Gab: Understanding an Alt-Right Social Network Sentiments and Users Interactions

Zhaolin Fang
2725173
LSDE Group 12
Vrije Universiteit Amsterdam
z2.fang@student.vu.nl

Leonardo Kuffo Rivero
2722539
LSDE Group 12
Vrije Universiteit Amsterdam
l.x.kufforivero@student.vu.nl

Xuan Zhou
2698592
LSDE Group 12
Vrije Universiteit Amsterdam
x9.zhou@student.vu.nl

ABSTRACT

Gab is a unique social network that do not moderate hate-speech content generated by their users. Hence, it has created an unprecedented opportunity to capture and analyze a different type of online speech regarding political news and world events. Existing research on Gab is focused on exploratory analysis of the platform. Currently, there is no research regarding Gab users speech such as discovering and analyzing topics sentiments and users interactions. On this work, using 20M messages from Gab and computation power from Databricks, we discovered what topics are Gab users talking about and how they change over time. At the same time, we analyzed the topics sentiments, linguistic dimensions and interactions between users. We discovered that topics in Gab are driven by real-world events and that the main topics and their sentiments change quickly over time. Furthermore, we compared Gab speech about certain topics to the speech of users from Twitter. The latter hinted Twitter messages are more emotionally charged than Gab's. And, in some topics, different polarization of sentiments was found between both social networks.

The following document is the final report of group 12 for Large Scale Data Engineering Assignment 2.

1 INTRODUCTION

Recently, social networks such as Facebook, Twitter and Youtube are doing huge efforts to remove *hate speech* from their platforms [1] [2]. These efforts have gone as far as banning thousands of users due to *hate speech*, including the accounts of the ex-president of the United States of America, Donald J. Trump [3]. The United Nations defines *hate speech* as "any kind of communication that attacks or uses discriminatory language with reference to a person or a group on the basis of who they are (i.e. religion, ethnicity, nationality, etc.)" [8]. Therefore, it is no surprise that in order to be advertiser-friendly, these platforms invest resources on detecting and removing hate speech within the huge amount of content produced everyday. However, some of these platforms users allege that this persecution against hate speech is an assault to *freedom of speech*. This controversy has led them to search for alternative social platforms in which they can have *freedom of speech*, even if that speech is categorized as *hate speech*. One of the most famous ones being Gab[4].

Gab is an open source social platform founded by Andrew Tolba and Ekrem Bjuyyukkaya in August 2016. At the date of writing this work, Gab has a total of 4 million users, from which 100K use the platform everyday. This popularity relies heavily on their support to freedom of speech. For this reason, Gab has specially attracted a demographic of alt-right users who have been banned or suspended from other platforms. This includes Donald J. Trump, which is currently an active user of Gab. Gab allows users to post messages of up to 3000 characters, called GABs. Since Gab does not moderate *hate speech*, it is common to find edgy messages containing xenophobia, sexism or racial slurs. Although it is important to say that Gab controls illegal content such as pornography, promotion of terrorism and selling weapons or drugs.

Up until now, research on Gab has been focused on exploratory analysis regarding posts content (i.e. What are the most hashtags used? How many hate words does posts contains?) and users interaction networks analysis (i.e. What are the most central users to Gab *who-follows-who* network?). However, little research has been done about the sentiments and in-depth speech analysis of Gab content. In addition to this, no work has been done on discovering what are Gab users talking about without using the already existing 'Topics' that users occasionally gives to their own posts.

On this work, we propose a framework to analyze the sentiments and linguistics dimensions of Gab messages. In order to do this, we use Unsupervised Machine Learning to discover what topics are Gab users talking about (i.e. topic discovery). On top of this, we analyze how users interact inside each topic (i.e. networks analysis). Finally, we compare Gab users speech with other social network (i.e. Twitter) speech on the same topic. On doing so we are eager to answer the following **research questions**:

- Q1.1: How the discovered topics differs from the already existing user-generated topics inside the social network?
- Q1.2: How did the major topics and the messages' sentiment change over time in the Gab network?
- Q1.3: Are topics in Gab correlated to real-world events or are they mainly driven by the narrative of the network?
- Q2.1: Is Gab users speech about topics any different from other social network users speech?
- Q3.1: How is Gab users interaction network at a platform-level?
- Q3.2: How are Gab users interaction networks at a topic-level?

This work is structured as follows: Section 2 describes the used dataset of Gab posts to conduct the research. Next, Section 3 presents previous work in which Gab was used as subject of analysis. Furthermore, this section introduces successful cases that uses techniques that we are going to use in our methodology. Section 4 describes the methodology and technology used to analyze Gab posts content

[1]https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/
[2]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[3]https://www.bbc.com/news/world-us-canada-57365628
[4]https://gab.com/

and to build the users interaction networks. Section 5 presents and discuss the results of our methodology. In addition to this, it also presents the computation times and encountered challenges of the methodology. Finally, in Section 6, we close the work by presenting conclusions and suggestions for future research.

2 DATASET

Our entire dataset consist on roughly 20M Gab posts authored between August 2016 and May 2018 (i.e. 21 months), ordered by their creation date. Posts can be of four different types: Standalone posts, replies (i.e. a response from one user to another user standalone post), quotes (i.e. a standalone post which quotes another user standalone post) and re-posts (i.e. a user sharing another user standalone post). We say there is a *conversation* when a standalone post receives one or more replies. Replies were implemented in Gab only since April, 2017. Standalone posts can be published on the user's own profile or on Gab groups. Gab groups, very similarly to Facebook groups, are closed communities of users that shares a common goal or discuss about certain specific topics.

For each post we can find the following information: a unique identifier, author information (i.e. user unique id, username), number of replies, number of likes, number of dislikes, score (i.e. number of likes minus number of dislikes), raw text content, media attachments information, Gab category, Gab topic, authored date and time, and the conversation tree (i.e. all the information of the posts inside the conversation). 29.80% of the posts are replies. Meanwhile 10.99% of them are quotes. Re-posts only represent 1.30% of the data. The rest of the posts (57.91%) are standalone posts. Posts text content have 17.63 words in average (i.e. 1 to 2 sentences in English). 50% of the posts have between 4 and 24 words with a median of 11 words, with very few posts exceeding 55 words. This is an indicator that the majority of posts in Gab are not syntactically complex. In fact, posts content length is comparable to the ones in Facebook and Twitter.

2.1 Gab Topics and Categories

Inside Gab, posts can be automatically classified within a topic or/and a category. This classification is not mandatory, hence, there are posts without this information. Topics are created by users based on events or situations that they are eager to discuss with the community (e.g. Yountville Shooting, Activist Martin Sellner Detention). Topics are publicly available, and users can classify their standalone posts as related to that topic. This mechanic was introduced in 2017-05. In our dataset, only 20.86% of posts have a topic assigned to them. Figure 1 shows the top 10 topics based on number of messages talking about that topic for a sample taken between March 10 and March 13, 2018. We can clearly see topics related to real news such as: Trump Visits Border Wall, Martin Sellner Arrested, Tillerson Ousted and Yountville Shooting. This hint us that the discussions inside the social network are partly driven by real-world events (Q1.3). There are also topics related to leisure such as MEME WARS, Introduce Yourself and Memes. Finally, we also discovered topics that are too broad to be considered for individual analysis such as: North Korea, BritFam, South Africa and Deutsch.

