


Article

Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse

Margarida Mendonça ^{1,†} and Álvaro Figueira ^{2,*,†} ¹ Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal² CRACS-INESCTEC and Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

* Correspondence: arfiguei@fc.up.pt

† Current address: DCC-FCUP, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal.

Abstract: As social media (SM) becomes increasingly prevalent, its impact on society is expected to grow accordingly. While SM has brought positive transformations, it has also amplified pre-existing issues such as misinformation, echo chambers, manipulation, and propaganda. 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 the potential negative effects. One such tool is BERTopic, a novel deep-learning algorithm developed for Topic Mining, which has been shown to offer significant advantages over traditional methods like Latent Dirichlet Allocation (LDA), particularly in terms of its high modularity, which allows for extensive personalization at each stage of the topic modeling process. In this study, we hypothesize that BERTopic, when optimized for Twitter data, can provide a more coherent and stable topic modeling. We began by conducting a review of the literature on topic-mining approaches for short-text data. Using this knowledge, we explored the potential for optimizing BERTopic and analyzed its effectiveness. Our focus was on Twitter data spanning the two years of the 117th US Congress. We evaluated BERTopic's performance using coherence, perplexity, diversity, and stability scores, finding significant improvements over traditional methods and the default parameters for this tool. We discovered that improvements are possible in BERTopic's coherence and stability. We also identified the major topics of this Congress, which include abortion, student debt, and Judge Ketanji Brown Jackson. Additionally, we describe a simple application we developed for a better visualization of Congress topics.

Keywords: Topic Mining; BERTopic; 117th Congress; Twitter; short-text data

Citation: Mendonça, M.; Figueira, Á. Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse. *Informatics* **2024**, *11*, 8. <https://doi.org/10.3390/informatics11010008>

Received: 16 November 2023

Revised: 1 February 2024

Accepted: 5 February 2024

Published: 17 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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 numerous—for example, driving civic engagement, increasing both early registration and voting, and mobilizing grassroots movements [1]. Social media also allowed politicians to campaign and communicate directly, informally, and at a lower cost, creating an opportunity for dialog between elected officials and those they represent [2]. The downsides of social media in general have also been well documented. Social media has proven to be a highly efficient platform for the spread of disinformation and misinformation, while simultaneously amplifying confirmation bias and exacerbating audience polarization [3].

Among the social media platforms available today, Twitter (or X, as of July 2023) is the leading social media platform in terms of adoption by politicians [4]. Created in 2006, as of December 2022 it boasts 368 million monthly active users worldwide [5]. Its current constraint of 280 characters per tweet fosters the 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 [6], and therefore an increasing source of user-generated textual data.

Since 2016, there have been rapid advancements in natural language processing (NLP), which can be attributed to the advent of deep learning and the availability of large-scale data [7]. One such development was the Transformer architecture [8]. 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 performance and capabilities. They include large-language models (LLMs), such as GPT-3 (Generative, Pre-trained Transformer, third iteration), which is an autoregressive model with a large number of parameters (about 175 billion), and the Bidirectional Encoder Representations from Transformers (BERT), which uses bidirectional context to understand the meaning of words in a sentence (it has between 110 and 340 million parameters). These models are applicable to a wide range of functions. Given the context of Twitter, our focus is on topic modeling. 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 [9].

This work was also motivated by the interest in obtaining a global perspective on Congress topics. Topic Mining is often performed either on an individual, focusing on a specific person [10], or on a global level, where a theme is under analysis [11,12]. This work positions itself on an intermediate level, where the subject is the whole of Congress—specifically, the 117th. This provides a comprehensive frame of reference on what were the concerns at this point in U.S. politics.

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 that employ 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 are more prevalent in terms of the 117th US Congress?

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

2. Background

Given the context of this work, this section aims to ensure that the reader has the essential knowledge with regard to 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, comprising Congress, which is responsible for law-making; executive, which implements laws and policies and is led by the U.S. President; and judicial, composed of the Supreme Court and smaller federal courts, which ensure that laws are abided by.

The U.S. political landscape contains over 400 different parties, but it is dominated by only two nationally recognized ones. The platform of the Democratic Party is that of

social liberalism [13] 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” [14]. The Republican Party, also known as GOP (“Grand Old Party”), abides by a conservative ideology and stands “for freedom, prosperity, and opportunity [...]. The principles of the Republican Party recognize God-given liberties while promoting opportunity for every American” [15].

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 [16]. Article I of the Constitution creates a Bicameral Congress consisting of a Senate and a House of Representatives. 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 with regard to how state representation should be defined. 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 [17].

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, determines the outcome of the 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 [18].

2.1.2. The 117th United States Congress

The 117th United States Congress was convened on 3 January 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 20 January, 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 117th Congress mandate include the January 6th 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 [19].

2.2. *American Politics and Twitter*

As of 2022, members of Congress have a total of 515 Twitter accounts [20], but not all accounts are used in the same way. 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 [21].

Some members have taken to Twitter better than others. Alexandra Ocasio-Cortez (D (Democrat)) is the most-followed House Representative, counting 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 (Republican)), Bernie Sanders (I (Independent)), and Mitt Romney (R) [22].

3. Review of Short-Text Topic Mining Techniques

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 [23]. 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 [24], which presents its own processing challenges. Textual data are unstructured, in the sense that these do not have an underlying schema. They are more ambiguous—making use of synonyms, slang, and abbreviations—and often require an understanding of the context. Spelling errors and lack of grammar further complicate their analysis. When studying textual data extracted from social media, their informal and short-text nature deepens these challenges.

Natural language processing (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 [25]. NLP has a diverse array of applications, each one contributing to how we interact with and process language data. Document summarization, for instance, involves distilling key information from extensive documents, and providing concise overviews that encapsulate the essence of the text. Information retrieval plays an important role in organizing and categorizing documents based on their central themes, an essential function for the efficiency of search engines and recommendation systems. Additionally, sentiment analysis delves into the emotional undertones within a text, offering insights into the sentiments and perceptions surrounding the discussed topics, which is invaluable in fields ranging from market analysis to social media monitoring. Each of these applications demonstrates the versatile and impactful nature of NLP in extracting, processing, and interpreting language data.

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, and locations, among others. Another NLP technique is Topic Mining (also known as Topic Extraction and Topic Discovery). 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 section. In terms of literature, we build upon the review of Murshed et al. [26]. We focus on publications from 2022 and 2023 that look into Twitter Topic Mining. In the cases where this was too restrictive, we looked into publications on other social media platforms.

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 refer 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 chronologically. Figure 1 offers a different perspective, with models organized according to the techniques used: advanced techniques encompass ML models, while traditional ones use classical, statistical approaches. This taxonomy is a simplified adaption of Murshed et al. [26].

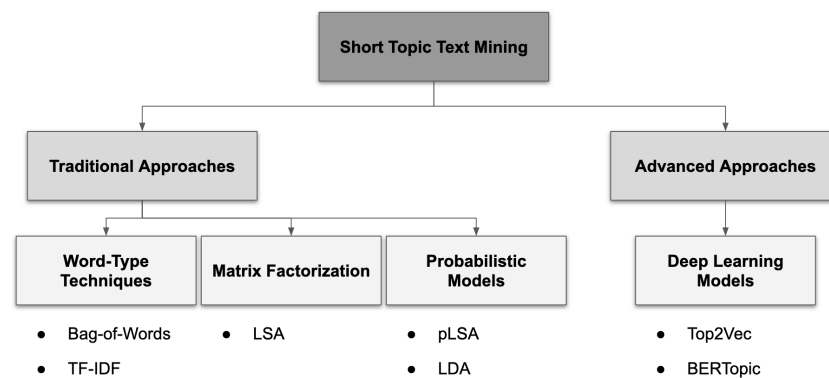


Figure 1. Taxonomy of Topic Mining models.

3.1. 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 and are instead suggested by a collection of words. Topic Modeling initially emerged as a method for representing text. One of the earliest approaches in this regard was the bag-of-words (BoW) model, where words are treated as features within a document. In the BoW model, the value assigned to each feature (term/word) can vary—it might be binary, indicating the presence or absence of a word, or numerical, reflecting the word's frequency [27]. Another significant approach is the term frequency-inverse document frequency (TF-IDF) method. Unlike BoW, which might be primarily used for word frequency, TF-IDF assigns weight to words based on their frequency in a particular document and inversely to their prevalence across the entire document corpus, favoring words that are frequent in a single document but infrequent in the overall dataset [28]. This distinction allows TF-IDF to highlight words that are uniquely significant to individual documents.

Topic Modeling, initially based on clustering techniques like Hierarchical Clustering [29] and K-Means clustering [30], evolved to incorporate more sophisticated models for representing and interpreting textual data. A notable advancement in this field was the Vector Space Model (VSM), where documents are represented as vectors not just of words, but of terms more broadly. This approach, echoing J.R. Firth's concept that "You shall know a word by the company it keeps", enables a nuanced understanding of language by capturing the context in which terms appear. VSM addresses the limitations of earlier models in handling polysemy and synonymy by allowing for different forms of term representation, including semantic forms like WordNet synsets or word senses. However, VSM itself does not inherently reduce dimensionality; this is achieved through additional methods such as principal component analysis or latent semantic analysis. Despite its advancements, VSM's treatment of words and terms still faced challenges in fully capturing the complexity of language, a subject further explored in the development of topic modeling techniques [31].

In the VSM, 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 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 . The Latent Semantic Analysis (LSA) model by Deerwester et al. [32] 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. This way, LSA reduces the dimensionality of the vector space while capturing the patterns in the data.

In Singular Value Decomposition (SVD) applied to topic modeling, we consider a term-document matrix X of size $m \times n$, where m is the number of terms and n denotes the

number of documents. The objective is to decompose X into three matrices: the term-topic matrix U of size $m \times r$, the diagonal matrix of singular values Σ , and the document-topic matrix V of size $r \times n$, where r is the number of selected topics or latent concepts. In this decomposition, U represents the relationship between terms and topics, with its elements indicating the strength of this relationship. The matrix V (or its transpose V^T), on the other hand, illustrates the association between documents and topics, with the rows of V^T specifying the strength of these associations. The matrix Σ contains the singular values, which correspond to how much each of the r latent topics explains the variability in the data. The decomposition can be formally represented as:

$$X = U\Sigma V^T \quad (1)$$

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. [33] used LSA to analyze tweets on the 2016 U.S. election, showing topics in parallel with the most frequent policy-related Internet searches at the time. More currently, Sai et al. [34] used LSA to guide the identification of fake news on Twitter, while Chang et al. [35] aimed to gain insights into public perception of the Russia–Ukraine conflict on Twitter through LSA. While works have shown that alternative models outperform LSA [36], others have tried to overcome LSA's limitations: Karami et al. [37] proposed Fuzzy LSA Topic Mining in health news tweets; Kim et al. [38] presented Word2Vec-based LSA for blockchain trend analysis.

In 1999, Hofmann introduced probabilistic LSA (pLSA) as an alternative to LSA under a probabilistic framework [39]. It assumes that topics are distributions of words and it aims to find a model of the latent topics that can generate the data in the document-term matrix. Formally, this distribution is defined as:

$$P(d, w_{di}) = P(d) \sum_{z=1}^K P(w_{di}|z) \cdot P(z|d) \quad (2)$$

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

Despite being a foundational method, the usage of pLSA probabilistic latent semantic analysis (pLSA) in NLP tasks has become infrequent, as it has been superseded by more advanced models. Kumar and Vardhan [40] used pLSA for the topic-based sentiment analysis of tweets, and Shen and Guo [41] used it to add context to a translation task. Though pLSA represents an advancement over the traditional LSA, it is notably susceptible to overfitting. Consequently, newer models have been developed to address this limitation, seeking to balance model complexity with generalization to broader datasets.

