Universiteit Utrecht

How does the language of corpora from radicalized communities discovered on Parler compare to online conversations on Twitter regarding the 2021 Capitol riots and election fraud?

Arne C. Esser

6211283

MSc. Applied Data Science

Faculty of Natural Science - Utrecht University

Thesis Committee:

Dr. Mirko Schäfer

Dr. Max Kemman

Supervisors:

Paul Verhaar

Mijke van den Hurk

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science.

Submitted on July 2nd, 2021

Abstract

In light of the unprecedented 2021 Capitol riots and the 2020 U.S. presidential election, there is no doubt that social media technologies like Parler facilitated online radicalization of individuals towards violence. As a result, the content on Parler, its lax positioning on content regulation, and heavily one-sided political leanings rendered it an attractive environment for empirically studying normative behaviors of a social network that is predominantly right-wing extremist. This thesis project considers characterizing the use of language by sub-communities detected on Parler composed of a conglomeration of conservatives, conspiracy theorists, and far-right extremists under a common goal to spread radical propaganda and disinformation on a mainstream social network. Through implementation of HDBSCAN clustering of Parler users into sub-communities based on the similarity of their hashtag patterns, it was possible to identify virtual social groups that subscribed to a wide assortment of far-right beliefs in varying degrees. Drawing on the theoretical framework of radicalism by Kruglanski et al. (2014) combined with literature insights on NLP applications to online extremism (Torregrosa et al. 2020, Hitkul et al 2021), it was also possible to establish clear measures of radicalization of a sub-community as well as differences across sub-communities on Parler. The results show that the most popular topics of conversations among Parler communities were centered around Trumpism, conspiracy theories, far-right extremist groups, voter fraud, and conservative in-group favoritism. The texts were subsequently subjected to corpus linguistics tools such as n-gram frequency measures, domain-specific vocabulary significance tests and concordance analysis in order to find patterns in the usage of words. In general, users in radicalized communities deployed a number of strategies in their discourse that is characteristic of perpetuating the far-right narrative thereby exposing more and more people to opportunities for radicalization. This entailed constructing a violent and aggressive discourse with warmongering qualities effectively creating an ‘us vs. them’ dichotomy that legitimizes their need for violence. It was also discovered that radicalized sub-communities used persuasive language in endorsing far-right beliefs to others among their community by leveraging the affordances of social media engagement features as a way to normalize extremist views.

Keywords: online extremism, data science, computational linguistics, social media, alt-tech platforms, community detection

Table of Contents

1 Introduction 4
1.1 Motivation and Context 4
2 Theoretical framework 6
2.1 Towards conceptualizing online extremism 6
2.1.1 Radicalization 6
2.1.2 Extremism 7
2.1.3 Extremist discourse 8
2.2 Literature Review 9
2.2.1 Community detection for extremism 9
2.3 Applications of NLP in extremism research 11
3 Methodology 12
3.1 Data acquisition 13
3.1.1 GDPR considerations 13
3.1.2 Parler 13
3.1.3 Twitter 14
3.2 Data pre-processing 15
3.3 Exploratory data analysis 16
3.3.1 Hashtag popularity 17
3.3.2 Term popularity 18
3.3.3 Density plots 19
3.4 Community detection system 20
3.4.1 Hashtag similarity 20
3.4.2 HDBSCAN clustering for community detection 21
3.4.3 Corpus similarity evaluation 22
3.5 Corpus linguistics 23
3.5.1 Descriptive statistics 23
3.5.2 Beyond descriptive statistics 24
4 Results 25
4.1 HDBSCAN clustering 25
4.2 Overview of final datsets 26
4.3 t-SNE visualization 27

4.4 Top hashtags in each cluster 27
4.5 Corpora similarity 29
4.6 N-gram frequency measures 30
4.7 Domain-specific vocabulary significance test 32
4.8 Concordance analysis 35
4.8.1 Cluster 0 36
4.8.2 Cluster 1 37
4.8.3 Cluster 2 39
4.8.4 Cluster 3 41
4.8.5 Twitter 43
5 Discussion and Limitatons 44
5.1 Discussion 44
5.2 Limitations 46
6 Conclusion and Future work 47
6.1 Conclusion 47
6.2 Future work 47
7 Acknowledgements 49
8 References 50
9 Appendix 62
9.1 Parler EDA 62
9.2 Twitter EDA 66

1 Introduction

1.1 Motivation and Context

The use of online social media (OSM) has grown significantly over the past decade. Social media are website and app-based digital communication tools that enable users to interact with each other by disseminating and consuming information (Lifewire 2021). As of 2021, within a world population of 7.9 billion human beings, there are 4 billion active social media users worldwide in which the average person has 8.6 various social media accounts (Backlink 2021). This number amounts to more than half of the world being active on social media. Moreover, according to Statista (2021), an average user spends 142 minutes daily on social media platforms, engaging in activities facilitating the broad exchange of ideas and content while interacting with friends and strangers alike. Evidently, these platforms play a fundamental role in daily life as anyone can easily express their thoughts, opinions and feelings through the internet and these ideas can spread garnering global attention in a matter of seconds. While this has a created an effective medium for organizations, consumers, businesses, and governments to easily communicate with each other; it has also created an ideal space for extremist groups and individuals to manipulate behavior and public opinion. Indeed, research done by Berger at al. (2013), Kwok & Wang (2013) and O'Callaghan et al. (2013) into this subject show that OSM platforms are being misused for radicalism and disinformation due to their low publication barrier, lack of rigorous moderation, anonymity, and mass reachability. In 2018, Twitter reported over 1.2 million accounts were suspended for terrorist content since August 2015 (Twitter 2018). While an impressive undertaking, the problem of extremism propagated through OSM still impacts the mobilization and uprising of violent events to this day (Stern, Linke & Holland 2016; Karrel, Linke, & Holland 2021).

A living example of this occurrence is the infamous January 6th 2021, storming of the United States Capitol by an angry mob composed of various right-wing extremist groups and supporters of Trump shortly following the 2020 U.S. presidential election (Farivar 2020). The riots were an attempt to overturn the 2020 presidential election by disrupting the joint session of Congress assembled to count electoral votes (Reeves, Mascaro & Woodward 2021). Experts on the matter revealed that rioters openly planned to disrupt the counting of electoral votes for several weeks prior to the event (Graff 2021). Plans were coordinated on 'alt-tech'[1] platforms such as the social networking service Parler among other instant messaging services (e.g., Telegram, Gab etc.) which were used to discuss previous Trump rallies and for storming the Capitol (Frenkel 2021). Evidently, the rise of online social media that cater to alt-right groups as digital tools for online radicalization and mobilization have notably played a pivotal role in the manifestation of the 2021 United States Capitol attack.