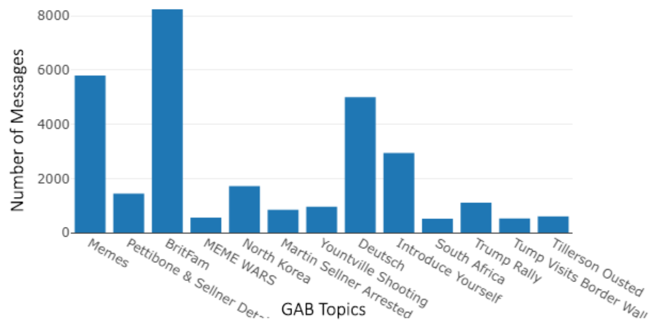

Figure 1: Bar chart showing the 10 most popular topics in Gab based on number of posts related to that topic authored between March 10 and March 13, 2018.

On the other hand, categories are defined by Gab itself. At the time of writing this work, there are currently 15 categories and the way Gab determines posts categories is entirely based on the category of a groups in which a post is published. Hence, only posts made in Gab groups have a category related to them. Standalone posts made in users profiles do not have a category. Figure 2 shows the messages distribution by category in a sample taken between March 10 and March 13, 2018. We can see how AMA (i.e. Ask Me Anything), News, Politics and Humor are the categories with the most active groups inside the network. **News and Politics** are the categories we are willing to focus for this work. We hypothesize that the discussed topics on groups of these categories are able to reveal patterns and behaviours inside Gab that may differ from other social networks such as Twitter or Facebook. It is important to say that Categories were introduced in Gab in 2016-11.

2.2 Gab Scoring System

Posts inside Gab can be liked or disliked by other users. By exploiting these metrics, Gab implemented a user reputation system. A user reputation is calculated based on how many times all of their posts have been liked minus how many times all of their posts have been disliked. As a matter of fact, only users with more than 250 overall score are able to dislike other users posts. The average user score is 30.86, and the median is 4, and 75% of the people have a score lower than 15. This user score is an important metric that can be exploited since it may reveal the *attention* of the network towards users and topics.

2.3 Dataset Filtering

We are interested on analysing standalone posts and quotes. Re-posts do not contains speech belonging to the users themselves. Furthermore, threaded replies are usually short, and can be out of context. Hence, we excluded re-posts and replies from our analysis. Furthermore, we are going to focus on posts categorized as **News or Politics**. We hypothesize that the discussed topics on groups of these categories are able to reveal behaviours inside Gab that may differ from other social networks. In addition to this, these categories are prone to contain controversial topics with polarized and edgy opinions. Next, we filter posts that contained less than

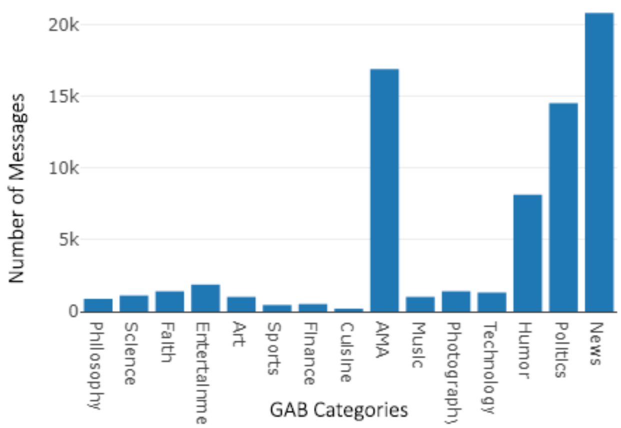

Figure 2: Bar chart showing the 10 most popular categories in Gab based on number of posts related to that category authored between March 10 and March 13, 2018.

5 words as a polluted-content filtering strategy. Afterwards we removed posts from May 2018 since we only have available a few posts from the first days of that month. Finally we removed posts from August, September and October 2016 since at that time Gab categories were not implemented.

We end up with 1'918,155 posts. Authored by 43,172 different users between November 1, 2016 and April 30, 2018 (i.e. 18 months). The top 10 users of the network based on their posts scores are presented in Table 1. Users average a total of 44.43 messages.

There are 6,011 Gab pre-existing topics. However, 50% of these topics have less than 19 messages. Due to the low quantity of messages, these topics are not suitable for analysis. Hence the importance of discovering the true topics of the platform (Q1.1). On each topic there is an average of 31.21 users which posted about that topic.

Username	User Score	# of Messages
TukkRivers	260186	3726
USMC-DevilDog	242440	5102
Don	189428	6164
Kek_Magician	152422	9743
RealTrumpTweets	144324	1523
genophilia	130243	2311
GuardAmerican	93984	8041
truthwhisper	86926	4356
SurvivorMed	85773	3207
KetzerHexe	83886	6223
a (Gab CEO)	83376	388

Table 1: Top 10 users in Gab based on their user score obtained on standalone posts inside the News and Politics category.

3 PREVIOUS WORK

One of the main goals of this project is to perform sentiment analysis on Gab posts text content. Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics to systematically identify, extract, quantify, and study affective states and subjective information[5], such as people's views, emotions, evaluations and attitudes towards products, services, organizations, individuals, problems, events, topics and their attributes.

Existing research has produced a large number of techniques that can be used in multiple tasks of sentiment analysis, including supervised and unsupervised methods. In the supervised methods, early papers used methods such as support vector machine, maximum entropy, naive Bayes, and feature combination. Unsupervised methods include different methods which uses affective dictionaries, grammatical analysis and syntactic patterns [19]. There are three kinds of granularity of sentiment analysis: document granularity, sentence granularity and aspect granularity:

- Document level sentiment analysis refers to marking the overall emotional tendency / polarity from the documents viewpoint, that is, determining whether the document conveys positive or negative views as a whole.
- Sentence level sentiment analysis is used to calibrate the expressed sentiment in a single sentence. The sentiment of sentences can be inferred by subjective classification and polarity classification.
- Aspect level is also called topic level. Each aspect represents a topic. Different from document level and statement level sentiment analysis, aspect level sentiment analysis considers both sentiment information and topic information (sentiment generally has a topic).

The analysis of posts in social network Gab belongs to document level sentiment analysis. The sentiment classification of film reviews based on supervised classification and multilingual features proposed by Pang et al.[16] in 2002 was the start of sentence-level sentiment analysis, followed by the CNN text classification proposed by Kim et al. [12] in 2003, which is also known as one of the important baselines of sentence-level sentiment classification tasks. In addition to the work of Kim et al., ISTM-RNN, FastText and other neural network models also perform well in sentence-level sentiment classification.

