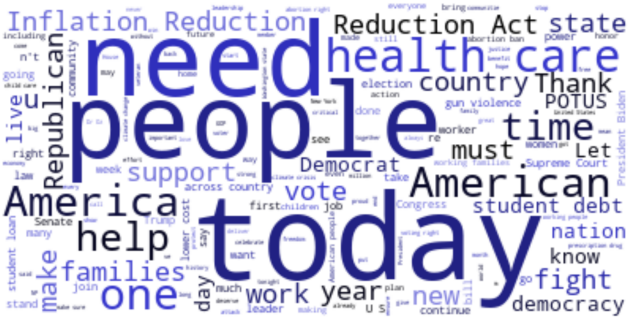


FIGURE 5.4: Length of tweets

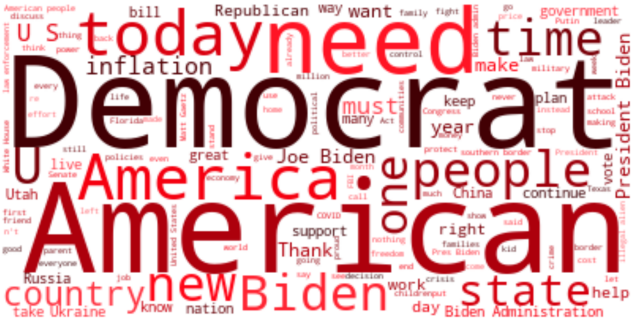
Fig. 5.4 depicts the distribution of the length of tweets. Tweets have an upper bound of 280 characters, which averages 40 to 70 words. Since this analysis is being performed on cleaned data, *length* considers only the relevant tokens that were not removed during preprocessing. This allows *length* to reach values over 250. The distribution of *length* of all tweets is colored in green. It shows a global maximum of around 110 tokens. This is the trend for Democrat tweets (in blue) and for Republican tweets (in red). A local maximum

is found around 280 tokens. Democrat tweets follow this pattern, but the same does not happen for Republicans. In the case of Bernie Sanders (in black), his tweets reach a global maximum of around 170 tokens, with no local maximum: this suggests his tweets tend to be longer than those of his fellow members.

Wordclouds consist of graphical representations of word frequency, where larger words appear more often in a text. As illustrated in Fig. 5.5, when wordclouds are generated for the set of tweets of each color, some differences are highlighted. Democrats tend to use the words 'need', 'people', 'today'. Republicans more often tweet 'Democrat' and 'American', followed closely by 'America', 'need' and 'today'. Finally, Independents' more frequent words are 'time', 'worker', 'must'. These differences in terms used suggest different positionings in the types of tweets published. For instance, Independent tweets are directed towards workers - namely, labour concerns -, with a call to action. Democrat tweets follow a similar trend, although they focus on 'people' in general. Finally, Republican tweets target their political opponents - which could be argued to be expected, given that Democrats were in power at the time - and appeal more to a sense of patriotism.



(A) Democrat tweets



(B) Republican tweets

(C) Independent tweets

FIGURE 5.5: Wordclouds according to color of tweets

N-grams are contiguous sequences of n words in a corpus. Besides looking at wordclouds, analysing n-grams helps further understand topics in a text. Fig. 5.6 collects the top five n-grams (at each level of n) present in all tweets.

N-grams of longer size are less common in a text. Bigrams (2-grams, in yellow) refer to 'health care' (which is repeated as a 4-gram), 'President Biden' and 'American people'. They also include mentions to the 'Inflation Reduction Act', which is the most common trigram, and which is also repeated as a 5-gram. Besides these, the tweets also often mention 'Roe v. Wade', 'student loan debt', acts such as 'Build Back Better Act', 'Women

Health Protection Act', 'Protect Children Innocence Act' and 'John Lewis Voting Rights Advancement Act'.

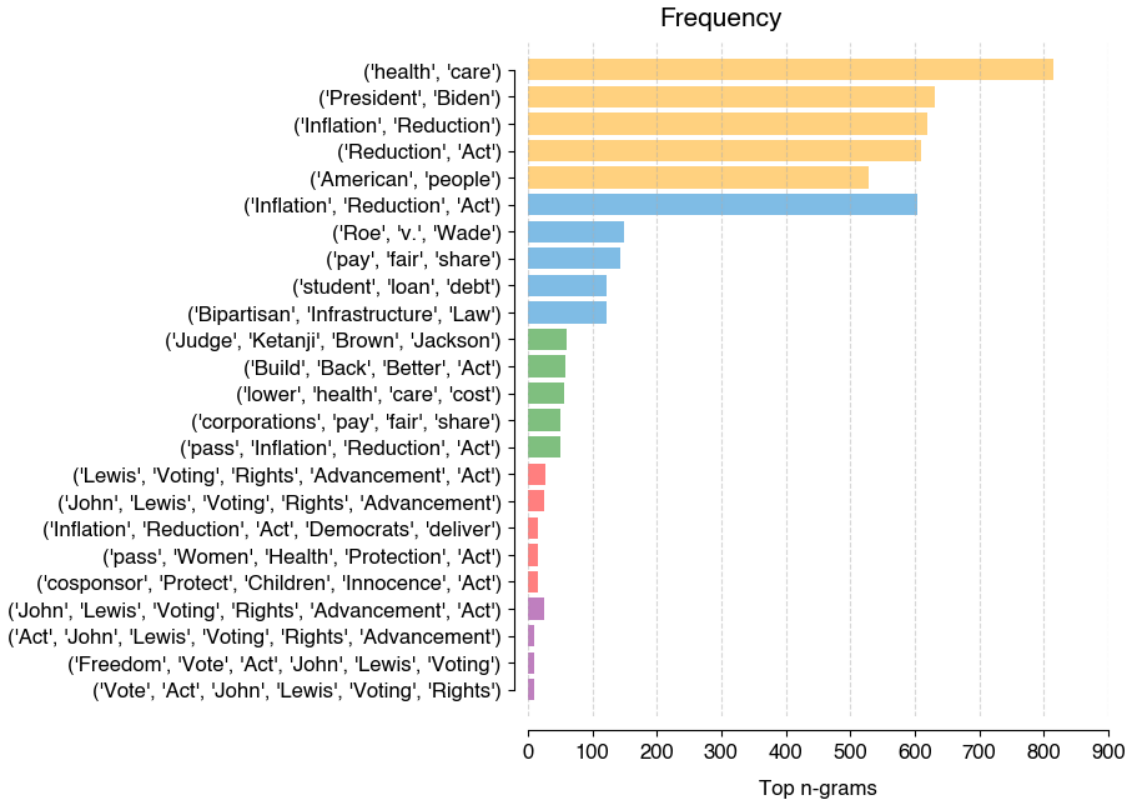


FIGURE 5.6: Top n-grams

5.2 Optimizing BERTopic

The optimization process of BERTopic is done assuming independence between the stages. While performance plots for each stage can be consulted in Appendix B, a summary of the optimized stages and its performance is found at the end of this section, in Fig. 5.7 and Table 5.1.

Fig. B.2 illustrates how SBERT, SpaCy and Word2Vec compare in terms of coherence, perplexity, diversity, and stability scores in the **Embeddings** stage. SpaCy coherence is more erratic and generally lower than the other options (0.61), which perform similarly to each other (0.75). Also, it shows lower perplexity (0.00018) and higher diversity (0.996) scores, although the actual difference in performance is less significant. Regarding stability, the three models perform similarly (0.595, 0.559, 0.586). Given these results, Word2Vec better optimizes BERTopic to our data, as it is less erratic than SBERT on diversity.

In the **Dimensionality Reduction** stage (Fig. B.3), we found that PCA has the lowest overall performance in average coherence (UMAP: 0.76, PCA: 0.43, None: 0.66) and diversity (UMAP: 0.97, PCA: 0.79, None: 0.92). While UMAP has the best perplexity result (UMAP: 0.0003, PCA: 0.0005, None: 0.0005), the distance to the other models is too small to justify its poor stability. Using an empty dimensionality model ('none') is the best option for this stage.

The **Clustering** stage has two algorithm options. Although HDBSCAN has a lower stability score (HDBSCAN: 0.51, k-Means: 0.59), it has significantly lower average perplexity (HDBSCAN: 0.0003, k-Means: 1) and similar average coherence (HDBSCAN: 0.75, k-Means: 0.74) and diversity (HDBSCAN: 1, k-Means: 1) to k-Means. The difference in average perplexity drives the choice of HDBSCAN in the current stage.

As illustrated by Fig. B.5, for the parameter that allows longer n-grams to be considered (*ngram_range*), the six values tested on the **Vectorizer** stage all perform similarly across perplexity (0.00032), diversity (0.98), and stability (0.65). Setting the parameter to (1, 1), meaning only single words are used, is the value that distinguishes itself on coherence (0.79), which justifies its choice. Still in this stage, Fig. B.6 shows how setting the minimum word frequency (*min_df*) at 5 allows the model to perform better on coherence (0.79), diversity (1), and stability (0.66) scores. For the *max_features* parameter, which limits the size of the term-topic matrix, all values analysed perform similarly (Fig. B.7). 17000 is the one with the highest stability (0.674), hence it being chosen.

Finally, the **Topic Representation** stage is optimized by manipulating two boolean parameters. *Bm25_weighting* (Fig. B.8) does not improve model performance, as both values behave similarly across metrics. On the other hand, reducing frequent words (Fig. B.9) increases average coherence (True: 0.79, False: 0.75) and average diversity scores (True: 1, False: 0.96) enough to account for the decrease in stability (True: 0.514 False: 0.614).

Fig. 5.7 illustrates how the optimized BERTopic performs against the default, across the four metrics. While the diversity scores are similar for both (1), the optimized model outperforms the default on average coherence (optimized: 0.65, default: 0.51) and stability (optimized: 1.0, default: 0.521). The difference in average perplexity is small enough to be disregarded (optimized: 0.00029, default: 0.00033).

Table 5.1 summarizes the options selected at each stage.

TABLE 5.1: Optimized BERTopic algorithms/ parameters

Stage	Selected algorithms / parameters
Embeddings	Word2Vec
Dimensionality reduction	Base dimensionality model
Clustering	HDBSCAN
Vectorizer	*ngram_range*: (1, 1)
	min_df: 5
	max_features: 170000
Topic representation	*bm25_weighting*: False
	reduce_frequent_words: True

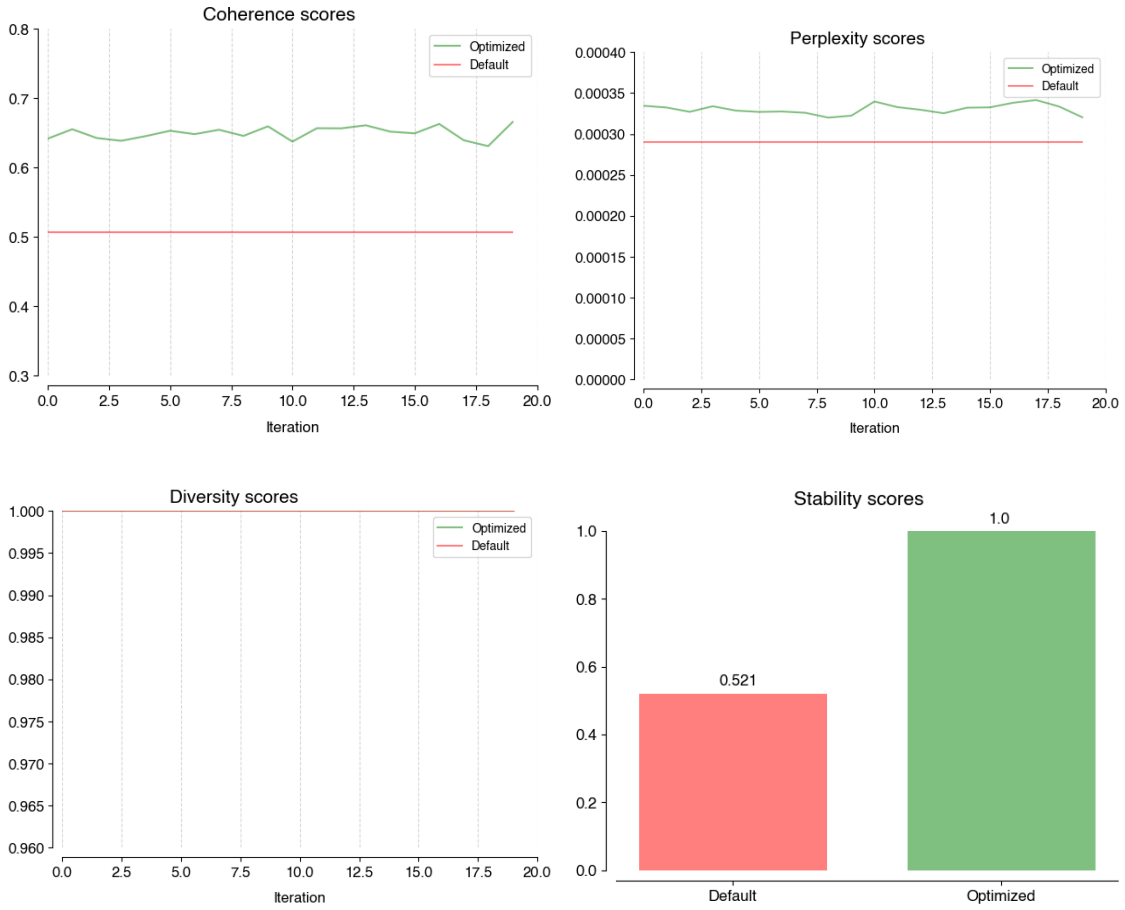


FIGURE 5.7: Performance of optimized BERTopic against the default

5.3 Topic Results

The optimized BERTopic algorithm is applied to the processed data. It obtained 182 unique topics, which were assigned to 4068 tweets. This means 23714 tweets were considered as outlier topics. Fig. 5.8 shows the frequency of each topic: topics 0 to 4 occur at least 100 times each, while topics 5 to 12 occur at least 50 times each. These topics represent 46% of the total frequency. Topic 0 is the leader, showing up 772 times. The least common topic (182) corresponds to 4 tweets.

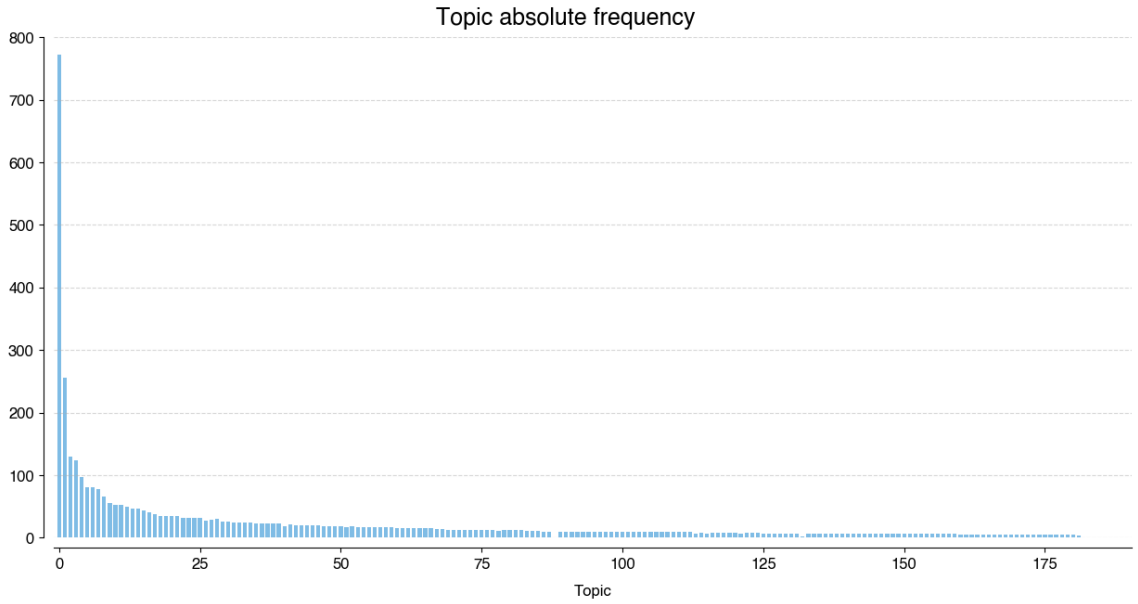


FIGURE 5.8: Absolute frequency of topics

Restricting this analysis to the top 13 topics allows for a deeper understanding. Table 5.2 lists the topic labels, as suggested by BERTopic, and the name given. Further information on each topic can be found in Appendix B (Table B.1).