First proposed by Blei et al. [42] in 2003, Latent Dirichlet Allocation (LDA) is a widely recognized topic modeling framework. Unlike traditional text classification algorithms that categorize documents into predefined classes, LDA identifies a range of topics from the data themselves. It then assigns a distribution of these topics to each document, reflecting the varying degrees to which different topics are represented in the text. 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:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\phi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi_{Z_{j,t}}) \quad (3)$$

Figure 2 (from Anastasiu et al. [43]) illustrates the LDA algorithm and how each parameter interacts with all others.

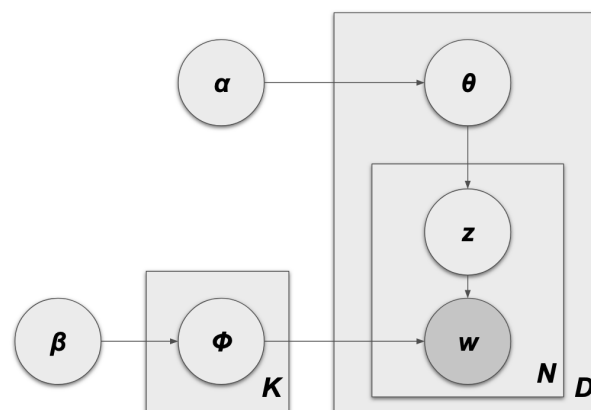


Figure 2. LDA algorithm.

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, $\text{Dir}(\alpha)$, 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 that the probability of the documents is higher near the individual topics/vertices than their combinations. $\text{Dir}(\beta)$ 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 closer to the individual words than to their combinations.

From the Dirichlet distributions, a new document, D , can be generated as set in Equation (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 Equation (3).

To account for documents with different lengths, a Poisson distribution is attached to the formulation in Equation (3). Gibbs sampling was later suggested as an efficient method for estimating θ and ϕ [44].

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 [45–47], health [48–50], and specifically COVID-19 [51–53], as well as other subjects such as climate change [54], layoffs [54], and hate speech [55].

As the limitations of LDA seem to impact performance in STTM, research has shifted to modifying the original LDA model. Table 1 lists some LDA variations, with a simple description of each one.

Table 1. LDA variations 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 above-listed models, Twitter-LDA has achieved particularly interesting results in performing STTM [56]. 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.

3.2. Topic Classification

The evolution of machine learning (ML) and neural networks has significantly advanced Topic Mining, particularly in the realm of Topic Classification. This advancement enables the precise automatic labeling of documents with predefined topics. Unlike Topic Modeling, which identifies latent topics within a corpus without prior definition, Topic Classification focuses on categorizing documents into specific, predetermined categories based on the topics they contain. This process leverages the semantic understanding facilitated by techniques such as word embeddings to accurately match documents with relevant topic labels. The development of sophisticated models and approaches in ML has improved the accuracy and efficiency of Topic Classification, making it possible to effectively handle complex and large-scale datasets.

A key concept applicable 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 the 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 numerically, they do so differently. In document-term matrices, each word from the corpus is represented and organized by its occurrence in individual documents. This results in a matrix that is typically large and sparse, reflecting the presence or absence of words across different documents. However, due to this format of representation, other aspects like the context and the nuanced meaning of words in their natural linguistic environment are not captured in the matrix. Additionally, vectorization is corpus-dependent, which complicates training in different datasets.

Word2Vec by Mikolov et al. [57] 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 [58]: It expanded on the ideas of Word2Vec: instead of words, the model is able to learn the 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 consist 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 the 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 [59]. It does so in five steps:

1. Creating embeddings: generates embedded document and word vectors, using Word2Vec and Doc2Vec.
2. Reducing dimensionality: reduces the number of dimensions in the documents using Uniform Manifold Approximation and Projection (UMAP). UMAP is a dimensionality reduction algorithm useful for complex datasets. It also has the advantage of clustering similar data points close to each other, 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 are user-defined). It then projects them into a low-dimensional graph and adjusts the distances between the data points according to their cluster [60]. 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. Segmenting 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, a 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 can 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 [61].
4. Computing centroids: For each dense area within the document vectors, the centroid is calculated, which serves as the topic vector.
5. Finding topic representations: the topic representations consist of the nearest word vectors to the topic vector.

In the context of short-text data, such as social media posts from patients with rare diseases, Top2Vec has been utilized for analysis. According to Karas et al. (2022), it has been

demonstrated that Top2Vec outperforms LDA in this application [62]. Zengul et al. [63] showed that Top2Vec results are more closely correlated to LDA's than to LSA's. Vianna and Silva De Moura [64] achieved successful results in extracting topics from the abstracts of legal cases—which are much shorter than the complete document, and Crijns et al. [65] managed to scrape web texts that identify innovative economic sectors or topics. Bretsko et al. [66] applied Top2Vec to the abstracts of academic papers.

Although Top2Vec was only introduced in 2020, the approach is now considered dated. It has been outdone by Transformers as the state-of-the-art models in NLP [67].

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 efficiently handle sequential data 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 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 the 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 the different aspects of the relationships.

Transformers have become the foundation for many state-of-the-art NLP models, such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), Text-to-Text Transfer Transformer (T5), 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 (Figure 3, from [68]). 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 [69]. The dimensionality of these embeddings tends to be quite high, with some models achieving ten thousand dimensions [70]. UMAP is used to reduce this to 2D or 3D since it preserves the local and global features of high-dimensional data better than alternatives such as PCA or t-SNE [60]. 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 relative to other Topic Modeling techniques [61]. The next step is to perform c-TF-IDF. This is a variation of the classical TF-IDF: firstly, it generates a bag-of-words at the cluster level, concatenating all documents in the same class. From this, TF-IDF is applied to each cluster bag-of-words, resulting in a measure for each cluster, instead of a corpus-wide one.

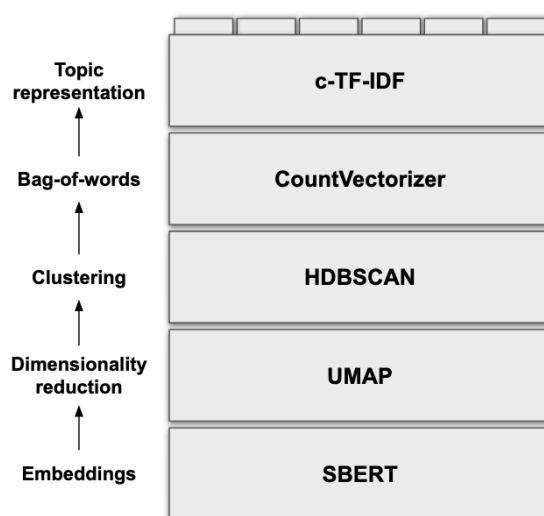


Figure 3. BERTopic algorithm.

BERTopic has been increasingly adopted in STTM. Hägglund et al. [71] aimed to identify the topics discussed on Twitter by autistic users; Li [72] applied it to examine how tweets can predict tourist arrivals to a given destination; Strydom and Grobler [73] analyzed COVID-19 misinformation on Twitter; Turner et al. [74] searched for concept drift on historical cannabis tweets; and Koonchanok et al. [75] attempted to track public attitudes towards ChatGPT on the same platform. Grigore and Pintilie [76] trained BERTopic to help identify potential patients with eating disorders based on their social media. Finally, with regard to hate speech on social media, Mekacher et al. [77] analyzed 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. [78] presented a pipeline for detecting hate speech and its targets on Parler.

While BERTopic is 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.

3.3. Other Related Work

In our study, we focus on the application of BERTopic for Twitter data analysis, which aligns with current research trends in topic modeling. Notably, [79] offers a comprehensive comparison of various topic modeling techniques, including BERTopic, specifically within the Twitter environment. This comparison is instrumental in contextualizing our choice of BERTopic, highlighting its effectiveness over traditional methods like LDA for short, unstructured social media texts. However, unlike our research, this work is much centered on a qualitative evaluation of the outcomes of the methods and BERTopic is essentially compared to Top2Vec.

Another work [80], employed BERTopic to analyze Twitter data, demonstrating its practical application in extracting coherent themes from social media conversations. While this study reinforces the utility of BERTopic in topic modeling, it differs from ours in its specific focus on the marketing implications of ChatGPT as opposed to our political discourse analysis.

On the other hand, [81] explored the challenges of topic modeling in crisis-related Twitter data using a different tool, “PASTA”. This study’s focus on ‘crisis situations’ presents a contrast to our focus on legislative discussion, highlighting the varying complexities and dynamics of Twitter data in different contexts. By comparing these works, we can better understand the landscape of topic modeling in social media analysis, validating

our approach and identifying areas where BERTopic particularly excels or requires further development for optimized application in diverse contexts.

4. Materials and Methods

Given the novelty and performance of BERTopic, and the research questions this work aims to answer, this section offers a description of the methods used to select and extract data, and the methodology followed to preprocess it. It also elaborates on BERTopic capabilities and offers a discussion on the techniques used for model evaluation.

4.1. Data Selection and Extraction

This work focuses on the Twitter activity of the members of the 117th U.S. Congress. This encompasses 102 Senators and 413 Representatives Twitter handles, totaling 515 accounts [20]. To reduce the number of accounts in the analysis, a first look was taken into how active they were. Tweet Congress [82] is a grassroots project that informed users about Congress's 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 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 the personal and professional accounts of a given selected name were included.

Table 2 lists the final selection. It consists of 27 members with 40 accounts, comprising 13 Democrats (D); 13 Republicans (R); and 1 Independent (I). These constitute 11 Representatives and 14 Senators, with the remaining 2 corresponding to the U.S. President and Vice-President.

Tweets from the above-mentioned accounts were extracted using the Twitter API (v.1.1). For each account, the tweets corresponding to the duration of the 117th Congress—from January 2021 to December 2022—were obtained, amounting to 27,782. The retrieved data included the tweet text, its publishing date and time, and the account handle.

Table 2. 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

Table 2. Cont.

Member Name	Twitter Handle(s)	Chamber	Party
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

Figure 4 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 with it.

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

- Replacing HTML character entities with their corresponding symbols;
- Removing hashtags, hyperlinks, and mentions, creating new tweet features with each one;
- Removing 'RT' from tweets that added no further information;
- Removing punctuation;
- Removing tweets of length 0 (not having associated text), which could result from the previous steps.



Figure 4. Preprocessing steps.

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 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 have small windows of context, and further removal would be ill-advised. Since removing stopwords can reduce tweets to a length of 0, those were once more dropped.

In handling tweets, which are often brief and filled with non-informative content, our primary objective was to distill meaningful information. The removal of stopwords enabled us to focus on significant words, thereby enhancing the relevance of our analysis. This approach was particularly effective in reducing data dimensionality, and it also improved the signal-to-noise ratio and preserved the coherency of the study by using the exact same dataset in the analyses.

4.3. Topic Extraction with BERTopic

BERTopic, known for its high-performance capabilities in Topic Mining, was selected for this study due to its recent development and state-of-the-art performance, as evidenced in recent research [79,83].

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 3 shows the options considered for each step of BERTopic. While it would be interesting to test all possible combinations, this 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.

Table 3. BERTopic personalization options considered.