In this work we are going to use Vader to analyse the sentiment of our Gab posts at a document level. Vader is a lightweight, fast and a widely used rule-based model for sentiment analysis specialized in short text data such as the one found in social media[11]. Vader has previously been used successfully in measuring the sentiment of content posted in social media such as Twitter[7],[4]. In addition to this, to perform a more detailed topics speech characterization of Gab users we are going to use the *Linguistic Inquiry and Word Count (LIWC)* software. LIWC is a text analyzer computer program based on word frequency counts that computes a score for psychologically meaningful categories, including: thoughts; feelings; personality; motivations; thinking styles; and social insights starting from raw text[19]. LIWC has been previously used successfully

[5]https://en.wikipedia.org/wiki/Sentiment_analysis

when analysing data from social media sites such as Twitter [20] and Facebook [9].

As shown in Section 2, Gab pre-existing topics can be a toss of a coin. They are created by users and only 20.86% of the post have a topic assigned. In addition to this, some topics are too broad or lack of meaning to be analyzed (e.g. North Korea, South Africa). Hence, we are eager to apply unsupervised learning in order to discover the true topics that are hidden among Gab posts. This is known as Topic Discovery. Topic discovery algorithms tries to identify groups of elements that share characteristics between them, without knowing a-priori the characteristics of these groups. Latent Dirichlet Allocation (LDA) is an implementation of Topic Discovery [3] that has been previously used successfully to discover topics in text corpus from social media such as Twitter [15]. In [21], LDA is used in a Twitter text corpus to discover topics within users discourse during the COVID-19 pandemic.

In terms of previous works on social network analysis, Kwak et al.[13] are among the first to study Twitter, aiming to understand its role on the Web. They show that Twitter is a powerful network that can be exploited to assess human behavior on the Web. Zannettou et al.[23] study how mainstream and alternative news propagate across multiple Web communities, measuring the influence that each community have on each other. With the same multi-platform point of view, Chandrasekharan et al.[5] propose an approach, called Bag of Communities, which aims to identify abusive content within a community. Hine et al.[10] study 4chan's Politically Incorrect board, and show that it attracts a high volume of hate speech.

Savvas et al.[22] study the Gab social network, analyzing what kind of users it attracts, what are the main topics of discussions, and to what extent Gab users share hateful content. The latter work also analyses users popularity in Gab by building a graph of *who-follows-who* and finding the most important nodes to the network using the PageRank algorithm. Metrics of centrality inside networks can also be used to reveal users that highly influence others with their speech[2]. In our work, we will follow a similar approach but using a *who-replies-who* network. Applying graphs communities detection algorithms, such as Label Propagation Algorithm (LPA) [17], on this kind of graphs can reveal how the network is distributed, why users interact between each other and what characteristics those communities of users share[1].

4 METHODOLOGY

In order to answer our research questions we divided our analysis in three tasks: A) Posts content analysis, B) Gab vs Twitter, C) Users interaction network analysis. Computing will be carried out in a Spark Cluster running on Databricks. Cluster's nodes were i3.xlarge instances provided by AWS (i.e. 4vCPU, 30.5 GiB of RAM). The cluster was composed of 1 driver node and 8 worker nodes.

4.1 Posts content analysis

First, we are going to partition our data into months to reduce our problem complexity. Second, we will use Vader and LIWC libraries to find overall sentiments and linguistic dimensions from our Gab posts. Next, we will find the topics that Gab users discusses using an unsupervised machine learning technique called Latent Dirichlet Allocation (LDA). However, to use LDA the text from our Gab posts needs to be suitable for analysis (i.e. text normalization and vectorization). Finally, the found topics were aggregated and metrics such as total number of likes, total number of dislikes and average sentiments, are computed in order to find topics insights. In the following subsections we will explain each pipeline step in detail.

4.1.1 Data Partitioning. Gab data is temporally distributed over 18 months. Detecting topics on these 18 months as a whole would be complex due to the amount of topics to be found. Furthermore, manually characterizing these topics would also have to be done in a temporal fashion (i.e. discover to which period of time the topics belong to). For this reason, we partitioned our dataset into monthly periods using the date the posts were published. The next steps were carried out for each month in our dataset.

4.1.2 Sentiment Analysis. In order to find the overall sentiments of each post we used Vader. Vader is a rule-based model which analyze if a piece of text is negative, positive or neutral. The latter is represented as a score ranging from -1 (i.e. negative) to 1 (i.e. positive). If the score is close to 0, the message is neutral from a sentimental point of view. In addition to this we used LIWC to measure the posts linguistic dimensions. We focus our analysis on the following dimensions: Negative emotions, positive emotions, social concerns, anger, sadness, anxiety, drives, death, religion and time concerns. For each message, LIWC represents dimensions score as a percentage of words in the message related to each dimension. Both Vader and LIWC libraries are going to be applied to the text of the posts in our dataset using a User Defined Function (udf)[6] in Spark.

4.1.3 Text Normalization and Vectorization. The informal writing style used in social networks can negatively affect the performance of machine learning algorithms to analyze text [18]. Hence, our text corpus must be normalized. In order to do this, we follow an approach commonly applied to content extracted from the web: Emojis, URLs, punctuation signs, symbols and stopwords (e.g. where, and, or, by) are removed. Furthermore, accents are brought to their canonical form (e.g. *à* becomes *a*).

Machine Learning algorithms only understand text when it is represented as a vector. Hence, we transformed our text corpus into a document-term matrix M, where each row represents one document, i.e. a Gab post text, each column represents a feature, i.e. a word/term present on the corpus, and M_{ij} represents the Term Frequency-Inverse Document Frequency (TF-IDF) score for the word$_j$ with respect to the document$_i$. TF-IDF is a metric that depicts how important a word is to a document inside a text corpus [14]. This score is comprised by two components: Term Frequency and Inverse Document Frequency, described in equations 1 and 2 respectively.

$$TF = \frac{ft_{w,d}}{\max\{ft_{w,d} : w \in d\}} \tag{1}$$

$$IDF = log\left(\frac{|D|}{n_w}\right) \tag{2}$$

where,

[6]https://spark.apache.org/docs/latest/sql-ref-functions-udf-scalar.html

- $ft_{w,d}$ is the number of times that *word w* occurs in *document d*.
- n_w is the number of documents in which *word w* is present.
- |D| is the total number of documents in the text corpus.

The Term Frequency (TF), captures the local frequency of a word in a document, meanwhile Inverse Document Frequency (IDF) captures the relevance of a word in the entire text corpus. The TF-IDF score is obtained computing the product of these two components.

Spark provides an implementation of text vectorization using TF-IDF as part of their Machine Learning package (Mlib)[7].