TABLE 5.2: Label and name of the top topics

Topic	Label	Name
0	0_reduction_prescription_cap_abortion	Abortion
1	1_cancel_student_debt_borrowers	StudentDebt
2	2_ketanji_jackson_brown_judge	JudgeBrown
3	3_poll_early_location_register	Voting
4	4_putin_kyiv_ukraine_ukrainian	UkraineWar
5	5_marijuana_cannabis_legalize_possession	Cannabis

6	6_commitmenttoamerica_housegop_replamalfa_built	#CommitmentToAmerica
7	7_defendourdemocracy_victory_congressman_congratulations	#DefendOurDemocracy
8	8_firebrand_episode_matt_feat	Firebrand
9	9_sacrifice_veteransday_memorialday_memorial	MemorialDay
10	10_vaccine_19_vaccinate_booster	CovidVaccine
11	11_utah_wildfires_infrastructure_mitigation	Utah
12	12_birthday_247th_usmc_wishing	Congratulations

Fig. 5.9 illustrates how these topics are distributed by color. The first conclusion is that only three topics of the top 13 are represented by both Democrats and Republicans: *UkraineWar*, *MemorialDay* and *Congratulations* (4, 9, 12, respectively). These topics include the handles shown in Fig. 5.10. Topic *UkraineWar* is dominated by @marcorubio, who tweeted 41 times about it between February and March 2022. Senator Rubio's tweets on this topic are strongly focused on his opinions on the developments of the Russia-Ukraine war, especially on the political figures' strategies and motivations:

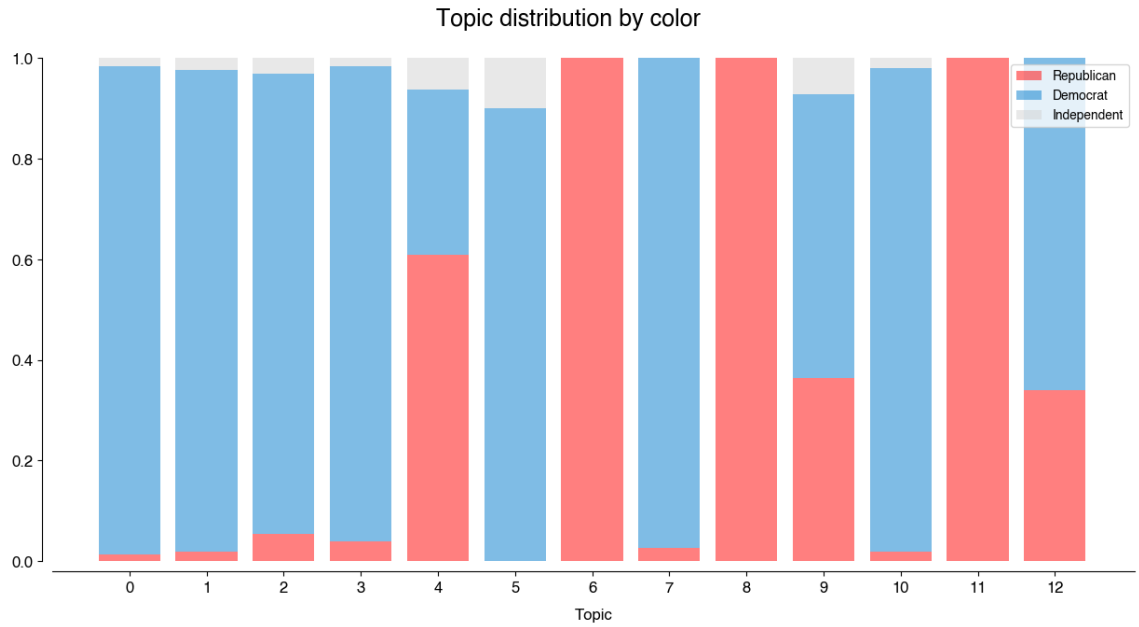


FIGURE 5.9: Top topics distribution by color

- *DANGER #Putin's legitimacy built on image as the strong leader who restored #Russia to superpower after the disasters of the 90's. Now the economy is in shambles & the military is being humiliated & his only tools to reestablish power balance with the West is cyber & nukes* (28-02-2022)

- *To force #Ukraine into deal he can claim as a victory #Putin needs battlefield momentum. The danger now is that he has no economic or diplomatic cards to play,his conventional forces are stalled & cyber,chemical,biological & non-strategic nukes are his only escalation options* (25-03-2022)

Similarly, the topic *Congratulations* main tweeter is @LeaderHoyer, with 32 publications between May and November 2022. This topic consists of the different times a user congratulates others on achievements/ events:

- *Happy birthday to my friend @RepDavids from #KS03, a champion for creating economic opportunity and giving working families the tools they need to #MakeItInAmerica as Vice Chair of @TransportDems and as a Member of @HouseSmallBiz.* (22-05-2022)

- *Wishing a happy birthday to my friend from #IL06, @RepCasten, a champion for climate as Co-Chair of the New Democrat Coalition's #ClimateChange Task Force and a Member of @ClimateCrisis.* (24-11-2022)

MemorialDay is more balanced, with the main accounts (@VP, @RepAdamSchiff) tweeting 4 times each. It celebrates both Memorial Day (last Monday of May) and Veterans Day (November 11).

- *We must always remember and honor those who served our country and those who gave their lives in protecting our freedoms. May we never forget their sacrifice this Memorial Day.* (@VP, 30-05-2022)

- *On Veterans Day, we honor the brave men and women who answer the call to serve. They represent the best of what America has to offer. We owe them the greatest debt of gratitude. Thank you for your service - and a special thank you to my favorite veteran, my father Ed, now 94.* (@RepAdamSchiff, 11-11-2022)

The remaining topics tend to have a clear majority of either party. *#CommitmentToAmerica*, *Firebrand* and *UtahWildfires* are dominated by Republican tweets. Contrary to the previous topics, these are essentially the responsibility of one account each (Fig. 5.11). @GOPLeader tweeted 79 times on the topic of *#CommitmentToAmerica*, from September to November 2022. Commitment to America consists of Republicans' guidelines of their vision of the U.S. politics:

- *An economy that's strong. A nation that's safe. A future that's built on freedom. A government that's accountable. This is the Republican Commitment to America.* (22-09-2022)

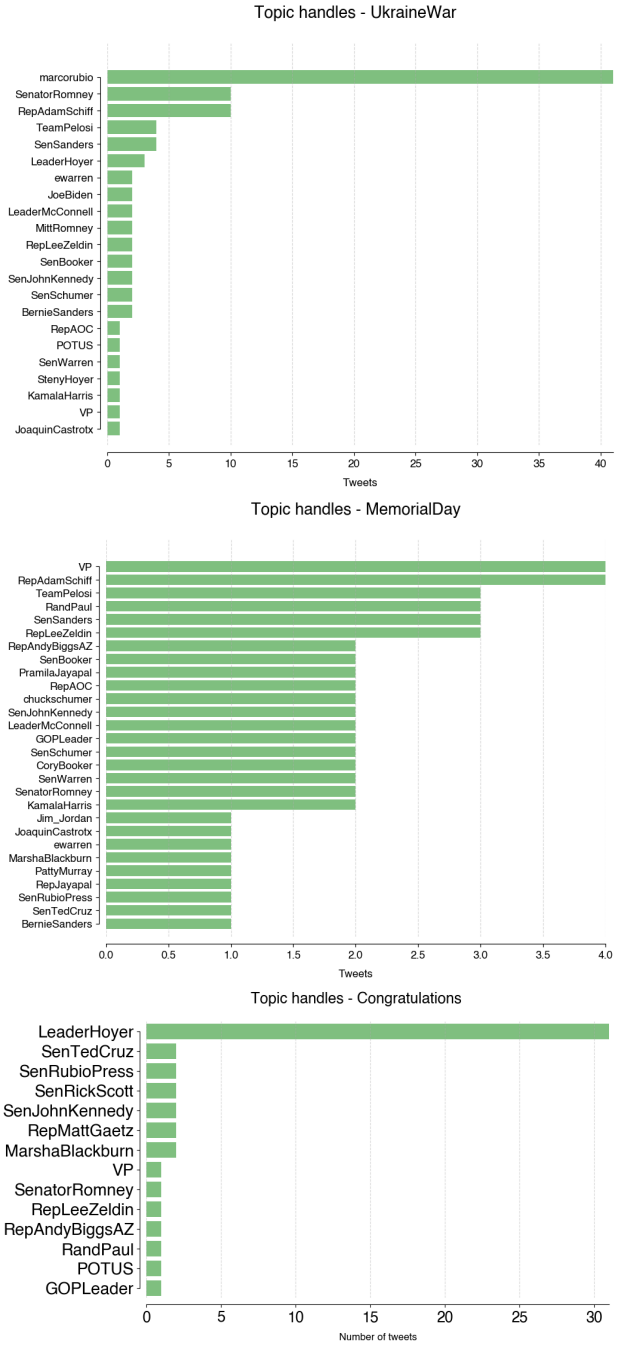


FIGURE 5.10: Topics *Ukraine*, *MemorialDay* and *Congratulations* handles

- *Republicans have made our #CommitmentToAmerica. Under the leadership of @GOPLeader, we will build: An Economy That's Strong, A Nation That's Safe, A Future That's Built on Freedom, A Government That's Accountable. Now let's get to work.* (17-11-2022)

@RepMattGaetz tweeted 67 times on his podcast Firebrand, from March to November 2022:

- *In today's episode of @Firebrand_Pod, Rep. Matt Gaetz brings us an exclusive report from the U.S.-Mexico border with @sherifflamb1, and discusses rising gas costs, Russian propaganda, men competing in women's sports, and more!* (24-03.2022)

- *RT @Firebrand_Pod: Episode 76 LIVE: Ban TikTok (feat. @GavinWax) – Firebrand with @RepMattGaetz https://t.co/KQdzDcSeHJ* (18-11-2022)

Finally, @SenatorRomney, who is Senator for Utah, tweeted 52 times, from July 2021 to November 2022 on the topic of *Utah*:

- *The needs of Utahns have been forefront as I've helped negotiate our bipartisan infrastructure plan. Our plan would provide Utah with funding to expand our physical infrastructure and help fight wildfires without tax increases or adding to the deficit.* (11-07-2021)

- *From funding water projects like the Central Utah Project to building transportation systems like High Valley Transit and modernizing wildfire policy through an expert commission, our infrastructure bill has been delivering for Utah since it was signed into law 1 year ago today.* (15-11-2022)

The remaining seven topics, dominated by Democrats, are distributed as illustrated in Fig. 5.12. On the *Abortion* topic, @PattyMurray (120), @RepJayapal (115) and @PramilaJayapal (87) represent 43% of tweets. As illustrated by the time series, interest on this topic began in May 2022, when a leaked draft suggested the Supreme Court's intention on overturning Roe v. Wade. The decision was taken in June 24, 2022 (in black), which saw an increase in tweeting activity on the topic.

StudentDebt top tweeters are @PramilaJayapal (69), @SenWarren (45) and @RepJayapal (35). Together, they account for 58% of tweets (Fig. 5.13). The topic began gaining traction in April 2022, when the Biden Administration announced changes to the student debt payment plans. The peak in tweeting activity occurs June 28, which fell on the week of the Supreme Court's decision on President Biden's student debt forgiveness plan.

Regarding *JudgeBrown*, @SenSchumer (29), @KamalaHarris (17), @JoeBiden (11), @SenBooker (11) and @CoryBooker (10) are responsible for 60% of topic tweets. While not a particularly intense topic, it gained relevance on April 7, 2022, which was the day Judge Brown Jackson was confirmed as the first black woman on the Supreme Court.

Voting is a balanced topic regarding its users. It refers to tweets about information on registering to vote, early voting and polls of the November 8 elections. This day coincides with the peak activity on the topic.

Cannabis was mostly tweeted by @CoryBooker (18 tweets, 23%). The peak in *Cannabis* topic usage was achieved on October 6, 2022, when President Biden announced his position on Marijuana Reform.

Finally, #DefendOurDemocracy is a hashtag that @TeamPelosi used when commenting on other party's actions, or when calling users to action on voting for Democrats.

The topic was first detected in July 2022 and has been in use up to November 2022:

- *#DefendOurDemocracy: Re-elect Democratic Rep. Tom O'Halleran in #AZ02.* (1-7-2022)
- *Democratic Victory: Congratulations to Congresswoman @MaryPeltola on your re-election in Alaska! -NP* (24-11-2022)

@VP contributed to topic *CovidVaccine* 19 times (36%), from April to November 2022. It mainly consists of tweets providing information of vaccines and appeals for population vaccination:

- *Yesterday I received my second COVID-19 booster shot. We know that getting vaccinated is the best form of protection from this virus and boosters are critical in providing an additional level of protection. If you haven't received your first booster—do it today* (2-4-2022)
- *Prepare for a healthy holiday season by getting your updated COVID vaccine. It's free, safe, and effective.* (13-11-2022)