Traditionally in online worlds, prevalence of extremism was once isolated to the fringes of anonymous message boards where hateful individuals would escalate extremist rhetoric and spill disinformation on current events while contently safeguarded by their anonymity and privacy (Lopez 2021). Despite this, the value of anonymity exercised in these message boards did not allow for mass organizing of extremist groups to enact social change in online and offline worlds.

[1] Alt-tech is a group of websites, social media platforms, and Internet service providers which have become popular among the alt-right, far-right, and others who espouse extreme or fringe opinions.

Over the years at increasingly accelerating rates, online conversations about extremist and radical ideas have shifted from the fringes of the internet and into the mainstream foreground (Lopez 2021). The boundaries between what constituted mainstream and non-mainstream/conspiracy-related ideas have dissolved as more and more extremist activity show up on our most common social media apps (Torok 2011). Lopez (2021) argues that with the rise of social media popularity, extremist beliefs and ideologies have been gradually normalized on mainstream platforms. Indeed, much of extremist discourse on the internet can now be found on common social media apps such as Parler (Floridi 2021).

Parler is an American microblogging and social networking service launched in 2018 (Cryst 2021). Registered users can share information instantly by posting short messages known as Parlays. Parler brands itself as a free speech-focused and 'unbiased' alternative to other mainstream OSM platforms such as Twitter and Facebook (Binder 2020). However, it is known to be a hyper-conservative platform (Munn 2021). Indeed, Parler attracts the interest of conservatives, Trump supporters, religious, and patriotic individuals (Aliapoulios 2021). Shortly after the 2020 U.S. presidential election, Parler drew the attention of many right-wing politicians and influencers (people with large online followings) as a platform where they can promote provocative ideas without concerns of having their content removed. Parler soon became a haven for far-right extremists and conspiracy theorists who were interacting with the radical right flocking to the platform when Donald Trump publicly denounced mainstream social media giants like Twitter and Facebook for targeting him and other conservatives (Newhouse 2020). Mass proliferation of OSM usage in recent years by right-wing extremist groups have produced decentralized digital communities whereby conservative folks and far-right extremists can now more easily coalesce to share and espouse radical beliefs and policies. Such alt-tech platforms enable the convergence and exposure of both extremists and non-extremists with shared attitudes and personal grievances to connect with each other in transformative ways that could lead to negative tangible societal outcomes as experienced in the Capitol riots.

While a large body of literature in recent years have examined the problem of extremism on social media, many of the studies published were conducted on the mainstream OSM platforms with very little academic attention on up-and-coming small-scale social media technologies, especially of ones adopted by far-right groups. Scientific research that delves into right-wing extremism on alt-tech OSM platforms is still early and recent which makes it an appealing and interesting material to learn more from. Nevertheless, due to research performed by Hitkul et al. (2021), it is now well understood that different social media platforms display clear differences in how political events and related discussions are covered by the public. The authors conducted a comparative study on the trending content and users surrounding the riots on Parler and Twitter that revealed a divisive contrast in the rhetoric between the two platforms. Hitkul et al. (2021) highlights well the notion that far-right groups are exerting greater influence of extremist ideologies on trending OSM platforms and normalizing them as mainstream views by leveraging real-time social network engagement features (e.g., Hashtags, sharing, commenting etc.) to further expand their political agenda.

What was largely left unexplored from the work by Hitkul et al. (2021), however, was a deeper investigation into studying latent radicalized communities with extremist agenda on Parler and understanding the language patterns that were being used within this social network that

could characterize them. This thesis project considers characterizing the use of language by sub-communities on Parler composed of a conglomeration of conservatives, conspiracy theorists, and far-right extremists under a common goal to spread radical propaganda and disinformation on mainstream channels.

Consequently, the research question of this thesis is:

How does the language of corpora from radicalized communities discovered on Parler compare to online conservations on Twitter regarding the 2021 Capitol riots and election fraud?

Corpus-based linguistic research into identifying normative behaviors of extremism based on quantitative analysis of language remains scarce (Prentice 2012). Despite this, it is believed that interpreting patterns in language will afford insights into how extremist discourse with its radical ideologies shapes itself on social media. As Litvinova (2018) critically points out, to develop effective counter-terrorism methods, it is important to analyze and understand extremist ideology reflected in texts. The importance of the analysis of extremist rhetoric produced immediately after a violent event is due to the fact that following a high-profile attack, not only does the character of extremist propaganda change but it also reaches its peak (Stepin 2015).

The remainder of this thesis is organized as follows. Section 2 will contextualize concepts of radicalization and extremism as well as review the current literature regarding the relevant analyses of extremism on social media. Section 3 will provide a detailed account of the approach and methodology taken to study language by Parler communities, including the datasets and data analysis techniques that were used. The results and observations from the computational corpus linguistics of Parler and Twitter data will be discussed in Section 4. Discussion and limitations of the work is elaborated upon in Section 5. The conclusion and suggestions for future work concludes this thesis in Section 6.

2 Theoretical framework

Section 2.1 addresses the concepts of radicalization, extremism and extremist discourse that provide the domain context important for this research. Section 2.2 reviews existing literature and techniques that have contributed to community detection methods of virtual communities that also engage in the study of online extremism. Section 2.3 reviews existing literature of NLP applications to online extremism.