4.1.4 Topic Discovery. Once text is suitable for analysis, we applied Latent Dirichlet Allocation, an unsupervised machine learning algorithm to find topics. By doing so we aim to answer **research question Q1.1**. LDA receives as input the document-term matrix M and assign each document of the matrix into one of the detected topics. It is important to highlight that this technique needs to know the number of topics a-priory. The latter would sound counter-intuitive since we are trying to discover topics from data. However, there are metrics to help us inferring the number of topics based on the topics distribution resulting from LDA. One of them is called Perplexity. Perplexity is a statistical metric that measures how well the LDA model predicts a sample[6]. Therefore, the process of finding the optimal number of topics requires to compute an LDA model for a range of possible number of topics. We performed this for each month of data. The optimal number of topics must be inferred by the researcher by plotting a Number of Topics vs Perplexity line plot. As a rule of thumb, the point in which the perplexity reaches a valley or stabilization is the optimal number of topics.

Spark provides an implementation of LDA as part of their Machine Learning package (Mlib)[8]. Furthermore, the same implementation includes a method to compute the model perplexity metric.

4.1.5 Topics Insights. In order to get insights for our topics, we group all their posts and compute aggregated metrics on them. The computed aggregated metrics are depicted in Table 2. Using this insights we aim to answer **research question Q1.2**.

4.1.6 Real world news correlation. In our aim to answer **research question Q1.3**, we performed a manual research of the top 3 ranked topics per month based on their aggregated score. Using Google advanced search we tried to find news related to these topics between the dates of the topic period.

4.2 Gab vs Twitter

We used Twitter as the social network to perform sentiments comparisons with Gab to answer **research question Q2.1**. We did not use Facebook due to their restrictions on data extraction. To extract data from Twitter, we used the Twitter API[9] and Tweepy Python library[10]. Using the API search parameters we are able to extract data from the specific time in which topics occurred. For each tweet we stored its creation date, and the tweet raw text. Finally, for each twitter topic text data we repeated the step 4.1.2 and 4.1.5 of the aforementioned pipeline (i.e. Sentiment Analysis and Topics insights). Using the insights of both Gab and Twitter messages we performed sentiments comparisons. For this analysis we only focused on three topics chosen from the most relevant topics based on their aggregated score and their nature.

Field Name	Description
Messages	Topic total number of messages
Score	Topic messages total score
Likes	Topic messages total number of likes
Dislikes	Topic messages total number of dislikes
Keywords	A map containing the top 50 terms that occurred in one topic
Sample	A list containing the 20 most representative messages for the topic
Users	A list containing the 5 most relevant users for the topic based on the score their posts received
Sentiment	Average Vader score for the topic messages
LIWC Sentiment	A map containing the average score for liwc dimensions for the topics messages (e.g. "anger:6.9,anxiety:4.1")

Table 2: Metrics aggregated to obtain topics insights.

4.3 Users interaction network analysis

In order to answer **research question Q3.1 and Q3.2** we built an interaction graph based on the *who-replies-who* relationship depicted in the posts replies. This relationship will help us build a graph G in which nodes (also called vertex) represent users and edges represent if an user have replied to another user post. It is important to say that due to data limitations, we are not able to built a *who-replies-who graph*. We built our graph G mainly at a topic-level. However, we also built a graph G at a platform-level to analyze the platform users interaction as a whole.

To analyze users interactions we first filter the graph giant component. A graph giant component is defined as the biggest connected component of a graph. In doing so we filter conversations between a small number of users outside of the topic conversation network. For example, two users that interacted with each other but did not interact with anyone else in the topic network are going to be filtered out from the analysis.

Next, we run the PageRank algorithm on the graph to find the most relevant users of the networks based on their interactions. PageRank algorithm gives weights to the users recursively depending on the weight of the users that connects with them. Therefore, PageRank does not only consider the number of interactions a user have had, but with *who* they interacted. For example, a user which interacted with two high-relevant users could obtain a *higher* score than a user that interacted with ten users with low relevancy.

Finally, we detected communities (i.e. group of users) in the network by applying Latent Propagation Algorithm (LPA) on the graph. LPA advantage is that is has a linear complexity. However, its biggest drawback is that it is prone to produce trivial solutions (i.e. individual nodes being identified as communities). By using LPA

[7]https://spark.apache.org/docs/latest/mllib-feature-extraction.html
[8]https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.LDA.html
[9]https://developer.twitter.com/en
[10]https://docs.tweepy.org/en/stable/

Period	Discovered topics	Gab Pre-existing topics
2017-05	Conspiracy Theory on the Murder of Seth Rich | Manchester Arena Bombing Terrorist Attack Victims | Manchester Arena Bombing Terrorist Attack"	Seth Rich | Manchester | James Comey
2017-06	Various International News | Longon Bridge Terrorist Attack | Dismissal of James Comey	London | International News | Alexandria
2017-07	Deutsch | USA News | 4th of July (USA Independence Day)	International News | CNN Blackmail | Deutsch
2017-08	North Korea vs Trump | Barcelona Terrorist Attacks | Deutsch	Hurricane Harvey | Charlottesville | Deutsch
2017-09	NFL (National Football League) Season Start | North Korea vs Trump | Hurricane Harvey	DACA | Regulate Big Tech | Torba on Tucker
2017-10	Las Vegas Shooting | Hillary Clinton vs Harvey Weinstein | Trump in Military Event	Las Vegas Terror Attack | Vegas Shooting | Weinstein
2017-11	Roy Moore sexual misconduct allegation | Trump Twitter Account Deactivation | San Antonio Baptist Church Shooting	Judge Roy Moore | San Antonio Baptist Church Shooting | Trump Asia Tour
2017-12	USA Tax Cuts and Jobs Act of 2017 | USA Alabama Elections Results | Trump in Jerusalem	Tax Cuts and Jobs Act | Flynn Pleads Guilty | Alabama Election Results
2018-01	Nunes Memo | DACA: The Trump Immigration Plan | Conspiracy Theory: The Storm	Release the Memo & DACA | SOTU LIVE Discussion | FISA Misconduct
2018-02	Stoneman Douglas High School shooting (Florida | USA) | Trump Proclaims February 2018 as National African American History Month | Social Media Platforms Tighten Rules	Florida School Shooting | FISA MEMO | Gun Control Push
2018-03	Racism | Youtube Censorship | President Trump News	Gun Control Push | Censorship | Facebook Probe
2018-04	Trump Attacks Syria | James Comey take on Hillary Clinton | Jews Immigration	Syria | Trump Strikes Syria | Central American Caravan

Table 3: Comparison between the top three ranked topics based on messages score of discovered topics and Gab topics.

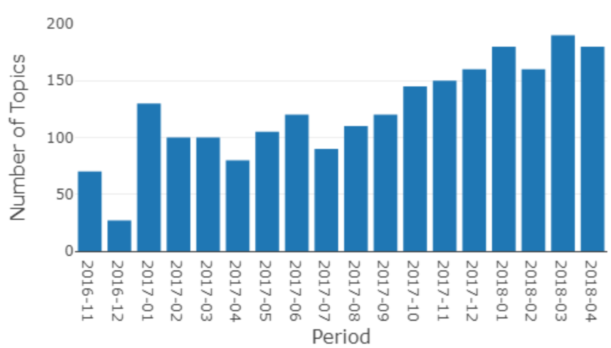

Figure 3: Optimal number of topics per period found using the perplexity criterion to our document corpus.