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

4.3.1. Embeddings

BERTopic's default algorithm for embedding is SBERT, which is transformer-based and has been shown to outperform other methods [69]. 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. Additionally, SBERT can capture the essence of text without stopwords.

SpaCy offers the option of using 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 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 the training of vector embeddings" [84] due to its efficient implementation [85].

4.3.2. Dimensionality Reduction

UMAP is the default dimensionality reduction technique used by BERTopic. Principal component analysis (PCA) can also be selected, and it is often found to be faster than UMAP. PCA [86] is a linear dimensionality reduction technique that assumes a 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 [87]. 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 [88] 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 randomly chosen 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 [89]. K-means is particularly useful when the number of clusters is known a priori, while HDBSCAN is more versatile and more adequate for 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. 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 its 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—reducing 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*:

$$W_{x,c} = \|tf_{x,c}\| \cdot \log\left(1 + \frac{A}{f_x}\right) \quad (4)$$

where $tf_{x,c}$ is the frequency of word *x* in class *c*, f_c is the frequency of word *x* across all classes, and *A* is the average number of words per class. Two parameters can be tuned in this definition of c-TF-IDF. *bm25_weighting* is a Boolean parameter that indicates what is the weighing scheme in use. The bm25 weighting scheme, defined as $\log\left(1 + \frac{A - f_x + 0.5}{f_x + 0.5}\right)$, is particularly effective in managing stopwords in smaller datasets. Its robustness in these scenarios stems from its formulation, which balances term frequency and document frequency, thereby reducing the impact of common but less informative words.

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:

$$W_{x,c} = \sqrt{\|tf_{x,c}\|} \cdot \log\left(1 + \frac{A - f_x + 0.5}{f_x + 0.5}\right) \quad (5)$$

4.4. Topic Evaluation with BERTopic

Selecting the optimal combination of BERTopic options for each stage requires well-defined evaluation metrics. However, evaluating topic modeling presents several challenges. One major challenge is the frequent absence of a ‘ground-truth’ model, which would serve as a definitive standard for comparison with new models. Additionally, assessing the performance of topic modeling can be somewhat subjective; the relevance of topic clusters can vary greatly depending on the specific objectives and context of the dataset analysis. Nevertheless, we aim to focus as much as possible on quantitative methods.

In line with the research of Abdelrazek et al. [90], 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 analyzing its predictive power, and it can therefore 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, these words must be 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 examines how different the topics are to each other. Low diversity is a result of redundant topics, implying difficulty in distinguishing them from what remains. A suggested technique for measuring diversity is counting the unique words present in the top words of all topics [91].
- Stability in topic modeling refers to whether the results obtained are consistent across runs. Since topic modeling applies concepts from statistical inference, some variation is expected 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 and across iterations. Jaccard similarity ranges from 0 to 1, where identical topics share a coefficient of 1.

5. Results and Discussion

The optimization of BERTopic was conducted with the assumption that each stage operates independently. While performance plots for each stage can be consulted in Appendix A, a summary of the optimized stages and their performances is found at the end of this section, in Figure 5 and Table 4.

Figure A1 illustrates how SBERT, SpaCy, and Word2Vec compare in terms of coherence, perplexity, diversity, and stability scores in the embedding stage. The 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. With regard to 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 (Figure A2), 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). However, 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 a 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 Figure A4, 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, Figure A5 shows how setting the minimum word frequency (*min_df*) to five 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 the analyzed values perform similarly (Figure A6). The value with the highest stability (0.674) is 17000; hence, it was chosen.

Finally, the Topic Representation stage is optimized by manipulating two Boolean parameters. *Bm25_weighting* (Figure A7) does not improve the model performance, as both values behave similarly across metrics. On the other hand, reducing frequent words (Figure A8) increases the 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).

Figure 5 illustrates how the optimized BERTopic performs against the default parameters, 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 4 summarizes the options selected at each stage.

Table 4. Final selection of Congress Twitter accounts.

Stage	Selected Algorithms / Parameters
Embeddings	Word2Vec
Dimensionality reduction	Base dimensionality model
Clustering	HDBSCAN
Vectorizer	<i>ngram_range</i> : (1, 1) <i>min_df</i> : 5 <i>max_features</i> : 170,000
Topic representation	<i>bm25_weighting</i> : False <i>reduce_frequent_words</i> : True

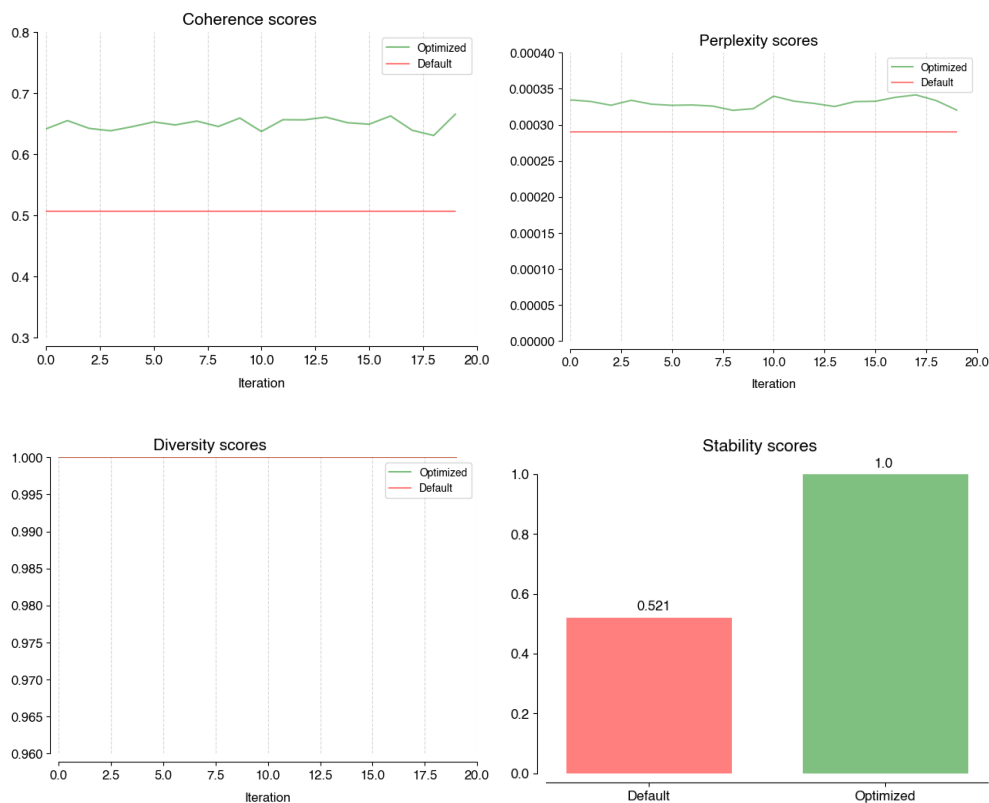


Figure 5. Performance of optimized BERTopic in coherence (top-left); perplexity (top-right); diversity (bottom-left); and stability (bottom-right) scores against the default parameters.

The optimized BERTopic model demonstrated superior performance over the default setup in terms of average coherence and stability, with similar diversity scores and only a marginal difference in average perplexity.

5.1. Topic Results

Applying the optimized BERTopic algorithm to the processed data generates 182 unique topics assigned to 4068 tweets. This means 23,714 tweets were considered as outlier topics. Figure 6 shows the frequency of each topic: topics 0–4 occur at least 100 times each, while topics 5–12 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 four tweets.

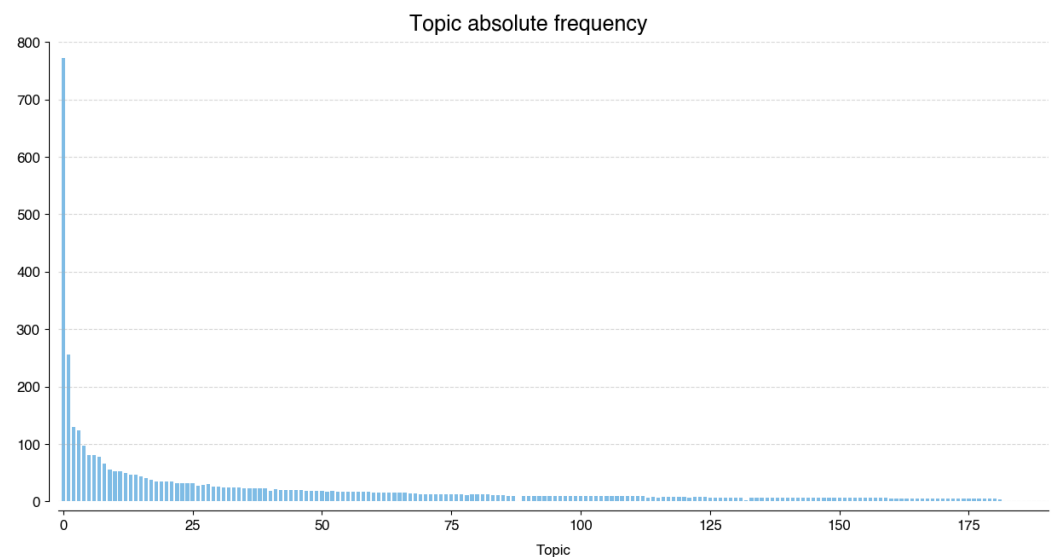


Figure 6. Absolute frequency of topics.

Limiting our analysis to the top 13 topics enables a more in-depth exploration. Table 5 presents the topic labels, which consist of the key terms identified by BERTopic for each topic, along with the corresponding names we have assigned to these topics for an easier understanding.

Table 5. Label and name of the top topics.

Topic	Label	Name
0	0_reduction_prescription_cap_abortion	<i>Abortion</i>
1	1_cancel_student_debt_borrowers	<i>StudentDebt</i>
2	2_ketanji_jackson_brown_judge	<i>JudgeBrown</i>
3	3_poll_early_location_register	<i>Voting</i>
4	4_putin_kyiv_ukraine_ukrainian	<i>UkraineWar</i>
5	5_marijuana_cannabis_legalize_possession	<i>Cannabis</i>
6	6_commitmenttoamerica_housegop_replamalfa_built	<i>#CommitmentToAmerica</i>
7	7_defendourdemocracy_victory_congressman_congratulations	<i>#DefendOurDemocracy</i>
8	8_firebrand_episode_matt_feat	<i>Firebrand</i>
9	9_sacrifice_veteransday_memorialday_memorial	<i>MemorialDay</i>
10	10_vaccine_19_vaccinate_booster	<i>CovidVaccine</i>
11	11_utah_wildfires_infrastructure_mitigation	<i>Utah</i>
12	12_birthday_247th_usmc_wishing	<i>Congratulations</i>

Figure 7 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 Figure A9. 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 their opinions on the developments of the Russia–Ukraine war, especially on the political figures' strategies and motivations:

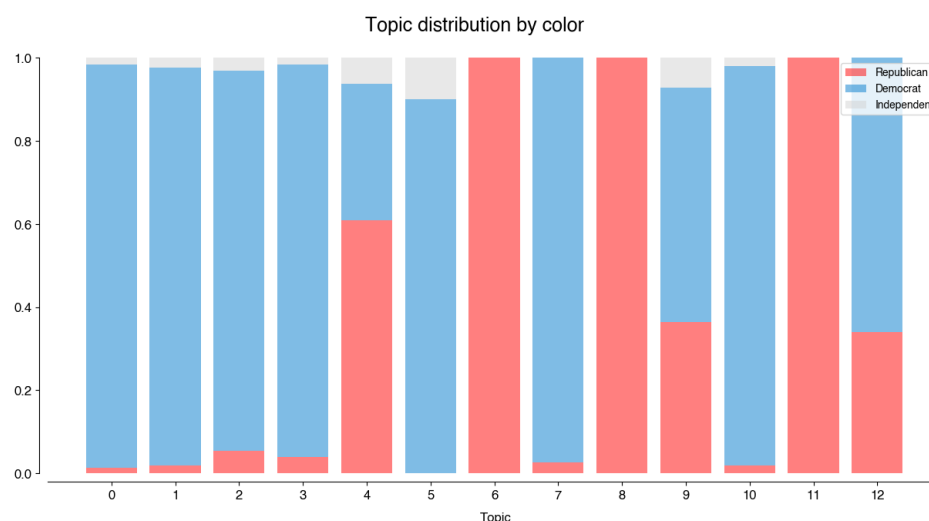


Figure 7. Distribution of the top topics by color.

