UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Jaan Kupri

Topics of Fact-Checking on Twitter Community Notes

Bachelor's Thesis (9 ECTS)

Supervisors: Uku Kangur, MSc
Roshni Chakraborty, PhD

Tartu 2024

Faktikontrolli teemad Twitter kogukonna märkmetes

Lühikokkuvõte:

Kogukonnamärkmed on *Twitter*-i platvormi funktsioon, mis võimaldab panustajatel lisada konteksti, näiteks faktikontrolle tehtud postituse alla. Bakalaureusetöö eesmärk on uurida Twitter-i kogukonna märkmete andmestikku ning nende märkmete sisu. Selleks on märkmed jagatud teemadeks kasutades teemamudelit ning seejärel analüüsitud kasutades kolme meetodit, analüüs poliitiliste erakondade baasil, mis hõlmab sagedusgraafikut ja oluliste märksõnade analüüsi, ning ajaline analüüs. Kokkuvõttes viitavad tulemused sellele, et enamik märkmeid käsitleb Donald Trump-i ja Joe Biden-i vastasseisu seoses käimasolevate 2024. aasta USA presidendivalimistega, kus mõlemad poliitikud on presidendikandidaadid.

Võtmesõnad:

Kogukonnamärkmed, teemamudel, sagedusgraafik, oluliste märksõnade analüüs, ajaline analüüs

CERCS: P175, Informaatika, süsteemiteooria

Topics of Fact-Checking on Twitter Community Notes

Abstract:

Community Notes is a feature on *Twitter* where contributors can add context such as fact-checks under a post. The objective of this bachelor's thesis is to explore Community Notes data and its content. For this, the notes are separated into topics using a topic model and then analyzed using three methods, political party-based analysis, which includes a frequency graph and keyphrase analysis, and temporal analysis. To summarize, the findings suggest that most of the notes are about Donald Trump's and Joe Biden's rivalry due to the ongoing 2024 US presidential elections where both of the politicians are presidential candidates.

Keywords:

Community Notes, topic model, frequency graph, keyphrase analysis, temporal analysis

CERCS: P175, Informatics, systems theory

Table of contents

1 Introduction

Information is used to communicate and share ideas on a variety of topics. The accuracy and credibility of textual information have been well-studied in fields such as psychology and journalism [1]. According to the *Canadian Centre for Cyber Security*, there are three types of information disorders, misinformation, disinformation, and malinformation [2]. In this thesis, the main focus is misinformation, which is false information not intended to cause harm. Although it is shared with non-malicious intent, it can still have a negative impact.

With the increase in social media platforms becoming a primary news source [3] and the amount of information they contain drastically increasing in recent years [4], the traditional approaches of managing the quality of information manually are no longer applicable. This has imposed more research into the credibility of information circulating through these platforms [5][6]. The problem regarding misinformation on social media has gone so far that the *World Economic Forum* or *WEF* has categorized it as one of the main threats to human society [6]. Misinformation has caused credibility issues for social media platforms such as *Facebook* and *Twitter* (currently known as *X*) since they are unable to control the low-quality or inaccurate content that will inevitably reach wide audiences due to the speed at which unchecked information is spread. During the *COVID-19* epidemic, social media platforms, along with *Twitter*, saw a big spike in misinformation about the virus [7]. This outbreak of false opinions and news caused confusion and further worsened public health outcomes [8]. It is also true that misinformation influenced the 2021 January 6th capitol riots [9], which led to the death of five people [10].

To battle misinformation, *Twitter* came out with Community Notes [11], formerly known as Birdwatch. It is a feature where contributors can add context, for example, a fact-check to a misinformative or otherwise misleading post, video, or image. It was initially released on January 25, 2021. Contributors must sign up to participate in writing these notes, and these users carry out the entire process, the moderators of *Twitter* only intervene when the note goes against the community guidelines [6]. In 2022, *Twitter* underwent a large shift, after Elon Musk fully bought the company on October 27th. According to Musk, he planned on introducing new features to the platform by making its algorithms open-source, mitigating the presence of spambot accounts, and promoting free speech.

This thesis aims to understand what kind of misinformation has been spread during the years 2021-2023 on *Twitter*. To accomplish this, publicly available Community Notes data [12] can be used, providing insight into prevalent content by analyzing notes on misinformative posts.

Therefore, a method is represented for content analysis on the Community Notes. The approach employs a topic model to discover abstract "topics" by finding hidden semantic structures in the notes. The primary aim of this work is to identify prevalent topics within the dataset, track their change through time, and explore what drives these changes. The objective is to analyze these changes through the scope of US politicians. This is important as misinformation about politicians can distort public perceptions of political parties and potentially sway election results. For that, the notes that contain the names of well-known US politicians were separated into Republican and Democratic subsets. They were then analyzed using a graph where the frequency of each topic for each subset was highlighted. Subsequently, keyphrase analysis was done, in which the most frequent and relevant keyphrases were highlighted. Lastly, temporal analysis was conducted in order to analyze how the note-taking of certain topics changes throughout time and to see which significant world events would impact the number of notes taken.

The thesis starts with the *Introduction*, followed by *Background* in Section 2, in which misinformation on social media, topic modeling, and Community Notes were reviewed. Then the whole implementation process and the methods for the analysis were described in the *Methodology* in Section 3. Lastly, the *Results and Discussion* in Section 4 was done, followed by the *Conclusion* in Section 5.

2 Background

In this section, information and previous works related to misinformation on social media, topic modeling, and Community Notes are introduced. In the first chapter, different approaches to detecting false information are explained. In the second chapter, the most common topic models are explained, some of which are also implemented in this thesis. The third chapter explains the effectiveness of Community Notes as a means to address misinformation.

2.1 Misinformation on Social Media

A *Massachusetts Institute of Technology* study suggested that false news spreads more rapidly on *Twitter* than real news [13]. For example, false news stories are 70% more likely to be shared amongst the platform than true stories are. The reason for that, as was found out in the paper, is the emotional response people have to false news compared to real news [13].

According to the *National Institutes of Health*, there are three approaches to detecting false information [14]. The first one is knowledge-based approaches [15], where the information is overlooked by either an expert, a team of fact-checkers, or an automatic system. The second one is features-based approaches, which are AI-based models that rely on several key features including content-based, network-based, or user-based features [16]. The third one is modality-based approaches, which are classified into unimodal and multimodal studies [17][18]. Unimodal studies use a single type of data, for example, news or visual content, whereas multimodal studies usually include both visual and textual data.

2.2 Topic Modeling

As previously mentioned, social media platforms continue to confront the pressing issue of misinformation. One method that addresses this challenge involves organizing large volumes of textual data into categories, thereby enhancing the efficiency of identifying and preventing misinformation. This strategy is known as topic modeling.

The most widely used models for topic modeling include *Latent Semantic Analysis* [19], *Non-negative Matrix Factorization* [20], *Latent Dirichlet Allocation* [21], *Top2Vec* [22]*, and BERTopic* [23].

Latent Semantic Analysis (*LSA*) [19] is a method used to organize and analyze large bodies of text. By processing extensive textual data, *LSA* represents words and passages in a

high-dimensional "semantic space." This approach, akin to factor analysis, enables *LSA* to determine the similarity of meaning between words and text segments. Despite its capability, *LSA* has limitations, including its independence from word order and syntactic relations.

Non-negative Matrix Factorization (*NMF*) [20] is an algorithm used to learn parts of objects from data. It decomposes a given matrix into two matrices, where one represents the parts of objects and the other their respective weights. In its nature, *NMF* has lower computational complexity, which is useful when working with larger datasets. On the other hand, *NMF* has difficulty with negative values, which limits the model's flexibility.

Latent Dirichlet Allocation (*LDA*) [21] is a generative probabilistic framework recognized as a three-level hierarchical Bayesian model. In *LDA*, each document in the collection is a finite mixture of different underlying topics, and each topic is represented as an infinite mixture of underlying topic probabilities. *LDA* assumes that documents are generated through a process where topics are first selected from a distribution over topics, and then words are selected from a distribution over words specific to those topics. This hierarchical structure allows *LDA* to capture the latent topic structure inherent in the documents.

Top2Vec [22] is an algorithm that uses word embeddings. It functions by finding semantically similar words, sentences, or documents and grouping them. *Top2Vec* uses a pre-trained model, which makes the process easier while also benefiting from the semantic knowledge encoded in these models. Being dependent on pre-trained embedding models can also be a con depending on the quality of these models.

BERTopic [23] is built upon the mechanisms of *Top2Vec*, making them similar in a lot of ways. It is an advanced topic modeling technique designed to uncover coherent topics within collections of documents. It builds upon recent advancements in natural language processing, particularly leveraging pre-trained transformer-based language models like *BERT*. *BERTopic* extracts coherent topics from collections of documents by first embedding the text using pre-trained transformer-based language models. Then, it clusters these embeddings to group similar documents, and finally generates representative topics using a class-based variation of *TF-IDF*, providing insightful and interpretable results.

LDA, *LSA*, and *BERTopic* were selected for the analysis due to their widespread use for Twitter content analysis and topic modeling [24][25]. While all topic modeling mechanisms have their pros and cons, it is imperative to identify the model that best suits the specific use case.

2.3 Community Notes

Introduced in 2021, Community Notes (an example can be seen in Figure 1) is a relatively new feature of Twitter [26]. Because of that, there have only been a few works regarding its functionality and operational procedures. One study [27] compares the effectiveness of Community Notes to "Snoping". "Snoping" is a popular strategy among social media users to combat misinformation, involving linking professional fact-checking articles from third-party fact-checking organizations to the original message. As a result, they concluded that each strategy focuses on different targets when fact-checking and that both approaches might work well together.

Another study [28] suggested that a community-based approach to fighting misinformation might cause opinion speculation and polarization among the user base, especially regarding influential user accounts.

Figure 1. An example of a Community Note

3 Methodology

In this section, the procedures of data collection, pre-processing, and analysis are introduced. The analysis related to methodology covers topic modeling, keyphrase analysis, and temporal analysis. All data and code[1] associated with the project are additionally released.

3.1 Dataset collection and pre-processing

This section explains the gathering and the pre-processing of the Community Notes data. The data consists of four separate files:

- *Notes,* which contains all notes written by *Twitter* users.
- *Ratings*, which contains all ratings about the notes. They are given by other users for whether the note was helpful or not and they are not aimed at the post itself.
- *Note Status History,* which contains metadata about the notes such as the statuses they were received in.
- *User Enrollment*, which contains metadata about each user's enrollment state.