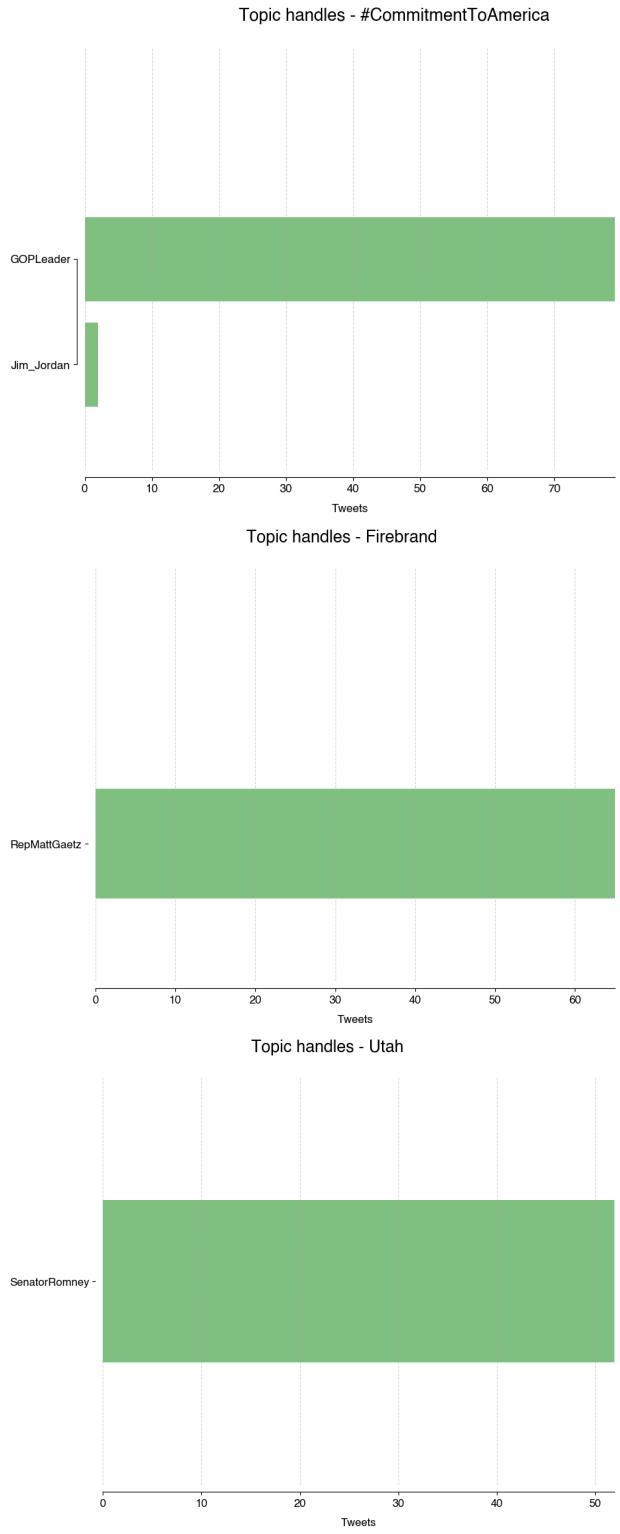


FIGURE 5.11: Topics #*CommitmentToAmerica*, *Firebrand* and *Utah* handles

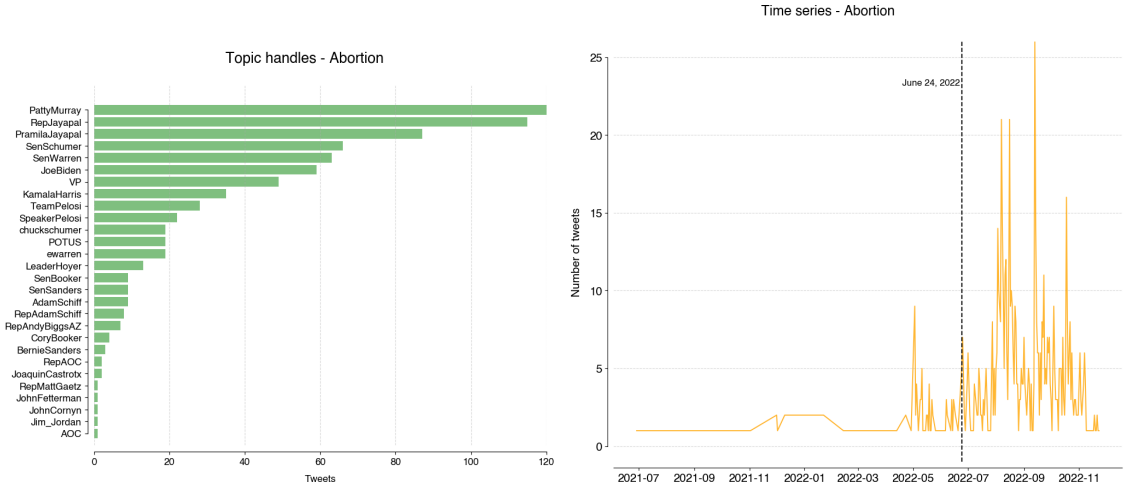


FIGURE 5.12: *Abortion* - handles and time series

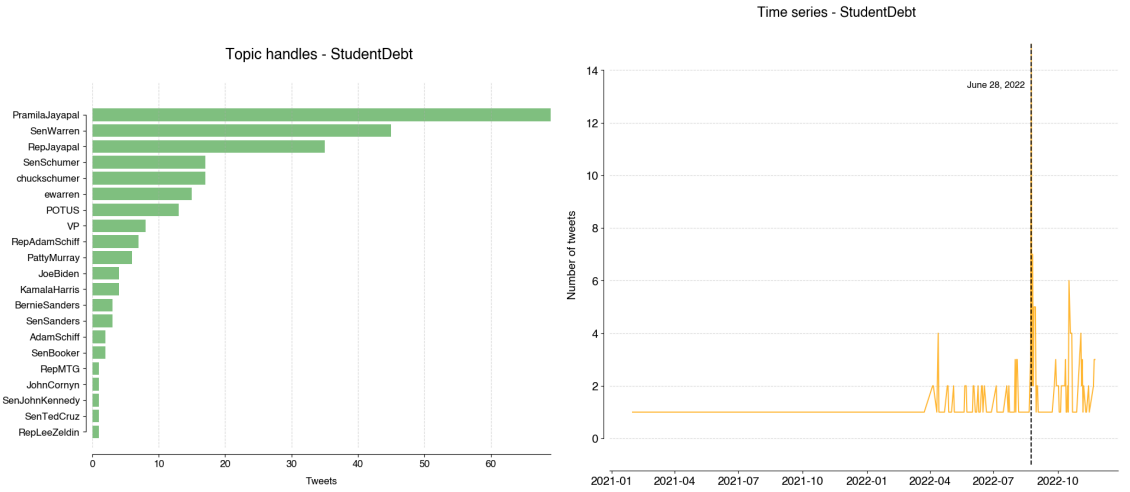


FIGURE 5.13: *StudentDebt* - handles and time series

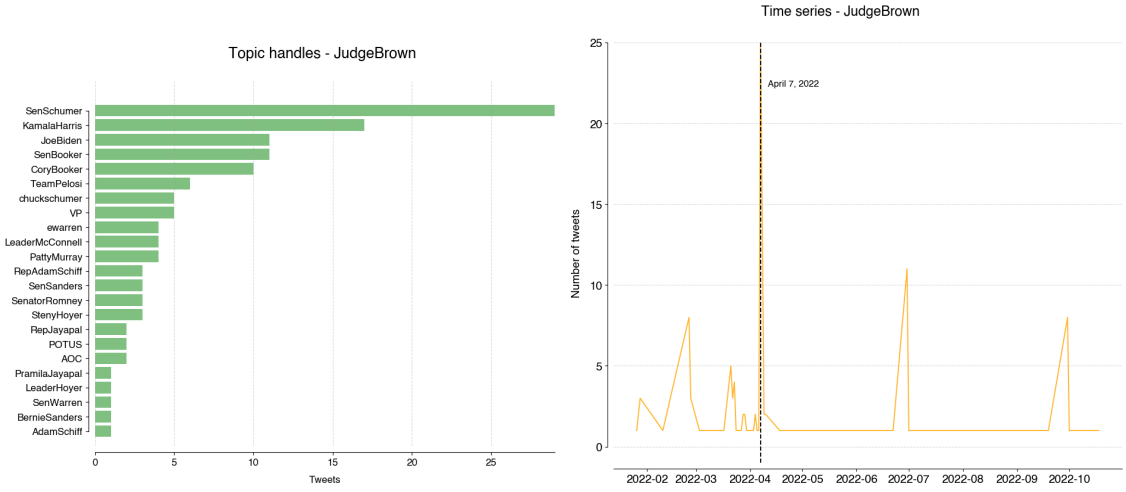


FIGURE 5.14: *JudgeBrown* - handles and time series

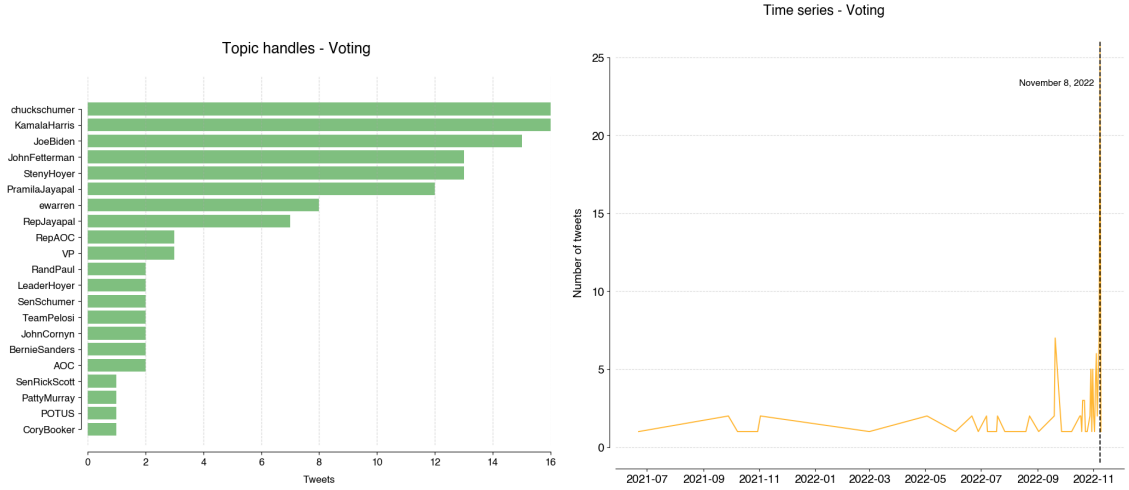


FIGURE 5.15: *Voting* - handles and time series

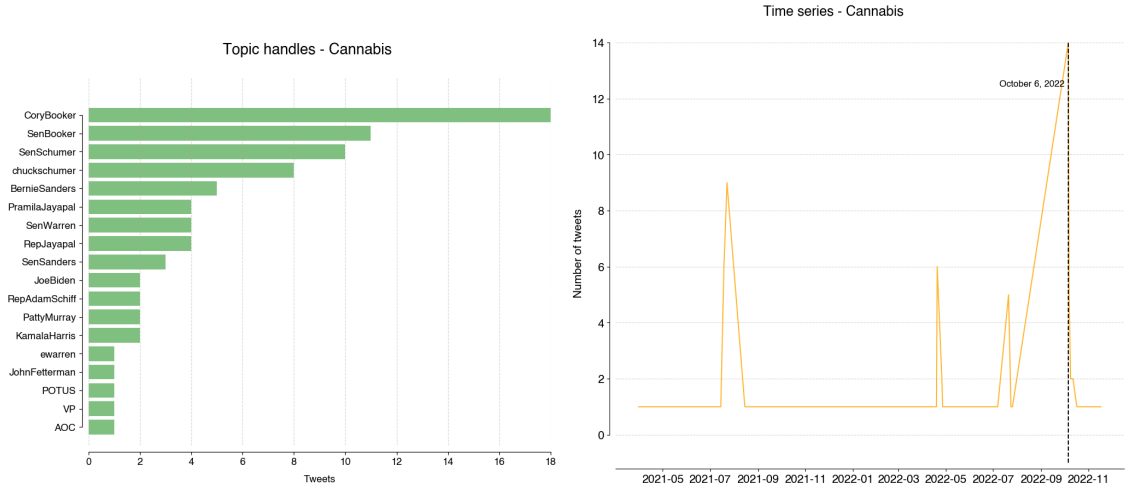


FIGURE 5.16: *Cannabis* - handles and time series

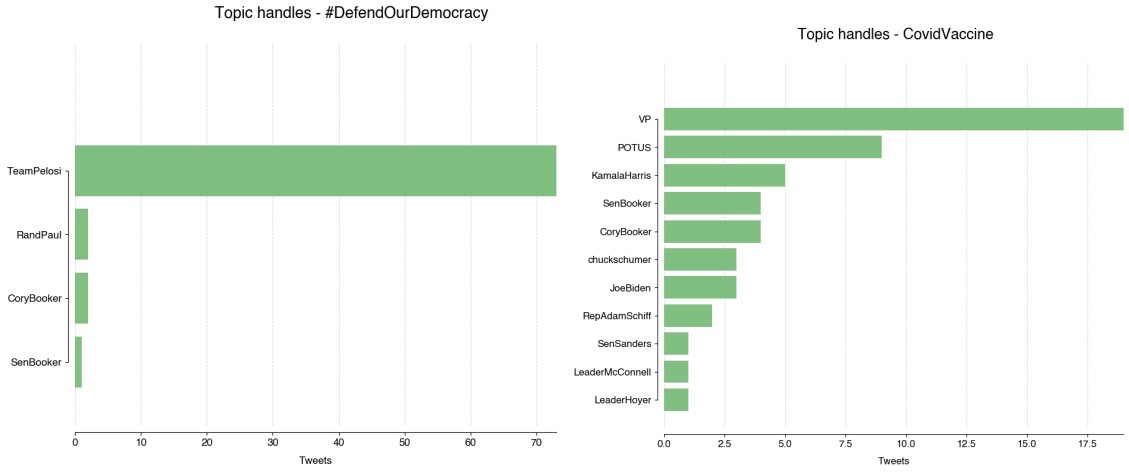


FIGURE 5.17: *#DefendOurDemocracy* (left) and *CovidVaccine* (right) handles

5.4 Visualizing Congress 117^{th} Topics

There are differences in tweeting activity between parties across several different metrics. Democrats tend to tweet more often, and their tweets are longer. They use more general terms - 'people' - than Republicans - 'American'. Independents focus on 'worker'. The top n-grams used provided a hint for the topics to be discovered: healthcare, Roe v. Wade and Judge Ketanji Brown Jackson are some examples. The analysis performed of global topic results focused on the top 13 topics discovered. The representation of different parties for each of these topics is not equal, further translating differences in political focus (e.g. *#CommitmentToAmerica*, *StudentDebt*).

To gain a better understanding of these topics, we developed a Shiny app. The app allows its users to select a Congress member, and visualize their activity on Twitter for each of the topics approached.

Fig. 5.18 illustrates a use-case of the app. The top banner presents two dropdown menus that allow the user to select a Congress member (left) and the topic to be visualized (right). Selecting 'Pramila Jayapal', the menu for Topics is automatically reduced to topics she mentioned. Below the banner, on the left the user is show information about Pramila: her top three topics, a photo, her party, her Twitter account handle, how many times she tweeted on the topic, and how she ranks overall.

On the center, a tweet by Pramila on the selected topic appears. This tweet changes every time the page is refreshed. The goal is to give the user an idea of how the Congress Member stands on this topic. In this particular case, Pramila defends abortion as a human right.

Below the tweet, a time series is shown (Fig. 5.19). The blue bars account for all tweets of all members on this topic, throughout 2021 and 2022. The red line shows Pramila's activity. Below this plot, an interactive barplot shows how Pramila ranks on this topic, in blue, and all the remaining members, in red. Being interactive, hovering over each bar allows us to see the name of the members. In this particular case, Pramila was the top tweeter on *Abortion*, followed by Patty Murray.

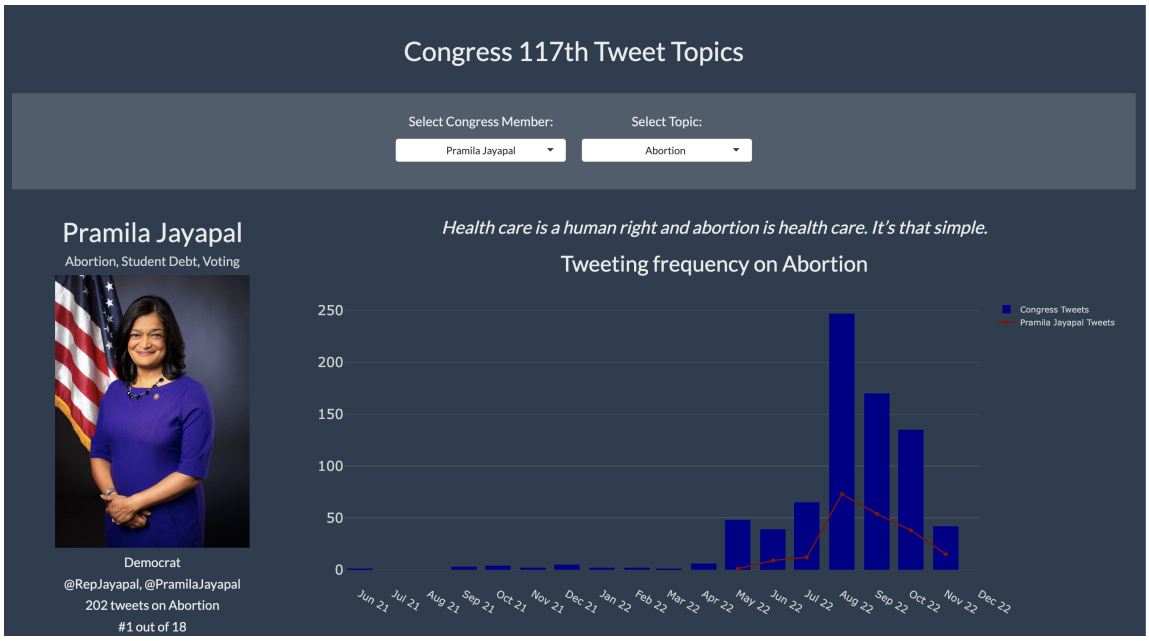


FIGURE 5.18: Use-case of app for member Pramila Jayapal and *Abortion* - part I

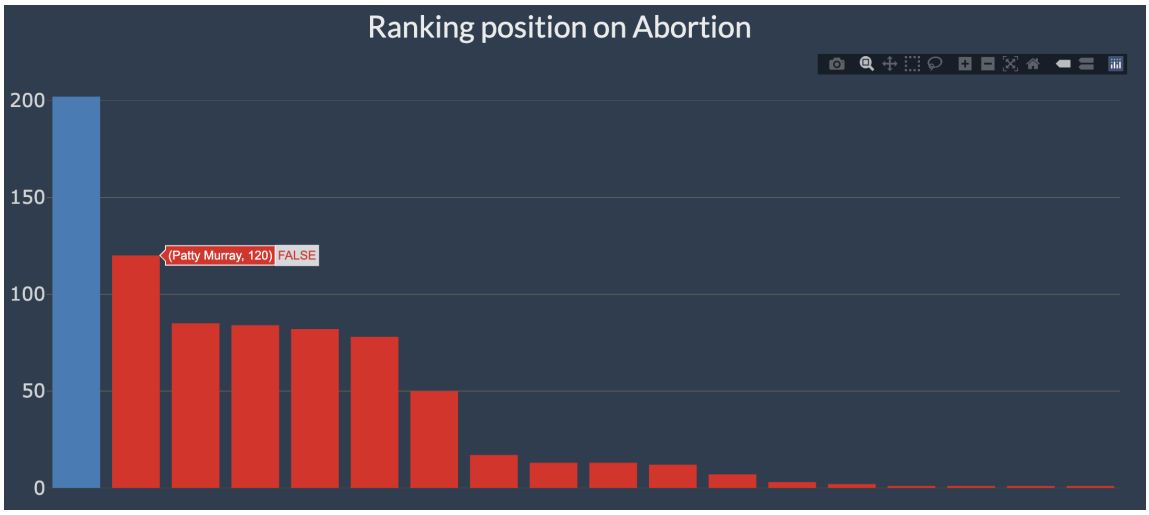


FIGURE 5.19: Use-case of app for member Pramila Jayapal and *Abortion* - part II

Chapter 6

Morality Foundations as Drivers of Topic Longevity: a Case Study

The analysis performed thus far is highly static and focuses on the topics found in the whole of the Twitter corpus from the 117^{th} Congress. Social media, however, is a fickle platform, and relevant topics mutate at a faster pace than every two years [108]. While analysing the whole of the topics is still important - it provides an overarching idea of what this Congress faced -, a second step is to decrease the interval for topic extraction.

6.1 Topic Evolution and Longevity

We chose a monthly scale, as it balances the need for grainier detail while having an adequate amount of tweets in each time window. For each of the time bins, BERTopic is run independently with the optimized stages, as defined previously. Table 6.1 lists the number of non-outlier topics found for each month, as well as the name of the two most common topics for each month. A total of 175 topics were extracted, which corresponds to 13436 tweets.