- *DANGER #Putin's legitimacy built on its image as the strong leader who restored #Russia to superpower after the disasters of the '90s. Now the economy is in shambles & the military is being humiliated & their only tools to reestablish power balance with the West are cyber & nukes* (28 February 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, their conventional forces are stalled & cyber,chemical,biological & non-strategic nukes are their only escalation options* (25 March 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 or 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 May 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 November 2022).

MemorialDay is more balanced, with the main accounts (@VP, @RepAdamSchiff) tweeting four times each. It celebrates both Memorial Day (the 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 May 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 November 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 (Figure A10). @GOPLeader tweeted 79 times on the topic of *#CommitmentToAmerica*, from September

to November 2022. Commitment to America (<https://www.speaker.gov/commitment/>) consists of Republicans' guidelines of their vision of the U.S. politics:

- *An economy that is strong. A nation that is safe. A future that is built on freedom. A government that is accountable. This is the Republican Commitment to America.* (22 September 2022);
- *Republicans have made our #CommitmentToAmerica. Under the leadership of @GOPLeader, we will build: An Economy That is Strong, A Nation That is Safe, A Future That is Built on Freedom, A Government That is Accountable. Now let us get to work.* (17 November 2022).

@RepMattGaetz tweeted 67 times on their 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 @sheriff1amb1, and discusses rising gas costs, Russian propaganda, men competing in women's sports, and more!* (24 March 2022);
- *RT @Firebrand_Pod: Episode 76 LIVE: Ban TikTok (feat. @GavinWax)—Firebrand with @RepMattGaetz <https://t.co/KQdzDcSeHJ>* (18 November 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 have 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 July 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 November 2022).

The remaining seven topics, dominated by Democrats, are distributed as illustrated in Figure A11. On the *Abortion* topic, @PattyMurray (120), @RepJayapal (115), and @PramilaJayapal (87) represent 43% of tweets. As illustrated by the time series, interest in this topic began in May 2022, when a leaked draft suggested the Supreme Court's intention to overturn *Roe v. Wade*. The decision was taken on 24 June 2022, 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 (Figure A12). 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 occurred 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 7 April 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 6 October 2022, when President Biden announced their position on Marijuana Reform (<https://www.whitehouse.gov/briefing-room/statements-releases/2022/10/06/statement-from-president-biden-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 July 2022);
- *Democratic Victory: Congratulations to Congresswoman @MaryPeltola on your re-election in Alaska! -NP* (24 November 2022).

@VP contributed to topic *CovidVaccine* 19 times (36%), from April to November 2022. It mainly consists of tweets providing information about 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 April 2022);
- *Prepare for a healthy holiday season by getting your updated COVID vaccine. It's free, safe, and effective.* (13 November 2022).

5.2. Visualizing 117th Congress Topics

There are differences in the 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 the ‘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 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. Our aim was not to create a comprehensive visualization tool but rather to provide a straightforward application that allows users to easily interpret and interact with the data, thereby facilitating a more accessible validation and exploration of our research findings.

Figure 8 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 the topics she mentioned. Below the banner, on the left, the user is presented with information about Pramila: her top three topics, a photo (if available), her party affiliation, her Twitter account handle, the number of times she has tweeted about the selected topic, and her ranking among Congress members based on tweet frequency for that topic.

In 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.

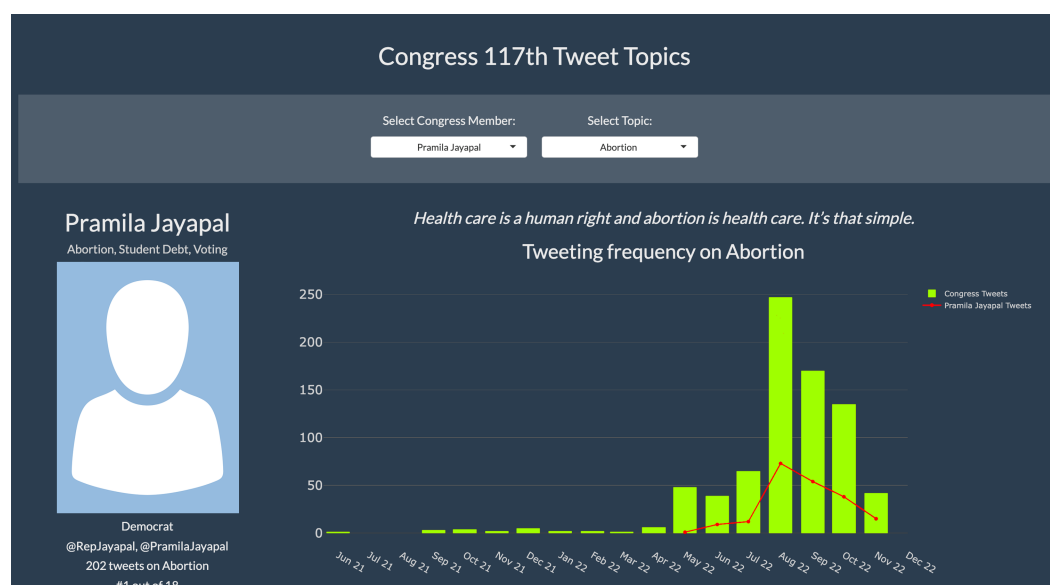


Figure 8. Use-case of app for member Pramila Jayapal and *Abortion*—part I.

Below the tweet, a time series is shown (Figure 9). The green (light) 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 names of the members. In this particular case, Pramila was the top tweeter on *Abortion*, followed by Patty Murray.

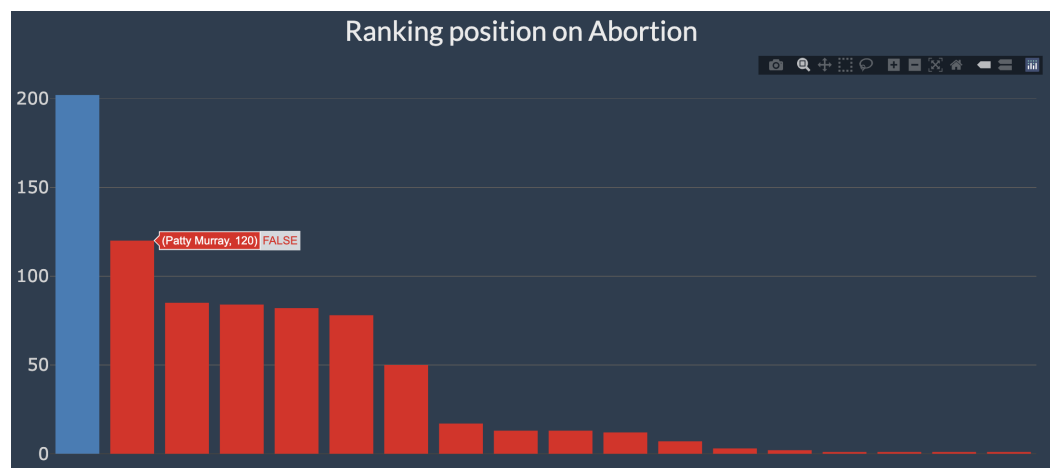


Figure 9. Use-case of app for member Pramila Jayapal and *Abortion*—part II.

6. Conclusions

It is now possible to answer the research questions raised at the outset of this work. Firstly, we show how BERTopic differs from traditional Topic Extraction techniques. Comparing it to traditional approaches such as LSA and LDA, BERTopic's use of the attention mechanism generates a higher performance, with the additional advantage of not requiring a pre-defined number of topics.

Secondly, BERTopic is optimized to handle Twitter data, achieving increases of 28% in coherence and 48% in stability scores, when compared with the default parameters of the algorithm.

Finally, the 117th Congress is shown to be defined by topics such as abortion, student debt, the election of Judge Ketanji Brown Jackson, and the Russia–Ukraine conflict. The contribution to the topics was not equal across parties, with only three of the top 13 having relevant participation from both Democrats and Republicans.

In terms of opportunities for expanding on this work, adding engagement metrics to the extracted data and incorporating that information into the analysis might provide richer insights into how topics behave over time. Future work could also expand on this quantitative analysis with qualitative evaluations of topic significance and impact.

Author Contributions: Conceptualization, M.M. and Á.F.; methodology, M.M. and Á.F.; software, M.M.; validation, Á.F.; formal analysis, M.M. and Á.F.; investigation, M.M. and Á.F.; resources, Á.F.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, Á.F.; visualization, M.M.; supervision, Á.F.; funding acquisition, Á.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on demand.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
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

Appendix A

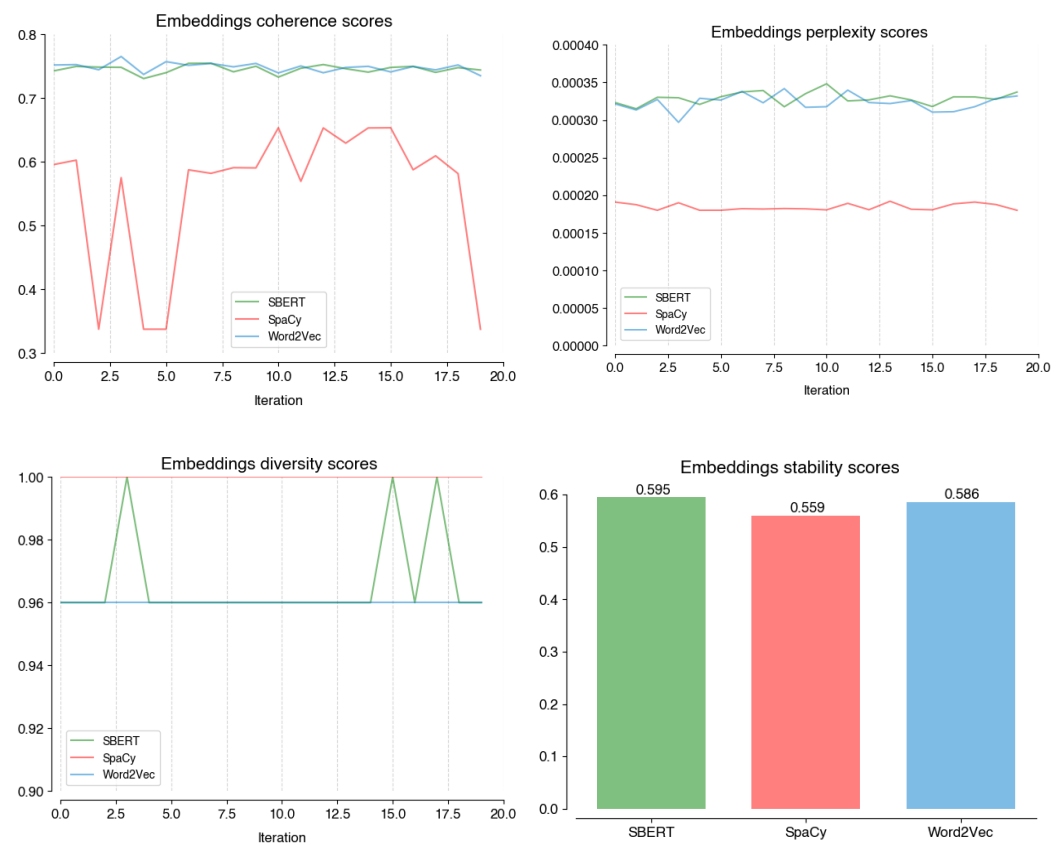


Figure A1. Performance scores on coherence; perplexity; diversity, and; stability of different models at the Embeddings stage.