In this thesis, the only dataset used was *Notes*. The notes are from 2021-01-23 until 2023-09-16 and they are written by 38 640 different users. In total, the dataset has 23 features and 225 985 notes. Another critical aspect to establish was the daily frequency of note-taking. Due to computational restrictions, a sample of the total dataset was used. To avoid a biased representation of the data, the same amount of notes have to be extracted from each day. For that, the *createdAtMillis* column had to be converted, which contained an integer representing a millisecond into a proper date. For that, *datetime* library was used [29]. As a result, it was found that on some days there are up to 2000 community notes made, with the average being 235 per day. It was established that the writing of the notes throughout time was sufficiently distributed for conducting the study. To address the memory restriction and time constraints imposed by the method, the final dataset was reduced to 100 notes per day, resulting in a total of 54 329 notes, or roughly 24% of the original dataset.

In the *Notes* dataset, the feature *summary* represents the content of the notes. Inside the summaries were also references in the form of *URL*s (as can be seen in Figure 1). As they would interfere with the topic modeling later on, they were removed using a *Regular Expressions Operations* library [30]. Another issue encountered with this dataset was notes that were either empty or duplicated. After examining a sample of the data, the notes that held no underlying significance were in the form of numbers. With that knowledge, all notes

[1] drive.google.com/drive/folders/1ZxyOfmRVS1_oC5X_oU7dnGqR508Vogxg?usp=sharing

containing less than 10 characters, subsequent digits, or ones that were identical to previous ones, were removed.

3.2 Topic Modeling

To analyze and make conclusions on the topics of fact-checking on community notes, firstly the topics have to be extracted using a topic modeling technique. To see which topic model works best with the given Community Notes data, three models were chosen: *Latent Semantic Analysis*, *Latent Dirichlet Allocation*, and *BERTopic*. In order to select the most suitable topic detection approach, the following was considered:

- Coherence: the topics are sufficiently different from each other.
- Representativeness: the topics match with the notes, meaning the topic represents what the note is about.
- Distribution: the frequency of topics is reasonably similar.

The following paragraphs describe the nature and operation of each of the models.

3.2.1 Implementation

Latent Semantic Analysis: In Python, the library used for getting the topics using *LSA* is called *TruncatedSVD* [31]. After experimenting with the model using different numbers of topics, a decision was made to continue with 10 topics because this resulted in the topics being the most coherent. As the model does not provide topic names, the topics were analyzed and labeled based on the most frequent words in each topic set. The topic model generated can be seen in Figure 2a.

Latent Dirichlet Allocation: In Python, the library used for the *LDA* model is *LatentDirichletAllocation* and *TruncatedSVD* [32]. On the basis of *LatentDirichletAllocation* library and its function to show the optimal parameters, the number of topics was decided as six. Although the number of topics was quite small, the topics were well distributed and sufficiently differentiable. The model did not assign topic names, instead, it provided scores for documents and words with the highest probability of occurrence, measured by the *TF-IDF* score. These words were then analyzed and used to name the topics. The topic model generated can be seen in Figure 2b.

BERTopic: A Python library called *BERTopic* was used for the topic extraction [33]. Initially, the model gave out nine different topics and also an additional outlier topic which included all notes that did not fit into the previously mentioned number of topics. With this approach,

the outlier topic contained most of the notes as can be seen in Figure 2c. This would lead to insufficient results, as a large portion of the notes would be uncategorized.

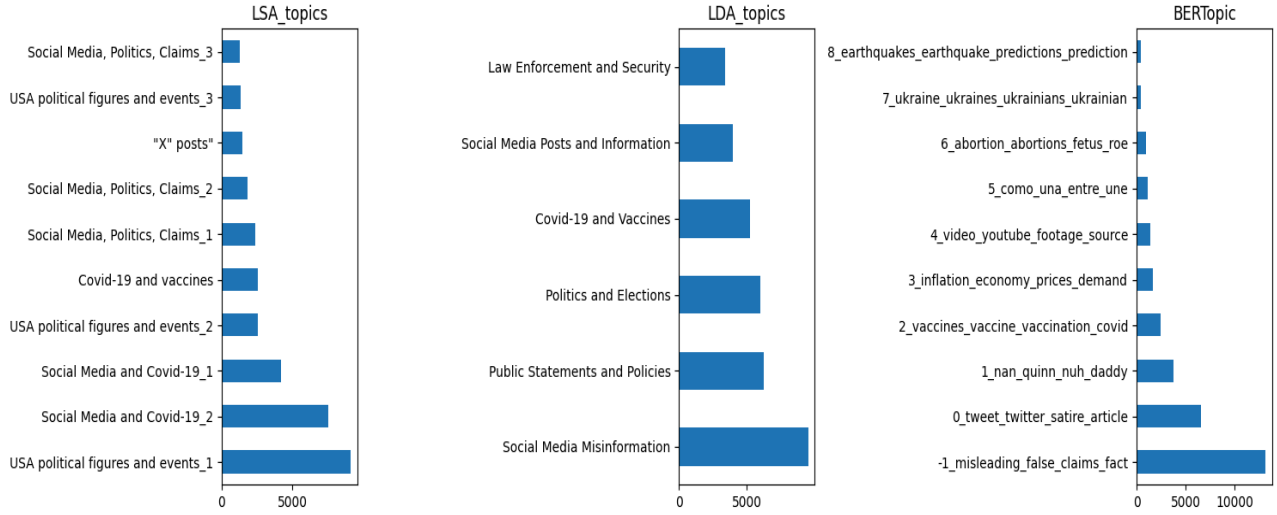

Figure 2. Results for LSA (2a), LDA (2b), and BERTopic (2c)

3.2.2 Selection

The results that *LSA*, *LDA*, and *BERTopic* gave were vastly different from each other, each having positive and negative aspects. The following table (Table 1. Comparison of topic models) demonstrates those aspects:

Table 1. Comparison of topic models

Model	**Coherence**	**Representativeness**	**Distribution**
LSA	✗	✗	✓
LDA	✓	✗	✓
BERTopic	✓	✓	✗

After analyzing the results of each model it was seen that *BERTopic* had the most distinguishable and wide-ranging topics, although the current distribution was not sufficient. The reason for this is that there were far too many uncategorized topics, as can be seen in Figure 2 (in the *BERTopic* model '-1' represents uncategorized instances). Although the *LDA* model had a substantial distribution and coherence of topics, an analysis of the notes revealed that the representativeness was inadequate. For example, when printing out a sample of notes from each topic, the content did not match the intended theme or purpose, demonstrating a

lack of representativeness throughout the model's output. For this reason, *BERTopic* was the optimal choice.

3.2.3 Fine-tuning BERTopic

BERTopic model also has a variety of parameters, which can be modified to fine-tune the model for better results. The next paragraphs are to explain which parameters or other natural language processing strategies are used and the reasoning behind their usage.

To address the issue of distribution, the outlier class must be considerably reduced. For optimal results, a threshold of 10% or less was sufficient. For this, a function was implemented to iteratively repeat the modeling process with uncategorized notes until the outlier class size was below the previously mentioned threshold. The result can be seen in Figure 3.

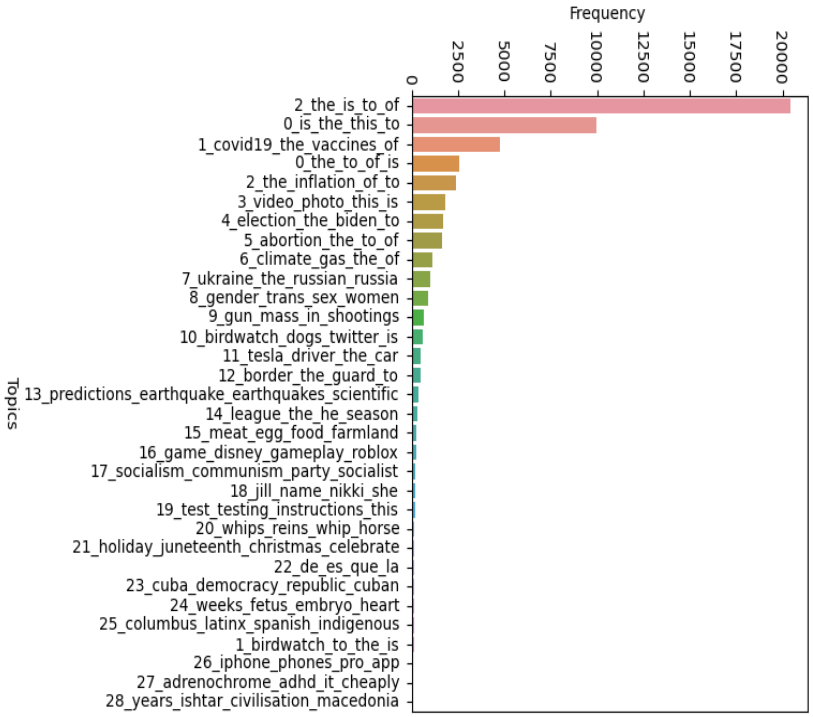

Figure 3. Fine-tuning 1

Although the previous step could ensure that the outlier topic was below a necessary threshold, it did not account for stopwords. For example, topics such as *2_the_is_to_of*, consist of words that add no value in terms of topic modeling. In natural language processing a popular strategy is to remove stopwords from the input, increasing the accuracy and reducing noise. Although *nltk* [34] library includes a corpus with the most used stopwords in the English language that needs to be removed from the model, additional dataset-specific

stopwords need to be added. In this thesis, these words were *tweet*, *twitter*, *birdwatch*, *video, photo, image, article, fact, real, fake, misleading, evidence, claim.* As can be seen in Figure 4, this step disposed of redundant topics that contained stopwords.

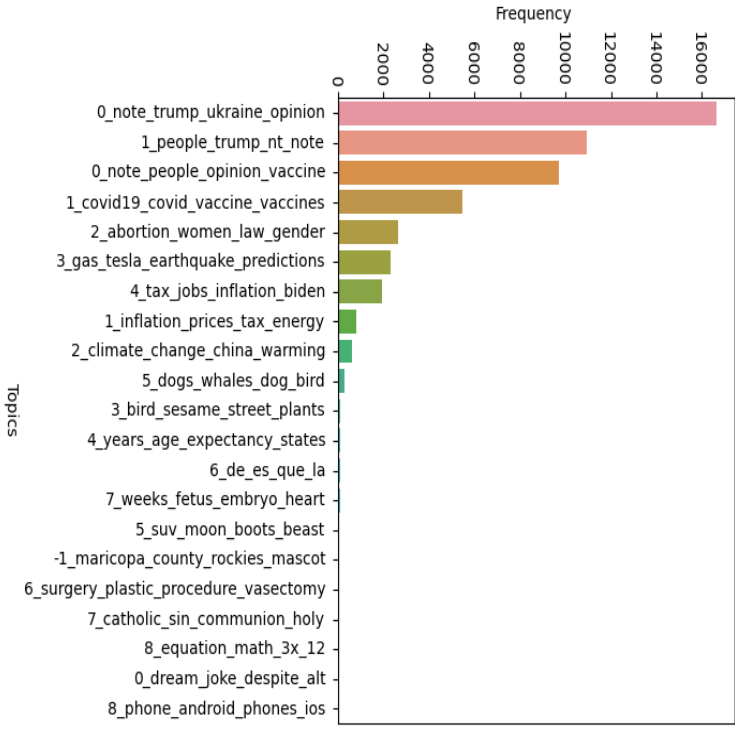

Figure 4. Fine-tuning 2

Although the topics containing stopwords were removed due to their redundancy, another factor contributing to the generation of redundant topics exists. This occurs when a topic is too small in size. To address this issue, a *BERTopic* parameter *min_topic_size* was utilized, as it filters out topics that fail to meet the specified threshold. Considering the number of documents and the current distribution of topics, a reasonable input would be 200, as it would approve smaller topics while reducing noise. The result can be seen in Figure 5.