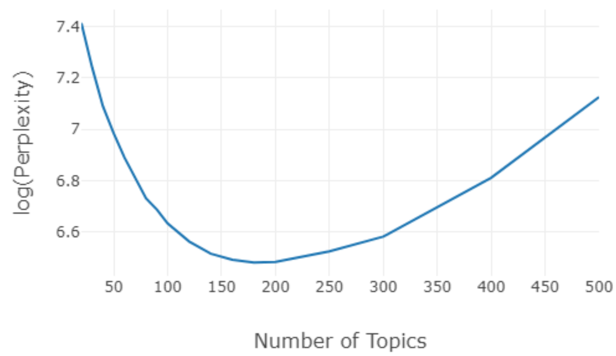

Figure 4: Perplexity evaluation for the 2018-04 period. At 180 topics, the LDA model probability distribution is the most optimal to describe the text corpus.

we are able to detect interactions between groups (i.e. communities) of users, but at the same time we can detect interactions between micro-comunities of users inside the topic discussion.

Graphs representations in Spark and algorithms (i.e. Giant Component, PageRank and LPA) are implemented in the GraphFrames package [11].

5 RESULTS

5.1 Posts content analysis

Figure 3 shows the optimal number of topics found for each period in our dataset. The optimal number of topics range from 27 to 190 with an average of 123.16 and a standard deviation of 43.05. Hence the importance of performing a perplexity analysis for each period.

[11]https://graphframes.github.io/graphframes/docs/_site/index.html

Furthermore, we can see a tendency of growth in the number of topics through time. This growth aligns with Gab user-base growth during these months. Figure 4 shows a plot of log(Perplexity) vs Number of Topics. This plot is useful to obtain the optimal number of topics for the 2018-04 period. The plot hint us that at around 180 topics the model performs the best. Using the optimal number of topics we proceeded to run LDA for each period of our data. Table 3 shows a comparison between the three most ranked topics for each period using Topic Discovery and the pre-existing topics created by Gab users. By dividing the number of topics that appears in both groups to the total number of topics we can compare how similar our found topics are to Gab topics. These comparisons can only be done from 2017-05 until 2018-04 (i.e. 12 months). The latter due to Gab topics mechanic being introduced in 2017-05.

	Aggregated Topic Metrics						
Period	**Topic**	**Score**	**Messages**	**Keywords**	**Top User**	**Sentiment**	**Linguistic**
2016-11	2016 United States Presidential Election	22,801	2,942	Trump, maga, speakfreely	BrittPettibone	-0.022	Social concerns: 9%, Drives: 8%
2016-12	Pizzagate Fake News in Social Media	70,402	9,603	News, pizzagate, fake	USSANews	-0.045	Social concerns: 9%, Drives: 8%
2017-01	President Trump News	39,557	7,273	Trump, news, speakfreely	USSANews	0.162	Social concerns: 8%, Drives: 8%
2017-02	President Trump News	27,821	7,553	Trump, news, draintheswamp	USSANews	-0.008	Social concerns: 10%, Drives: 8%
2017-03	President Trump News	34,946	5,930	Trump, news, speakfreely	USSANews	-0.028	Social concerns: 8%, Drives: 7%
2017-04	USA Inmigration	14,876	2,246	people, good, trump	RaviCrux	0.083	Social concerns: 10%, Drives: 9%
2017-05	Conspiracy Theory on the Murder of Seth Rich	16,848	2,039	Seth, Rich, murder	Fedeliskrieg	-0.018	Social concerns: 9%, Drives: 8%
2017-06	Longon Bridge Terrorist Attack	11,026	1,332	london, bridge, terror	shorty	-0.326	Drives: 10%, Social Concerns: 10%
2017-07	USA News*	10,550	1,836	trump, cnn, healthcare	RaviCrux	-0.074	Social concerns: 9%, Drives: 8%
2017-08	North Korea vs Trump	14,135	2,053	North, korea, trump	JoshC	-0.129	Social concerns: 10%, Drives: 9%
2017-09	NFL (National Football League) Season Start	16,857	2,280	nfl, anthem, players	RealTrumpTweets	-0.077	Social concerns: 10%, Drives: 8%
2017-10	Las Vegas Shooting	14,697	2,649	Vegas, shooting, police	Sperg	-0.170	Social concerns: 9%, Drives: 8%
2017-11	Roy Moore sexual misconduct allegation	17,795	1,616	Moore, Roy, judge	PatDollard	-0.079	Social concerns: 10%, Drives: 8%
2017-12	USA Tax Cuts and Jobs Act of 2017	14,420	1,810	taxs, cuts, senate	RealTrumpTweets	-0.018	Social concerns: 9%, Drives: 9%
2018-01	Nunes Memo	17,979	2,905	memo, fbi, state	GuardAmerican	-0.115	Social concerns: 9%, Drives: 8%
2018-02	Stoneman Douglas High School shooting (Florida, USA)	17,699	2,748	guns, people, teachers	RealTrumpTweets	-0.141	Social concerns: 11%, Drives: 10%
2018-03	Racism	13,992	2,018	south, africa, white	genophilia	-0.145	Social concerns: 10%, Drives: 9%
2018-04	Trump Attacks Syria	19,742	3,554	syria, attack, chemical	PNN	-0.370	Drives: 9%, Social concerns: 8%

Table 4: The most ranked discovered topics based on their messages aggregated score for each period in our dataset. For each topics the aggregated metrics are presented. Topics marked with an *, are not ranked first in their period. They are outranked by topics in other languages (e.g. Deutsch). These topics results are not reported since the used text analysis tools are optimized for the English language.

Among the top one most ranked topic we found a similarity of 66.67% (8 out of 12 topics). Among the top three ranked topics we found a similarity of 50.00% (16 our of 36 topics). These differences hint us that Gab existing topics do not depict the real narrative of the network. The latter could be due to users writing their opinion about topics without categorizing their posts into topics created by other users. These results answer our **research question Q1.1**.

Table 4 shows the most ranked discovered topics for each period and their aggregated metrics based on their messages score. During the 18 months the main topics of Gab changed constantly. These topics mainly covered themes of political news and terrorist attacks. It is important to highlight that from 2017-01 until 2017-03 the most ranked discovered topic was "President Trump News". These months correspond to the first months of mandate of the U.S.A. Ex-president Donald J. Trump. This was the only time in which a topic discussion extended for more than two months. Hence, demonstrating that Gab topics do not have a tendency to extend for long periods of times. Overall sentiment greatly vary between topics. Most of the topics resulted in neutral or negative sentiments. The only topic with positive sentiment is found in 2017-01 (i.e. President

Trump News). Interesting enough, this same topic sentiment was reduced on the next periods. The most negative topic can be found in 2018-04 (Trump Attacks Syria). Regarding linguistic dimensions, all of the topics speech predominately contains social concerns and drives related terms. These results answer our **research question Q1.2**.