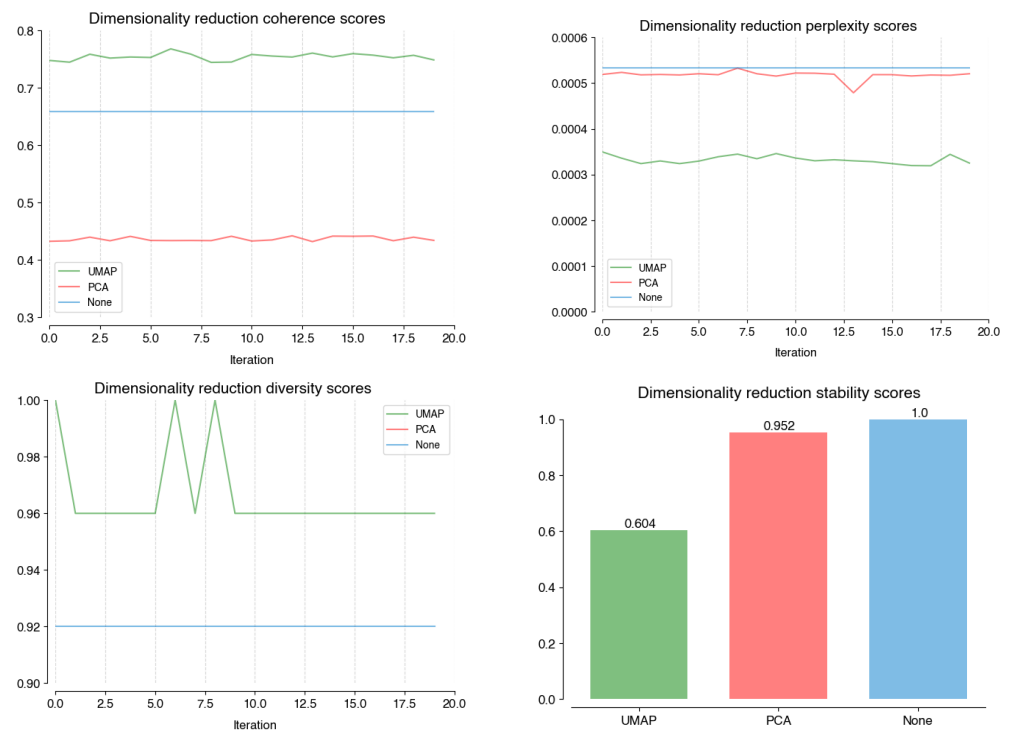


Figure A2. Performance scores on coherence; perplexity; diversity, and; stability of different models at the Dimensionality Reduction stage.

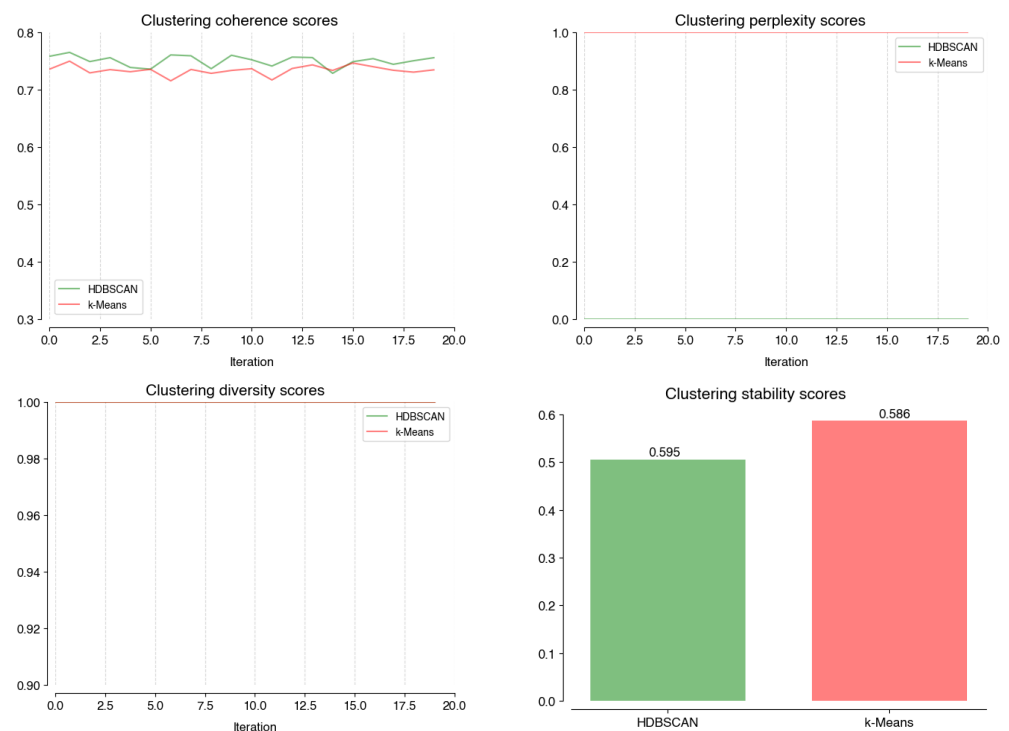


Figure A3. Performance scores on coherence; perplexity; diversity, and; stability of different models at the Clustering stage.

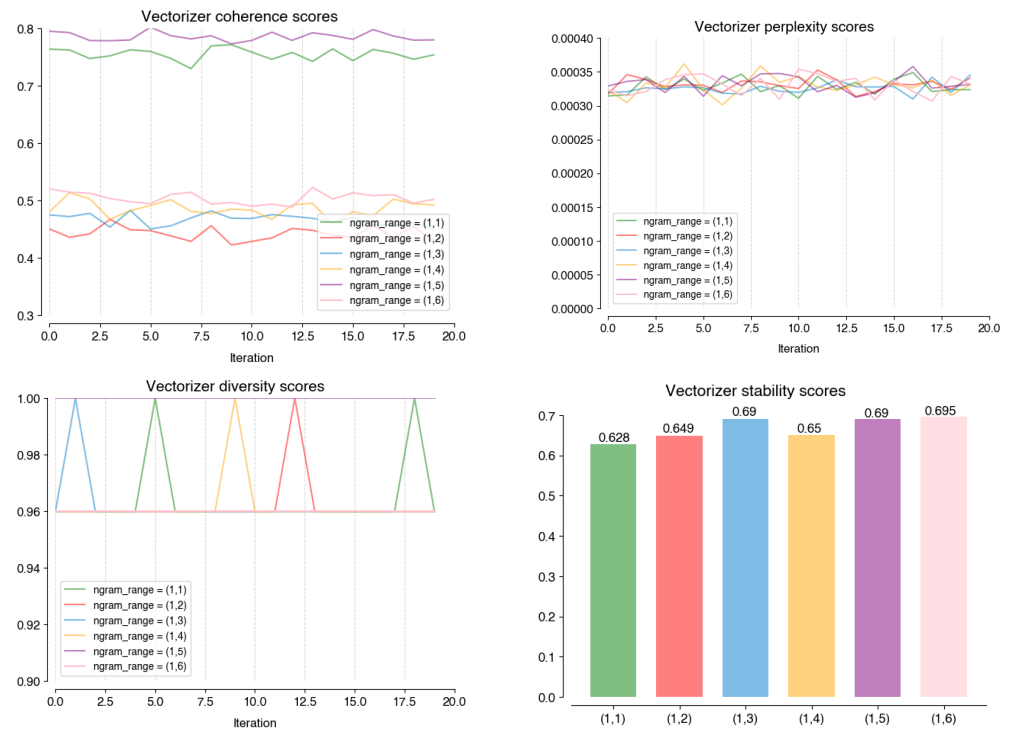


Figure A4. Performance scores on coherence; perplexity; diversity, and; stability of different values of *ngram_range* at the Vectorizer stage.



Figure A5. Performance scores on coherence; perplexity; diversity, and; stability of different values on the *min_df* at the Vectorizer stage.

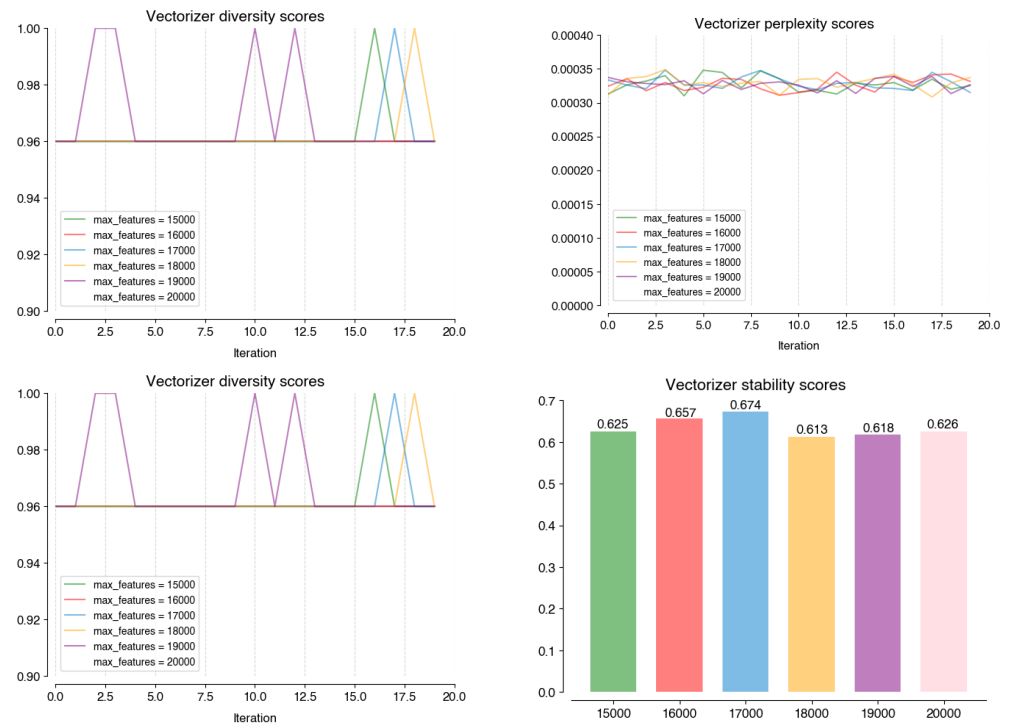


Figure A6. Performance scores on coherence; perplexity; diversity, and; stability of different values on the *max_features* at the Vectorizer stage.



Figure A7. Performance scores on coherence; perplexity; diversity, and; stability of different values on the *bm25_weighting* at the Topic Representation stage.