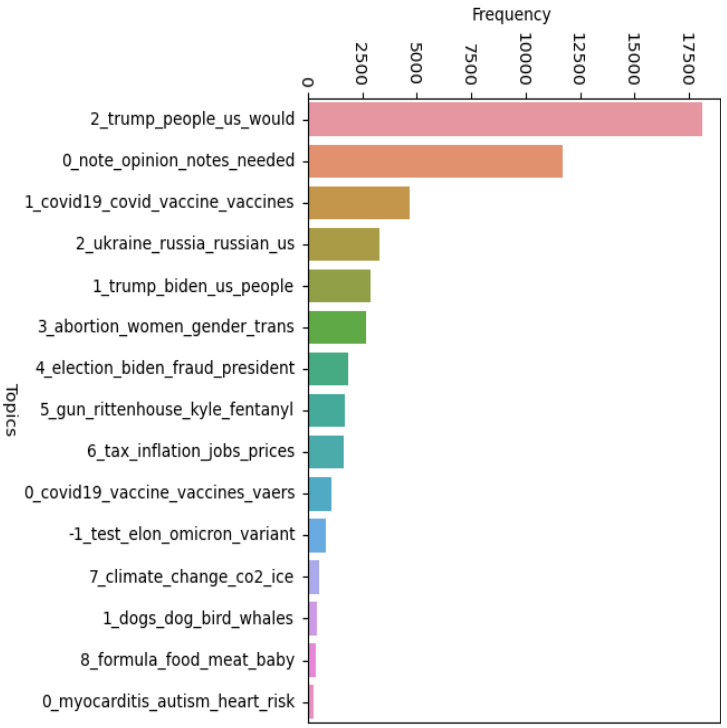

Figure 5. Fine-tuning 3

As can be seen in Figure 5, redundant topics that were too small in size were removed, resulting in the number of topics being reduced from 21 to 15. However, topics such as *2_trump_people_us_would* and *1_trump_biden_us_people* or *1_covid19_covid_vaccine_vaccines* and *0_covid19_vaccine_vaccines_vaers* were too similar to each other. To address this issue, custom *BERTopic* parameters for dimensionality reduction can be used. *Uniform Manifold Approximation and Projection* (*UMAP*) is a graph that helps display many types of data. In *BERTopic* it can be used to reduce the dimensionality of the document embedding for *HBDSCAN* to create good clusters more efficiently. *UMAP* also has a parameter called *random_state,* which can be included to generate the model with the same seed every time. This is optimal for this thesis because the results should be the same every time the code is run [35]. To implement the *UMAP* model, the initialization of the *UMAP* with *n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine'* and *random_state=42* has to be done. Then the initialization of *BERTopic* with an extra parameter umap_model can be done. Additionally, attempting to employ *HDBSCAN* in the *BERTopic* model did not enhance the quality of the topics, therefore it was disregarded.

After that, the outlier topic was removed due to it being redundant and the rest of the topics were renamed to best match the notes they contain and the information they convey. The final result can be seen in Figure 6.

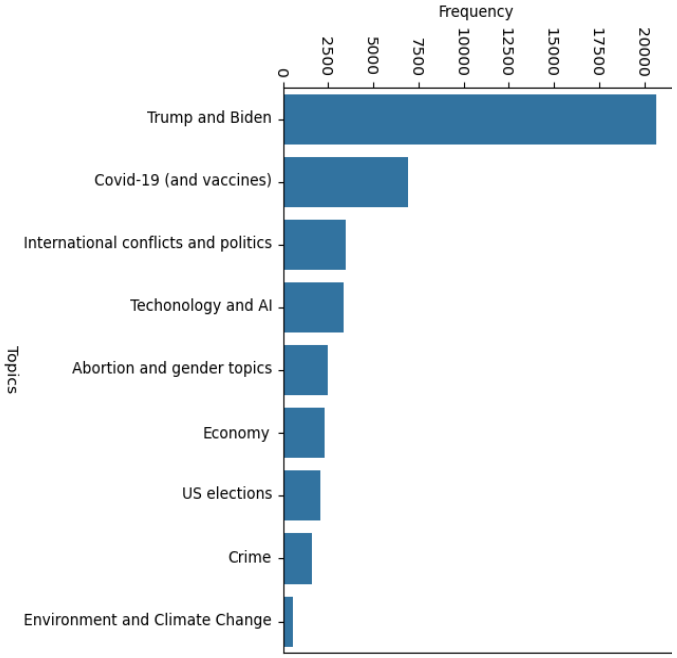

Figure 6. BERTopic final result

3.3 Political Party-Based Analysis

One primary goal outlined in this thesis is to examine the contents of the Community Notes from the perspective of US politicians. This examination is crucial as misinformation concerning politics has the potential to distort public perceptions and, consequently, influence election outcomes. As the US is a two-party system [36] consisting of the Republican Party and the Democratic Party, the focus of the analysis is between these two. For both political parties, the ten most popular and well-known politicians have been selected, as they would have the most impact on note-taking. The selection was made based on the popularity and fame of those politicians [37]. Using their names as keywords, the dataset was thoroughly examined to find how many notes contained these keywords. Job titles and acronyms such as President or *AOC* were also included as keywords. Below is a table (Table 2. Extraction of politicians) containing the politician's name, their political party, keywords used to find notes, frequency of the notes found, and the percentage of these notes compared to the whole dataset.

Table 2. Extraction of politicians [37]

Politician's name	Political party	Keywords	Absolute frequency	Relative frequency
Donald J. Trump	Republican	"Donald Trump", "President Trump", "Trump", "Donald J. Trump"	2371	4.36%
George W. Bush	Republican	"George W. Bush", "President Bush", "G.W. Bush"	767	1.41%
Ted Cruz	Republican	"Ted Cruz", "T. Cruz", "Senator Cruz"	186	0.34%
Ron DeSantis	Republican	"DeSantis", "Ron DeSantis", "R. DeSantis"	396	0.73%
Kevin McCarthy	Republican	"Kevin McCarthy", "K. McCarthy"	125	0.23%
Mike Pence	Republican	"Mike Pence", "M. Pence", "Pence"	72	0.13%
Nikki Haley	Republican	"Nikki Haley", "N. Haley"	27	0.05%
Marco Rubio	Republican	"Marco Rubio", "M. Rubio"	23	0.04%
Greg Abbott	Republican	"Greg Abbott", "G. Abbott"	55	0.10%
Vivek Ramaswamy	Republican	"Vivek Ramaswamy", "Ramaswamy"	9	0.02%
Joe Biden	Democrat	"Joe Biden", "J. Biden", "President Biden", "Biden"	2483	4.57%
Hillary Clinton	Democrat	"Hillary Clinton", "H. Clinton"	276	0.51%
Kamala Harris	Democrat	"Kamala Harris", "K. Harris", "Harris"	193	0.36%
Barack Obama	Democrat	"Barack Obama", "President Obama", "Obama"	257	0.47%

Bernie Sanders	Democrat	"Bernie Sanders", "B. Sanders"	48	0.09%
Bill Clinton	Democrat	"Bill Clinton", "B. Clinton", "President Clinton"	276	0.51%
Elizabeth Warren	Democrat	"Elizabeth Warren", "E. Warren"	54	0.10%
Jimmy Carter	Democrat	"Jimmy Carter", "J. Carter", "President Carter"	16	0.03%
Alexandria Ocasio-Cortez	Democrat	"Alexandria Ocasio-Cortez", "AOC"	170	0.31%
Al Gore	Democrat	"Al Gore"	22	0.04%

As can be seen in Table 2, most of the politicians were mentioned in less than 1% of the notes. Donald Trump and Joe Biden both had significantly more notes made about them than any other politician in their respective political party.

With the collected notes, two separate subsets were generated: Republican subset and Democratic subset.

3.3.1 Keyphrase Analysis

Keyphrase analysis helps to find frequent patterns in a large set of documents. In this case, using this analysis can aid in understanding which phrases or words are frequent in note-taking and making conclusions based on them. The analysis was done by a combined system of implementing N-grams [38] and *KeyBERT* [39].

N-grams are N-character slices of a longer string, meaning they represent all combinations of length N words while maintaining the order of occurrence. In this thesis, N-grams are used to identify keyphrases that occur multiple times in the notes.

KeyBERT is a keyword extraction technique that can generate keywords and keyphrases from a given N-gram range. Here it was used to get the relevancy score for each N-gram.

To perform the analysis, all summaries from each dataset (republican, democratic) were concatenated into two variables. They were then split into words and lemmatized using the *WordNetLemmatizer* function in *NLTK* [40]. After that, a function was used to transform them

into N-grams. This function provides all N-grams and their frequencies in the text. Subsequently, *KeyBERT* was used to generate relevant keyphrases. Following that, the keyphrases were filtered using a threshold for both the frequency and the relevancy score (>0.25). The threshold for the frequency depended on the topic size, and after experimenting with various thresholds, a manual evaluation determined 0.25 to be the optimal choice. Keyphrases containing stopwords were also disregarded.

3.4 Temporal Analysis

To get an adequate overview of how *Twitter*'s topics of discussion can vary throughout time, temporal analysis was done. For this, all dates were split into periods, and for each period the frequency of a particular topic was calculated. In this thesis, the period chosen was 15 days. This analysis aims to highlight if any events cause a spike in the frequency of notetaking. For that, 10 events have been chosen, they are as follows:

1) The US withdraws troops from Afghanistan (30.08.2021) **(E_1)** [41]
2) Russia invades Ukraine (24.02.2022) **(E_2)** [42]
3) Overturning of Roe vs Wade (24.06.2022) **(E_3)** [43]
4) Queen Elizabeth II dies (08.09.2022) **(E_4)** [44]
5) Elon Musk buys Twitter (27.10.2022) **(E_5)** [45]
6) Trump announces his presidential campaign (15.11.2022) **(E_6)** [46]
7) ChatGPT was released to the public (30.11.2022) **(E_7)** [47]
8) Donald Trump's first indictment (30.03.2023) **(E_8)** [48]
9) Biden announces his presidential campaign (25.04.2023) **(E_9)** [49]
10) First GOP debate (24.08.2023) **(E_{10})** [50]

4 Results and Discussion

In this section, an in-depth analysis of Community Notes was done. Firstly, political party-based analysis, which consists of analyzing a frequency graph and common keyphrases using N-grams, was conducted. After that, timeline graphs were generated to observe which events caused a spike in which topics.