Finally, 100% of the top three ranked topics on each period are strongly related to real-world events. All of them relate to news covered by digital media. Thus, answering our **research question Q.1.3**.

5.1.1 Computation. Table 5 describes the computation needed to perform the posts content analysis. The most expensive task was the perplexity analysis to determine the optimal number of topics for each one of our periods.

Task	**Time taken (hours)**	**Estimated Price (USD dollars)**
Data Decompression	0.3	2.808
Data Filtering and Partitioning	1.4	5.616
Sentiment Analysis	0.1	2.808
Perplexity Analysis (+ Text Normalization and Vectorization)	108.4	304.39
Topic Discovery (+ Text Normalization and Vectorization)	5.4	16.848
Topics Insights	0.1	2.808
TOTAL	**115.7**	**$335.28**

Table 5: Computation of Posts Content Analysis methodology. Prices estimation assumes 9 i3.xlarge instances (i.e. 1 driver and 8 workers) running per task (i.e. 2.808 USD per hour or fraction).

5.1.2 Challenges. Vader and LIWC were implemented as user defined functions in Spark (udf). This function computation per row was crashing after hours of being idle for an unknown reason. After further inspection, we found that the field containing the replies was the root of the problem. The latter contained a nested recursive JSON structure with more than 1,000 levels of nesting in some cases due to replies being stored recursively. Hence, the process was crashing when trying to create a new dataframe containing the results of the sentiment analysis udf's in addition to the replies field. Once we filtered out this field from the original dataframe, sentiment analysis completed in a few minutes.

It is important to highlight that the perplexity computing to find the optimal number of topics for each period took a considerable amount of time to be completed. The latter being almost the 90% of the computation time spent on the project.

5.2 Gab vs Twitter

The three selected topics to carry out a comparison against Twitter were: Seth Rich (Conspiracy Theory on the Murder of Seth Rich), Manchester (Manchester Arena Bombing Terrorist Attack Victims), and USA Alabama Elections Results. The three of them being chosen from Gab pre-existing topics. Furthermore, the three of them are related to real-world events. We extracted Twitter data in a JSON format using the Twitter Search API in a span of three days after the real-world event related to that topic occurred.

Table 6 shows the results of these topics sentiments comparison between Gab and Twitter. We found that Twitter messages are more emotionally loaded. Negative emotions and positive emotions load is higher in Twitter in every topic. On the other hand, social concerns speech is higher for all the topics in Gab. The latter hint us that Gab speech is less emotionally driven than Twitter. Religion related speech is lower in Gab for all the topics.

Overall sentiments have a similar tendency in Seth Rich and Manchester topic. A difference can be seen in the Manchester Topic for which in Gab it is much more negative than in Twitter. Although, the same polarization is maintained. On the other hand, in Alabama Election Results we can see a clear difference in the polarization of the overall sentiments. In Twitter being positive and in Gab being negative. This could be due to the election winner being a democrat candidate (left wing). The latter further demonstrate the predominance of alt-right users inside Gab. These results answer **research question Q2.1**.

Topic	**Dimension**	**Gab**	**Twitter**
Seth Rich	Overall sentiment	0.00	0.06
	Negative Emotions	5.99%	7.27%
	Positive Emotions	5.79%	6.39%
	Social Concerns	9.98&	9.98%
	Religion	4.83%	6.15%
	# of Messages	3,339	57,456
Manchester	Overall sentiment	-0.20	-0.12
	Negative Emotions	7.11%	10.18%
	Positive Emotions	5.52%	7.75%
	Social Concerns	10.76&	10.56%
	Religion	5.49%	6.95%
	# of Messages	1,456	36,336
Alabama Election Results	Overall sentiment	-0.07	0.03
	Negative Emotions	5.70%	7.42%
	Positive Emotions	5.40%	6.26%
	Social Concerns	9.44&	9.13%
	Religion	4.60%	5.02%
	# of Messages	2,528	46,467

Table 6: Gab vs Twitter sentiment analysis comparison on three topics.

5.2.1 Computation. Table 7 describes the computation needed to perform the Gab vs Twitter analysis. Data extraction was performed outside of the databricks environment and uploaded to Databricks File System through the Databricks CLI.

5.2.2 Challenges. No challenges were encountered on this methodology.

5.3 Users interaction network analysis

The entire dataset graph contains 17,028 nodes (i.e. users) and 1'092,180 edges (i.e. interactions). Filtering the giant component

Task	Time taken (hours)	Estimated Price (USD dollars)
Data Aggregation	0.1	2.808
TOTAL	**0.1**	**$2.81**

Table 7: Computation of Gab vs Twitter Analysis methodology. Prices estimation assumes 9 i3.xlarge instances (i.e. 1 driver and 8 workers) running per task (i.e. 2.808 USD per hour or fraction).

we are left with 16,813 users and 1'091,995 interactions. The most relevant users for the platform based on PageRank were: BGKB, shuhari, jachinglaplume, AStormsABrewin and darulharb. All of these users are very active inside the network. With BGKB having over 240K posts. The most relevant users based on their degree were: WolverineTongue (67,180 interactions), bbeeaann (43,047 interactions), TheRealDonaldTrump45 (38,913 interactions), OccamsEpilady (31,829 interactions) and PaesurBiey (26,844 interactions). Users have an average of 5.87 closed triangles.

LPA on the platform network resulted in 16,096 communities discovered. However, only 13 communities have more than 4 users. These 4 communities are comprised of 699 users and 30,363 interactions. Table 8 summarizes the number of users found by communities. Figure 5 shows the degree distribution of these 13 communities network. The latter follows a power-law distribution. These results answer **research question Q3.1.**

Community	# of Users
1	321
2	134
3	89
4	40
5	34
6	22
7	17
8	12
9	10
10	5
11	5
12	5
13	5
TOTAL	**699**

Table 8: Communities with more than 4 users found on the entire network graph.