2.1 Towards conceptualizing online extremism

2.1.1 Radicalization

The term ‘radicalization’ and ‘extremism’ is widely used online but within literature, there are different definitions given for these terms equipped with a number of overlapping views. As such, there is no universally accepted definition in academia or government for radicalization and extremism (Aldera et al. 2021). Even in academic research (e.g., Correa & Sureka 2013), extremism and radicalization are often used as exchangeable terms to refer to the same phenomenon, which can lead to a misunderstanding that both terms share the same conceptual meaning (van de Weert & Eijkman 2019). Instead, van de Weert & Eijkman (2019) posit that while both terms are synonyms, they are conceptually different, and radicalization is part of extremism. Thus, online radicalization is defined as the process by which an individual gets exposed to

ideological messages and belief system that encourages movement from mainstream beliefs towards extreme views, principally through social media (COPS 2021). McCauley & Moskalenko (2008) combines a political dimension to this notion that these extremity in beliefs and opinions are in support of intergroup conflict and violence. Although radicalization is an extremely multi-faceted concept (Trip et al. 2019), researchers (e.g., Doosje et al. 2013, Koehler 2013) have attempted to explain critical triggers for the phenomenon. While studies like United Nations Office on Drugs and Crime (2012) have linked socio-economic and demographic conditions as vulnerability factors towards radicalization, these explanations generally lack insight as to how psychological factors are involved. Instead, Lara-Cabrera (2017) asserts that in addition to the socio-demographic factors, personal motives such as feelings, basic needs, emotions as well as personal life situations and experiences are much more important triggers for radicalization. The psychological model of radicalization by Kruglanski et al. (2014) attributes these types of personal motivations to a fundamental universal human need for significance, an abstract need in which people strive for a deeper meaning to their existence. In their theoretical framework, Kruglanski et al. (2014) conceptualizes radical behavior as a motivational imbalance whereby the individual becomes so deeply invested in reaching a focal goal that this simultaneously undermines other goals that matter to other people. In this manner, the greater the imbalance between one's commitment to a focal goal and commitment to alternative goals, the greater the degree of radicalization. According to Van den Hurk & Dignum (2021), looking at radical behavior in this way implies that radicalism can be measured by the difference in commitment on the focal goal. As people radicalize from low motivational imbalance to high motivational imbalance, their commitment towards the focal goal becomes stronger and people are more willing to perform increasingly extreme or violent actions (Van den Hurk & Dignum 2021).

2.1.2 Extremism

Often in literature, extremism has been loosely compared with other similar concepts that tend to co-occur such as terrorism, racism, nationalism, and Jihadism (Olteanu et al. 2018; Rowe & Saif 2016; Fuchs 2016). Consequently, this term can take on different theoretical perspectives depending on the problem domain considered by researchers and explains why it is relevant to establish a clear definition to work with. Nevertheless, governments and other agencies have offered definitions of extremism that often share the following characteristics:

1. An emphasis on the political, social, and religious dimensions of extremism (Berger 2018).
2. A positioning of the 'extremist' as an individual or entity that stands in active opposition to a context-dependent set of norms and values (UK Home Office 2015).
3. A recognition that to some extent, radicalization can lead to acts of violence and in particularly severe cases, acts of terrorism (El-said & Barrett 2011).

According to Bergen et al. (2015), different types of radical groups share common elements that are important to pay attention to when contextualizing extremist groups. Firstly, all radical groups perceive a serious problem or grievance in society that is largely context-dependent (i.e., political, religious, social). Secondly, radical groups are strongly dissatisfied with the manner in which the current institution handles their problem. They may argue that the institutions do not pay enough attention to their grievance or claim that the institutions do not show enough responsibility to

handle their grievance (Moghaddam 2005). This generates a low institutional trust and a perception that authorities are not legitimate (Doosje et al. 2013). An influential third characteristic of radical groups is that they consider their own group's norms and values superior to those of others (Doosje et al. 2016). This establishes a strong us versus them dichotomy, which might form the foundation of the use of violence (Doosje et al. 2012). The fourth element of radical groups is of particular importance. Most extremist groups embrace an ideology that legitimizes violence to address their concerns, and this violence is often directed towards an out-group viewed as the perpetrators responsible for creating the grievance; a social phenomenon that is clearly articulated in the application of social identity theory to radicalization (Richer & Haslam 2016). Awareness of the elements that characterize a radical group is essential towards recognizing how a radical group thinks and operates which can help identify when radical/extremist discourse manifests.

The research carried out in this thesis focuses on the subset problem domain of right-wing extremism (RWE) as the attacking mob that stormed the capitol was largely comprised of this nature (Farivar 2021). Torregrosa et al. (2021) refers to this type of extremism as "an ideological movement, contrary to the democratic and ethical values of a society, that uses different methods, including violence (physical or verbal) to achieve its objectives". Uncovering what these different methods are is what this thesis aims to elucidate with the aid of community detection and computational linguistic tools. RWE draws on more political and social dimensions of extremism than religious aspects and is associated with 'racism, xenophobia[2], conspiracy theories and authoritarianism[3]' (Jupskas & Segers 2020).

2.1.3 Extremist discourse

Since the research question deals with analyzing the language use of virtual radicalized communities, it is also then necessary to operationalize and clarify what is understood as 'extremist discourse' engaged by people with extremist views. Doing so will help to discern and realize how extremist speech manifests on OSM platforms. Torregrosa et al. (2021) provides a comprehensive summary of several key features that characterizes and distinguishes extremist narrative from regular discourse gathered from existing literature. These characteristics, derived from different authors, are briefly described as follows:

- **Types of extremist narrative:** There exists several dimensions on which extremist narratives employ to legitimize their vision and objectives. Ashour (2016) grouped these narratives into five categories: historical, political, instrumental, sociopsychological and theological/moral.

- **Use of discursive rhetoric such as otherness, war narrative and hate speech:** extremist messages tend to use discursive rhetoric to justify their actions and notions towards others. Some of these techniques have been closely examined such as otherness (Sakki & Pettersson 2016), hate speech (Fortuna & Nunez 2018), the use of war

[2] dislike of or prejudice against people from other countries.

[3] authoritarianism is a form of government characterized by the rejection of political plurality, the use of a strong central power to preserve the political status quo, and reductions in the rule of law, separation of powers, and democratic voting.

terminology to frame “enemies” and incite violence onto others (Furlow & Goodall Jr. 2011).

- **Linguistic style:** The narrative types for extremism previously mentioned are constructed based on a specific vocabulary and style, that helps extremists formulate their discourse. Several papers (Cohen et al. 2014, Thorburn et al. 2019) have found stark differences in the linguistic style from radical and extremist texts compared to a regular sample of texts. For instance, extremist narratives: tend to have higher use of third-person plural pronouns and less first person singular and second person pronouns, have a more negative tone and more frequent words related to negative topics such as death and anger than a regular narrative.

In essence, the extremist discourse is characterized by the use of specific narratives, an aggressive and polarized linguistic style coupled with several techniques oriented to justify a framing of superiority towards another group. These notable features that are highlighted would serve as the basis and set criteria in which extremist content shall be diagnosed from.