TABLE 6.1: Topics found on monthly intervals

Year	Month	Nbr. topics	Name of two most frequent topics
2021	January	0	N/A
	February	2	0_trumpcovid_relief_need 1_robinhood_trade_customers_business
	March	3	0_asian_american_relief_check 1_voter_vote_georgia_democracy

	April	3	0_infrastructure_republicans_democrats_climate 1_officer_capitol_evans_police
	May	2	0_vote_trump_democracy_republicans 1_israel_terrorist_hamas_palestinians
	June	2	0_vote_democrats_infrastructure_democracy 1_lgbtq_pride_month_carl
	July	6	0_marijuana_marijuanajustice_federal_war 1_child_families_payments_cut
	August	4	0_infrastructure_bill_tax_bipartisan 1_afghanistan_afghan_administration_biden
	September	2	0_need_democrats_tax_must 1_tovah_mar_kippur_yom
	October	6	0_bill_americans_american_democrats 1_midnight_book_washington_democracy
	November	3	0_infrastructure_bipartisan_bill_families 1_hanukkah_happy_thanksgiving_family
	December	2	0_workers_work_back_get 1_contempt_meadows_mark_january6thcmte
2022	January	6	0_vote_act_right_democracy 1_kroger_wealth_workers_greed
	February	13	0_putin_russia_ukraine_war 1_starbucks_workers_organize_corporate
	March	3	0_ukraine_putin_russia_work 1_lord_psalms_praise_matthew
	April	21	0_jackson_ketanji_brown_judge 1_starbucks_amazon_workers_amazonlabor
	May	3	0_right_abortion_must_fight 1_cuba_de_en_la
	June	17	0_abortion_roe_gun_right 1_student_debt_cancel_potus
	July	2	0_right_biden_get_act 1_abe_japan_minister_prime
	August	3	0_inflation_act_reduction_cost 1_gaetz_matt_firebrand_episode
	September	40	0_abortion_ban_reduction_care 1_strong_housegop_economy_commitment

	October	2	0_biden_make_democrats_republicans 1_israel_isaac_herzog_lebanon_breakthrough
	November	30	0_poll_voice_location_early 1_victory_np_congratulations_democratic

In January 2021, no topics were found. In general, topics 0 (the most common) of each month intersect with those identified in the previous analysis. Looking at the second most common topics, it is possible to see some new themes. April 2021 sees mentions of the Capitol car attack, which killed Officer Evans (1_officer _capitol_evans_police); In May 2021 there 11 days of conflict between Israel and Palestine (1_israel_terrorist _hamas_palestinians). The push for unions and labor rights gained momentum in 2022: Kroger workers went on strike in January (1_kroger_wealth_workers_greed), while Starbucks' did so in February (1_starbucks_amazon_workers_amazonlabor), and on July 8^{th} 2022, the former Japanese prime-minister Shinzo Abe was assassinated (1_abe_japan_minister_prime).

To evaluate how topics evolve, it is first necessary to develop a measure of linkage between topics of consecutive months. As mentioned in Chapter 3, topics are characterized by their representations, i.e., words that strongly identify them. As topics change monthly, the representations vary accordingly. Measuring the similarity of consecutive months' topics can be achieved by comparing the corresponding representations. This comparison was done using cosine similarity. Cosine similarity is the cosine of the angle between two vectors; a cosine similarity of 0 suggests no similarity between documents, while a value of 1 says the documents are the same. Concretely, in the context of document comparison this metric measures the angle of each document's vector. This vector is n-dimensional, where n is the number of words present in the document.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{6.1}$$

Topic evolution is determined by calculating the cosine similarity for all consecutive months and reducing the topic selection to those with non-negative similarity. The final result for our data consists of 54 topics and is illustrated in Fig. 6.1. Each vertical bar corresponds to one month. Topics are represented by one of three symbols: circles indicate topic emergence, squares are for topic stagnation, and triangles are for topic disappearance. Similar colors of consecutive topics represent a similarity above 0.5; topic splits or mergers are highlighted by changes in colors [109].

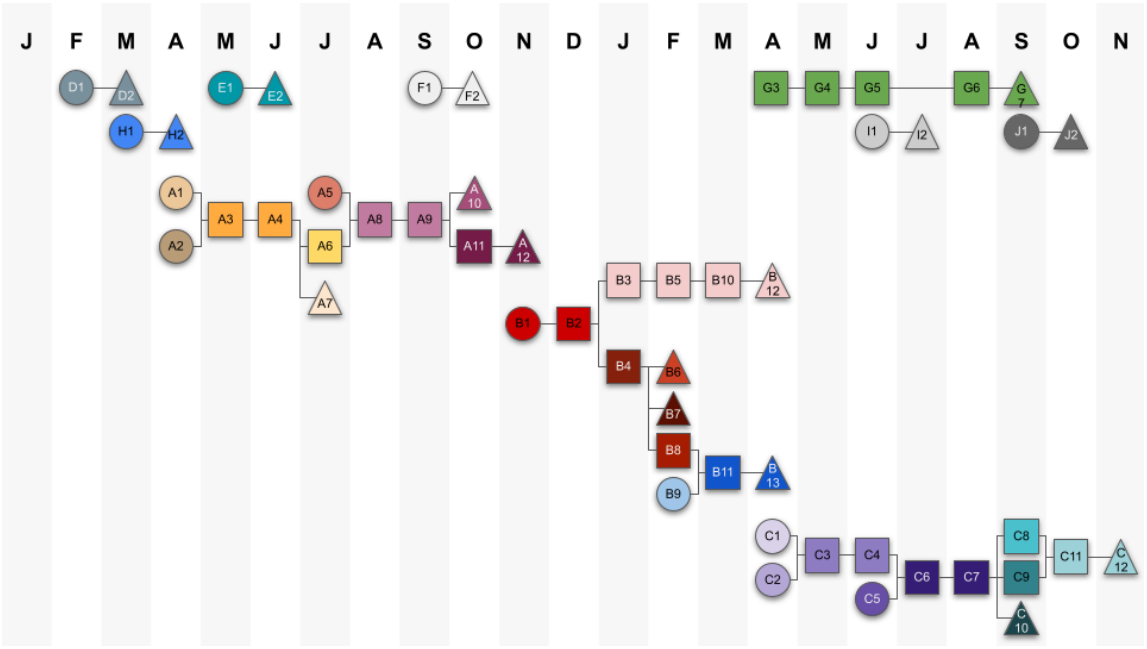


FIGURE 6.1: Timeline of topic evolution

A1. 0_infrastructure_republicans_democrats
A2. 2_genocide_armenian_recognize_truth
A3. 0_vote_trump_democracy_republicans
A4. 0_vote_democrats_infrastructure_democracy
A5. 0_marijuana_marijuanajustice_federal_war
A6. 1_child_families_payments_cut
A7. 4_democracy_wherever_facts_january
A8. 0_infrastructure_bill_tax_bipartisan
A9. 0_need_democrats_tax_must
A10. 0_bill_americans_american_democrats
A11. 5_register_vote_day_november
A12. 2_veterans_service_day_sacrifice

B1. 0_infrastructure_bipartisan_bill_families
B2. 0_workers_work_back_get
B3. 2_mask_booster_covid_n95
B4. 1_kroger_wealth_workers_greed
B5. 8_mask_kathy_ mandate_hochul
B6. 1_starbucks_workers_organize_corporate
B7. 2_drug_prescription_ pharma_insulin
B8. 6_economy_wealth_billionaires_bottom
B9. 0_putin_russia_ukraine_war
B10. 2_days_ago_write_ exactly
B11. 0_ukraine_putin_russia_work
B12. 19_china_well_economically_lockdown
B13. 6_putin_ukraine_war_deter

C1. 8_maternal_women_health_systemic
C2. 14_oklahoma_unconstitutional_abortion_reprod
C3. 0_right_abortion_must_fight
C4. 0_abortion_roe_gun_right
C5. 12_child_affordable_by_quality
C6. 0_right_biden_get_act
C7. 0_inflation_act_reduction_cost
C8. 0_abortion_ban_reduction_care
C9. 11_debt_student_borrowers_cancellation
C10. 12_registration_register_nationalvoter
C11. 0_biden_make_democrats_republicans
C12. 0_poll_voice_location_early

D1. 0_trump_covid_relief_need
D2. 0_asian_american_ relief_check

E1. 1_israel_terrorist_hamas_palestinians
E2. 1_lgbtq_pride_month_carl

F1. 1_tovah_mar_kippur_yom
F2. 1_midnight_book_washington_democracy

G1. 11_episode_firebrand_matt_feat
G2. 2_matt_firebrand_ episode_gaetz
G3. 6_firebrand_episode_ matt_gaetz
G4. 1_gaetz_matt_firebrand_episode
G5. 28_firebrand_episode_matt_gaetz

H1. 1_voter_vote_georgia_democracy
H2. 1_officer_capitol_evans_police

I1. 3_birthday_friend_daca_wishing
I2. 1_abe_japan_minister_prime

J1. 36_israel_nuclear_usa_axis
J2. 1_israel_isaac_herzog_lebanon_breakthrough

The longevity of a topic is the number of months a topic lasted without fully disappearing. It is possible to identify three particular cases of Longevity in the data:

- **High Longevity** (more than six months), such as groups A and C. Group A begins in April 2021 discussing *infrastructure*, *democrats*, and *republicans*. In July the topics split into *marijuana*, *family payments*, and *democracy, january, facts*. The former two then merge into *tax*. In October a new split occurs: *americans* and *register, vote, day*. Group C begins on April 2022 with *maternal, women, health* and *oklahoma, unconstitutional, abortion*. This merges into *right, abortion*, which in July also merges with *child, affordable*, resulting in *right, biden, act*. This splits into *abortion*, as well as *student debt* and *national, vote, registration*. The group disappears in November with *poll, location, early*

- **Medium Longevity** (five to six months) such as groups B and G. Group B begins in November 2021 with *infrastructure, bills, families*. It splits into *workers, organize, corporate*, as well as *drug, prescription* and *economy, billionaires*. Topic emergence occurs simultaneously to the split: *russia, ukraine, war*. This stagnates until April 2022. Group G encompasses the tweets where Matt Gaetz promotes his podcast Firebrand.

- **Short Longevity** (less than five months). Given their short duration, it follows that topics in this group are unlikely to merge or split. Group D mentions *relief checks*, and group E evolves from *israel, palestinians* to *lgbtq, pride, month*. Group F mentions Jewish celebrations and evolves to the release of Adam Schiff's book *Midnight in Washington*. Group H talks about *vote, georgia, capital, evans*; group I about *prime, minister, abe* and J about *israel, nuclear, lebanon*

6.2 Longevity as a Function of Moral Foundations

The Moral Foundations Theory (MFT) is a framework designed by Jonathan Haidt and Jesse Graham [12]. Its primary aim is to explore the presence of common moral themes across cultures. MFT suggests the existence of inherent psychological systems that form the basis of intuitive ethical judgments. Cultures then build upon these systems to develop virtues and societal structures, resulting in the diverse moral beliefs observed globally. This paradigm of human morality is descriptive, meaning it refrains from making normative claims about the moral goodness of these systems. The original MFT framework identified five foundational pillars:

1. **Care/Harm**: This foundation is related to humans' evolution as mammals, with attachment systems and an ability to feel the pain of others. It underlies the virtues of kindness, gentleness, and nurture.

2. **Fairness/Proportionality**: Related to the evolutionary process of reciprocal altruism, which is a concept that suggests individuals in a community can benefit from cooperating with one another. In the context of morality, it suggests that humans have evolved to value principles of justice and rights because, historically, cooperating and treating others fairly led to mutual benefits and enhanced their survival and well-being.

3. **Loyalty/Disloyalty**: This foundation is related to Human history as tribal creatures that created coalitions. It underlies the principles of patriotism and self-sacrifice for the greater good.

4. **Authority/Subversion**: This foundation was shaped by the history of hierarchical social interactions. It underlies virtues of leadership, including deference to prestigious authority figures and respect for traditions.

5. **Purity/Degradation**: It is influenced by the psychological concepts of disgust and contamination and it is related to the human inclination to aspire to a less primal, more dignified way of living, a notion often present in religious narratives. This foundation supports the belief that the body is sacred and can be defiled by immoral actions and impurities.

Moral Foundations Theory is broadly adopted, in part due to the development of the Moral Foundations Dictionary (MFD) [12]. The MFD defines a taxonomy of values

together with a term dictionary, making it an important resource for NLP applications: since Moral Foundations are focused on intrinsic psychological systems, they can be assigned to groups by their written documents - specifically, party members' tweets. The purpose of this analysis is to understand whether Moral Foundations impact Longevity. Do topics that appeal more to a certain Foundation last longer? The hypotheses are as follows:

FIGURE 6.2: Hypothesis sequence

H1. Longevity of a tweet is related to an author's party. Democrats' tweets Longevity is statistically significantly different from the Longevity of Republican tweets. For a simpler analysis, we are discarding Independents.

H2. Affiliation to different parties translates into different morals. This has been suggested in the literature, and it is expected that the result is consistent in the data [110].

H3. Longevity is driven, at least in part, by Moral Foundations detected in tweets. Assuming the confirmation of H1 and H2, it follows that there are statistically significant differences in Longevity according to morals.

To analyse Moral Foundations present in the data, the package *moralstrength* was used, as developed Araque et al. [111]. In this paper, the authors expand on the MFD to overcome its limitations, namely, increasing its variety and diversity of lemmas, as well as developing a scale of strength of Foundations, instead of only a binary metric.

H1: Longevity of a tweet is related to an author's party

To explore whether there is a relationship between Longevity and Color, the first step is to look at each variable's distribution. Fig. 6.3 shows the violin plots of Longevity for each group. They have similar profiles, with Democrats having a slightly larger frequency around Short Longevity. The second step is choosing the correct statistical test of independence. Longevity is a categorical, ordinal variable, and color is categorical. Both populations Democrats and Republicans are assumed to be independent. Knowing this,

a $\chi 2$ test will be used. In a $\chi 2$ test, the null hypothesis says that the two variables are independent. To reject this hypothesis, a p-value must be below the significance level, α. Running the test for the data, a p-value of 1. This is the highest p-value possible, concluding that at no level of α is it possible to reject the null hypothesis; specifically, H1 cannot be verified.

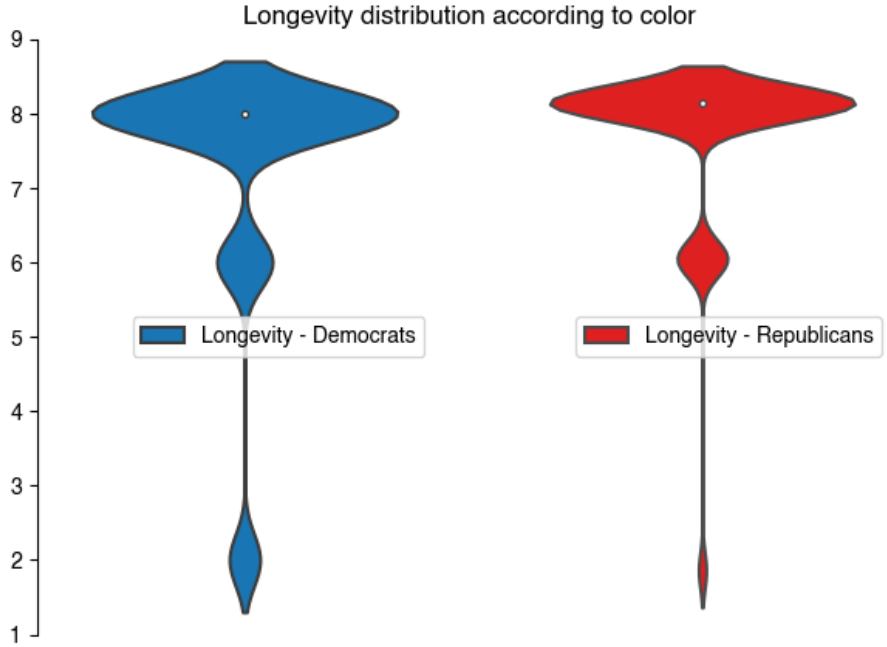


FIGURE 6.3: Distribution of Longevity for Democrats and Republicans

H2: Affiliation to different parties translates into different Moral Foundations

A similar strategy is followed to investigate H2. Firstly, Fig. 6.4 illustrates the distributions of MF scores according to color. It suggests Republican tweets tend to score lower in Care, Authority, and Purity. Fairness scores are similar. Loyalty has a slightly larger frequency of lower scores in Republicans.

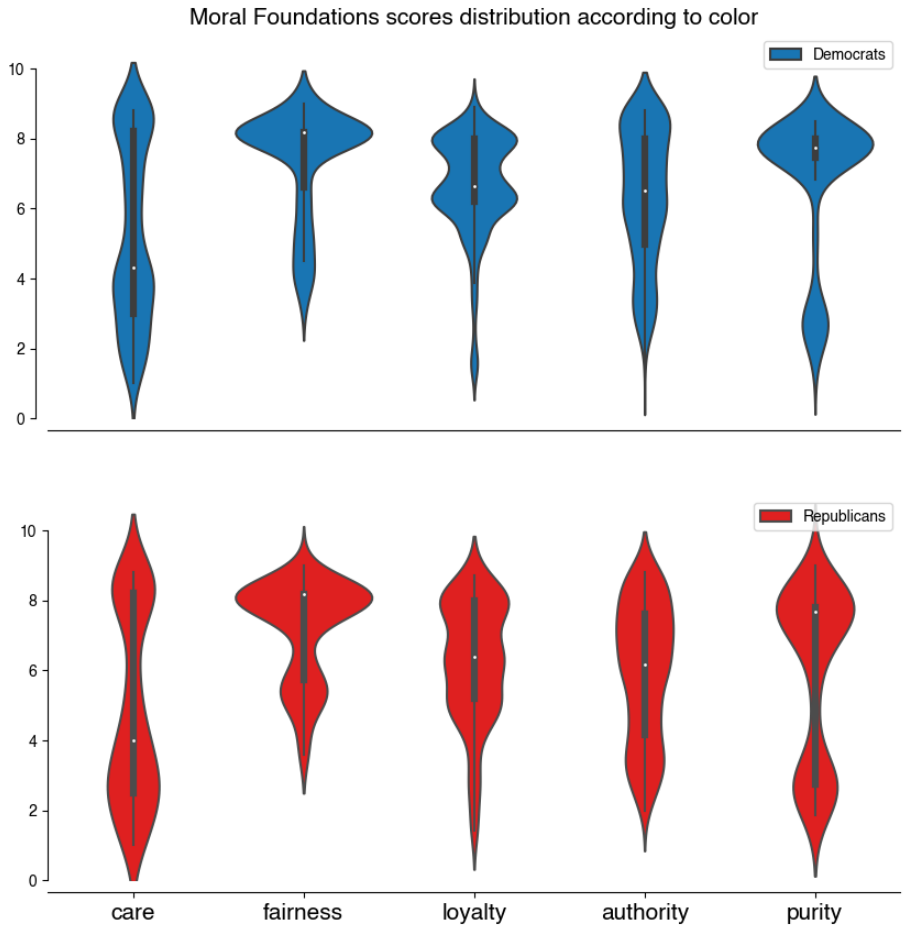


FIGURE 6.4: Distribution of Moral Foundation scores for Democrats and Republicans

Moral Foundations are continuous variables, while color is categorical, and the two populations are assumed to be independent. Since the goal is to understand whether these two variables are correlated, the statistical test for this situation is Mann-Whitney U. Table 6.2 shows the p-values of running the Mann-Whitney U-test for each pair of MF of different parties. It is possible to conclude that, for a significance level of 95%, the null hypothesis of no relationship between the variables can be rejected; at a higher significance level of 99%, this is also true for all except Fairness.