Figure A8. Performance scores on coherence; perplexity; and; stability of different values on the `reduce_frequent_words` at the Topic Representation stage.

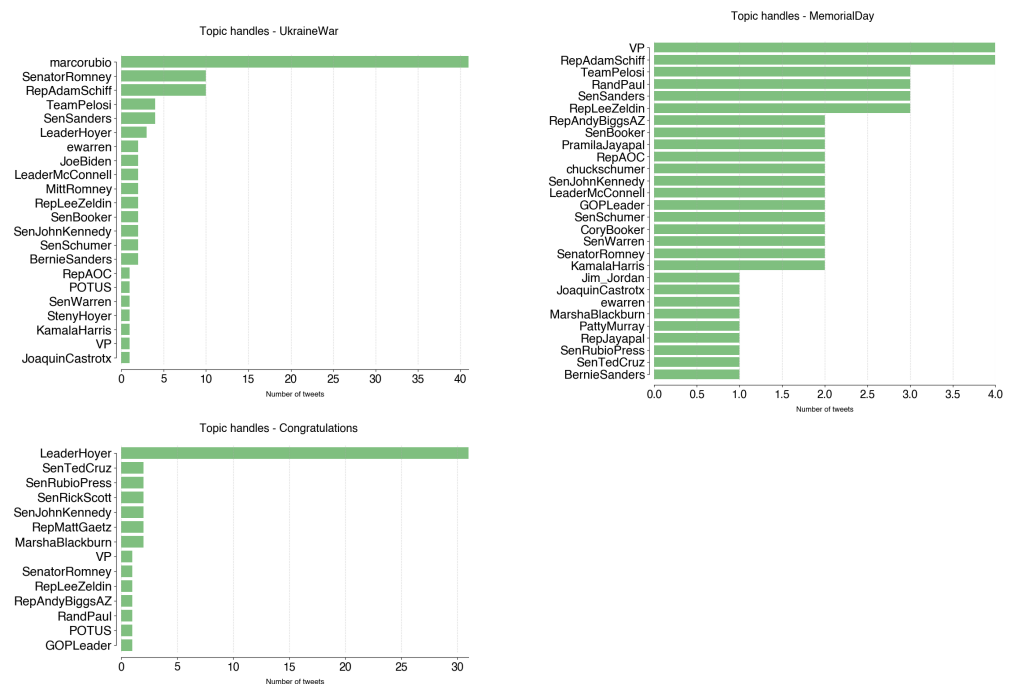


Figure A9. Topics: *Ukraine*; *MemorialDay*; and *Congratulations* handles.

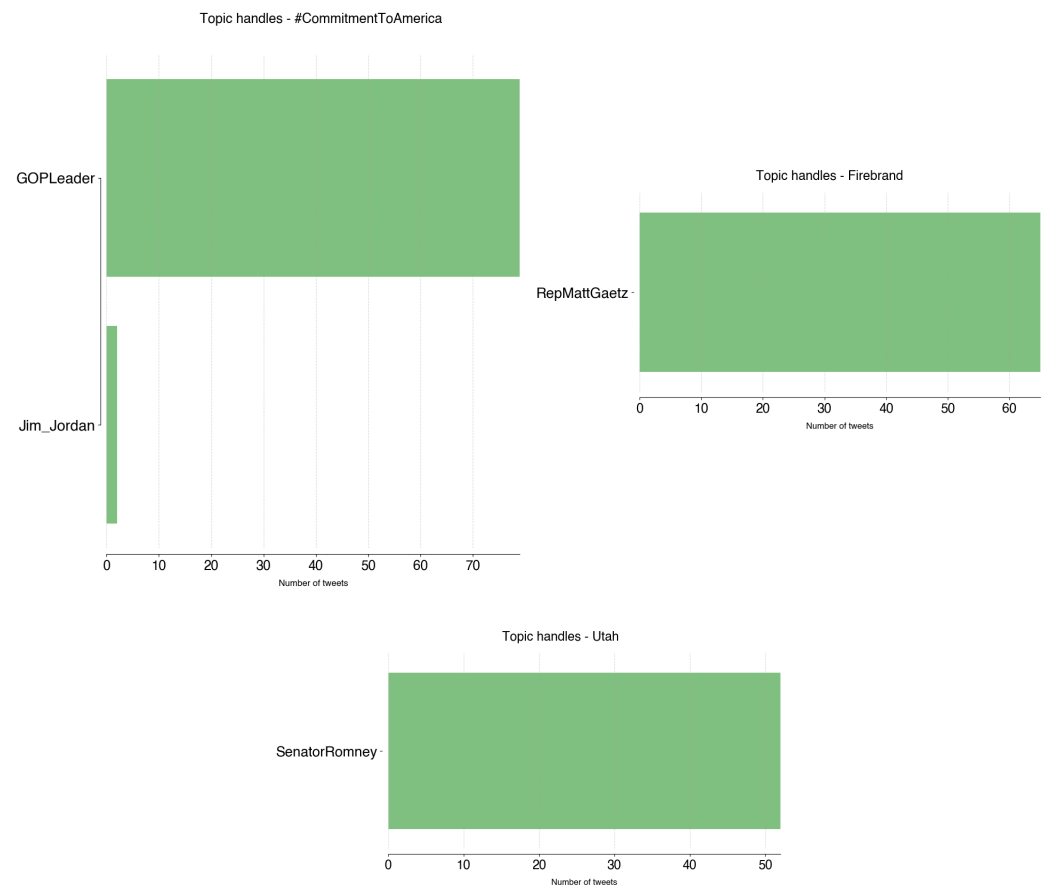


Figure A10. Topics: #CommitmentToAmerica; Firebrand; and Utah handles.

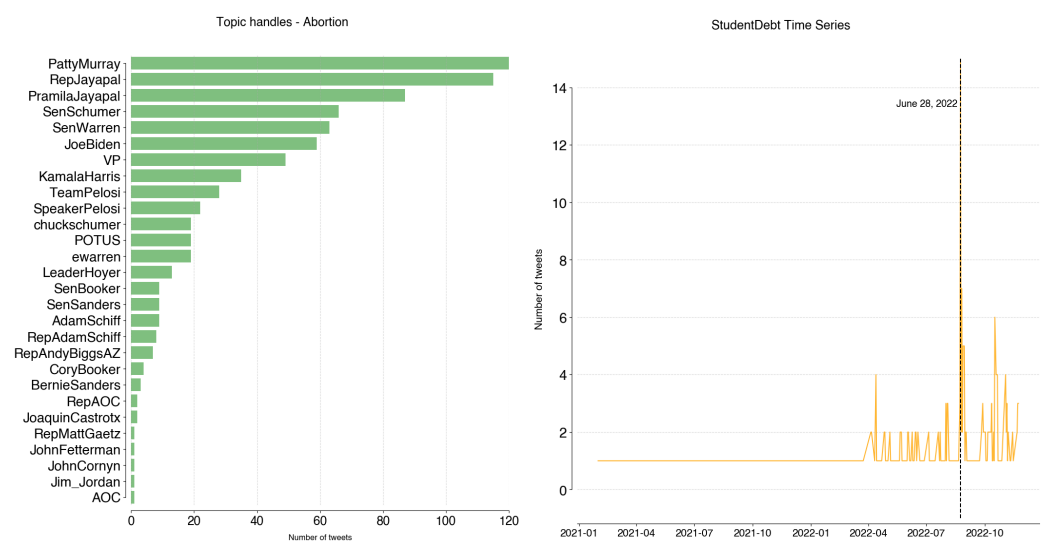


Figure A11. Abortion: handles and time series.

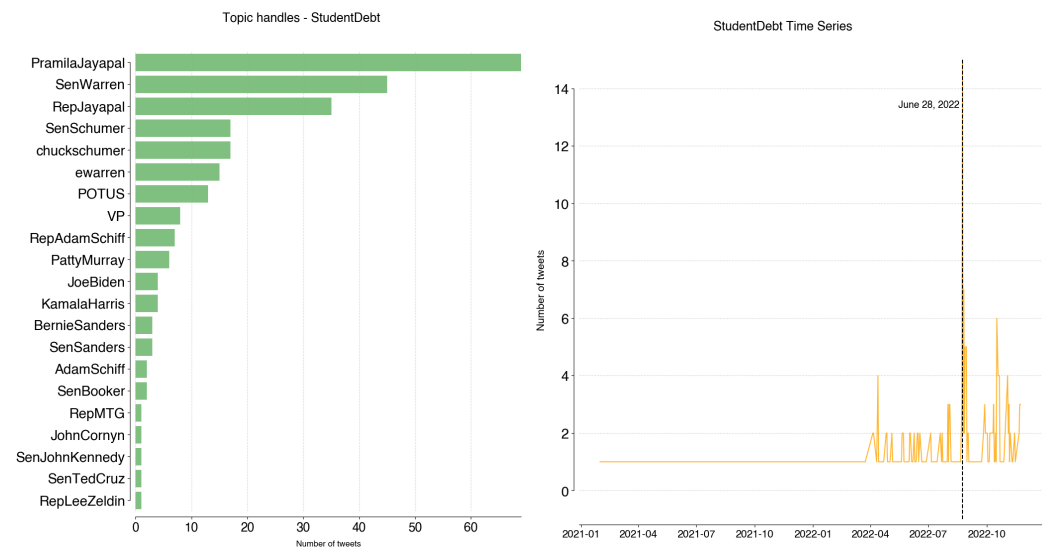


Figure A12. *StudentDebt*: handles and time series.

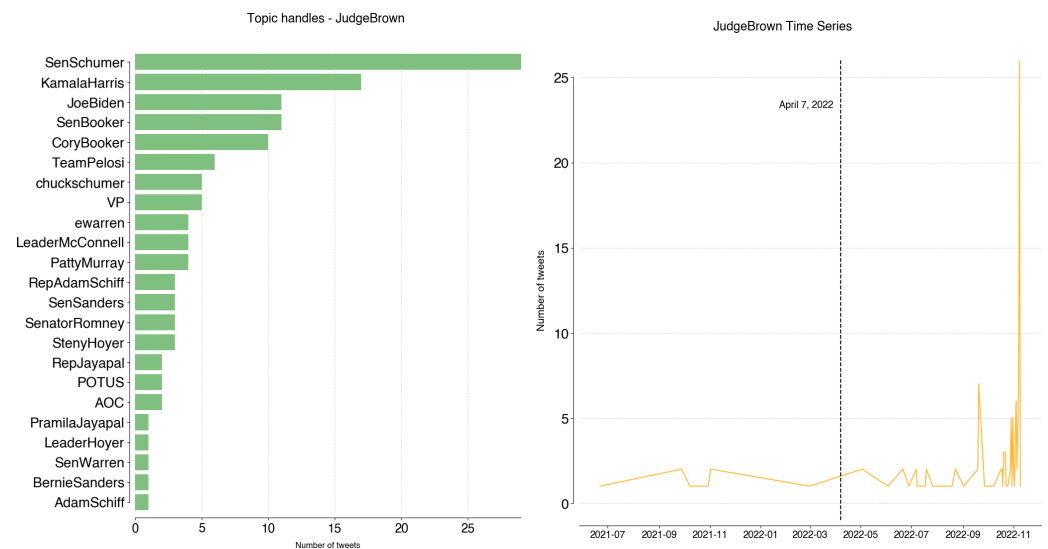


Figure A13. *JudgeBrown*: handles and time series.