4.1 Political Party-Based Analysis

For each political party, the frequency of each topic was now known. With that information, a bar plot distribution was made, where the x-axis represents the name and the y-axis represents the relative frequency of the topic. For each topic, two frequencies are measured, the first one (red) is for the Republican party and the second one (blue) is for the Democratic party. The distribution can be seen in Figure 7.

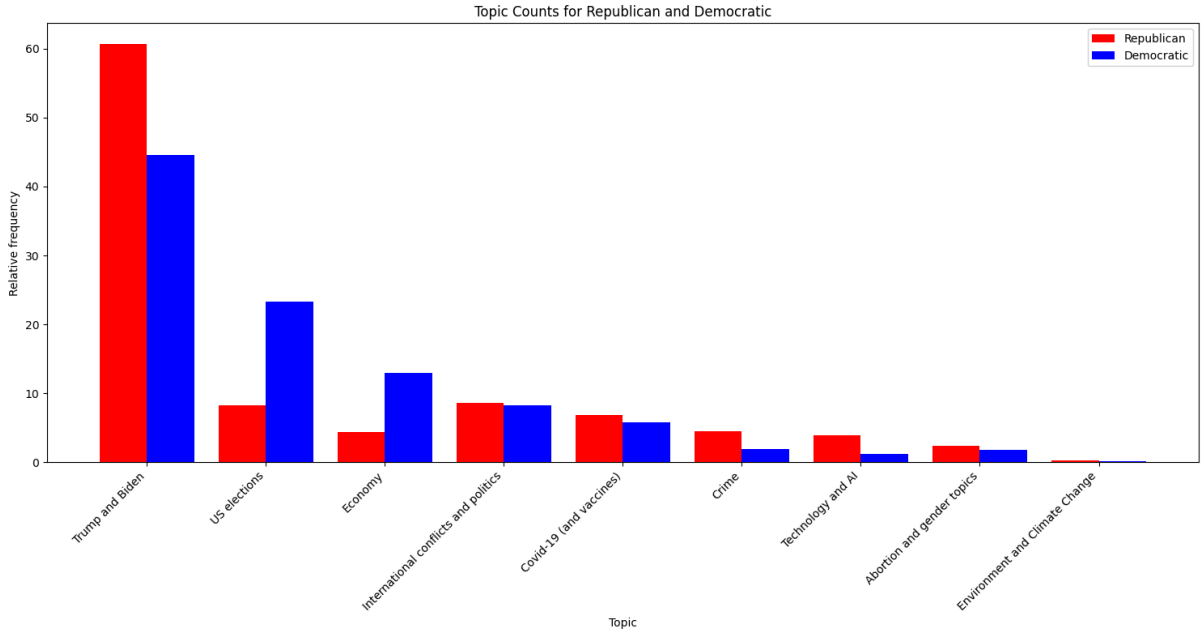

Figure 7. Frequency graph

Using the bar plot as a source, the following conclusions were made:

1) The main talking point for Republicans and Democrats was the “rivalry” between Donald Trump and Joe Biden. This was expected as both have served as the President of the United States. When analyzing the notes that talk about Democrats, some recurring talking points include political allegations, particularly regarding the Biden administration, such as investigations into the family's business dealings, claims of corruption, and critiques of policy decisions. There were also references to the 2020

presidential elections and disputes regarding its integrity and policy debates, including infrastructure plans, student loan forgiveness, energy policies, and immigration reforms. For Republicans, some recurring talking points were also about the 2020 presidential elections, although from the angle of disputing the claims of a rigged election. There were also many references to legal issues, court rulings, and indictments of Donald Trump and mentions of conspiracy theories, disinformation campaigns, and false narratives, particularly regarding topics like *QAnon*, election fraud, and media manipulation.

2) Considerably more notes about the US elections referred to Democrats. Upon a closer inspection of the notes of both political parties, it was seen that most of them are about the 2020 elections and the discussion regarding their integrity. The reason there were more notes about Democrats in this case was the wording and structure of the sentences, meaning that more notes confirm that Joe Biden won the election than those that say Donald Trump lost the election.
3) Considerably more notes about the economy were referring to Democrats, mostly Joe Biden. This is logical because he was the residing President of the United States during the period outlined in our data. Additionally, Biden is blamed by many for economic issues, such as gas prices and inflation. This was proved by the notes, which were overwhelmingly about these issues. Many notes also compared historical economic indicators under previous administrations to contextualize current discussions. When inspecting notes where Republicans were mentioned, some recurring talking points included job growth and loss, economic recovery, and policies regarding oil and energy.

4.1.1 Keyphrase Analysis

To perform the analysis, two graphs were generated for each topic, containing two different measurements. On the x-axis, there was the *KeyBERT* relevancy score [39]. On the y-axis, there was the frequency of the n-grams that were generated separately. In this thesis, 2-grams were used as these are the optimal balance between frequency and information they convey. 3-gram graphs were also inspected, but they did not give any additional information and thus were considered redundant.

Using the 2-gram graphs as a source, the following conclusions were made:

1) The most mentioned politicians across all topics remained Donald Trump and Joe Biden. However, for the 2-grams it can be seen that for the five largest topics Joe Biden was mentioned considerably more than Donald Trump (this can be seen in Figures 8-12). In the *Trump and Biden* (see Figure 8) topic, Hunter Biden (Joe Biden's son) and Senator Ted Cruz were also mentioned. When analyzing the summaries, most notes about Hunter Biden were allegations about his potential criminal violations such as money laundering, and references to his laptop controversy [51]. For Ted Cruz, the notes were mostly statements asserting that Senator Cruz did not incite or cause any violence during the Capitol riots on January 6, 2021 [9].

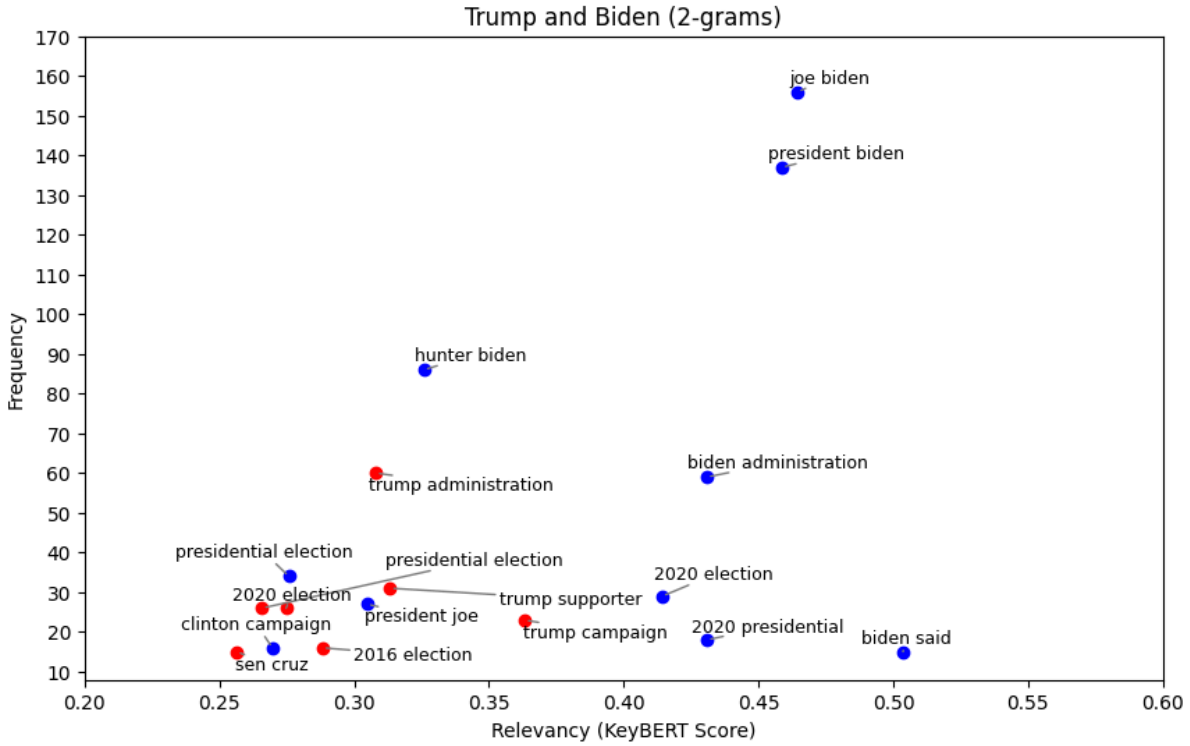

Figure 8. N-grams for topic *Trump and Biden*. Blue dots represent notes about democrats and red represents notes about republicans.

2) Topic *US elections* (see Figure 9) was also overwhelmingly about democratic keyphrases, with the main talking point being the 2020 presidential elections and issues/speculations regarding this event.