TABLE 6.2: Test statistics for pairs of Moral Foundations of distinct parties

Moral Foundation	**p-Value**	**Democrat Avg. Score**	**Republican Avg. Score**
Care	0.00022	5.27	4.80
Fairness	.0137	7.36	7.17
Loyalty	.0006	6.65	6.22
Authority	.0028	6.34	5.95

Purity	.00008	6.76	5.91

H3: Longevity is driven, at least in part, by Moral Foundations detected in tweets

Similarly to before, this hypothesis compares a categorical variable (Longevity) with a continuous one - Moral Foundations. Fig. 6.5 illustrates the distribution of MF scores for the three levels of Longevity. Medium and Short Longevity share similar distributions of all foundations except Loyalty. Since Longevity has more than two levels, Mann-Whitney U test was run for each pair of MF, for each pair of Longevity levels. Table 6.3 shows the p-values of running the Mann-Whitney U-test for each pair of MF, combining Low and Medium Longevity into one (the remaining combinations can be found in Appendix B, Table C.1). This is the result that yielded the lowest p-values: for a significance level of 99%, the null hypothesis of no relationship between the variables can be rejected for Care and Loyalty, but not for the remaining.

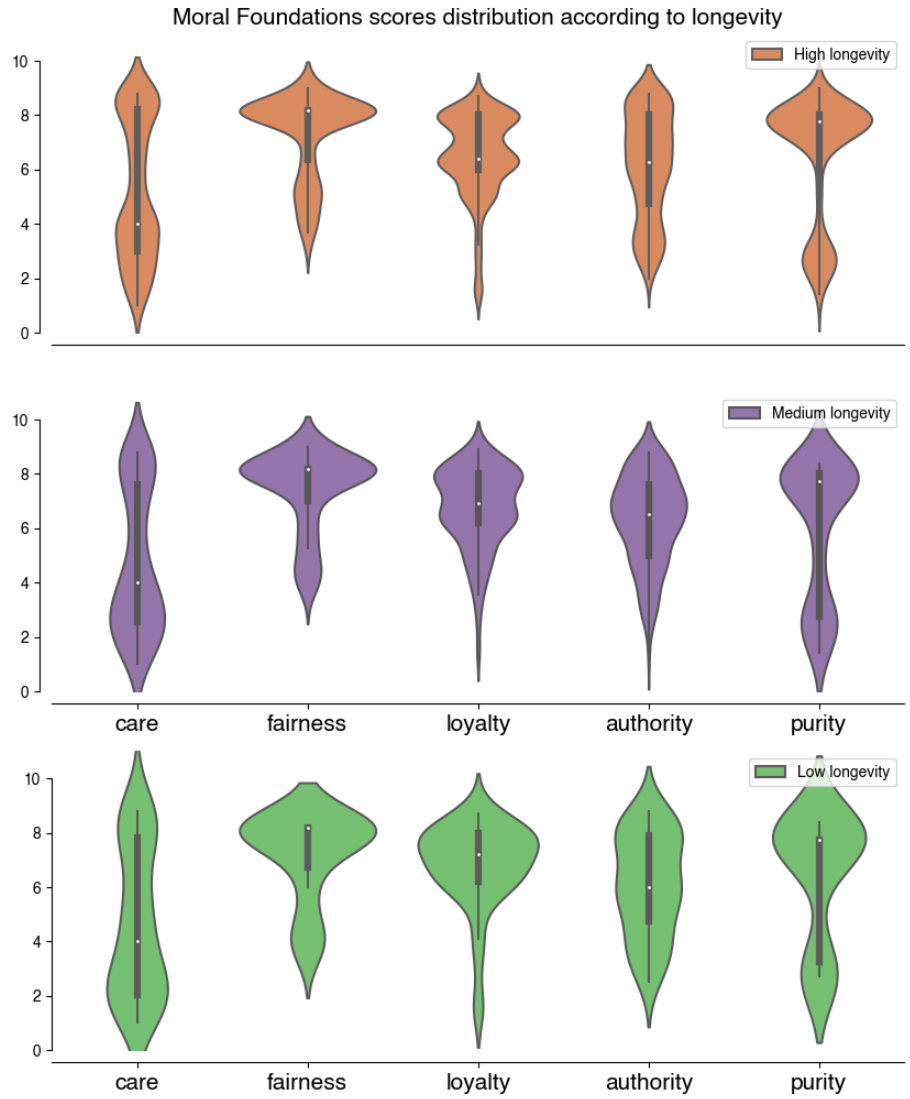


FIGURE 6.5: Distribution of Moral Foundation scores for different levels of Longevity

TABLE 6.3: Test statistics for pairs of Moral Foundations of distinct Longevity

Moral Foundation	p-Value	High Long. Avg. Score	Medium/Short Long. Avg. Score
Care	.0009	5.21	4.69
Fairness	.7817	7.27	7.37
Loyalty	.0072	6.48	6.76
Authority	.7023	6.15	6.27
Purity	.2059	6.47	6.03

6.3 Discussion

Reducing the time interval for topic extraction of tweets is an important step since it is closer to the reality of the quick pace of social media's zeitgeist. In this case, the monthly binning allowed for an increase of 230% in tweets assigned to a topic. This implies that, while the most common topics found for each month tend to align with the overarching ones, there is more granularity found for the subsequent topics. In essence, the most common topics highlight the deeper, structural topics Congress faced (Abortion, Student-Debt), while the remaining relate to singular events such as the Capitol Car Attack or the Assassination of Former Prime Minister Abe Shinzo. While a weekly binning might arguably bring even more detail, the data was not enough for topic extraction. Broader topics also evolve through time, but not simultaneously. Data suggests that High Longevity topics occur one at a time and that the end of one comes with the beginning of another.

Regarding the investigation of Moral Foundations and Topic Longevity, the statistical analysis indicates tweeters of different parties tend to score differently in Moral Foundations. As previously mentioned, this framework is merely descriptive, not providing any sense of superiority of one system over the other. On average, Republicans score lower in Care than Democrats. Care's characteristic emotion is compassion [112] and an example of tweets with distinct levels are:

- *The Senate just passed a bipartisan infrastructure bill. That's a big deal. But the next step must be an investment in our people, too. We need to address climate, health care, housing, education and childcare. Investing in our people, our country, and our planet. Let's do it.* (@RepAndySchiff, 10-08-2021, Democrat. Care = 8.8)

- *More than 3 million people have entered our country illegally on Biden's watch. Some are on the terrorist watch list and others are trafficking fentanyl, which is now the greatest killer of Americans aged 18-45* (@GOPLeader, 03-10-2022, Republican. Care = 1.0)

Loyalty scores of Democrats are higher on Loyalty than Republicans. This Foundation is related to the in-group/out-group perspective. This result differs from those found by previous literature, suggesting a change of attitudes from Democrats:

- *Under Trump and DeVos, the victims of predatory for-profit colleges were abandoned. Now @POTUS and @SecCardona are helping student loan borrowers get debt relief faster and simpler—and forcing colleges to cover the cost of fraud. Special interests lost.* (@SenWarren, 01-11-2022, Democrat. Loyalty = 1.6)

- *The illegitimate January 6 Committee's vote to subpoena President Trump is a political hatchet job read by a political hatchet committee. This committee is illegitimately formed, in violation of House rules, and is organized to search and destroy perceived political enemies.* (@RepAndyBiggsAZ, 13-10-2022, Republican. Loyalty = 8.71)

Republicans score lower in Authority than Democrats. This Foundation applies to traditions, institutions, and values; typically Social Conservatives tend to score higher than Social Liberals. However, this lower score can be interpreted as against the Authority of the ruling party:

- *No amnesty for COVID bureaucrats. No amnesty for illegal aliens* (@RepAndyBiggsAZ, 05-11-2022, Republican. Authority = 3.2)
- *'The former president is still trying to stonewall subpoenas. But this time, we have a Justice Department devoted to the rule of law. This time, lawbreaking witnesses must weigh the prospect of criminal prosecution. Americans deserve answers. We will make sure they get them.* (@RepAdamSchiff, 07-20-2021, Democrat. Authority = 8.67)

Purity is correlated with a sense of disgust, and Social Conservatives particularly rely on this Foundation when discussing the sanctity of life (in the abortion debate), the sanctity of marriage (in the gay rights debate), and the sanctity of self (in the contraception debate). Republicans scored lower than Democrats:

- *The fight to protect voting rights is a long march and an uphill battle. But nothing is more important than securing the most sacred right in our democracy. We will not give up.* (@chuckschumer, 20-01-2022, Democrat. Purity = 8.13)
- *These sick individuals should have never been in our country in the first place. The Biden Administration should be enforcing the immigration laws on our books instead of giving a free pass to those unlawfully entering our country* (@RepAndyBiggsAZ, 22-11-2022, Republican. Purity = 2.71)

The statistical analysis also confirmed that there is a relationship between Longevity and Moral Foundations Care and Loyalty. High Longevity tweets tend to score higher in Care and lower in Loyalty.

Finally, while only two of the hypotheses were confirmed, and the middle connection in Fig. 6.2 was not found, tweets that appeal to Care and Loyalty seem to belong to longer-lasting topics. Fig. 6.6 illustrates the final conclusions.

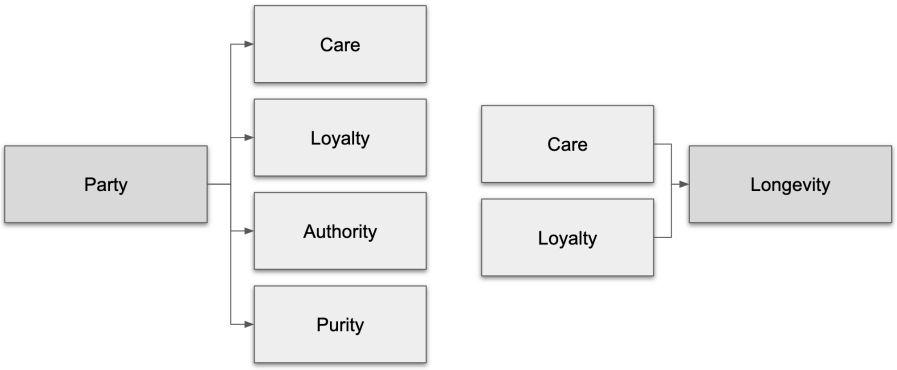


FIGURE 6.6: Summarizing the hypotheses verified

Chapter 7

Conclusions

This chapter contains a general summary of the work performed, a review of the research questions, and the conclusions that emerged. It delves into the limitations and challenges found, and identifies possible opportunities for future work.

7.1 General Summary

This work was carried out in several phases. Firstly, we investigated the different approaches currently available for Topic Extraction, their behavior, capabilities, and drawbacks. Recent years have witnessed the growth of Deep Learning methods, which in the NLP context have focused on the Transformer architecture. Today, one of the best-performing Topic Extraction algorithms built on this framework is BERTopic.

Having identified this algorithm, the next phase was understanding how it could be optimized for the data we selected. BERTopic is highly modular, meaning at each stage, it was necessary to find the best combination of parameters/techniques. Simultaneously, we began treating the data, performing the necessary preprocessing steps and an exploratory analysis.

With an optimized BERTopic and data prepared, the third phase was extracting the topics and investigating their findings. This was done, at first, disregarding timeframes, and only then considering monthly intervals. It was then possible to examine these results under a Moral Foundations framework, in order to find further relations between tweeting activity and user characteristics.

With this work, it is possible to answer the research questions raised in Chapter 1, as summed below:

R.Q.1. How does BERTopic differ from traditional Topic Extraction techniques?

In Chapter 3, we compared the traditional approaches such as LSA and LDA to BERTopic. We showed BERTopic's use of the attention mechanism generates a higher performance, with the additional advantage of not requiring a pre-defined number of topics.

R.Q.2. How can BERTopic be optimized to handle Twitter data?

Chapter 4 offered a discussion on the most relevant techniques at each stage of BERTopic, and Chapter 5 showed how they compared. For our data, BERTopic was optimized by using the techniques listed in 5.1. Concretely, the optimized version achieved increases of 28% in coherence and 48% in stability, when compared with the default algorithm.

R.Q.3. What topics define Congress 117^{th}?

As discussed in Chapter 5, Congress 117^{th} was defined by topics such as Abortion, Student Debt, the election of Judge Ketanji Brown Jackson, and the Russian-Ukrainian War. The contribution to the topics was not equal across parties, with only three of the top 13 having relevant participation from Democrats and Republicans.

R.Q.4. Is there a correlation between the duration of a topic and Moral Foundations it displays?

The monthly analysis of topics in Section 6.1 revealed that some topics evolve throughout time, and others disappear. The Moral Foundations Theory suggests these behaviors are related to an increased use of Care and Loyalty Foundations. It also hints that differences in political ideology are revealed in distinct uses of Care, Loyalty, Authority, and Purity Foundations.

7.2 Contributions

In the course of answering these research questions, this study offers the following contributions:

1. A review of the recent literature in the most common methods of Topic Mining of Short-Text Data

2. An additional document using a state-of-the-art Topic Extraction algorithm, helping to further build its body of use-cases

3. A case-study in the combination of an NLP segment of work - Topic Mining - with a Social Psychology theory - Moral Foundations Theory

4. An instrument for visualization Congress 117^{th} topics of conversation on Twitter, providing an overview of the defining themes approached during its session (Section 5.4), available on GitHub bertopic-twitterverse repository.

7.3 Limitations and Future Work

One of the limitations of this work concerns the difference in the amount of data available between 2021 and 2022. The latter year saw a substantial increase in tweeting activity from the selected members of Congress. This might lead to an overestimation of the importance of 2022 topics in the total of those extracted.

Another limitation is the lack of engagement metrics on the data. Incorporating features such as the number of likes and retweets could have provided a more comprehensive difference between topics, as well as parties.

At the time of the analysis, the package used to determine the presence of Moral Foundations ('moralstrength') in the tweets only featured the five original Foundations. The authors have since developed their Theory to split Fairness into Equality - intuitions about equal treatment and equal outcomes for individuals - and Proportionality - intuitions about individuals getting rewarded in proportion to their merit or contribution. This was a result of this Foundation being skewed toward concerns about equality - a preoccupation more strongly endorsed by politically left-leaning individuals across cultures. Consequently, this skewness might be represented in our study. Other potential new Foundations are also under consideration, namely, Liberty, Honour, and Ownership [113, 114].

Two opportunities for expanding this work are overcoming these limitations. On the one hand, adding engagement metrics to the extracted data and incorporating that information in the analysis - for example, as an additional parameter of Longevity - might provide richer insights into how topics behave over time. However, the new API restrictions defined by X might be an obstacle to this. On the other hand, managing to include the alterations in MFT might yield new results. Moreover, at the time of this writing, the package 'moralstrength' has developed to include Liberty. Given its relationship with libertarianism, it stands as an interesting avenue of investigation [114–116].

Finally, another opportunity would be obtaining data across several different Congresses to analyse how they differed on defining topics, and how similar topics have been approached.

Appendix A

Additional data from Chapter 4

TABLE A.1: Description of selected Congress members

Member name	Description
Adam Schiff	Born in 1960, Schiff is a Harvard lawyer that has served as California Representative since 2001. Schiff chaired the House Intelligence Committee and was the lead impeachment manager in the first trial of President Donald Trump.
Alexandria Ocasio-Cortez	Born in 1989, has served as the Representative for New York since 2019, when she defeated the 10-term incumbent. Taking office at age 29, Ocasio-Cortez is the youngest woman ever to serve in Congress.
Andy Biggs	Born in 1958, is an American attorney and politician serving as Representative for Arizona. In September 2019, Biggs became chairman of the Freedom Caucus, which includes the House Republican's most conservative members.
Bernie Sanders	Born in 1941, serves as Senator from Vermont since 2007. Sanders is the longest-serving independent in U.S. congressional history, and has a close relationship with the Democratic Party. He sought the Democratic Party nomination for President in 2016 and 2020.
Charles Schumer	Also known as Chuck Schumer, he was born in 1950, has served as New York Senator since 1999, and as Senate Majority Leader since 2021. Schumer is the longest-serving senator from New York.