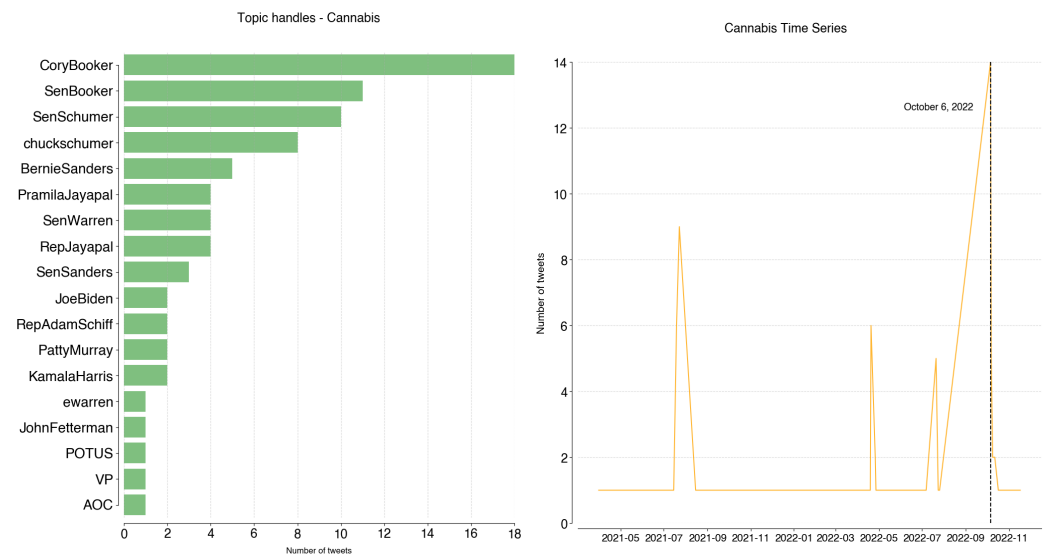


Figure A14. Cannabis: handles and time series.

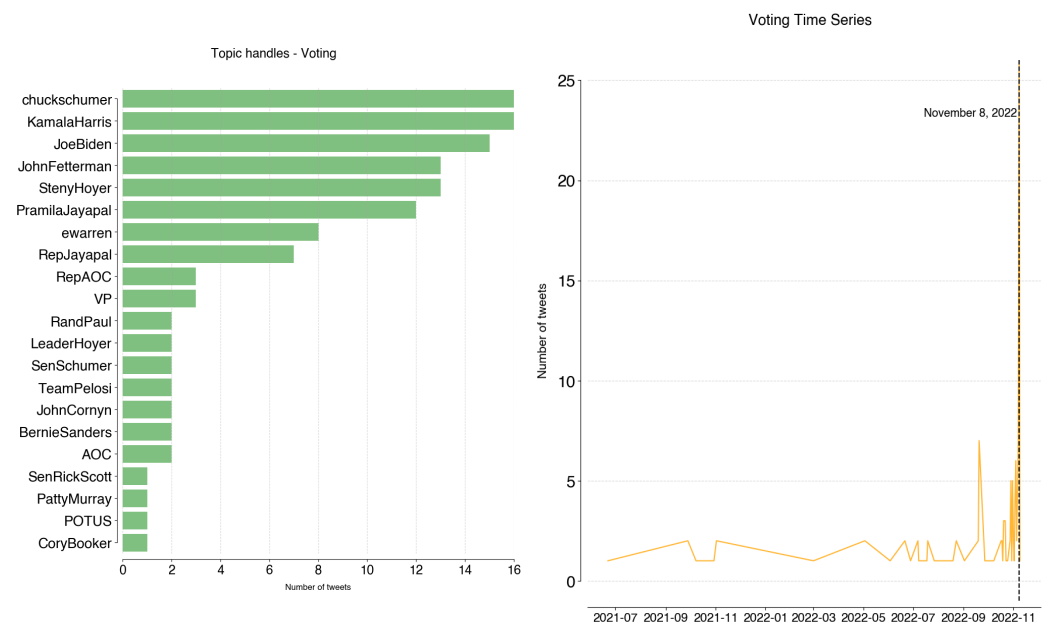


Figure A15. Voting: handles and time series.

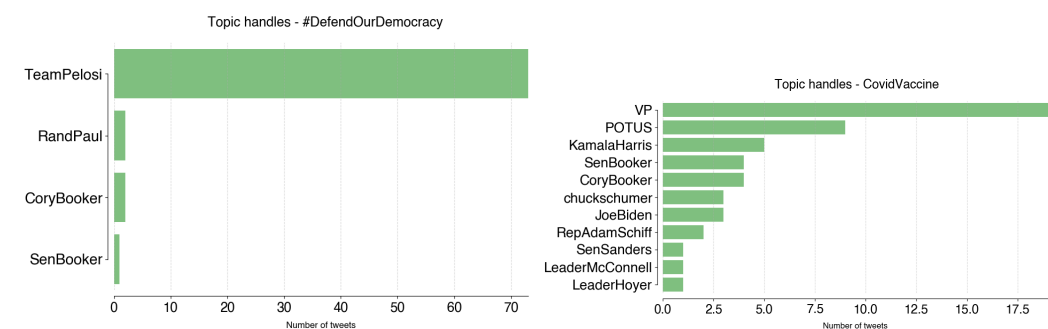


Figure A16. #DefendOurDemocracy and CovidVaccine handles.

References

1. Satterfield, H. How Social Media Affects Politics. 2020. Available online: <https://www.meltwater.com/en/blog/social-media-affects-politics> (accessed on 1 September 2023).
2. Bonney, V. How Social Media Is Shaping Our Political Future. 2018. Available online: <https://www.youtube.com/watch?v=9Kd99IIWJUw> (accessed on 3 August 2023).
3. Center for Humane Technology. How Social Media Polarizes Political Campaigns. 2021. Available online: <https://www.youtube.com/watch?v=1GRxORsQhY4> (accessed on 3 August 2023).
4. Statista. Social Media and Politics in the United States. 2023. Available online: <https://www.statista.com/topics/3723/social-media-and-politics-in-the-united-states/> (accessed on 26 September 2023).
5. Statista. X/Twitter: Number of Users Worldwide 2024. Available online: <https://www.statista.com/statistics/303681/twitter-users-worldwide/> (accessed on 26 September 2023).
6. Reveilhac, M.; Morselli, D. The Impact of Social Media Use for Elected Parliamentarians: Evidence from Politicians' Use of Twitter During the Last Two Swiss Legislatures. *Swiss Political Sci. Rev.* **2023**, *29*, 96–119. [CrossRef]
7. Anand, A. Timeline of Advances in the Field of NLP that Led to Development of Tools like ChatGPT. 2020. Available online: <https://dev.to/amananandrai/recent-advances-in-the-field-of-nlp-33o1> (accessed on 3 September 2023).
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6000–6010.
9. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
10. Hajjej, A. Trump Tweets: Topic Modeling Using Latent Dirichlet Allocation. 2020. Available online: <https://medium.datadriveninvestor.com/trump-tweets-topic-modeling-using-latent-dirichlet-allocation-e4f93b90b6fe> (accessed on 26 September 2023).
11. Abadah, M.S.K.; Keikhosrokiani, P.; Zhao, X. 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: Hershey, PA, USA, 2023; pp. 156–188. [CrossRef]
12. Zhou, S.; Kan, P.; Huang, Q.; Silbernagel, J. A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. *J. Inf. Sci.* **2023**, *49*, 465–479. [CrossRef]
13. Knoll, B. President Obama, the Democratic Party, and Socialism: A Political Science Perspective. Available online: https://www.huffpost.com/entry/obama-romney-economy_b_1615862 (accessed on 3 August 2023).
14. DemocraticParty. Where We Stand. Available online: <https://democrats.org/where-we-stand/> (accessed on 3 August 2023).
15. Republican National Committee. GOP—About Our Party. Available online: <https://gop.com/about-our-party/> (accessed on 3 August 2023).
16. U.S. Senate. Constitution of the United States. Available online: <https://www.senate.gov/about/origins-foundations/senate-and-constitution/constitution.htm> (accessed on 3 August 2023).
17. Benzine, C. The Bicameral Congress: Crash Course Government and Politics 2. 2015. Available online: <https://www.youtube.com/watch?v=n9defOwVWS8> (accessed on 3 August 2023).
18. Benzine, C. Congressional Elections: Crash Course Government and Politics 6. 2015. Available online: <https://www.youtube.com/watch?v=qxiD9AEX4Hc&list=PL8dPuuaLjXtOfse2ncvffeelTrqvhrz8H&index=6> (accessed on 3 August 2023).
19. Binder, S. Goodbye to the 117th Congress, Bookended by Remarkable Events. 2022. Available online: <https://www.washingtonpost.com/politics/2022/12/29/congress-year-review/> (accessed on 3 August 2023).
20. PressGallery. Members' Official Twitter Handles. Available online: <https://pressgallery.house.gov/> (accessed on 27 August 2023).
21. Lee, S.; Panetta, G. Twitter Is the Most Popular Social Media Platform for Members of Congress—However, Prominent Democrats Tweet More Often and Have Larger Followings than Republicans. 2019. Available online: <https://www.businessinsider.com/democratic-republican-congress-twitter-followings-political-support-2019-2> (accessed on 27 August 2023).
22. Mills, B.R. Take It to Twitter: Social Media Analysis of Members of Congress. 2021. Available online: <https://towardsdatascience.com/take-it-to-twitter-sentiment-analysis-of-congressional-twitter-in-r-ee206a5b05bc> (accessed on 27 August 2023).
23. Marr, B. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. 2018. Available online: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/> (accessed on 3 August 2023).
24. Ma, L.; Goharian, N.; Chowdhury, A.; Chung, M. Extracting Unstructured Data from Template Generated Web Documents. In Proceedings of the Twelfth International Conference on Information and Knowledge Management, New York, NY, USA, 3–8 November 2003; pp. 512–515. [CrossRef]
25. Defined.ai. The Challenge of Building Corpus for NLP Libraries. Available online: <https://www.defined.ai/blog/the-challenge-of-building-corpus-for-nlp-libraries/> (accessed on 3 August 2023).
26. Murshed, B.A.H.; Mallappa, S.; Abawajy, J.; Saif, M.A.N.; Al-ariqi, H.D.E.; Abdulwahab, H.M. Short text topic modelling approaches in the context of big data: Taxonomy, survey, and analysis. *Artif. Intell. Rev.* **2023**, *56*, 5133–5260. [CrossRef] [PubMed]
27. Harris, Z.S. Distributional Structure. *WORD* **1954**, *10*, 146–162. [CrossRef]
28. Jones, K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]

29. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [\[CrossRef\]](#)
30. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1986; pp. 281–297.
31. Xia, L.; Luo, D.; Zhang, C.; Wu, Z. A Survey of Topic Models in Text Classification. In *Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 25–28 May 2019; pp. 244–250. [\[CrossRef\]](#)
32. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [\[CrossRef\]](#)
33. Valdez, D.; Pickett, A.; Goodson, P. Topic Modeling: Latent Semantic Analysis for the Social Sciences. *Soc. Sci. Q.* **2018**, *99*. [\[CrossRef\]](#)
34. Sai, T.V.; Lohith, K.; Sai, M.; Tejaswi, K.; Ashok Kumar, P.; Karthikeyan, C. Text Analysis On Twitter Data Using LSA and LDA. In *Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 23–25 January 2023; pp. 1–6. [\[CrossRef\]](#)
35. Chang, P.; Yu, Y.T.; Sanders, A.; Munasinghe, T. Perceiving the Ukraine-Russia Conflict: Topic Modeling and Clustering on Twitter Data. In *Proceedings of the 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, Athens, Greece, 17–20 July 2023; pp. 147–148. [\[CrossRef\]](#)
36. Qomariyah, S.; Iriawan, N.; Fithriasari, K. Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. *Proc. AIP Conf.* **2019**, *2194*, 020093. [\[CrossRef\]](#)
37. Karami, A.; Gangopadhyay, A.; Zhou, B.; Kharrazi, H. Fuzzy Approach Topic Discovery in Health and Medical Corpora. *Int. J. Fuzzy Syst.* **2018**, *20*, 1334–1345. [\[CrossRef\]](#)
38. Kim, S.; Park, H.; Lee, J. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Syst. Appl.* **2020**, *152*, 113401. [\[CrossRef\]](#)
39. Hofmann, T. Probabilistic Latent Semantic Indexing. In *Proceedings of the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57. [\[CrossRef\]](#)
40. Kumar, P.; Vardhan, M. Aspect-Based Sentiment Analysis of Tweets Using Independent Component Analysis (ICA) and Probabilistic Latent Semantic Analysis (pLSA). In *Advances in Data and Information Sciences*; Springer: Singapore, 2019; pp. 3–13. [\[CrossRef\]](#)
41. Shen, Y.; Guo, H. Research on high-performance English translation based on topic model. *Digit. Commun. Netw.* **2023**, *9*, 505–511. [\[CrossRef\]](#)
42. Blei, D.; Ng, A.; Jordan, M.; Lafferty, J. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
43. Anastasiu, D.; Tagarelli, A.; Karypis, G. Document Clustering: The Next Frontier. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2013; pp. 305–338. [\[CrossRef\]](#)
44. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Mehrpour, F. Analyzing Twitter Sentiment and Hype on Real Estate Market: A Topic Modeling Approach. 2023. Available online: <https://dr.library.brocku.ca/handle/10464/17848> (accessed on 29 September 2023).
46. Fakhri, M.I.; Irawan, H. Analyzing Sentiment and Topic Modelling of iPhone Xs Post Launch Event through Twitter Data. *AIP Conf. Proc.* **2023**, *2646*, 040030. [\[CrossRef\]](#)
47. Strydom, I.F.; Grobler, J.; Vermeulen, E. Investigating the Use of Topic Modeling for Social Media Market Research: A South African Case Study. In *Proceedings of the 23rd International Conference*, Athens, Greece, 3–6 July 2023; pp. 305–320. [\[CrossRef\]](#)
48. Kaur, J.; Hussain, I.Z.; Lotto, M.; Butt, Z.; Morita, P. Preventing public health crises: An expert system using Big Data and AI in combating the spread of health misinformation. *Popul. Med.* **2023**, *5*, A631. [\[CrossRef\]](#)
49. Praveen, S.V.; Ittamalla, R.; Mahipalan, M.; Mahitha, M.; Priya, D.H. What Do Veterans Discuss the Most about Post-Combat Stress on Social Media?—A Text Analytics Study. *J. Loss Trauma* **2023**, *28*, 187–189. [\[CrossRef\]](#)
50. Lyu, A.; Liu, C.; Ding, Z.; Li, J.; Zhang, W. Analysis of gender sentiment expression in network based on TF-LDA algorithm. *Adv. Eng. Technol. Res.* **2023**, *5*, 322–322. [\[CrossRef\]](#)
51. Bheema, S.T.; Kotha, S.K. Insights from COVID-19 #Vaccine Twitter analytics. In *Proceedings of the 19th Annual Symposium on Graduate Research and Scholarly Projects*; Wichita State University: Wichita, KS, USA, 2023.
52. Comito, C. How Do We Talk and Feel About COVID-19? Sentiment Analysis of Twitter Topics. In *Proceedings of the 12th International Conference, Held as Part of the Services Conference Federation, SCF 2023*, Honolulu, HI, USA, 23–26 September 2023; pp. 95–107. [\[CrossRef\]](#)
53. Anchal, N.G.; Sriram, A.; Mathew, J.J.; Iyer, L.S.; Mahara, T. Analyzing the role of Indian media during the second wave of COVID using topic modeling. In *Hybrid Computational Intelligent Systems*; CRC Press: Boca Raton, FL, USA, 2023; Chapter 11.
54. Meier, F.; Fugl Eskjær, M. Topic Modelling Three Decades of Climate Change News in Denmark. *SSRN* **2023**. [\[CrossRef\]](#)
55. Rathod, R.G.; Barve, Y.; Saini, J.R.; Rathod, S. From Data Pre-processing to Hate Speech Detection: An Interdisciplinary Study on Women-targeted Online Abuse. In *Proceedings of the 2023 3rd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 23–25 June 2023; pp. 1–8. [\[CrossRef\]](#)
56. Zhao, W.X.; Jiang, J.; Weng, J.; He, J.; Lim, E.P.; Yan, H.; Li, X. Comparing Twitter and Traditional Media Using Topic Models. In *Advances in Information Retrieval*; Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6611, pp. 338–349. [\[CrossRef\]](#)

57. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013. Available online: <http://arxiv.org/abs/1301.3781> (accessed on 3 August 2023).
58. Le, Q.V.; Mikolov, T. Distributed Representations of Sentences and Documents. 2014. Available online: <http://arxiv.org/abs/1405.4053> (accessed on 3 August 2023).
59. Angelov, D. Top2Vec: Distributed Representations of Topics. 2020. Available online: <http://arxiv.org/abs/2008.09470> (accessed on 3 August 2023).
60. StatQuest with Josh Starmer. UMAP Dimension Reduction, Main Ideas!!! 2022. Available online: <https://www.youtube.com/watch?v=eN0wFzBA4Sc> (accessed on 28 September 2023).
61. StatQuest with Josh Starmer. Clustering with DBSCAN, Clearly Explained!!! 2022. Available online: <https://www.youtube.com/watch?v=RDZUdRSDOok> (accessed on 28 September 2023).
62. Karas, B.; Qu, S.; Xu, Y.; Zhu, Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Front. Artif. Intell.* **2022**, *5*, 948313. [CrossRef] [PubMed]
63. Zengul, F.D.; Bulut, A.; Oner, N.; Ahmed, A.; Ozaydin, B.; Yadav, M. A Practical and Empirical Comparison of Three Topic Modeling Methods using a COVID-19 Corpus: LSA, LDA, and Top2Vec. In Proceedings of the 56th Hawaii International Conference on System Sciences, Maui, HI, USA, 3–6 January 2023.
64. Vianna, D.; Silva De Moura, E. 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, 11–15 July 2022; pp. 3388–3392. [CrossRef]
65. Crijns, A.; Vanhullebusch, V.; Reusens, M.; Reusens, M.; Baesens, B. Topic modelling applied on innovation studies of Flemish companies. *J. Bus. Anal.* **2023**, *6*, 243–254. [CrossRef]
66. Bretsko, D.; Belyi, A.; Sobolevsky, S. Comparative Analysis of Community Detection and Transformer-Based Approaches for Topic Clustering of Scientific Papers. In Proceedings of the 23rd International Conference, Athens, Greece, 3–6 July 2023; pp. 648–660. [CrossRef]
67. Von Der Mosel, J.; Trautsch, A.; Herbold, S. On the Validity of Pre-Trained Transformers for Natural Language Processing in the Software Engineering Domain. *IEEE Trans. Softw. Eng.* **2023**, *49*, 1487–1507. [CrossRef]
68. Grootendorst, M.P. The Algorithm—BERTopic. Available online: <https://maartengr.github.io/BERTopic/algorithm/algorithm.html> (accessed on 29 September 2023).
69. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
70. Briggs, J. BERTopic Explained. 2022. Available online: <https://www.youtube.com/watch?v=fb7LENb9eag> (accessed on 3 August 2023).
71. Hägglund, M.; Blusi, M.; Bonacina, S. *Caring Is Sharing—Exploiting the Value in Data for Health and Innovation: Proceedings of MIE 2023*; IOS Press: Amsterdam, The Netherlands, 2023.
72. Li, Y. Insights from Tweets: Analysing Destination Topics and Sentiments, and Predicting Tourist Arrivals. Doctoral Dissertation, Durham University, Durham, UK, 2023.
73. Strydom, I.F.; Grobler, J. Topic Modelling for Characterizing COVID-19 Misinformation on Twitter: A South African Case Study. In Proceedings of the 23rd International Conference, Athens, Greece, 3–6 July 2023; pp. 289–304. [CrossRef]
74. Turner, J.; McDonald, M.; Hu, H. An Interdisciplinary Approach to Misinformation and Concept Drift in Historical Cannabis Tweets. In Proceedings of the 2023 IEEE 17th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 1–3 February 2023; pp. 317–322. [CrossRef]
75. Koonchanok, R.; Pan, Y.; Jang, H. Tracking public attitudes toward ChatGPT on Twitter using sentiment analysis and topic modeling. *arXiv* **2023**, arXiv:2306.12951. [CrossRef]
76. Grigore, D.N.; Pintilie, I. Transformer-based topic modeling to measure the severity of eating disorder symptoms. In Proceedings of the CLEF 2023: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 18–21 September 2023; pp. 18–21.
77. Mekacher, A.; Falkenberg, M.; Baronchelli, A. The Systemic Impact of Deplatforming on Social Media. *arXiv* **2023**, arXiv:2303.11147. [CrossRef]
78. Schneider, N.; Shouei, S.; Ghantous, S.; Feldman, E. Hate Speech Targets Detection in Parler using BERT. *arXiv* **2023**, arXiv:2304.01179. [CrossRef]
79. Egger, R.; Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef]
80. Zhou, W.; Zhang, C.; Wu, L.; Shashidhar, M. ChatGPT and marketing: Analyzing public discourse in early Twitter posts. *J. Mark. Anal.* **2023**, *11*, 693–706. [CrossRef]
81. Di Corso, E.; Ventura, F.; Cerquitelli, T. All in a twitter: Self-tuning strategies for a deeper understanding of a crisis tweet collection. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 3722–3726. [CrossRef]
82. Libit, D. Website that Helped Bring Down Anthony Weiner Is Coming Back. 2016. Updated May 20, 2016. Available online: <https://www.cnbc.com/2016/05/19/website-that-helped-bring-down-anthony-weiner-is-coming-back.html> (accessed on 28 September 2023).
83. de Groot, M.; Aliannejadi, M.; Haas, M.R. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *arXiv* **2022**, arXiv:2212.08459. [CrossRef]

84. Gensim: Topic Modelling for Humans. Available online: <https://radimrehurek.com/gensim/> (accessed on 28 September 2023).
85. Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 22 May 2010; pp. 45–50.
86. Gewers, F.L.; Ferreira, G.R.; Arruda, H.F.; Silva, F.N.; Comin, C.H.; Amancio, D.R.; Costa, L.D. Principal Component Analysis: A Natural Approach to Data Exploration. *arXiv* **2018**, arXiv:1804.02502. [[CrossRef](#)]
87. Shlens, J. A Tutorial on Principal Component Analysis. *arXiv* **2014**, arXiv:1404.1100. [[CrossRef](#)]
88. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [[CrossRef](#)]
89. Jin, X.; Han, J. K-Means Clustering. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2010; pp. 563–564. [[CrossRef](#)]
90. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131. [[CrossRef](#)]
91. Dieng, A.B.; Ruiz, F.J.R.; Blei, D.M. Topic Modeling in Embedding Spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.