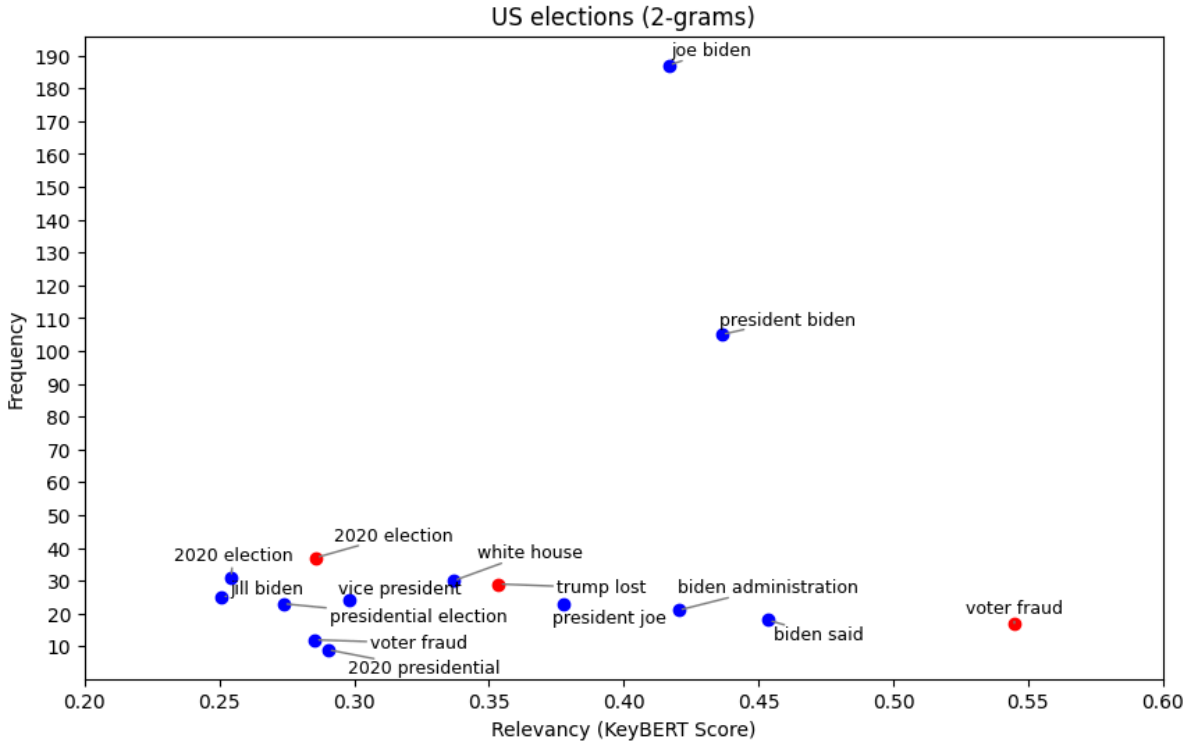

Figure 9. N-grams for topic *US elections*. Blue dots represent notes about democrats and red represents notes about republicans.

3) Topic *Economy* (see Figure 10) was dominated by keyphrases about Democrats. It was also confirmed previously that these notes are mainly talking about economic issues that the Biden administration was getting blamed for.

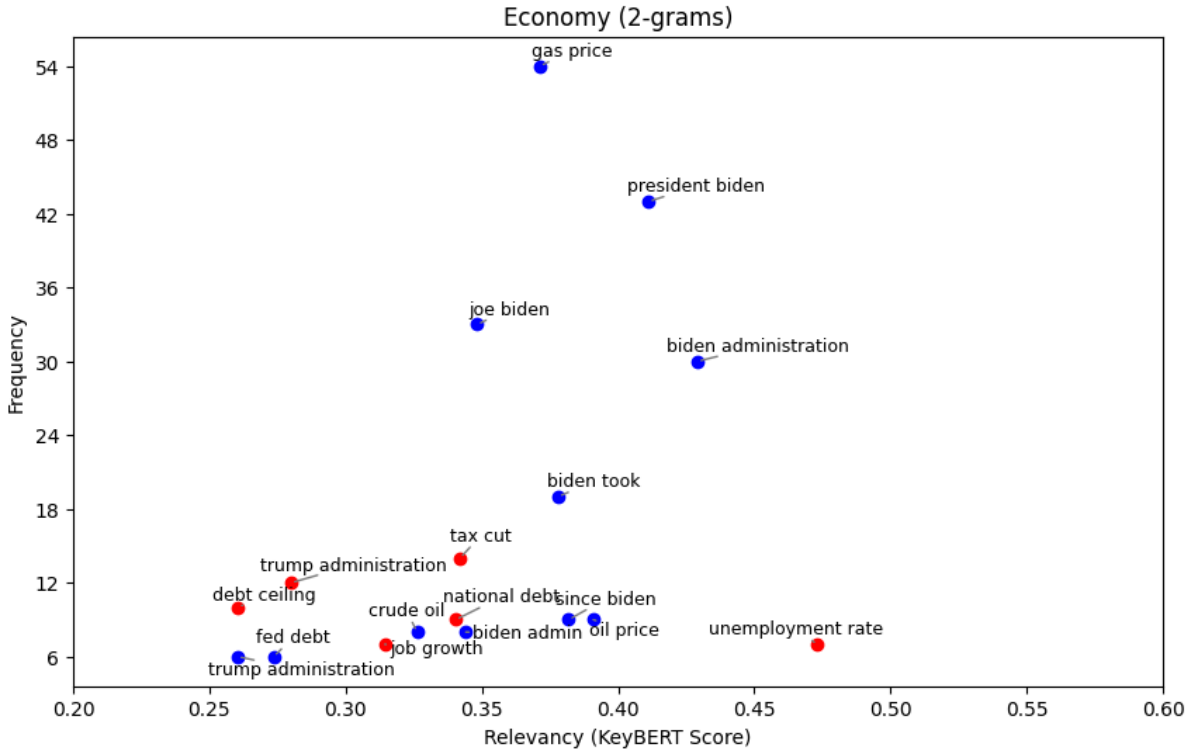

Figure 10. N-grams for topic *Economy*. Blue dots represent notes about democrats and red represents notes about republicans.

4) Topic *International conflicts and politics* (see Figure 11) consisted mostly of Joe Biden's stances on issues such as immigration, foreign affairs, and domestic policies.

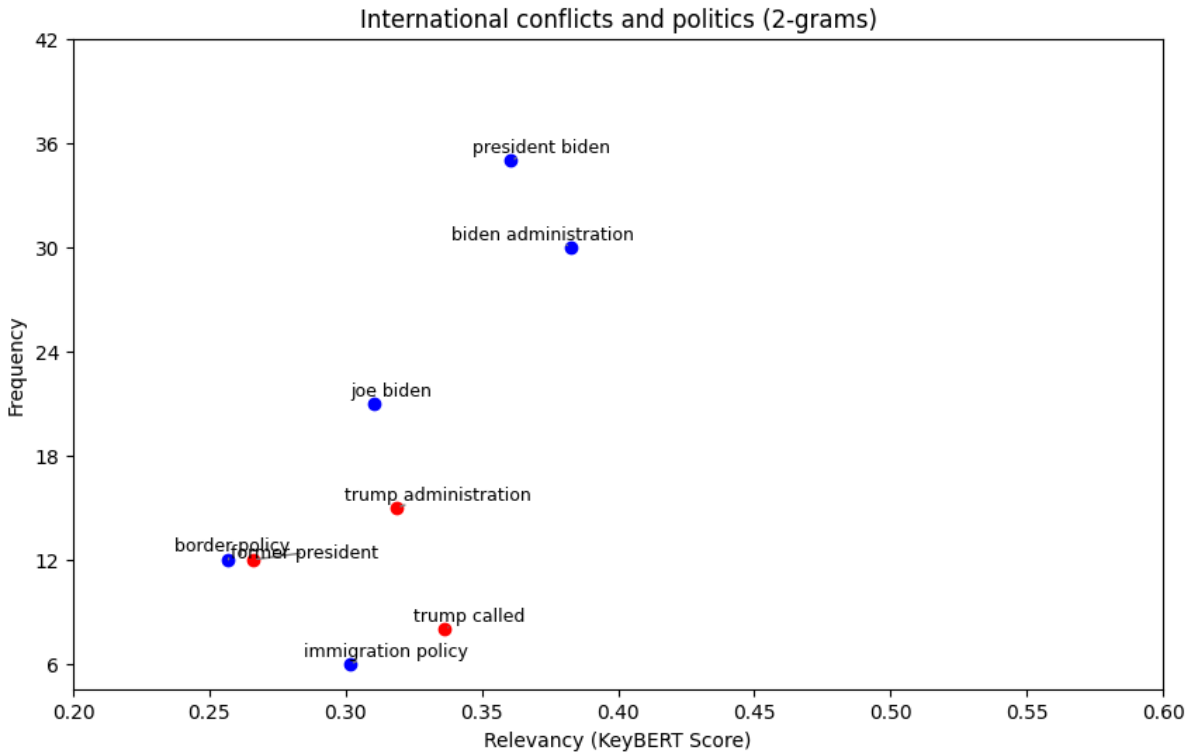

Figure 11. N-grams for topic *International conflicts and politics*. Blue dots represent notes about democrats and red represents notes about republicans.

5) For topic *Covid-19 (and vaccines)* (see Figure 12) the keyphrases for both parties were mostly similar and there were no differences other than mentions of Joe Biden and Donald Trump.

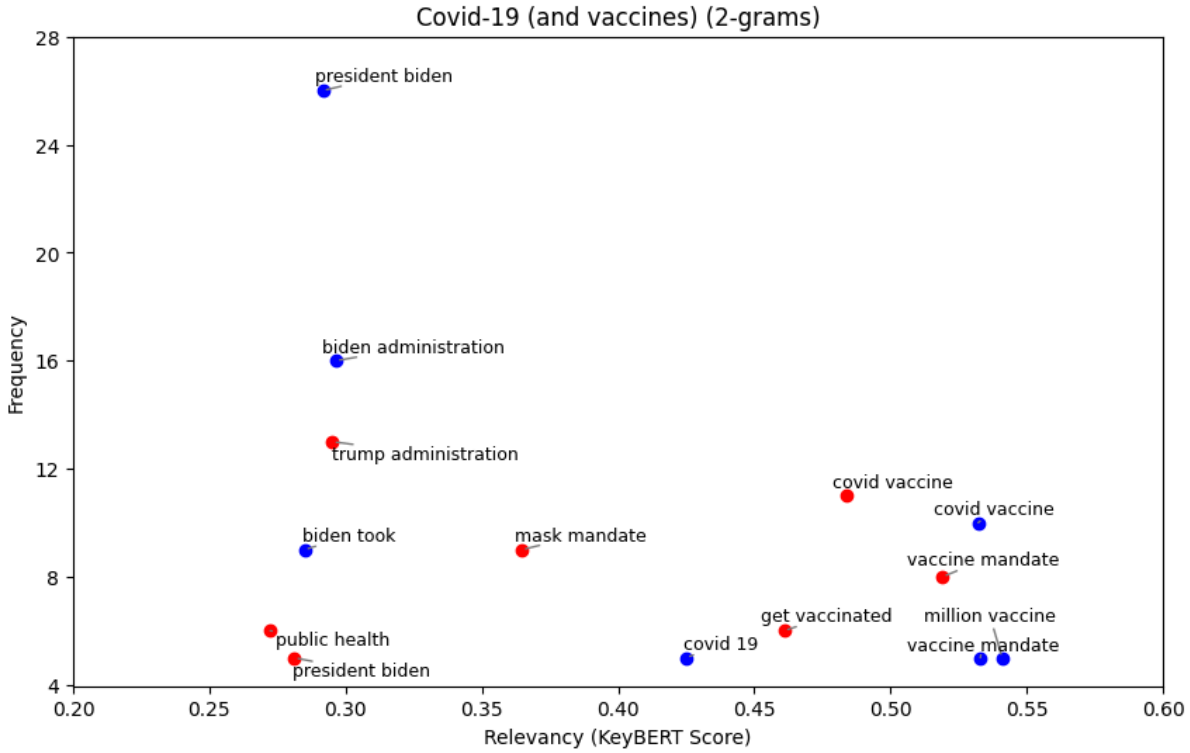

Figure 12. N-grams for topic *Covid-19 (and vaccines)*. Blue dots represent notes about democrats and red represents notes about republicans.