2.2 Literature Review

2.2.1 Community detection for extremism

Realizing the dangers of violent extremism and how it is becoming a pressing challenge to societies worldwide, many researchers have investigated radicalism and their associated behaviors online. Radicalization can be a heavily social and interactive process that involves numerous participants influencing each other to form same groupthink and ideologies (Doosje 2016). Due to this reason, an important concern when examining radicalization in the online world is the ability to detect or discover virtual communities within a social network. The problem of community detection has been widely studied within the context of social media and is well documented in works like Papadopoulos et al. (2011) and Fortunato (2010). Community detection refers to the task of identifying subgroups in a network where entities are more densely connected to another than to the rest of the network. Social networks extracted from social media present unique challenges due to their extensive size and high user-user interaction possibilities (Benigni et al. 2017). Similarly, ties in online social networks like Facebook are commonly known to represent different types of relationships (Boccaletti et al. 2006). While this may increase the complexity of mapping out social processes in a system, it offers wide possibilities to investigate the problem from an array of theoretical perspectives. Literature shows that techniques for virtual community discovery rely on two distinct and overarching approaches: network-based analysis (also referred to as graph-based) or clustering-based. Furthermore, there has been considerable research effort into investigating online radicalization in a variety of similar problem domains such as jihad, anti-Islam, and ISIS but less so in right-wing extremism. The lack of computational approaches rooted in community detection that is dedicated to the study of RWE necessitates reviewing literature from other relevant domains.

2.2.1.1 Network analysis for online extremism

Network-based analysis exploits the structure of links between agents in a network to evaluate the topological characteristics of the network (Correa & Sureka 2013). In terms of network-based analysis, the works done by (Benigni et al. 2017), Gialampoukidis et al. (2017), and Rios and Munoz (2012) can be highlighted.

Benigni et al. (2017) proposed an ensemble of graph-based clustering techniques capable of extracting online extremist communities from social media networks that leverages the rich data structures common to many OSMs such as users' friends, mentions and hashtag patterns. The novel approach developed is called 'iterative vertex clustering and classification' (IVCC) and was applied towards detecting an ISIS supporting community on Twitter. Within the community, Benigni et al. (2017) was able to observe a wide range of actor types affiliated with ISIS such as fighters, recruiters, and propagandists. IVCC has shown to outperform two pre-existing approaches on the classification task of identifying ISIS supporters by a significant margin. Despite this, the authors remark however, that their method cannot account for changes in group dynamics over time.

Gialampoukidis (2017) applied a novel framework that combines community detection with key-player identification to retrieve communities of terrorism-affiliated social media users. Their findings showed that most of the members of each retrieved community were already suspended by Twitter for having violated their terms, a good indication that the technique is making appropriate inferences.

Another intriguing study done by Rios and Munoz (2012) combined the analytical prowess of social network analysis and text mining as a strategy for community detection in a Dark Web portal. Indeed, the authors proposed a novel approach for detection of overlapping sub-communities in a network by leveraging traditional network analysis methods combined with topic-modeling (LDA). The study highlights well that both network analysis and text mining techniques are useful tools in the detection of hidden communities of social networks.

2.2.2.2 Cluster analysis

Clustering algorithms involves the machine learning task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (Alkarkhi & Alqaraghuli 2020). Concerning the studies that have used clustering-based algorithms for community detection, the works of Sanchez et al. (2016), Kelly (2017) and Panizo-Lledot et al. (2019) are worthy of attention.

Sanchez et al. (2016) introduced a novel agent-based platform for Twitter user clustering. This system tracks the activity for a given topic in the social network and is able to detect communities of users with similar political preferences by means of the Louvain modularity. Interestingly, the authors attribute the method of clustering users to a network analysis metric that extracts information of following relations between users to map these communities out. This method is also claimed to be applicable to detecting communities of users with extremist ideologies.

In another study done by Kelly (2017), an algorithm was developed that combines the results of two unsupervised clustering algorithms to find sub-communities in a given network, namely clique percolation and NMF clustering on the user's posts. One algorithm uses the structure of the network, and the other algorithm employs the text data associated with the nodes. Moreover, the combined algorithm was tested on a hand labeled ground-truth ISIS twitter recruitment network that generated a slightly higher score of detecting sub-communities than with just the clique percolation or NMF textual clustering methods alone.

Similarly, Panizo-Lledot (2019) used community finding and topic extraction algorithms to analyze the behavior of alt-right supporters on Twitter. The detection of relevant groups of users was addressed by the application of community detection algorithms in a retweet network. Then, the most similar users were grouped together by way of Clauset-Newman-Moore greedy modularity maximization method. Once the separate communities were obtained for each re-tweet network, a topic modeling technique (LDA) was applied to identify the topics covered by a collection of texts. The results revealed interesting details about alt-right norms. They found that the sample commonly used racist language and anti-immigration themes, criticized mainstream media, and endorsed alternative media sources. A sub-section of alt-right supporters were also found to focus heavily on white supremacist themes.

The above-mentioned studies reflect the importance of using community detection approaches to studying the extremist phenomenon in online social networks. Such methods serve as powerful tools that help experts better realize the dynamic relationships and normative behaviors of radical groups within an ever-changing digital landscape.

2.3 Applications of NLP in extremism research

In the past decade, Natural Language Processing (NLP) techniques have made great strides in knowledge discovery in nearly all domains of study. The applications of NLP for research on extremism is no exception. According to Torregrosa (2021), the main objective of NLP is to transform free text into structured data by capturing its lexical, syntactic and/or semantic information to acquire or infer new knowledge. Some of the previously mentioned studies (Kelly 2017, Panizo-Lledot 2019, Rios & Munoz 2012) have utilized NLP in their research pursuits. This section highlights several important studies that have used lexical approaches towards examining extremist discourse and their associated behaviors.

To date, only one work of comparative corpus linguistic analysis between extremists and non-extremists' texts is known to have been carried out. Litvinova (2018) subjected Islamic extremist texts and ones by common internet users on the same topic to word frequency measures, word clusters, concordances and collocates as linguistic tools towards the analysis of a Russian-language extremist chat forum. Doing so allowed them to identify a number of features characteristic of extremist rhetoric. Litvinova (2018) reports that in their rhetoric, extremists make use of the *we vs. you, them* contrast and build up a 'war' discourse that avoids addressing the event as a terrorist attack. Their corpus-based study revealed that extremist discourse differs significantly from the regular discourse on the same topic with respect to most frequent words, word clusters and collocations which reflects different views on the same event by extremists and non-extremists.