Cory Booker	Born in 1969, has served as Senator from New Jersey since 2013, being its the first African-American to do so. Booker was a candidate for the Democratic nomination in the 2020 U.S. Presidential election, but suspended his campaign due to lack of funding.
Elizabeth Warren	Born in 1949, is the first female Senator from Massachusetts, serving since 2013. Warren was a candidate in the 2020 Democratic Party presidential primaries, finishing third.
Jim Jordan	Born in 1964, has served Ohio Representative since 2007. Jordan is a close ally of former president Donald Trump. After Biden's 2020 election, Jordan supported lawsuits to challenge the election results and voted not to certify the Electoral College results.
Joaquin Castro	Born in 1974, he has served as Texas Representative since 2013. He currently serves on the House Committee on Foreign Affairs and the House Permanent Select Committee on Intelligence.
Joe Biden	Born in 1942, he is the current U.S. President, and served as V.P. under President Barack Obama. Biden is the oldest president in U.S. history, the first to have a female vice president, and the first from Delaware. In April 2023, he announced his candidacy for the Democratic Party nomination in the 2024 election.
John Cornyn	Born in 1952, has served as Texas Senator since 2002. Cornyn chaired the National Republican Senatorial Committee from 2009 to 2013, and served as the Senate majority whip for the 114th Congress.
John Kennedy	Born in 1951, has served as Senator from Louisiana since 2017. Kennedy was an unsuccessful Republican Senate candidate in 2004 and 2008. In 2007, he switched parties and became a Republican.
Kamala Harris	Born in 1964, she is the current and first U.S. female VP. She is also the first African-American and first Asian-American vice president. Harris gained recognition for her questioning of Trump administration officials during Senate hearings.

Kevin McCarthy	Born in 1965, he is the current speaker of the House of Representatives. McCarthy and the Biden administration negotiated to resolve the 2023 debt-ceiling crisis and prevent what would have been a first-ever national default. To resolve the crisis, the parties negotiated the Fiscal Responsibility Act of 2023.
Lee Zeldin	Born in 1980, he served as New York Representative from 2015 to 2023. Zeldin prominently defended Trump during his first impeachment hearings in relation to the Trump–Ukraine scandal.
Marco Rubio	Born in 1971, he has served as Floride Senator since 2011. He unsuccessfully sought the Republican nomination for President in 2016. Due to his influence on U.S. policy on Latin America during the Trump administration, he was described as a "virtual secretary of state for Latin America".
Marjorie Taylor Greene	Born in 1974, has served as Georgia Representative since 2021. Greene is a controversial figure, being considered a far-right conspiracy theorist. She was expelled from the conservative House Freedom Caucus after insulting Congresswoman Lauren Boebert.
Marsha Blackburn	Born in 1952, she has served as Senator from Tennessee since 2018, being the first woman to do so. She has been rated among the House's most conservative members.
Matt Gaetz	Born in 1982, he has served as Representative from Florida since 2017. He is regarded as a staunch proponent of far-right politics, and gained a national profile for defending Florida's controversial "stand-your-ground law".
Mitt Romney	Born in 1947, he has served as Utah Senator since 2019. He won the 2012 Republican presidential nomination, becoming the first Mormon to be a major party's nominee. Romney is considered a moderate Republican: in 2020, he was the only Republican to vote to convict Trump in his first impeachment trial, and also voted to convict in the second trial.
Nancy Pelosi	Born in 1940, she has served as Representative Speaker from 2007 to 2011 and again from 2019 to 2023, being the first woman to hold the office. During her second speakership, the House twice impeached President Donald Trump.

Patty Murray	Born in 1950, she has served as Senate President since 2023, and as Washington Senator since 1993. She was Washington's first female U.S. senator and is the first woman in American history to hold the position of president.
Pramila Jayapal	Born in 1965, she has served as Representative from Washington since 2017. She is the first Indian-American woman to serve as a Representative; and also the first Asian-American to represent Washington at the federal level.
Rand Paul	Born in 1963, he has served as Senator for Kentucky since 2011. He was one of Trump's top defenders in the Senate during his first impeachment trial.
Rick Scott	Born in 1952, he has served as Florida Senator since 2019. He was also the 45^{th} Governor of Florida from 2011 to 2019.
Steny Hoyer	Born in 1939, he has served as Maryland Representative since 1981. He was also a House Majority Leader from 2007 to 2011 and again from 2019 to 2023. Hoyer first attained office through a special election.
Ted Cruz	Born in 1970, has served as Texas Senator since 2013. Cruz ran for President in the 2016 elections against Donald Trump. Cruz received backlash for objecting to Biden's election and giving credence to fraudulent election claims.

Appendix B

Additional data from Chapter 5

FIGURE B.1: Wordcloud of tweets from 28-09-2022

TABLE B.1: Description of the top thirteen topics

Topic name	Description
Abortion	While the name selected was abortion, this topic encompasses several themes, as its representation suggests: 'reduction' shows up often in context of the Inflation Reduction Act; 'prescription', 'drug' and 'pharma' relate to tweets on decreasing drug prices and the monopoly of pharmaceutical companies. Nonetheless, 'abortion' is the top representation, and the topic use correlates strongly to discussions on overturning Roe v. Wade. Roe v. Wade is a U.S. legal case from 1973, where the Supreme Court ruled that restrictive state regulation of abortion is unconstitutional and violated a woman's constitutional right of privacy. Roe v. Wade was overturned by the Supreme Court in 2022. Abortion rights are now decided on a state-level, with some conservative states such as Texas, Alabama and Idaho making it illegal.
StudentDebt	Student debt is a common topic in U.S. politics, where the average student debt is $25,921 and totals $1.73 trillion, the second-highest consumer debt. One of the Biden Administration's key goals was student debt relief, which was blocked by the Supreme Court on June 2022. This topic representations includes 'cancel', 'student', 'debt', 'relief'.
JudgeBrown	In the U.S. Supreme Court, justices have lifetime tenure, meaning they remain on the court until they die, retire, resign, or are impeached and removed from office. When a vacancy occurs, the president appoints a new justice. Following Judge Breyer's retirement, Judge Ketanji Brown Jackson was elected, becoming the first Black woman and the first former federal public defender to serve on the Court. Topic representations include 'ketanji', 'jackson', 'brown', 'judge', 'confirm'.
Voting	In preparation for the midterm election of November 8, 2022, Congress members called Twitter users to vote, sharing information on voting locations, early registration and polls. Topic representations include 'poll', 'early', 'location', 'register'.

UkraineWar	The Russo-Ukrainian war began in 2014, and escalated in 2022, with the Russian invasion of Ukrainian territory. Topic representations include 'putin', 'kyiv', 'ukraine', 'zelensky'.
Cannabis	Biden's statement on marijuana reform, that aims to decrease imprisonment due to cannabis possession brought the topic back to the limelight. Topic representations include 'marijuana', 'cannabis', 'legalize', 'possession', 'prohibition'.
#CommitmentToAmerica	Commitment to America are the Republican Party guidelines to political positioning and strategy. Topic representations include 'commitmenttoamerica', 'housegop','replamalfa'. Rep. LaMalfa is a Republican Representative.
#DefendOurDemocracy	#DefendOurDemocracy is @TeamPelosi's hashtag used to urge voters, celebrate Democrat achievements and comment on other parties. Topic representations include 'defendourdemocracy', 'victory', 'congressman', 'congratulations'.
Firebrand	Firebrand is Rep. Matt Gaetz podcast whose aim is to give *the listener full access, behind-the-scenes look into the Swamp of Washington without the spin of Fake News and Deep State influence*. Topic representations include 'firebrand', 'episode', 'matt', 'feat', 'gaetz'.
MemorialDay	Memorial Day (last Monday of May) and Veterans Day (November 11) are yearly holidays that celebrate U.S. veterans. Topic representations include 'sacrifice', 'veteransday', 'memorialday'.
CovidVaccine	This topic consists mainly of calls-to-action regarding COVID vaccination and booster shots. Topic representations include 'vaccine', '19', 'vaccinate', 'booster', 'update', 'appointment'.
Utah	Essentially driven by Mitt Romney, Utah's Senator, it consists of tweets regarding Utah news; Utah also has a large focus on wildfires. Topic representations include 'utah', 'wildfires', 'infrastructure', 'mitigation', 'drought'.
Congratulations	This topic encompasses tweets celebrating events and achievements, namely the 247^{th} birthday of the U.S. Marine Corps. Topic representations include 'birthday','247th', 'usmc', 'wishing'.

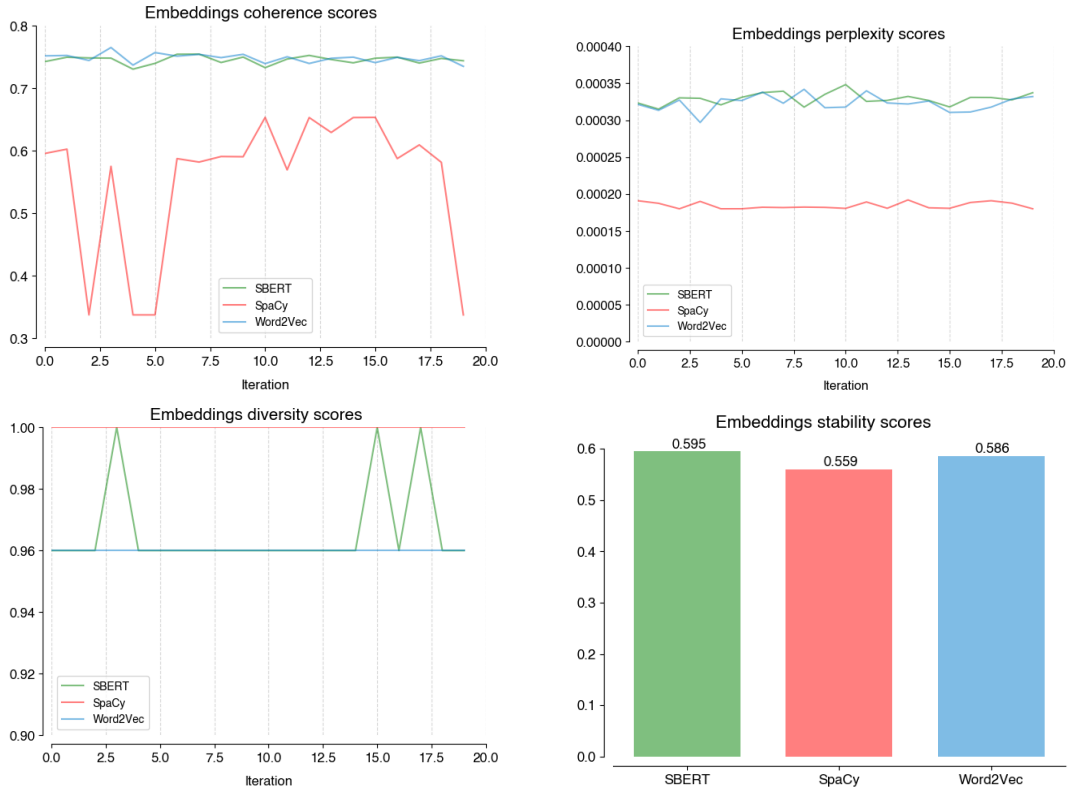


FIGURE B.2: Performance of different models on the **Embeddings** stage

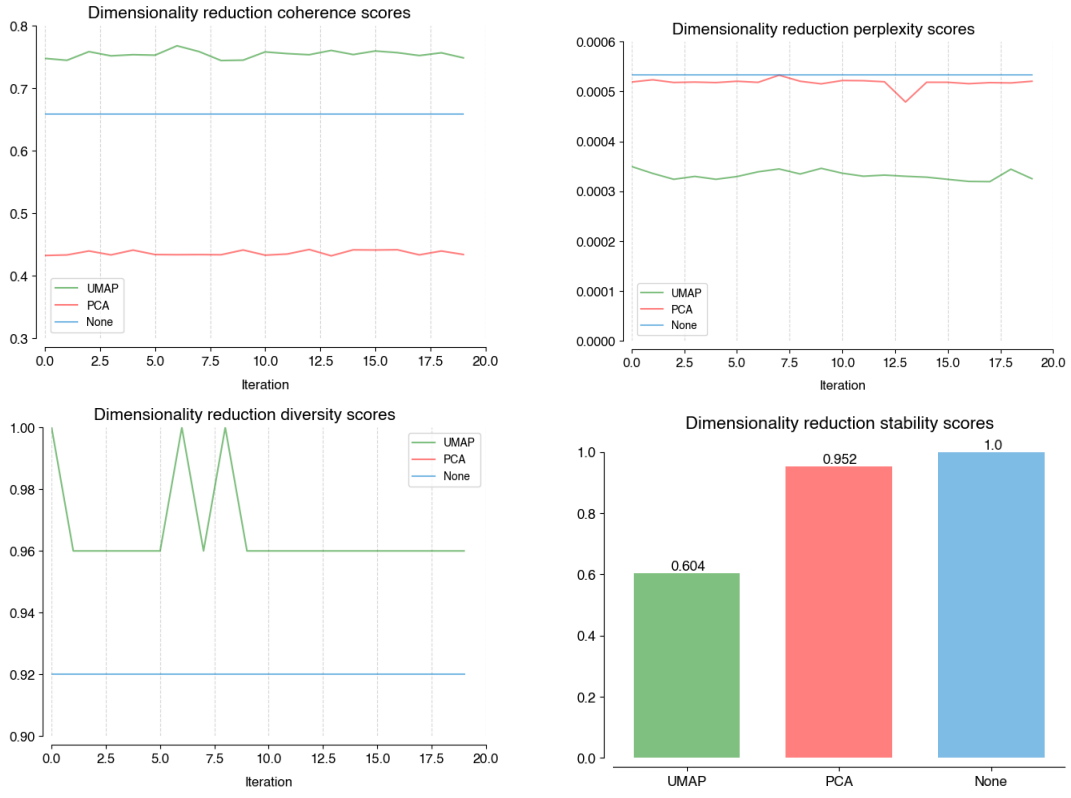


FIGURE B.3: Performance of different models on the **Dimensionality Reduction** stage

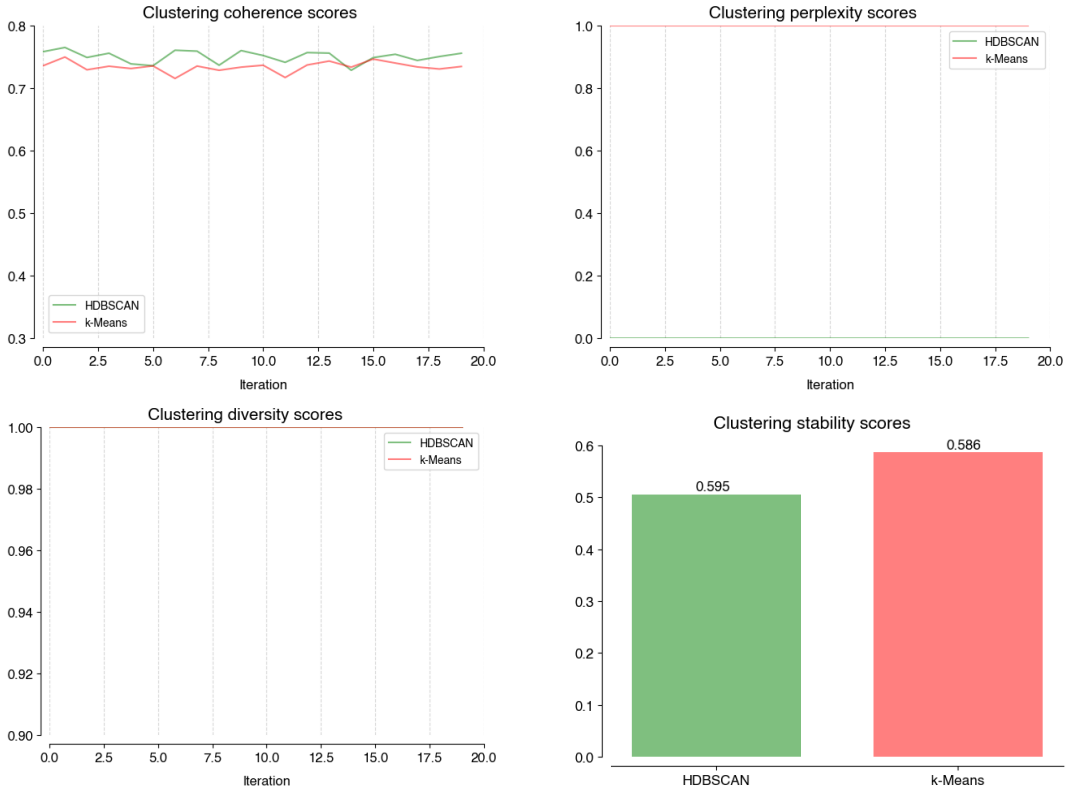


FIGURE B.4: Performance of different models on the **Clustering** stage

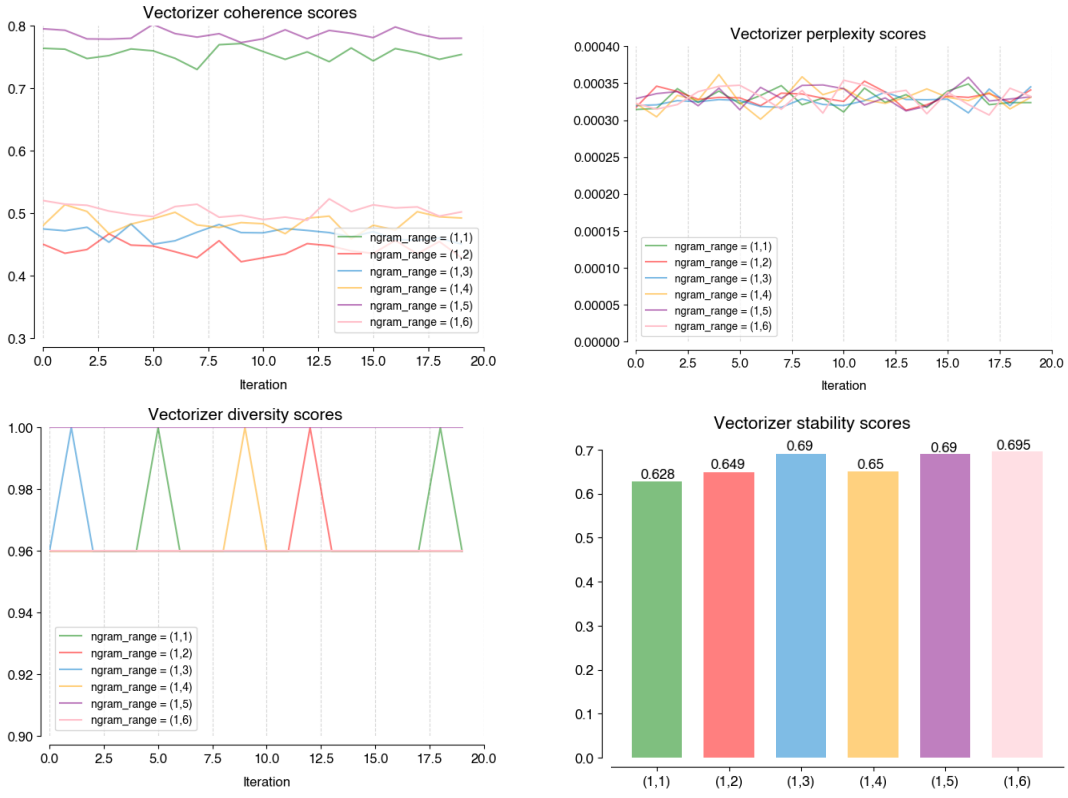


FIGURE B.5: Performance of different values of *ngram_range* on the **Vectorizer** stage