6) Topic *Crime* (see Figure 13) was mostly about Donald Trump, his current ongoing criminal investigations, or other allegations, that were debunked.

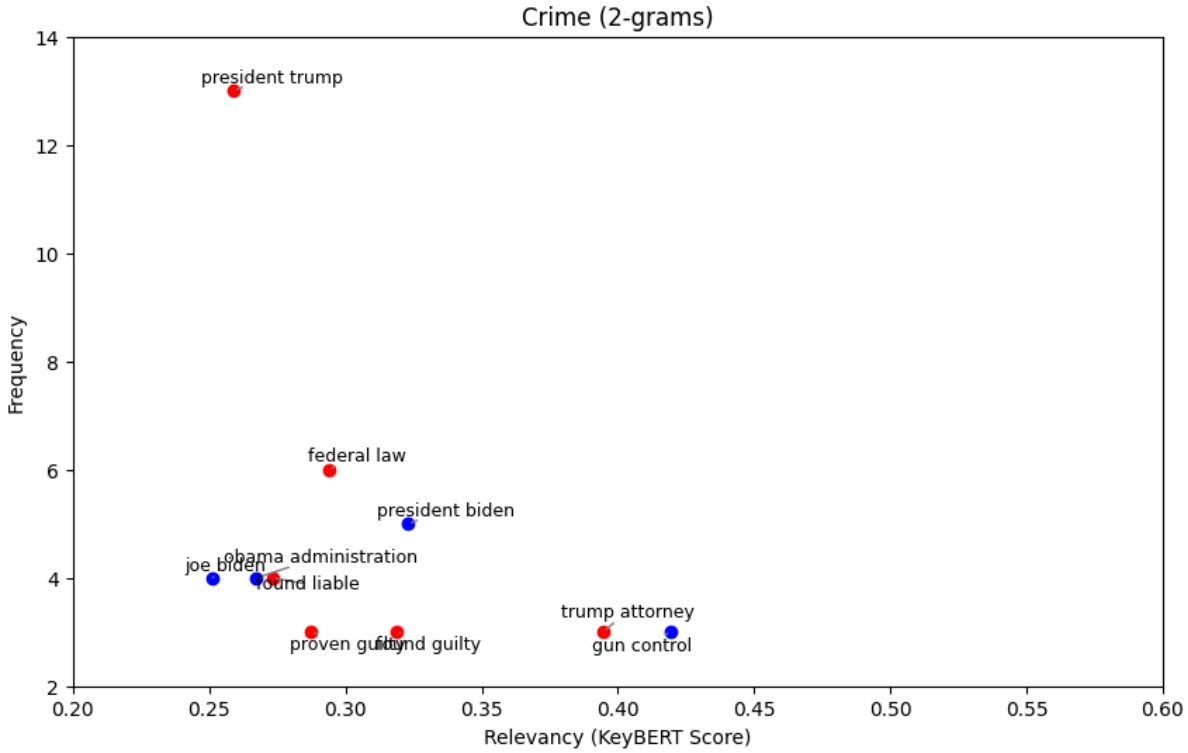

Figure 13. N-grams for topic *Crime*. Blue dots represent notes about democrats and red represents notes about republicans.

7) Topic *Technology and AI* (see Figure 14) was largely about content related to Donald Trump, which was in most instances misrepresented, digitally altered, or falsely attributed to him.

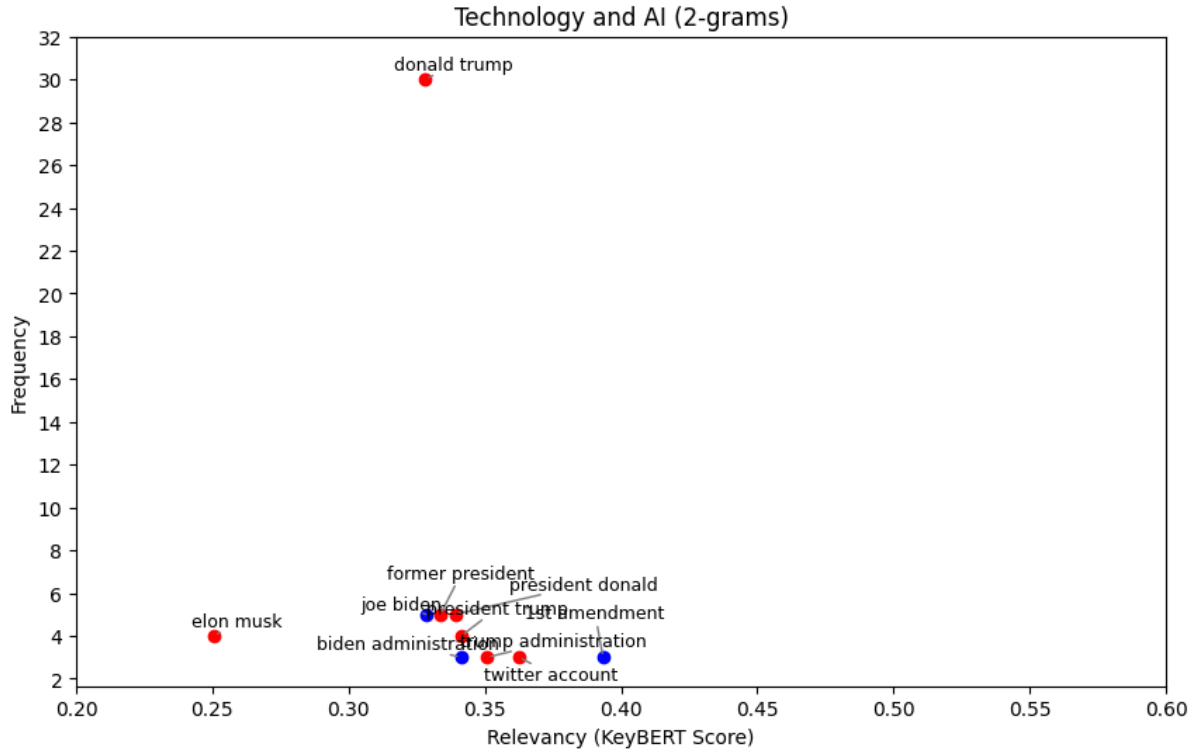

Figure 14. N-grams for topic *Technology and AI*. Blue dots represent notes about democrats and red represents notes about republicans.

8) Topic *Abortion and gender topics* (see Figure 15) was largely about discussions on abortion laws and LGBTQ+ rights. The notes were mostly similar across both parties and graphs.

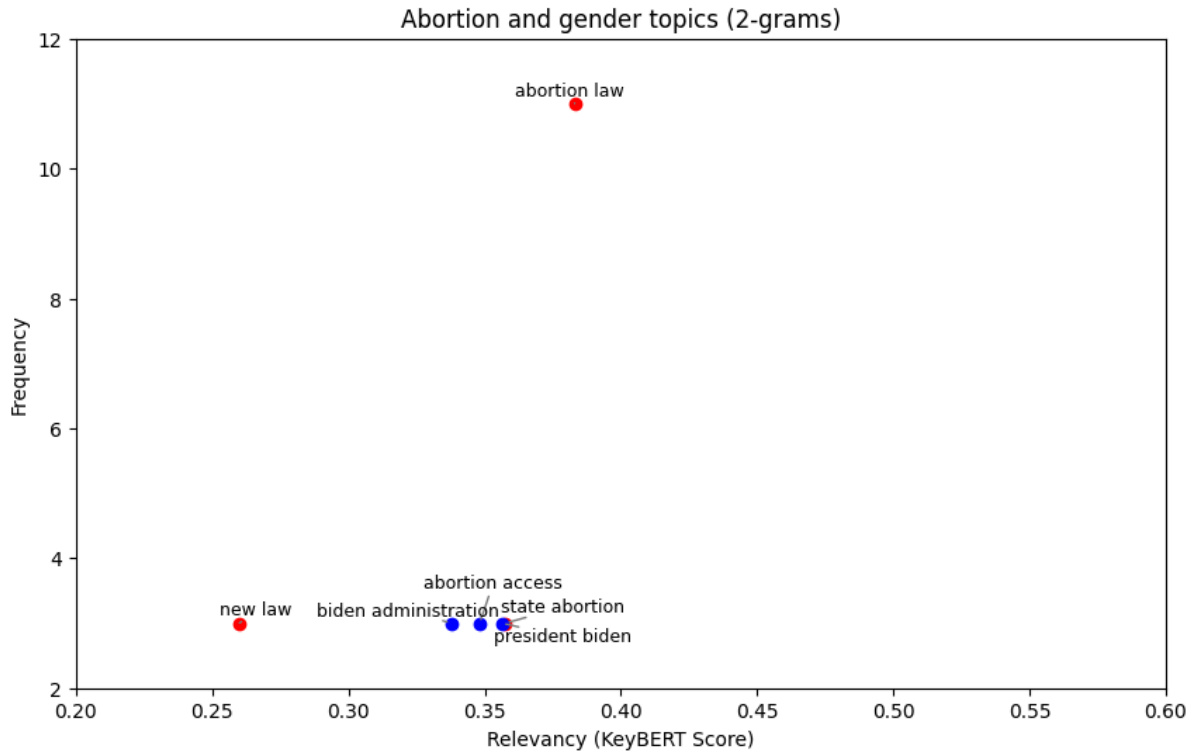

Figure 15. N-grams for topic *Abortion and gender topics*. Blue dots represent notes about democrats and red represents notes about republicans.

9) Topic *Environment and Climate Change* was missing due to there not being any phrases that bypass the necessary threshold highlighting that this topic did not discuss the same events or people.

4.2 Temporal Analysis

To perform the analysis, a graph was generated, where on the x-axis there was time and on the y-axis the frequency of notes on a given period. Each event was examined separately and the analysis aimed to establish whether these events were single-topic events (meaning they created spikes for a single topic), many-topic events (meaning they created spikes for multiple topics), or non-topical events (meaning they didn't create any spikes). To confirm whether the spikes correlate to that specific event, the notes were analyzed during the period in which it took place. The following table (Table 3. Analysis of events) contains the label of the event, what type of event it was, and additional comments about this event and its correspondence to different topics.