The research efforts of Rehman et al. (2017) aimed at identifying radical text in Twitter data demonstrated that incorporating religious text in model training significantly improves the performance metrics of the machine learning classifiers. In addition to this, it was observed that violence and bad words are creating a differentiating factor in distinguishing between radical and random users. Moreover, an in-depth analysis of new and old datasets indicates a variation in extremist group narrative highlighting that the extremist discourse may change over time. While indeed intriguing, their research follows the language narrative of ISIS supporters, so it is

uncertain whether their findings can be generalized to other forms of extremist ideologies held by other social groups.

Social media generates various ways of facilitating social interaction by individuals and groups alike in order to let ideas and messages spread quickly to the masses of their community. One of these tools involves the use of hashtags. Hashtags are an essential convention frequently used in microblogging sites to follow or create a discussion thread denoted by a '#' symbol in front of the topic name as one word (e.g., #trump2020). Hashtags can serve as the symbol of a community by linking users with similar ideas (Zhang et al. 2012). Moreover, textual content such as text, URLs, and hashtags closely correlate with user's interests. As put forth by Hapal (2015), hashtags have become the extremists' choice for propaganda. Analysis of topics used by political extremist groups found that the more common topics discussed were of racial topics (Ottoni et al. 2018, Rehman et al. 2021), war (Ottoni et al. 2018) and immigration (Ben-David 2016), while being very aggressive with these topics. An article (Torregrosa et al. 2020) analyzing a far-right community on Twitter found that they used racist, anti-immigration, and anti-left terms in their tweets. This work also noticed far-right members to use specific slang to refer to other racial minorities or political doctrines, such as 'libtards' to refer to liberals. Descriptive analyses such as these provide evidence that extremists on social media use hashtags as a marker of inclusion for engaging in radical conversations that they find common ground in.

Evidently, online radicalization and extremism has been extensively studied in several domains such as jihad, ISIS, nationalism, and religion. What is more, existing research have utilized a wide array of computational approaches based on relevant data transformations and feature extractions (e.g., topic modeling, keywords) for community detection and NLP within their respective subjects. The diversity of approaches in these highlighted works illustrate the multiplicity of ways online radicalization, in its different modalities, have been investigated in depth. No de-facto or status quo approach to detecting virtual communities exists as of yet. Therefore, existing research suggests that the method of approach to use is entirely open to the researcher's devices and motivations on how community detection should be carried out in order to solve the problem at hand.

3 Methodology

The methodology section explains the process carried out to conduct a comparative computational linguistic analysis concerning two social media technologies that covered the capitol riot incident. One is Parler, and the other is Twitter. Based on the empirical findings of Hitkul et al. (2021), it is assumed that the discourse expressed on Parler will exhibit a more right-wing extremist linguistic profile than its Twitter counterpart when analyzing language patterns and content posted on the day of the Capitol riots. Studying and comparing these two datasets will help gain insight into the radical rhetoric and domain-specific lexical components that underlie the discovered communities to uncover how the far-right narrative and discourse shapes itself around political issues on mainstream channels.

The approach consists of two main phases which will be described in more detail under the following sections:

Phase 1: User clustering or community detection, where Parler data is input into this step to perform a series of computational tasks for discovery of potentially radicalized communities. In the first task, a pairwise hashtag similarity score is computed for all users as data point features for HDBSCAN clustering. The next task seeks to generate clusters of users based on their hashtag similarities. The third task concerns identifying the cluster or set of clusters that best represent users who post on controversial topics regarding right-wing ideologies and election fraud. Once these communities have been identified, corpus linguistics is conducted on them to uncover the nature of their rhetoric.

Phase 2: Corpus linguistics involves studying the collection of posts to find emergent themes and contextual semantics associated with lexical and grammatical patterns that comprises of two sequential tasks. The first task concerns conducting corpus linguistics on the Parler corpora with the analysis of n-grams as lexical features, in addition to, applying domain-specific vocabulary significance tests comparing Parler corpora against the twitter corpus. Next, the lexical features that are of substantial contextual relevance are taken and explored further in concordance analysis. Doing so facilitates a closer examination at the natural language use within these communities.

3.1 Data acquisition

3.1.1 GDPR considerations

Under the GDPR (General Data Protection Regulation), researchers and organizations alike must conduct a DPIA (Data Protection Impact Assessment) when engaging in a project or other personal data processing activity (hereafter referred to as an 'initiative'). The DPIA describes and addresses important elements of data processing for an initiative. The DPIA is designed to help data controllers systematically analyze, identify, and minimize the data protection risks of an initiative while ensuring that these practices comply with data protection guidelines (ICO 2021). This structured approach helps towards identifying potential privacy risks associated with the initiative. A description of the DPIA for the acquired Parler and Twitter data will include the nature, scope, context, and purpose of the processing. Of course, some of the answers presented may be similar or fall under a common reason.

3.1.2 Parler

Parler data was collected and downloaded from a publicly available and online source on GitHub (Smith 2021). The Web scraping of the entire data dump was undertaken by @donk_enby done through Parler's API (Application Programming Interface) (Clary 2021). The Parler dataset[4] used in this work is a subset of that data dump. This dataset contains roughly 2.3 million posts published on January 6 with the user's name and username as metadata. From an ethical perspective, it is worth noting that this is a legal method used to obtain text from publicly accessible Web pages through batch requests (Munn 2021). It should also be noted that this thesis never identifies individuals attributed to particular posts. For these reasons, and to avoid directing traffic to alt-tech platforms like Parler, posts are quoted without author's names. Additionally, this dataset does not contain any immediate sensitive personal information of individuals other than potentially their social media handle tied to their content. As the dataset only contains information on an

[4] Parler Data & Tools · GitHub

individual's posting content and their associated handle, the potential risks associated with the data subjects are minor. Parler data is especially useful for the study of extremism on social media. Data processing and analysis of the data was carried out between the months of May 2021 and July 2021 involving a total of 489 data subjects.