FIGURE B.6: Performance of different values on the *min_df* of the **Vectorizer** stage

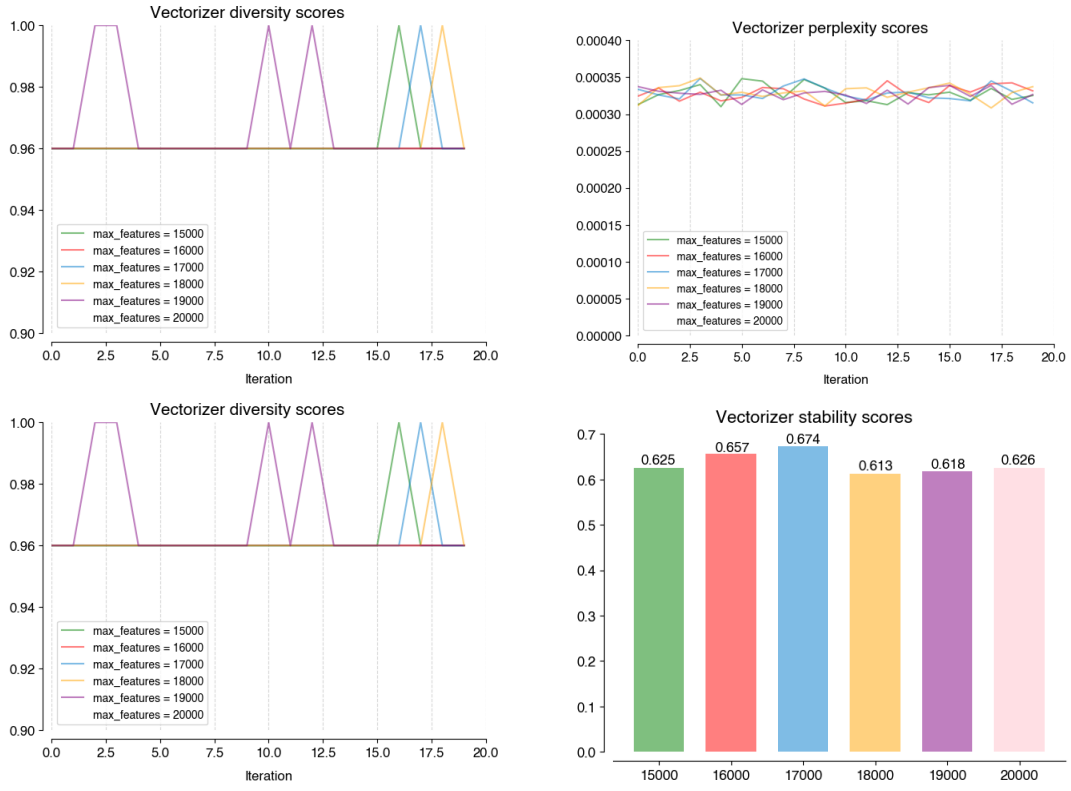


FIGURE B.7: Performance of different values on the *max_features* of the **Vectorizer** stage

FIGURE B.8: Performance of different values on the *bm25_weighting* of the **Topic Representation** stage

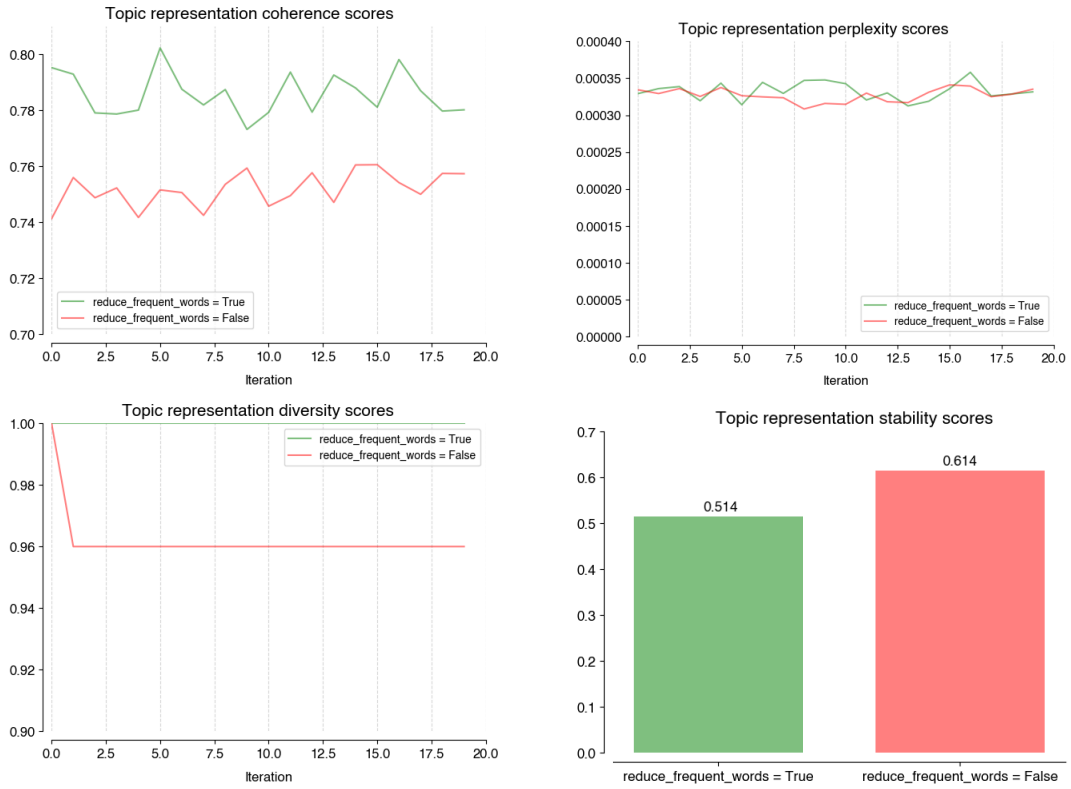


FIGURE B.9: Performance of different values on the *reduce_frequent_words* of the **Topic Representation** stage

Appendix C

Additional data from Chapter 6

TABLE C.1: Test statistics for pairs of Moral Foundations of distinct Longevity

Longevity	Moral Foundation	p-Value
High v. Medium	Care	.0009
	Fairness	.7817
	Loyalty	.0072
	Authority	.7023
	Purity	.2059
Medium v. Low	Care	.0009
	Fairness	.7817
	Loyalty	.0072
	Authority	.7023
	Purity	.2059
High v. Low	Care	.0009
	Fairness	.7817
	Loyalty	.0072
	Authority	.7023
	Purity	.2059
High/Medium v. Low	Care	.0009
	Fairness	.7817
	Loyalty	.0072
	Authority	.7023
	Purity	.2059

Bibliography

[1] H. Satterfield, "How social media affects politics," *Meltwater*, 3 2020. [Cited on page 1.]

[2] TEDx Talks, "How Social Media is Shaping Our Political Future | Victoria Bonney | TEDxDirigo," Dec. 2018. [Online]. Available: https://www.youtube.com/watch?v=9Kd99IIWJUw [Cited on page 1.]

[3] S. J. Dixon, "Topic: Social media and politics in the United States." [Online]. Available: https://www.statista.com/topics/3723/social-media-and-politics-in-the-united-states/ [Cited on page 1.]

[4] "X/Twitter: number of users worldwide 2024." [Online]. Available: https://www.statista.com/statistics/303681/twitter-users-worldwide/ [Cited on page 1.]

[5] M. Reveilhac and D. Morselli, "The Impact of Social Media Use for Elected Parliamentarians: Evidence from Politicians' Use of Twitter During the Last Two Swiss Legislatures," *Swiss Political Science Review*, vol. 29, no. 1, pp. 96–119, 2023, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/spsr.12543. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/spsr.12543 [Cited on page 1.]

[6] "One year inside Trump's monumental Facebook campaign | Donald Trump | The Guardian." [Online]. Available: https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election [Cited on page 2.]

[7] Center for Humane Technology, "How Social Media Polarizes Political Campaigns," Dec. 2021. [Online]. Available: https://www.youtube.com/watch?v=1GRxORsQhY4 [Cited on page 2.]

[8] A. C. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, H. C. Alcântara Paulo, Y. Zhang, J. S. Erickson, and K. P. Bennett, "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse," *AMIA Summits on Translational Science Proceedings*, vol. 2021, pp. 555–564, May 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378598/ [Cited on page 2.]

[9] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136–147, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585317304999

[10] D. Bor, B. S. Lee, and E. J. Oughton, "Quantifying polarization across political groups on key policy issues using sentiment analysis," Mar. 2023, arXiv:2302.07775 [stat]. [Online]. Available: http://arxiv.org/abs/2302.07775 [Cited on page 2.]

[11] "Papers." [Online]. Available: https://www.mft-nlp.com/papers.html [Cited on page 2.]

[12] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations." *Journal of Personality and Social Psychology*, vol. 96, no. 5, pp. 1029–1046, May 2009. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0015141 [Cited on pages 2 and 56.]

[13] L. M. Galli and J. G. Modesto, "Polarization about Human Rights: Political Orientation, Morality, and Belief in a Just World," *Trends in Psychology*, Feb. 2023. [Online]. Available: https://doi.org/10.1007/s43076-023-00260-4 [Cited on page 2.]

[14] R. S. Abeywickrama, J. Rhee, D. L. Crone, and S. M. Laham, "Why Moral Advocacy Leads to Polarization and Proselytization: The Role of Self-Persuasion," Jan. 2020, publisher: Deakin University. [Online]. Available: https://dro.deakin.edu.au/articles/journal_contribution/Why_Moral_Advocacy_Leads_to_Polarization_and_Proselytization_The_Role_of_Self-Persuasion/20639154/1

[15] J. Wronski, "Intergroup Identities, Moral Foundations, and Their Political Consequences: A Review of Social Psychology of Political Polarization by Piercarlo

Valdesolo and Jesse Graham (Eds),” *Social Justice Research*, vol. 29, no. 3, pp. 345–353, Sep. 2016. [Online]. Available: https://doi.org/10.1007/s11211-016-0271-0 [Cited on page 2.]

[16] “Timeline of advances in the field of NLP that led to development of tools like ChatGPT,” Jun. 2020. [Online]. Available: https://dev.to/amananandrai/recent-advances-in-the-field-of-nlp-33o1 [Cited on page 2.]

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762 [Cited on page 2.]

[18] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” 3 2022. [Online]. Available: http://arxiv.org/abs/2203.05794 [Cited on page 3.]

[19] A. Geiger, “Political Polarization in the American Public,” Jun. 2014. [Online]. Available: https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/ [Cited on page 3.]

[20] A. Hajjej, “Trump tweets: Topic Modeling using Latent Dirichlet Allocation,” Jun. 2020. [Online]. Available: https://medium.datadriveninvestor.com/trump-tweets-topic-modeling-using-latent-dirichlet-allocation-e4f93b90b6fe [Cited on page 3.]

[21] M. S. K. Abadah, P. Keikhosrokiani, and X. Zhao, “Analytics of Public Reactions to the COVID-19 Vaccine on Twitter Using Sentiment Analysis and Topic Modelling,” in *Handbook of Research on Applied Artificial Intelligence and Robotics for Government Processes*. IGI Global, 2023, pp. 156–188. [Online]. Available: https://www.igi-global.com/chapter/analytics-of-public-reactions-to-the-covid-19-vaccine-on-twitter-using-sentiment-analysis-and-topic-modelling/www.igi-global.com/chapter/analytics-of-public-reactions-to-the-covid-19-vaccine-on-twitter-using-sentiment-analysis-and-topic-modelling/312626 [Cited on page 3.]

[22] S. Zhou, K. Pengyu, H. Qunying, and J. Silbernagel, “A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura,” 2023. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/01655515211007724 [Cited on page 3.]

[23] B. Knoll, "President Obama, the Democratic Party, and Socialism: A Political Science Perspective," Jun. 2012, section: Politics. [Online]. Available: https://www.huffpost.com/entry/obama-romney-economy_b_1615862 [Cited on page 7.]

[24] DemocraticParty, "Where we stand." [Online]. Available: https://democrats.org/where-we-stand/ [Cited on page 7.]

[25] R. N. Committee, "Gop - about our party." [Cited on page 7.]

[26] "U.S. Senate: Constitution Day." [Online]. Available: https://www.senate.gov/about/origins-foundations/senate-and-constitution/constitution-day.htm [Cited on page 8.]

[27] U. S. Senate, "Constitution of the united states." [Cited on page 8.]

[28] C. Course and C. Benzine, "The bicameral congress: Crash course government and politics 2," 1 2015. [Cited on page 8.]

[29] ——, "Congressional elections: Crash course government and politics 6," 2 2015. [Cited on page 8.]

[30] S. Binder, "Goodbye to the 117th congress, bookended by remarkable events," *Washington Post*, 12 2022. [Cited on page 9.]

[31] U. H. of Representatives Press Gallery, "Members' official twitter handles." [Cited on pages 9 and 23.]

[32] S. Lee and G. Panetta, "Twitter is the most popular social media platform for members of congress — but prominent democrats tweet more often and have larger followings than republicans," *Business Insider*, 2 2019. [Cited on page 9.]

[33] B. R. Mills, "Take it to twitter: Social media analysis of members of congress," *Towards Data Science*, 8 2021. [Cited on page 9.]

[34] B. Marr, "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read," section: Enterprise & Cloud. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/ [Cited on page 11.]

[35] "What is the size (ratio) of unstructured data (text) on the web?" [Online]. Available: https://www.researchgate.net/post/What_is_the_size_ratio_of_unstructured_data_text_on_the_web [Cited on page 11.]

[36] F. P, "The Challenge of Building Corpus for NLP Libraries," Jul. 2020. [Online]. Available: https://www.defined.ai/blog/the-challenge-of-building-corpus-for-nlp-libraries/ [Cited on page 11.]

[37] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5133–5260, Jun. 2023. [Online]. Available: https://link.springer.com/10.1007/s10462-022-10254-w [Cited on page 12.]

[38] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, Aug. 1954. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/00437956.1954.11659520 [Cited on page 13.]

[39] K. S. Jones, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL." [Cited on page 13.]

[40] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, Mar. 1963. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845 [Cited on page 13.]

[41] J. Macqueen, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS," *MULTIVARIATE OBSERVATIONS*. [Cited on page 13.]

[42] L. Xia, C. Zhang, D. Luo, and Z. Wu, "A Survey of Topic Models in Text Classification." [Cited on page 13.]

[43] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, Sep. 1990. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9 [Cited on page 13.]

[44] D. Valdez, A. Pickett, and P. Goodson, "Topic Modeling: Latent Semantic Analysis for the Social Sciences*," *Social Science Quarterly*, vol. 99, Sep. 2018. [Cited on page 14.]

[45] T. V. Sai, K. Lohith, M. Sai, K. Tejaswi, P. Ashok Kumar, and K. C, "Text Analysis On Twitter Data Using LSA and LDA," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Jan. 2023, pp. 1–6, iSSN: 2473-7577. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10128417?casa_token=6W1hWExHs0AAAAAA:-feVE4MDRHAmwSX52eRiX92XOX4562tFuNNuYwXxBFuMtKYayOPmUQmPcKiAG4Onc-n0X14KAao [Cited on page 14.]