In addition to this, we created one graphs for every topic. This resulted in 36 networks for Gab pre-existing topics and 36 networks for the discovered topics. Taking the topic "Seth Rich" from May 2017 as an example, we found that it contained 308 users and 16 communities of more than 2 users. The most relevant users on this network based on PageRank were TukkRivers, a (Gab CEO), Fedelishkrieg, HighPriestess, USMC-DevilDog and Spnn2732. Which

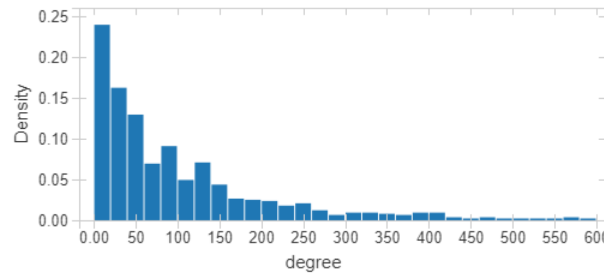

Figure 5: Degree distribution of Gab platform top communities users interaction

aligned with the most relevant users based on their messages scores. Figure 6 presents the network as a graph visualization. The latter clearly shows communities of users found on the topic interactions. Most of the users interact with central users that then interacts with other relevant users from other communities. These results answer **research question Q3.2**

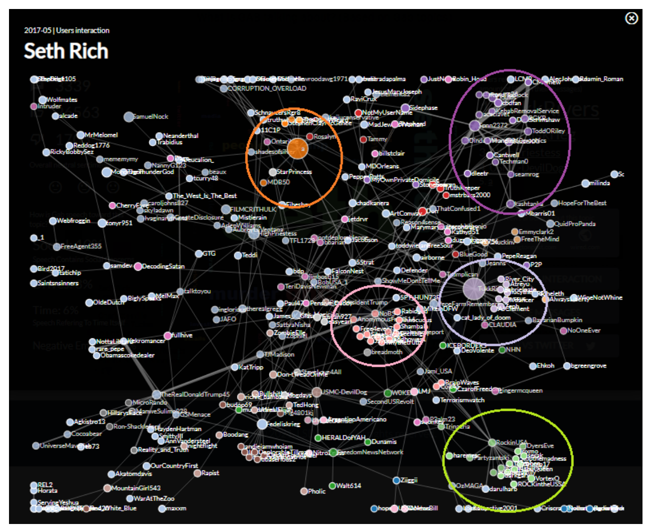

Figure 6: Communities of users visually found on users interaction inside the 'Seth Rich' topic.

5.3.1 Computation. Table 9 describes the computation needed to perform the users interaction network analysis. It is important to highlight that the graph pipeline was done for the top 3 ranked topics of each of our periods on Gab topics and discovered topics.

5.3.2 Challenges. Users interactions were stored in the recursive schema that we mentioned on Subsection 5.1.2. Hence, we needed to parse this recursive field to obtain the *who-replies-who* information. In order to do this we created a User Defined Function (udf) which recursively parsed each level of the object. On each level, the username of the reply is obtained. We limited the access to a maximum of 2000 levels (i.e. replies).

In addition to this, in all the generated graphs around 90% of the communities found by the LPA algorithm were comprised of 2 or

Task	Time taken (hours)	Estimated Price (USD dollars)
Replies parsing	0.3	2.808
Graph building	0.1	2.808
Filter Giant component	10.3	30.888
Finding most central nodes	3.2	11.232
Finding communities	0.6	2.808
Graphs storage	0.8	2.808
TOTAL	**15.3**	**$53.36**

Table 9: Computation of users interaction network analysis methodology. Prices estimation assumes 9 i3.xlarge instances (i.e. 1 driver and 8 workers) running per task (i.e. 2.808 USD per hour or fraction).

less users. This is a known limitation of the algorithm. We did not filter these communities of users since that could change the graph original structure.

5.4 Website & Visualization

Using the aforementioned results we built a website which contained visualizations that presented these results in an interactive fashion. First we built a stacked bar chart depicting the top three ranked topics based on their posts score (Figure 7). The x-axis represents each one of the analyzed periods in our dataset. The y-axis represents the topic score. This visualization can be shown using Gab pre-existing topics or discovered topics. When a topic bar is hovered a tool-tip appears with the topics names and scores. When a topic bar is clicked, a container with detailed information appears in front of the visualization. The latter includes: number of messages, likes, dislikes, overall average sentiment as an icon (i.e. Vader result), speech analysis (i.e. LIWC results), a wordcloud depicting the most used terms in the topic messages, the top five ranked users of that topic (based on their messages scores), news articles hyperlinks, users interaction network (if available), messages sample and Gab vs Twitter comparison (if available) (Figure 8). This visualization further supports **research question Q1.2**.

Users interactions are shown as an interactive graph in which nodes size represents the importance of the node inside the network (based on PageRank) and the color of the nodes represents the community to which a node belongs to (Figure 9). We limit the number of communities shown with a different color to the 10 biggest communities. The rest of the communities are shown with only one color to avoid visual saturation. In some topics, the users interaction network is too complex to be analyzed in a visualization (Figure 10). This visualization further supports **research question Q3.2**.

The stacked bar chart and the wordclouds were generated using Highcharts.js[12]. Highcharts is a Javascript package to generate interactive charts for HTML5 projects. The users interactions graphs were generated using D3.js[13]. D3.js is also a Javascript package to

[12]https://www.highcharts.com/
[13]https://d3js.org/

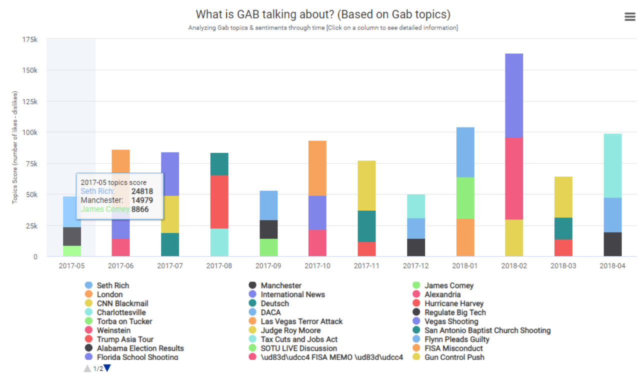

Figure 7: Results visualization: Stacked bar chart depicting the top three ranked topics based on their posts score. The x-axis represents each one of the analyzed periods in our dataset. The y-axis represents the topic score.

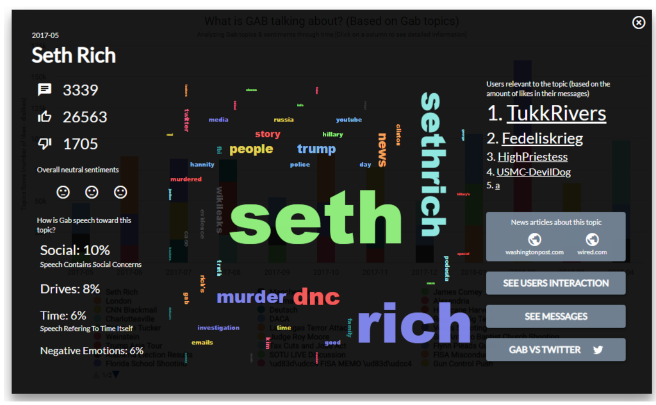

Figure 8: Results visualization: 'Seth Rich' Topic detail accessed when the topic bar is clicked.

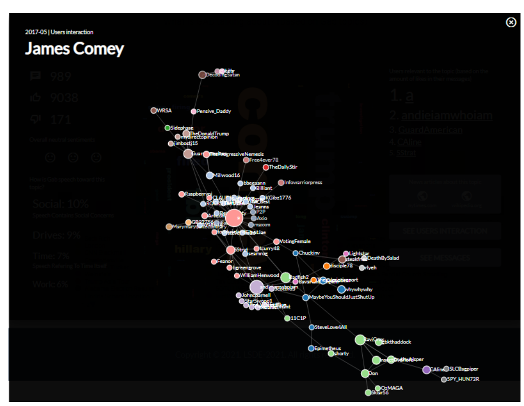

Figure 9: Results visualization: 'James Comey' topic users interaction graph. Nodes sizes represent the importance of the node inside the network. Nodes color represents the community to which a node belongs to.