3.1.3 Twitter

A novel dataset of historical Twitter data was manually created for this thesis project as there are no good publicly available twitter datasets that exemplify coverage of the capitol riots on the same day. The Twitter dataset was crawled using a scraping tool[5] called 'snscrape' via the Twitter API. Snscrape can scrape from major social network sites for data like user profiles, hashtags, or searches and returns the discovered items (e.g., relevant posts). In building a comprehensive Twitter dataset that matches public coverage on the same issues, a list of 8 candidate hashtags[6] related to Trump, the Capitol riots and election fraud (e.g., #capitolriot, #electionfraud, #stopthesteal, #trump2020 etc.) were provided as query terms that define the social context of the tweets. Half of the hashtags pertain to election fraud and Trump ideology while the other half focuses on conversations about the Capitol riots. The motivation behind this approach was to specifically gather a dataset that covered topics of conversations on both the Capitol riots and election fraud evenly that is fairly comparable. A request of 10000 tweets per seed hashtag since 2021-01-07 until 2021-01-8 was sent to the Twitter API which was then parsed and written to a .csv file for further analysis. The scraped dataset contained just above 15,000 instances with datetime, tweet, id, username, and language as metadata. It should be noted too that none of the hashtags as query terms returned the maximum specified limit of relevant tweets. A possible explanation for this is Twitter API rate limits while archiving for tweets with the scraping tool.

Historical twitter data was collected using a publicly available scraping tool on GitHub that scraped for tweets without requiring personal API keys. Twitter advocates and supports the usage of Twitter content for non-commercial research purposes (Twitter 2020). The metadata of this dataset includes datetime, tweet, id, username, and language. None of these features match Twitter's characteristics of sensitive information listed on their Developer terms page (Twitter 2020). Thus, the data does not contain any private information as defined by Twitter guidelines and is therefore considered public data. As such, the potential risks associated with the data subjects can also be considered minor. In the same manner of reporting results as with Parler data, the author's name is completely omitted. Data processing and analysis of Twitter data was carried out between the months of May 2021 and July 2021 involving a total of 8141 data subjects. Twitter and Parler data are used for the purposes of carrying out non-commercial empirical research for a Master's thesis in the Applied Data Science university program of Utrecht University. The work and findings acquired from these datasets involve corpus linguistics as an explorative tool into contemporary digital culture. The final data will be surrendered to appropriate data stewards of Utrecht University and secured in the institution's online environment at the end of the thesis period.

[5] GitHub - JustAnotherArchivist/snscrape: A social networking service scraper in Python

3.2 Data pre-processing

Corpus linguistics involve analysis of large amounts of text in data. Before analyzing the bodies of text in terms of keywords and n-gram features, the data must undergo preprocessing steps. These steps are aimed at cleaning the data and removing redundant information. This involved a number of tasks such as the removal of mentions (@username), URLs, stop words and tokenization within the original posts. Preprocessing began immediately after the data was loaded into Python[7] as dataframes using Pandas[8] library. It is worth noting too that Python is the main programming language to carry out all the computational tasks in this research methodology. Firstly, 125 duplicate entries in the dataframe were dropped. Secondly, the corpus was parsed through a tweet preprocessor tool[1] that effectively removed any mentions, URLs and hashtags from the text while simultaneously converting all characters to lowercase form for tokenization. Tokenization is the act of breaking down a piece of text into small units called tokens. Using tokens are helpful for finding patterns in language within text data and facilitates similar analyses of linguistic components across texts. In this case, tokens are classified as words. Next, the Gensim library was used to remove Gensim's collection of stop words from the bag of tokens (Teja 2020). Gensim is a widely used python tool that facilitates a variety of tasks for NLP (Natural Language Processing) purposes[2]. Gensim's default collection of stop words includes a total of 337 stop words. Stop words are the words in any language that do not add much meaning to a sentence. Thus, they can be safely ignored when digitally processing natural language. These typically include the most common words, short function words, and sometimes pronouns. However, every NLP toolkit have their own custom stop words dictionary. A custom list[9] of 5 stop words were added to the collection after it was found that these words appeared in many of the parlays as they refer to a video URL and were not of particular significance to the research question. After stop word removal, regular expressions with the function `findall()` were used to remove any extra white space and punctuations from the bag of tokens. A regular expression is a sequence of characters that specifies a search pattern (Goyvaerts 2019). The remaining tokens were lemmatized which resolves a word to their base dictionary form, disregarding grammatical changes such as tense and plurality. When studying a word, it is often useful to consider the different forms of the word collectively (Biber, Conrad & Reppen 1998). The goal of lemmatization is to group together the inflected forms of a word so that they can be analyzed as a single item. For example, the lemma for the tokens 'cries' and 'cry' is 'cry'. This helps minimize text ambiguity and reduces the word density in the given text allowing for sharper comparisons to be made. The process of tokenization and subsequently lemmatization was done with Spacy's toolkit. Spacy[10] is another popular NLP library that provides many underlying text processing capabilities similar to Gensim. In addition to lemmatizing the tokens, the hashtags within posts were extracted as a column feature using the regex function `findall()` for further data analysis.

In conjunction to text preprocessing, a certain set of criteria were applied for reshaping the dataset to a smaller size. The criterion for dropping an instance is listed as follows:

[7] https://www.python.org/
[8] https://pandas.pydata.org/
[9] 'browser','support','video', 'tag', 'com'
[10] https://spacy.io/

- Any post with less than 10 characters.
- Any post with less than five hashtags.
- Posts from users that have less than five post submissions.
- Any post with NaN values.
- Any duplicate posts made by the same user (reduce spam).
- All posts from ‘Private User’.

The data re-sizing criteria proposed is required to narrow the bandwidth of data that allows broadly selecting for users who actively publish content on social media that is rich in text data. Data from ‘Private User’ was filtered out as the username is a pseudonym for a collection of hidden accounts whose identities have been redacted due to privacy preferences. Therefore, it is impossible to figure out the number of unique entities associated with this pseudonym and discriminate what they are publishing online. After successfully applying these dimensionality reduction steps; the reshaped dataset now contains just over 54,000 posts from a set of 2327 users.

A near exact procedure to the steps above was performed for preprocessing of the Twitter dataset. In addition to the essential tasks of tokenization and lemmatization as well as removing noisy text data (e.g., @mentions, URLs). Only English tweets were retained by filtering for ‘en’ on the language column attribute. The only criterion considered for data compression was filtering out any post with less than 10 characters. After successfully applying the preprocessing steps, the Twitter dataset was reduced to roughly 11,600 posts from a set of 8141 users.