Table 3 - Analysis of events

Event	**Type**	**Comment**
E_1	Single-topic event	This event caused a meaningful spike in topic *International conflicts and politics*. As can be seen in Figure 16, the spike was slightly before the event itself. The reason was that this was a peace agreement that spanned many months, although becoming a large topic of discussion right before the conclusion. Spikes in other topics had no meaningful relation to this event.
E_2	Many-topic event	As can be seen in Figures 16 and 17, this topic caused a spike in topics *Trump and Biden, International conflicts and politics*, and *Economy*. After examining the notes, the first topic contained some notes that referred to this conflict while also mentioning Donald Trump or Joe Biden. Meanwhile, the topic of *Economy* contained notes about why the gas prices increased, being directly related to Russia's invasion of Ukraine.
E_3	Single-topic event	As can be seen in Figure 17, this topic caused a massive spike in the topic of *Abortion and gender topics*. An analysis of the notes reveals that most of the notes were about that particular event.
E_4	Non-topical event	This event caused no spikes in any of the topics. Moreover, no notes were found during that period that mentioned Queen Elizabeth II.
E_5	Single-topic event	This event caused an increase in notes regarding Elon Musk in topic *Technology and AI* (can be seen in Figure 17). When analyzing the notes, many of them referred to people impersonating Elon Musk on *Twitter* and others were about his announcements as the new owner of the platform. In conclusion, this event made Elon Musk more popular and

		therefore more users were posting about him, increasing the amount of notes written on this topic.
E_6	Non-topic event	When analyzing the notes on topics such as *Trump and Biden* and *US elections*, very few had mentioned his announcement of running for President of the United States in 2024.
E_7	Non-topic event	After analyzing the notes during that time period, it can be said that this event did not produce any significant changes in note-taking.
E_8	Single-topic event	Although the spike in topic *Crime* (as can be seen in Figure 18) started slightly before the indictment happened, analysis of the notes revealed that many notes were about this event. This prompts the conclusion that this event caused a small spike in the mentioned topic.
E_9	Non-topic event	Analysis of the notes revealed that this event was not a significant topic of discussion in the Community Notes.
E_{10}	Non-topic event	Analysis of the notes revealed that this event was not a significant topic of discussion in the Community Notes.

In conclusion, events E_4, E_6, E_7, E_9, and E_{10} did not emerge as significant talking points across these topics. The reason for this could be that events such as presidential announcements, debates, and deaths didn't produce much controversy and by extension had an absence of misinformative posts related to them. The only event that caused a spike in multiple topics is E_2. This is logical because this event was indisputably the most significant global incident during that time. E_1, E_3, E_5, and E_8 were all single-topic events. This thesis posits that these events produced controversy and discussion among the users, which also led to an increase in misinformative posts. For example, the indictment of Donald Trump sparked a contentious exchange between Republicans and Democrats.

The chosen events did not produce any noticeable spikes in topics *US elections*, *Covid-19 (and vaccines)*, and *Environment and Climate Change*. Although in the first topic, there was a large spike in October-November 2022, however an analysis of the notes did not reveal any

prevalent common event associated with them. The absence of spikes correlated to events related to *Covid-19 (and vaccines)* (see Figure 16) can be attributed to the fact that none of these events were associated with the pandemic or the vaccines. This can also be said about topic *Environment and Climate Change* (see Figure 18), although the frequency of this topic also plays a pivotal role, meaning that a topic with very few notes was not discussed nearly as much as others.

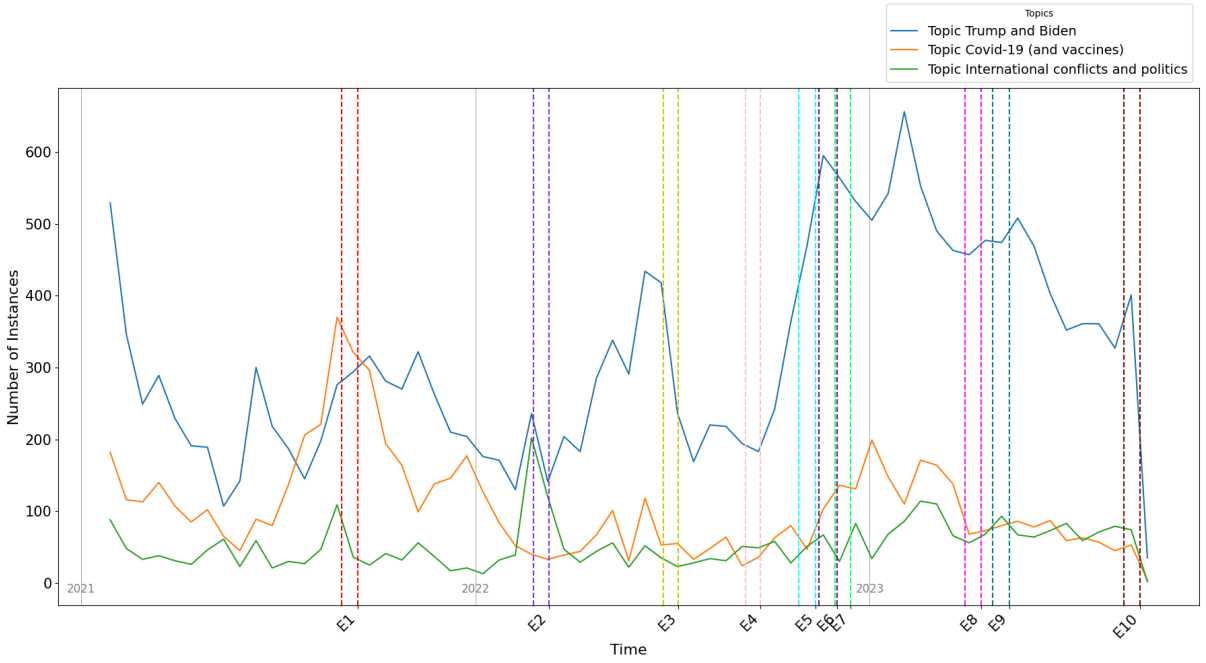

Figure 16. Timeline graph for topics *Trump and Biden*, *Covid-19 (and vaccines)*, and *International conflicts and politics*

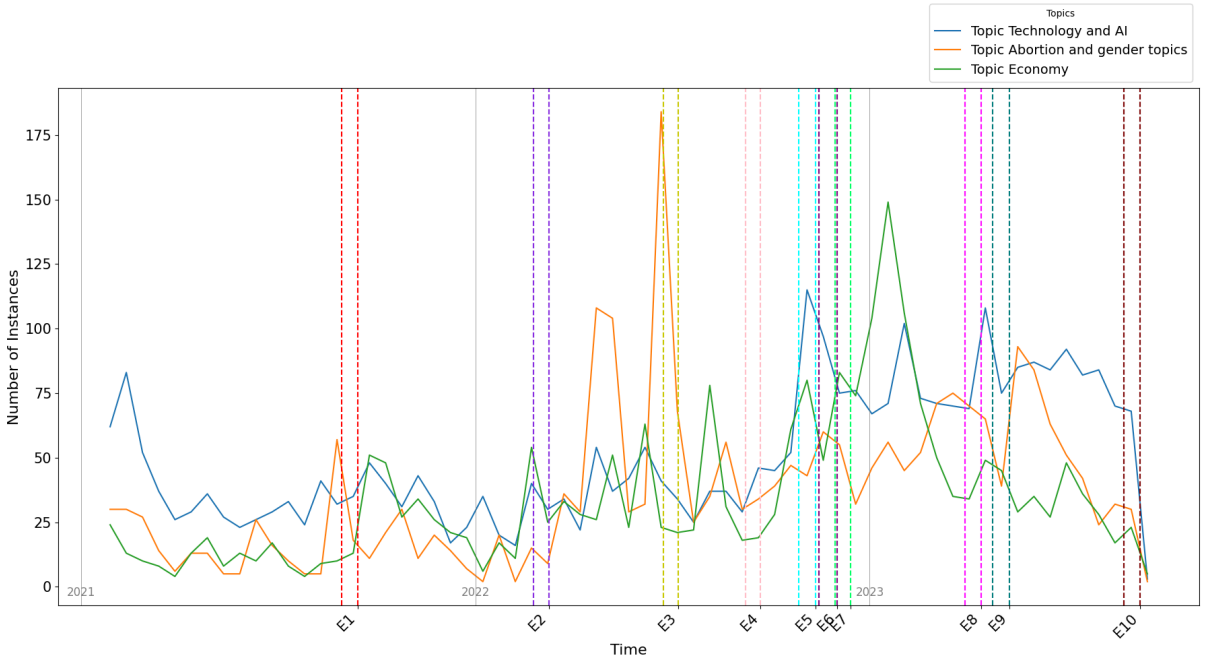

Figure 17. Timeline graph for topics *Technology and AI*, *Abortion and Gender topics*, and *Economy*

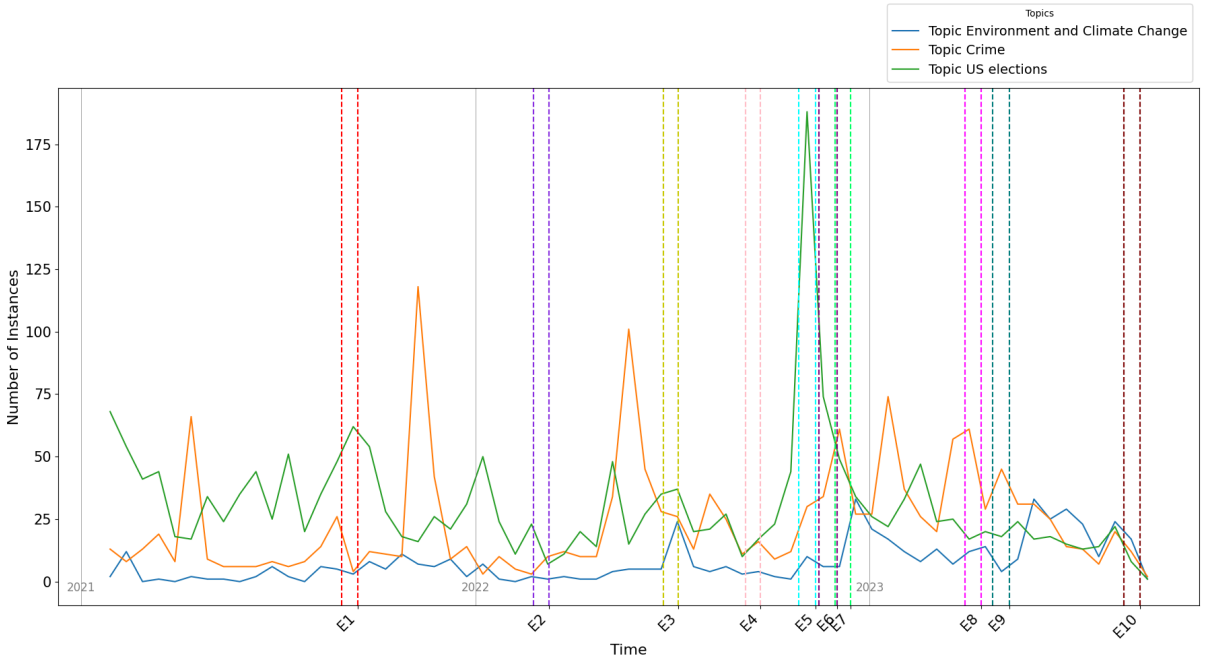

Figure 18. Timeline graph for topics *Environment and Climate Change*, *Crime*, and *US elections*

5 Conclusion

In this thesis, a study was conducted to analyze the content of *Twitter*'s fact-checking project called Community Notes. To achieve this, a method was represented to discover abstract "topics" by finding hidden semantic structures in the notes. This method involved an evaluation of three different topic models, from which one (*BERTopic*) was ultimately selected. During the implementation process, numerous challenges arose, such as filtering out noisy or redundant notes, implementing multiple models and comparing them, and fine-tuning a model using different techniques and parameters. With the topic model results, two subsets were made containing notes about US politicians to conduct a political party-based analysis along with a keyphrase analysis. Additionally, a temporal analysis with ten significant events was conducted. The main findings from each analysis are discussed next.