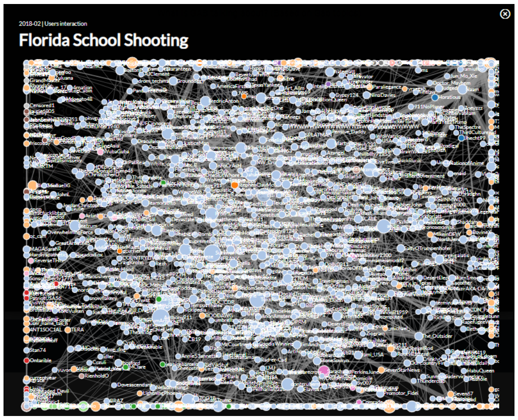

Figure 10: Results visualization: 'Florida School Shooting' topic users interaction graph. In this case, the graph is not easily readable due to its size and complexity.

generate interactive charts. However, it uses SVGs for rendering charts. The latter makes it extremely fast and provides smooth animations even in computers with low graphical power.

The aforementioned visualization is hosted in a Website. In addition to this, the website provides context into the project with relevant information about Gab, our Dataset and our project team members. The website can be found in the following URL: https://lsde-2021-gab.herokuapp.com/. In addition to this, it can be found in the following path on the DBFS: dbfs:/mnt/group12/visualization

6 CONCLUSIONS AND FUTURE WORK

Gab platform have created an unprecedented opportunity to perform research on users from an alt-right political alignment. On this work we have successfully developed a pipeline to discover Gab topics and analyze their sentiments and users interactions, overcoming difficulties inherent to Big Data and complex data formats. In doing so we discovered how Gab major topics and their sentiments changed overtime. Furthermore, we demonstrated that the topics in Gab are driven by real-world events. In addition to this, we compared Gab users speech with users from other social network (i.e. Twitter) and discovered significant differences between them. Finally, we characterize topics in terms of the interactions between users that discussed about those topics.

The developed pipelines in Spark are reproducible and they have proven to be low-complexity and low-cost inside the Databricks environment (with the exception of the perplexity analysis). Hence, the use of these methodologies is recommended for similar future research.

Our work hinted differences between Gab and Twitter users speech. Future work should include other social network platforms such as Facebook in order to further prove the differences of Gab users speech. On the other hand, Twitter has been previously used successfully to predict election results from users speech (e.g. Venezuela, USA). However, given the recent bans and restrictions imposed in Twitter, Gab could be an alternatively to predict election results more accurately by weighting measurements from both social networks. Experiments could be carried out using data prior to the USA elections of November, 2016. Carrying out a new crawling of Gab to extend the dataset to recent years would open the door to more in-depth analysis. For example, comparisons can be done between elections results predictions models from the 2016 and 2020 USA elections. Finally, users speech toward COVID-19 could be compared between Twitter and Gab. The latter would reflect alt-right position towards COVID-19 topics such as vaccination or restrictions.

7 WORK DISTRIBUTION

The work distribution of our group is shown in Table 10.

REFERENCES

[1] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho. Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, 66:125–136, 2017.

[2] M. C. Benigni, K. Joseph, and K. M. Carley. Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PloS one*, 12(12):e0181405, 2017.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] A. Borg and M. Boldt. Using vader sentiment and svm for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746, 2020.

[5] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *CHI 2017 - Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems - Proceedings, pages 3175–3187. Association for Computing Machinery, May 2017.

[6] S. F. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. 1998.

[7] S. Elbagir and J. Yang. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 122, page 16, 2019.

[8] A. Guterres et al. United nations strategy and plan of action on hate speech. *no. May*, pages 1–5, 2019.

[9] M. Hall and S. Caton. Am i who i say i am? unobtrusive self-representation and personality recognition on facebook. *PloS one*, 12(9):e0184417, 2017.

[10] G. E. Hine, J. Onaolapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. *CoRR*, abs/1610.03452, 2016.

[11] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[12] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[13] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *WWW '10*, 2010.

[14] C. D. Manning, P. Raghavan, and H. Schütze. Scoring, term weighting and the vector space model. *Introduction to Information Retrieval*, 100:2–4, 2008.

[15] D. A. Ostrowski. Using latent dirichlet allocation for topic modelling in twitter. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 493–497. IEEE, 2015.

[16] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July 2002.

[17] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[18] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[19] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[20] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[21] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu. Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PloS one*, 15(9):e0239441, 2020.

Activity	Responsible	Description
Data Decompression	Leonardo	Decompressed data from ZST to a binary JSON format into the DBFS.
Data Filtering & Partitioning	Xuan	Applied Section 2.3 and 4.1.1 pipelines
Sentiment Analysis	Leonardo	Applied Section 4.1.2 pipeline
Topic Discovery	Leonardo	Applied Section 4.1.3 and 4.1.4 pipelines
Data Aggregation	Xuan	Applied Section 4.1.5 pipeline
Real World Correlation	Xuan	Applied Section 4.1.6 pipeline
Gab vs Twitter: Data Extraction	Leonardo	Applied Section 4.2 pipeline
Gab vs Twitter: Data Aggregation	Xuan	Applied Section 4.1.5 pipeline on Twitter Data
Build network per topic	Zhaolin	Applied Section 4.3 pipeline
Pagerank of networks	Zhaolin	
Communities Detection of networks	Zhaolin	
Website Template	Xuan	Built website HTML
Topics Visualization	Leonardo	Built website topics visualization (Figure 7 and Figure 8)
Graph Visualization	Zhaolin	Built website graphs visualization (Figure 9)
Report Writing	Leonardo	Section 1 (Introduction), Section 2 (Dataset), Section 4.1 (Methodology: Posts content analysis), Section 4.2 (Methodology: Gab vs Twitter), Overall document quality (Ensure the documents sections are connected between each other, grammar check, writing style consistency).
	Zhaolin	Section 3 (Previous Work), Section 4.3 (Methodology: Users interaction network analysis), Section 5.3 (Results: Users interaction network analysis), Section 6 (Future Work), Overall document quality (grammar check, references check, format check)
	Xuan	Section 5.1 (Results: Posts content analysis), Section 5.2 (Gab vs Twitter), Section 5.4 (Website & Visualization), Section 6 (Conclusions)

Table 10: Responsibilities distributions by group member

[22] S. Zannettou, B. Bradlyn, E. D. Cristofaro, M. Sirivianos, G. Stringhini, H. Kwak, and J. Blackburn. What is gab? A bastion of free speech or an alt-right echo chamber? *CoRR*, abs/1802.05287, 2018.

[23] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtelris, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 405–417, New York, NY, USA, 2017. Association for Computing Machinery.