3.3 Exploratory data analysis

Exploratory data analysis (EDA) is an important process used by data scientists to analyze and explore various aspects about the data by summarizing their main characteristics, often employing data visualization methods. EDA involves generating summary statistics for numerical data of relevant features in the dataset and creating various graphical representations to understand parts of the data better.

3.3.1 Hashtag popularity

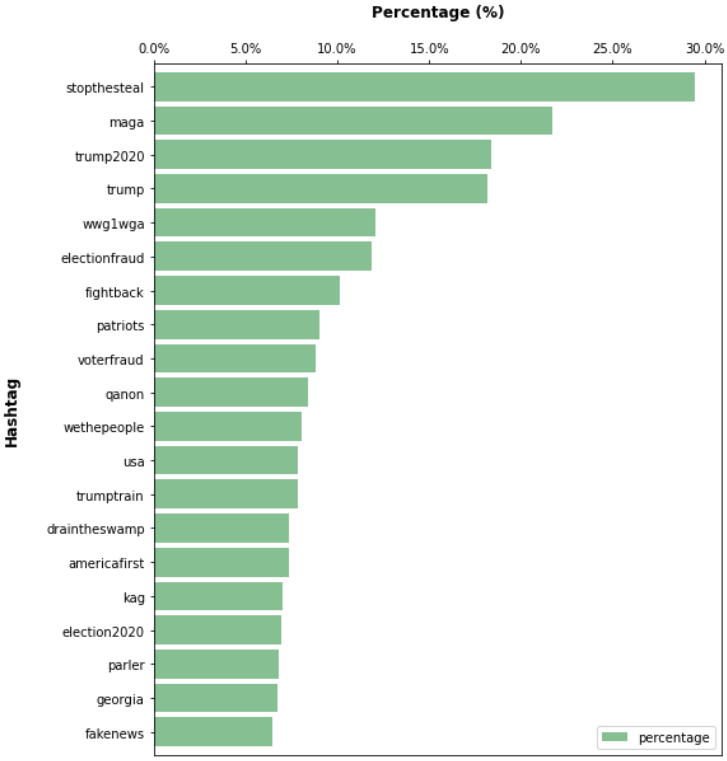


Figure 1A: Top 20 hashtags on Parler

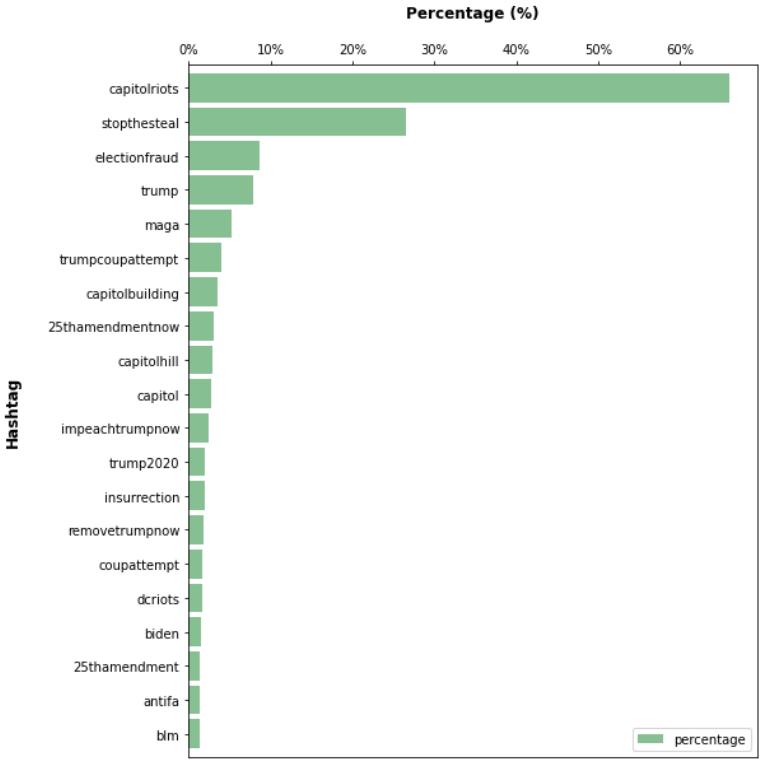


Figure 1B: Top 20 hashtags on Twitter

To understand each platform's topics of trending conversation, the top 20 hashtags on respective platforms are examined upon. The percentage value for a particular hashtag used has been calculated by counting the number of posts containing that hashtag at least once, normalized by the total number of posts on that platform that mention at least one hashtag. Figure 1A shows the top hashtags on Parler. Among the most popular hashtags are *trump, maga, and trump2020*, which suggests that many of Parler's users are indeed Trump supporters and discuss political issues. There are also hashtags that refer to conspiracy theories such as *wwg1wga* and *qanon* which relates to the Qanon conspiracy (BBC News 2021). Furthermore, several hashtags (e.g., *electionfraud, voterfraud* and *stopthesteal*) are related to the alleged election fraud that Trump and his supporters claimed transpired during the 2020 US elections. Figure 1B shows the top hashtags on Twitter. Any hashtag visible that was not part of the original search terms are co-occurring hashtags. Seven out of the eight original search hashtags appeared in the top list. The graph reveals an imbalanced distribution of counts by candidate hashtags. Out of all the candidate hashtags, *capitolriots* had the highest count by a large margin compared to the others. Outside of hashtags such as *stopthesteal*, *trump* and *maga* that are affiliated with Trump ideology, popular hashtags like *trumpcoupattempt*, *removetrumpnow* and *impeachtrumpnow* reveal a stark contrast vs. Parler that advocate the removal of Trump from social media and from office. Other hashtags such as *insurrection, dcriots* and *capitolriots* infers strong disapproval towards the rioters' actions in the Capitol.