The analysis of the distribution graph between Democrats and Republicans revealed that the main talking point for both political parties was Joe Biden's and Donald Trump's rivalry. However, Democrats compared to Republicans had considerably more notes about topics *US elections* and *Economy*, which would refer to the claims made by some Republicans that the 2020 elections were dishonest and that Joe Biden was to blame for economic issues relevant at the time. Additionally, keyphrase analysis confirmed the findings from the distribution graph, while conveying further information about smaller topics, such as *Technology and AI* and *Abortion and gender topics*.

For the temporal analysis research was done to gather the 10 most significant events during the time period in which the notes were from. After analyzing whether the events correlated to spikes in the timeline, it was found that half of the events were not reflected in the notes. The other four events produced a spike in a topic related to that event, and only one event (The invasion of Ukraine) produced a spike in multiple topics. These findings were confirmed by analyzing the notes during that time period.

Although this study has engaged in a thorough exploration of the Community Notes, additional steps can be taken to further research on this topic. Firstly, notes about politicians could be contextually filtered for better and more accurate results. Secondly, significant events could be extracted from the notes. This approach would be more robust and potentially produce more meaningful findings. Thirdly, large language models could be explored to annotate topics instead of modeling them.

References

[1] Memon A. A, Vrij A, Bull R. Psychology and law: Truthfulness, accuracy and credibility. John Wiley & Sons, 2003.

[2] How to identify misinformation, disinformation, and malinformation (ITSAP.00.300) https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300 (08.03.2024)

[3] Duch Guillot J, Corlett N. TV still main source for news but social media is gaining ground. Press Releases, 2023. https://www.europarl.europa.eu/news/en/press-room/20231115IPR11303/tv-still-main-source-for-news-but-social-media-is-gaining-ground

[4] Dean B. Social Media Usage & Growth Statistics. 2024. https://backlinko.com/social-media-users (08.03.2024)

[5] Muhammed TS, Mathew SK. The disaster of misinformation: a review of research in social media. Int J Data Sci Anal. 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8853081/

[6] Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. The spreading of misinformation online. Proceedings of the National Academy of Sciences. 2016. https://www.pnas.org/doi/10.1073/pnas.1517441113

[7] Shahi GK, Dirkson A, Majchrzak TA. An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media. 2021. https://www.sciencedirect.com/science/article/pii/S2468696420300458

[8] Ferreira Caceres MM, Sosa JP, Lawrence JA, et al. The impact of misinformation on the COVID-19 pandemic. AIMS Public Health. 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9114791/

[9] Ng LHX, Cruickshank IJ, Carley KM. Cross-platform information spread during the January 6th capitol riots. Soc Netw Anal Min. 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9461432/

[10] The New York Times. These Are the 5 People Who Died in the Capitol Riot. https://www.nytimes.com/2021/01/11/us/who-died-in-capitol-building-attack.html (19.03.2024)

[11] Twitter. About Community Notes on X. https://help.twitter.com/en/using-x/community-notes (19.03.2024)

[12] Twitter Community Notes. Under the Hood: Download Data. https://communitynotes.twitter.com/guide/en/under-the-hood/download-data (18.10.2023)
[13] Dizikes P. Study: On Twitter, false news travels faster than true stories [Internet]. MIT News Office. 2018. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308 (25.02.2024)
[14] Hamed SK, Ab Aziz MJ, Yaakub MR. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon. 2023. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10539669/
[15] Mayank, M., Sharma, S., & Sharma, R. DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection. 2021. https://arxiv.org/pdf/2107.10648.pdf
[16] Akhtar, P., Ghouri, A.M., Khan, H.U.R. et al. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. Ann Oper Res 327, 633–657. 2023. https://doi.org/10.1007/s10479-022-05015-5
[17] Abdali S, Shaham S, Krishnamachari B. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. 2024. https://arxiv.org/pdf/2203.13883.pdf
[18] Ying Q, Zhou Y, Qian Z, Zeng D, Ge S. Multimodal Fake News Detection with Adaptive Unimodal Representation Aggregation. 2022. https://arxiv.org/pdf/2206.05741v1
[19] Landauer TK, Foltz PW, Laham D. Introduction to Latent Semantic Analysis. Discourse Processes. 1998. http://wordvec.colorado.edu/papers/Landauer_Foltz_Laham_1998.pdf
[20] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature. 1999. https://doi.org/10.1038/44565
[21] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. In: The Journal of Machine Learning Research. 2001. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
[22] Angelov, D. Top2Vec: Distributed Representations of Topics. 2020. https://arxiv.org/pdf/2008.09470
[23] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022. https://arxiv.org/pdf/2203.05794
[24] Antypas D, Ushio A, Camacho-Collados J, Neves L, Silva V, Barbieri F. Twitter Topic Classification. 2022. https://arxiv.org/pdf/2209.09824.pdf
[25] Khan Q, Chua HN. Comparing Topic Modeling Techniques for Identifying Informative and Uninformative Content: A Case Study on COVID-19 Tweets. In: 2021 IEEE

International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). 2021. https://ieeexplore.ieee.org/document/9573878
[26] Twitter Community Notes. Introduction. https://communitynotes.x.com/guide/et/about/introduction (27.02.2024)
[27] Pilarski M, Solovev K, Pröllochs N. Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. 2023. https://arxiv.org/pdf/2305.09519.pdf
[28] Pröllochs N. Community-Based Fact-Checking on Twitter's Birdwatch Platform. Proceedings of the International AAAI Conference on Web and Social Media. 2022. https://ojs.aaai.org/index.php/ICWSM/article/view/19335/19107
[29] Python. datetime — Basic date and time types. https://docs.python.org/3/library/datetime.html (18.10.2023)
[30] Python. re — Regular expression operations. https://docs.python.org/3/library/re.html (18.10.2023)
[31] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html (25.10.2023)
[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html (25.10.2023)
[33] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022. https://arxiv.org/pdf/2203.05794
[34] NLTK. Documentation. https://www.nltk.org/ (28.12.2023)
[35] BERTopic. Getting Started. Dimensionality Reduction. https://maartengr.github.io/BERTopic/getting_started/dim_reduction/dim_reduction.html (28.12.2023)
[36] U.S. Embassy & Consulate in the Kingdom of Denmark. Presidential Elections and the American Political System.

https://dk.usembassy.gov/usa-i-skolen/presidential-elections-and-the-american-political-system/ (02.05.2024)
[37] Yougov. The Most Popular Politicians (Q1 2024). https://today.yougov.com/ratings/politics/popularity/politicians/all (05.03.2024)
[38] Cavnar W, Trenkle J. N-Gram-Based Text Categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval. 2001. https://www.let.rug.nl/vannoord/TextCat/textcat.pdf
[39] KeyBERT. https://maartengr.github.io/KeyBERT/ (09.04.2024)
[40] NLTK. WordNet. Documentation. https://www.nltk.org/_modules/nltk/stem/wordnet.html (09.04.2024)
[41] The White House. U.S Withdrawal from Afghanistan. Washington, DC. 2021. https://www.whitehouse.gov/wp-content/uploads/2023/04/US-Withdrawal-from-Afghanistan.pdf (09.05.2024)
[42] Ray M. Russia-Ukraine War. Encyclopedia Britannica [Internet]. 2024. https://www.britannica.com/event/2022-Russian-invasion-of-Ukraine (09.05.2024)
[43] Totenberg N, McCammon S. Supreme Court overturns Roe v. Wade, ending right to abortion upheld for decades. NPR. 2022. https://www.npr.org/2022/06/24/1102305878/supreme-court-abortion-roe-v-wade-decision-overturn (09.05.2024)
[44] BBC. Queen Elizabeth II has died. 2022. https://www.bbc.com/news/uk-61585886 (09.05.2024)
[45] The New York Times. Elon Musk Completes $44 Billion Deal to Own Twitter. 2022. https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html (09.05.2024)
[46] Kurtzleben D. Trump announces 2024 presidential run. 2022. https://www.npr.org/2022/11/15/1137052704/trump-2024-president-campaign (09.05.2024)
[47] OpenAI. ChatGPT: A conversational AI model. 2021. https://openai.com/index/chatgpt/ (09.05.2024)
[48] O'Kruk A, Merrill C. Donald Trump's criminal cases, in one place. CNN politics. 2024. https://edition.cnn.com/interactive/2023/07/politics/trump-indictments-criminal-cases/ (09.05.2024)
[49] Montanaro, D., Keith, T., & Bustillo, X. Biden warns of rights under threat from Trump and 'MAGA extremists' in reelect launch. 2023.

https://www.npr.org/2023/04/25/1145679856/biden-president-announcement-2024-running-re election (09.05.2024)
[50] Burnett, S., Colvin, J., & Cooper, J. J. Republican candidates fight each other, and mostly line up behind Trump, at the first debate. AP News. 2023. https://apnews.com/article/first-republican-debate-2024-elections-gop-89812d5aa1ed6a4ebe7373ff36858250 (09.05.2024)
[51] Judiciary House. Testimony Reveals FBI Employees Who Warned Social Media Companies about Hack and Leak Operation Knew Hunter Biden Laptop Wasn't Russian Disinformation. Press Release. 2023. https://judiciary.house.gov/media/press-releases/testimony-reveals-fbi-employees-who-warned-social-media-companies-about-hack (09.04.2024)

Appendix

I. License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Jaan Kupri,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Topics of Fact-Checking on Twitter Community Notes,

supervised by Uku Kangur and Roshni Chakraborty.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Jaan Kupri
16/05/2024