[46] P. Chang, Y.-T. Yu, A. Sanders, and T. Munasinghe, "Perceiving the Ukraine-Russia Conflict: Topic Modeling and Clustering on Twitter Data," in *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, Jul. 2023, pp. 147–148. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10234002?casa_token=ffoDYLyDBckAAAAA:hNcolK05ClImGYtdgmCw_JV5v14FWFj3PnvuuxZW0JWKTkkuvXoAjT---60Wvxxfa8tUSuxaKl0 [Cited on page 14.]

[47] S. Qomariyah, N. Iriawan, and K. Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis," vol. 2194, Dec. 2019, p. 020093. [Cited on page 14.]

[48] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy Approach Topic Discovery in Health and Medical Corpora," *International Journal of Fuzzy Systems*, vol. 20, no. 4, pp. 1334–1345, Apr. 2018, arXiv:1705.00995 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1705.00995 [Cited on page 14.]

[49] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis," *Expert Systems with Applications*, vol. 152, p. 113401, Aug. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417420302256 [Cited on page 14.]

[50] T. Hofmann, "Probabilistic Latent Semantic Indexing," *ACM SIGIR Forum*, vol. 51, no. 2, 2017. [Cited on page 14.]

[51] P. Kumar and M. Vardhan, "Aspect-Based Sentiment Analysis of Tweets Using Independent Component Analysis (ICA) and Probabilistic Latent Semantic Analysis (pLSA)," in *Lecture Notes in Networks and Systems*, Jun. 2019, pp. 3–13, journal Abbreviation: Lecture Notes in Networks and Systems. [Cited on page 14.]

[52] Y. Shen and H. Guo, "Research on high-performance English translation based on topic model," *Digital Communications and Networks*, vol. 9, no. 2, pp. 505–511, Apr. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864822000360 [Cited on page 14.]

[53] D. M. Blei, "Latent Dirichlet Allocation." [Cited on page 14.]

[54] D. Anastasiu, A. Tagarelli, and G. Karypis, *Document Clustering: The Next Frontier*, 01 2013, pp. 305–338. [Cited on page 15.]

[55] M. Griffiths, T. L.; Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences 2004-feb 10 vol. 101 iss. Supplement 1*, vol. 101, feb 2004. [Online]. Available: libgen.li/file.php?md5=e438e887b2eb5c499cfdf89786901480 [Cited on page 16.]

[56] F. Mehrpour, "Analyzing Twitter Sentiment and Hype on Real Estate Market: A Topic Modeling Approach," May 2023, accepted: 2023-05-26T15:54:33Z Publisher: Brock University. [Online]. Available: https://dr.library.brocku.ca/handle/10464/17848 [Cited on page 16.]

[57] M. I. Fakhri and H. Irawan, "Analyzing sentiment and topic modelling of iPhone Xs post launch event through Twitter data," *AIP Conference Proceedings*, vol. 2646, no. 1, p. 040030, Apr. 2023. [Online]. Available: https://doi.org/10.1063/5.0139392

[58] I. F. Strydom, J. Grobler, and E. Vermeulen, "Investigating the Use of Topic Modeling for Social Media Market Research: A South African Case Study," in *Computational Science and Its Applications – ICCSA 2023*, ser. Lecture Notes in Computer Science, O. Gervasi, B. Murgante, D. Taniar, B. O. Apduhan, A. C. Braga, C. Garau, and A. Stratigea, Eds. Cham: Springer Nature Switzerland, 2023, pp. 305–320. [Cited on page 16.]

[59] J. Kaur, I. Z. Hussain, M. Lotto, Z. Butt, and P. Morita, "Preventing public health crises: an expert system using Big Data and AI in combating the spread of health misinformation," *Population Medicine*, vol. 5,

no. Supplement, Apr. 2023, publisher: E.U. European Publishing. [Online]. Available: http://www.populationmedicine.eu/Preventing-public-health-crises-an-expert-system-using-Big-Data-and-AI-in-combating,165379,0,2.html [Cited on page 16.]

[60] P. S. V., R. Ittamalla, M. Mahipalan, M. Mahitha, and D. H. Priya, "What Do Veterans Discuss the Most about Post-Combat Stress on Social Media? – A Text Analytics Study," *Journal of Loss and Trauma*, vol. 28, no. 2, pp. 187–189, Feb. 2023, publisher: Routledge _eprint: https://doi.org/10.1080/15325024.2022.2068662. [Online]. Available: https://doi.org/10.1080/15325024.2022.2068662

[61] A. Lyu, C. Liu, Z. Ding, J. Li, and W. Zhang, "Analysis of gender sentiment expression in network based on TF-LDA algorithm," *Advances in Engineering Technology Research*, vol. 5, no. 1, pp. 322–322, May 2023, number: 1. [Online]. Available: https://madison-proceedings.com/index.php/aetr/article/view/1103 [Cited on page 16.]

[62] S. T. Bheema and S. K. Kotha, "_USInsights from Covid-19 #Vaccine Twitter analytics," Apr. 2023, accepted: 2023-05-11T18:30:12Z Publisher: Wichita State University. [Online]. Available: https://soar.wichita.edu/handle/10057/25264 [Cited on page 16.]

[63] C. Comito, "How Do We Talk and Feel About COVID-19? Sentiment Analysis of Twitter Topics," in *Big Data – BigData 2023*, ser. Lecture Notes in Computer Science, S. Zhang, B. Hu, and L.-J. Zhang, Eds. Cham: Springer Nature Switzerland, 2023, pp. 95–107.

[64] J. J. M. L. S. I. T. N. G. A. Mahara, Akshay Sriram, "Analyzing the role of Indian media during the second wave of COVID using topic modeling," in *Hybrid Computational Intelligent Systems*. CRC Press, 2023, num Pages: 11. [Cited on page 16.]

[65] F. Meier and M. Fugl Eskjær, "Topic Modelling Three Decades of Climate Change News in Denmark," Rochester, NY, Jul. 2023. [Online]. Available: https://papers.ssrn.com/abstract=4513921 [Cited on page 16.]

[66] R. G. Rathod, Y. Barve, J. R. Saini, and S. Rathod, "From Data Pre-processing to Hate Speech Detection: An Interdisciplinary Study on Women-targeted Online Abuse," in *2023 3rd International Conference*

on Intelligent Technologies (CONIT), Jun. 2023, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10205571?casa_token=3ZUAO97rFt4AAAAA:NmHKQ0osy2PfkT8JAHOH_61YMvLdslgTfIEZ37Mi-6Jlr2eSLGutZ6yxWxtLCeaHdxUVFQS4mPU [Cited on page 16.]

[67] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," in *Advances in Information Retrieval*, P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6611, pp. 338–349, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-20161-5_34 [Cited on page 16.]

[68] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013, arXiv:1301.3781 [cs]. [Online]. Available: http://arxiv.org/abs/1301.3781 [Cited on page 17.]

[69] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," May 2014, arXiv:1405.4053 [cs]. [Online]. Available: http://arxiv.org/abs/1405.4053 [Cited on page 18.]

[70] D. Angelov, "Top2Vec: Distributed Representations of Topics," Aug. 2020, arXiv:2008.09470 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2008.09470 [Cited on page 18.]

[71] StatQuest with Josh Starmer, "UMAP Dimension Reduction, Main Ideas!!!" Jul. 2022. [Online]. Available: https://www.youtube.com/watch?v=eN0wFzBA4Sc [Cited on pages 18 and 21.]

[72] ——, "Clustering with DBSCAN, Clearly Explained!!!" Oct. 2022. [Online]. Available: https://www.youtube.com/watch?v=RDZUdRSDOok [Cited on pages 19 and 21.]

[73] B. Karas, S. Qu, Y. Xu, and Q. Zhu, "Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis," *Frontiers in Artificial Intelligence*, vol. 5, p. 948313, Aug. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.948313/full [Cited on page 19.]

[74] F. D. Zengul, A. Bulut, N. Oner, A. Ahmed, B. Ozaydin, and M. Yadav, "A Practical and Empirical Comparison of Three Topic Modeling Methods using a COVID-19 Corpus: LSA, LDA, and Top2Vec." [Cited on page 19.]

[75] D. Vianna and E. Silva De Moura, "Organizing Portuguese Legal Documents through Topic Discovery," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid Spain: ACM, Jul. 2022, pp. 3388–3392. [Online]. Available: https://dl.acm.org/doi/10.1145/3477495.3536329 [Cited on page 19.]

[76] A. Crijns, V. Vanhullebusch, M. Reusens, M. Reusens, and B. Baesens, "Topic modelling applied on innovation studies of Flemish companies," *Journal of Business Analytics*, vol. 6, no. 4, pp. 243–254, Oct. 2023, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/2573234X.2023.2186274. [Online]. Available: https://doi.org/10.1080/2573234X.2023.2186274 [Cited on page 19.]

[77] D. Bretsko, A. Belyi, and S. Sobolevsky, "Comparative Analysis of Community Detection and Transformer-Based Approaches for Topic Clustering of Scientific Papers," in *Computational Science and Its Applications – ICCSA 2023*, ser. Lecture Notes in Computer Science, O. Gervasi, B. Murgante, D. Taniar, B. O. Apduhan, A. C. Braga, C. Garau, and A. Stratigea, Eds. Cham: Springer Nature Switzerland, 2023, pp. 648–660. [Cited on page 19.]

[78] J. Von Der Mosel, A. Trautsch, and S. Herbold, "On the Validity of Pre-Trained Transformers for Natural Language Processing in the Software Engineering Domain," *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 1487–1507, Apr. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9785808/ [Cited on page 19.]

[79] M. P. Grootendorst, "The Algorithm - BERTopic." [Online]. Available: https://maartengr.github.io/BERTopic/algorithm/algorithm.html [Cited on page 20.]

[80] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, arXiv:1908.10084 [cs]. [Online]. Available: http://arxiv.org/abs/1908.10084 [Cited on pages 20 and 27.]

[81] James Briggs, "BERTopic Explained," Nov. 2022. [Online]. Available: https://www.youtube.com/watch?v=fb7LENb9eag [Cited on page 20.]

[82] M. Hägglund, M. Blusi, and S. Bonacina, *Caring is Sharing — Exploiting the Value in Data for Health and Innovation: Proceedings of MIE 2023*. IOS Press, Jun. 2023, google-Books-ID: rObGEAAAQBAJ. [Cited on page 21.]

[83] Y. LI, "Insights from Tweets: Analysing Destination Topics and Sentiments, and Predicting Tourist Arrivals," Doctoral, Durham University, 2023. [Online]. Available: http://etheses.dur.ac.uk/15088/ [Cited on page 21.]

[84] I. F. Strydom and J. Grobler, "Topic Modelling for Characterizing COVID-19 Misinformation on Twitter: A South African Case Study," in *Computational Science and Its Applications – ICCSA 2023*, ser. Lecture Notes in Computer Science, O. Gervasi, B. Murgante, D. Taniar, B. O. Apduhan, A. C. Braga, C. Garau, and A. Stratigea, Eds. Cham: Springer Nature Switzerland, 2023, pp. 289–304. [Cited on page 21.]

[85] J. Turner, M. McDonald, and H. Hu, "An Interdisciplinary Approach to Misinformation and Concept Drift in Historical Cannabis Tweets," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, Feb. 2023, pp. 317–322, iSSN: 2325-6516. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10066734?casa_token=c_1E3jXmvW0AAAAA:kbF2O6RMAFmQgIQ0C3irxCx2sG2XnzuxF33E7gSLgkwUXR65RkBHpeM9oNvWI4e0CQFhii3dD [Cited on page 21.]

[86] R. Koonchanok, Y. Pan, and H. Jang, "Tracking public attitudes toward ChatGPT on Twitter using sentiment analysis and topic modeling," Jun. 2023, arXiv:2306.12951 [cs]. [Online]. Available: http://arxiv.org/abs/2306.12951 [Cited on page 21.]

[87] D.-N. Grigore and I. Pintilie, "Notebook for the eRisk Lab at CLEF 2023." [Cited on page 21.]

[88] A. Mekacher, M. Falkenberg, and A. Baronchelli, "The Systemic Impact of Deplatforming on Social Media," Mar. 2023, arXiv:2303.11147 [physics]. [Online]. Available: http://arxiv.org/abs/2303.11147 [Cited on page 21.]

[89] N. Schneider, S. Shouei, S. Ghantous, and E. Feldman, "Hate Speech Targets Detection in Parler using BERT," Apr. 2023, arXiv:2304.01179 [cs]. [Online]. Available: http://arxiv.org/abs/2304.01179 [Cited on page 21.]

[90] A. Amato and V. D. Lecce, "Data preprocessing impact on machine learning algorithm performance," *Open Computer Science*, vol. 13, no. 1, Jan. 2023, publisher: De Gruyter Open Access. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/comp-2022-0278/html [Cited on page 25.]

[91] A. Burchfiel, "What is NLP (Natural Language Processing) Tokenization?" May 2022. [Online]. Available: https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/ [Cited on page 25.]

[92] "Stop words list." [Online]. Available: https://countwordsfree.com/stopwords [Cited on page 25.]

[93] S. Huston and W. B. Croft, "Evaluating verbose query processing techniques," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '10. New York, NY, USA: Association for Computing Machinery, Jul. 2010, pp. 291–298. [Online]. Available: https://doi.org/10.1145/1835449.1835499 [Cited on page 25.]

[94] "Porter Stemmer algorithm," May 2019. [Online]. Available: https://iq.opengenus.org/porter-stemmer/ [Cited on page 25.]

[95] "Snowball." [Online]. Available: https://snowballstem.org/ [Cited on page 25.]

[96] E. Zvornicanin, "Differences Between Porter and Lancaster Stemming Algorithms | Baeldung on Computer Science," Jul. 2022. [Online]. Available: https://www.baeldung.com/cs/porter-vs-lancaster-stemming-algorithms [Cited on page 25.]

[97] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," Dec. 2022, arXiv:2212.08459 [cs]. [Online]. Available: http://arxiv.org/abs/2212.08459 [Cited on page 27.]

[98] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, p. 886498, May 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full [Cited on page 27.]

[99] M. P. Grootendorst, "1. Embeddings - BERTopic." [Online]. Available: https://maartengr.github.io/BERTopic/getting_started/embeddings/embeddings.html [Cited on page 27.]

[100] "Gensim: topic modelling for humans." [Online]. Available: https://radimrehurek.com/gensim/ [Cited on page 28.]

[101] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en. [Cited on page 28.]

[102] J. Shlens, "A tutorial on principal component analysis," *CoRR*, vol. abs/1404.1100, 2014. [Online]. Available: http://arxiv.org/abs/1404.1100 [Cited on page 28.]

[103] X. Jin and J. Han, "K-Means Clustering," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 563–564. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_425 [Cited on page 28.]

[104] M. P. Grootendorst, "4. Vectorizers - BERTopic." [Online]. Available: https://maartengr.github.io/BERTopic/getting_started/vectorizers/vectorizers.html [Cited on page 29.]

[105] ——, "5. c-TF-IDF - BERTopic." [Online]. Available: https://maartengr.github.io/BERTopic/getting_started/ctfidf/ctfidf.html [Cited on page 29.]

[106] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102131, Feb. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437922001090 [Cited on page 30.]

[107] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," Jul. 2019, arXiv:1907.04907 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.04907 [Cited on page 30.]

[108] K. Leetaru, "Is Twitter Really Faster Than The News?" section: AI & Big Data. [Online]. Available: https://www.forbes.com/sites/kalevleetaru/2019/02/26/is-twitter-really-faster-than-the-news/ [Cited on page 51.]

[109] M. Geller, V. V. Vasconcelos, and F. L. Pinheiro, "Toxicity in Evolving Twitter Topics," in *Computational Science – ICCS 2023*, ser. Lecture Notes in Computer Science, J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra,

and P. M. Sloot, Eds. Cham: Springer Nature Switzerland, 2023, pp. 40–54. [Cited on page 53.]

[110] S. P. Koleva, J. Graham, R. Iyer, P. H. Ditto, and J. Haidt, "Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes," *Journal of Research in Personality*, vol. 46, no. 2, pp. 184–194, Apr. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656612000074 [Cited on page 57.]

[111] O. Araque, L. Gatti, and K. Kalimeri, "MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction," *Knowledge-Based Systems*, vol. 191, p. 105184, Mar. 2020, arXiv:1904.08314 [cs]. [Online]. Available: http://arxiv.org/abs/1904.08314 [Cited on page 57.]

[112] J. Schuman, "Moral Foundations Theory Explained by Jonathan Haidt," Jul. 2018. [Online]. Available: https://dividedwefall.org/the-righteous-mind-moral-foundations-theory/ [Cited on page 62.]

[113] M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani, "Morality beyond the WEIRD: How the nomological network of morality varies across cultures," *Journal of Personality and Social Psychology*, pp. No Pagination Specified–No Pagination Specified, 2023, place: US Publisher: American Psychological Association. [Cited on page 67.]

[114] "Moral Foundations Theory | moralfoundations.org." [Online]. Available: https://moralfoundations.org/ [Cited on page 67.]

[115] O. Araque, L. Gatti, and K. Kalimeri, "LibertyMFD: A Lexicon to Assess the Moral Foundation of Liberty." in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, ser. GoodIT '22. New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 154–160. [Online]. Available: https://doi.org/10.1145/3524458.3547264

[116] O. A. Kalimeri, Lorenzo Gatti and Kyriaki, "moralstrength: A package to predict the Moral Foundations for a tweet or text." [Online]. Available: https://github.com/oaraque/moral-foundations/ [Cited on page 67.]