3.3.2 Term popularity

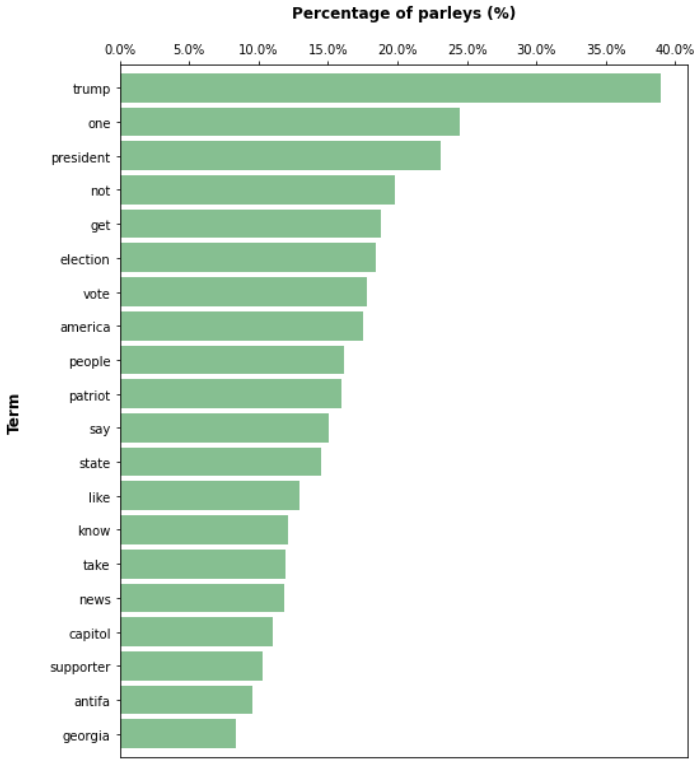


Figure 2A: Top 20 terms on Parler

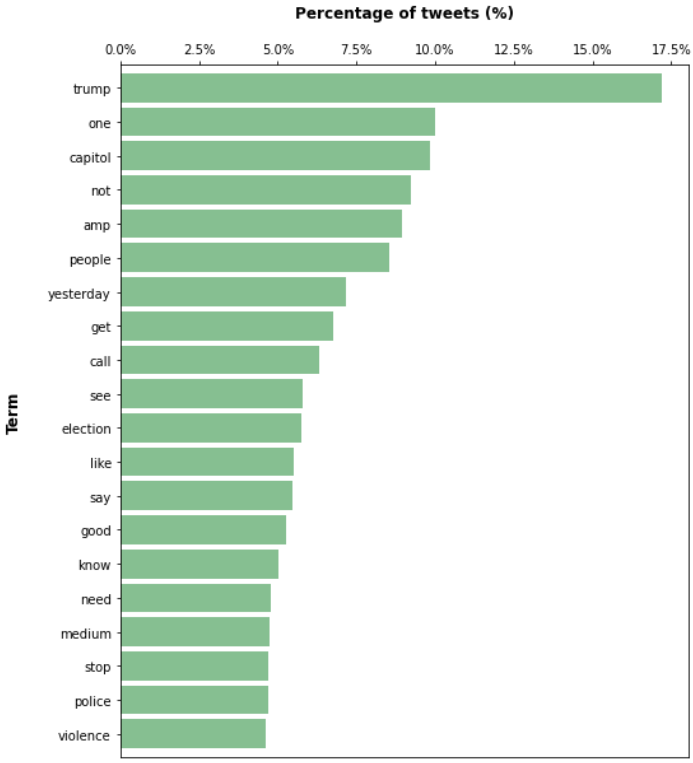
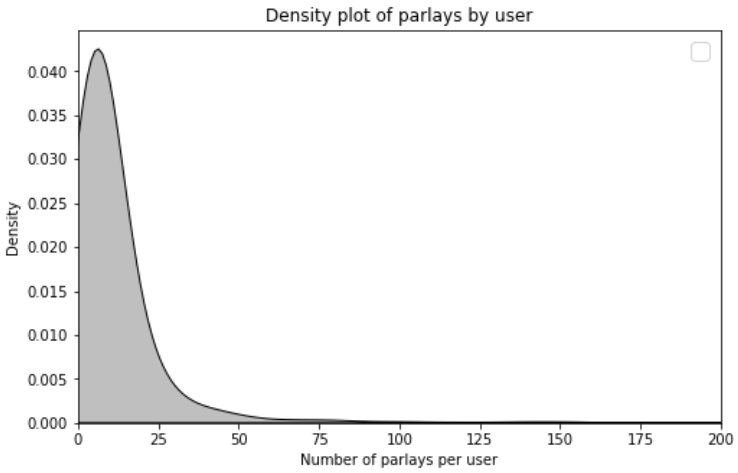


Figure 2B: Top 20 terms on Twitter

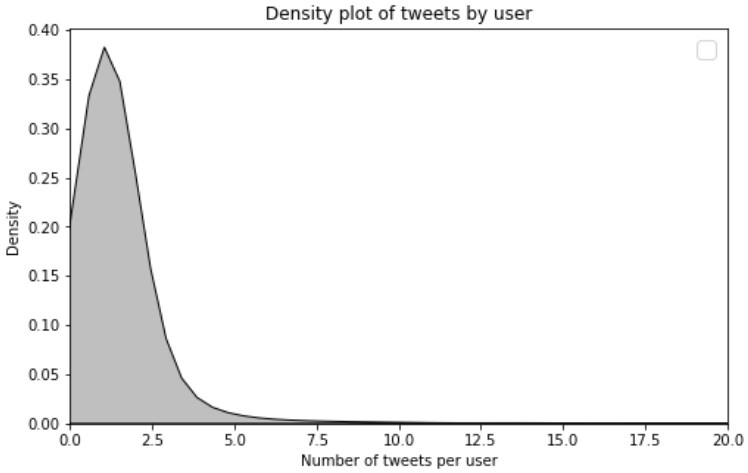
To learn more about the content being posted on both platforms, the frequency plots are repeated for the terms as shown in figure 2A and 2B. The percentage value of a particular term used has been calculated by counting the number of posts containing that term at least once, normalized by the total number of posts on that platform. The term *Trump* appeared more than twice as much on Parler than on Twitter portraying that Parler exhibits stronger political ties to the Trump party than Twitter. This is also corroborated by the observation that the term *president* is salient on Parler but not on Twitter indicating that users on Parler revere Trump as a head figure whereas this is not the case on Twitter. Frequent terms on Twitter such as *violence, need, stop* and *police* signal a sense of exigency and call to action against the rioters.

3.3.3 Density plots

A density plot visualizes the distribution of data over a continuous interval (Data Visualisation Catalogue 2021). The peak of a density plot help displays where observations are concentrated over the interval. Density plots are useful as they clearly illustrate the shape of the distribution curve as opposed to histograms.

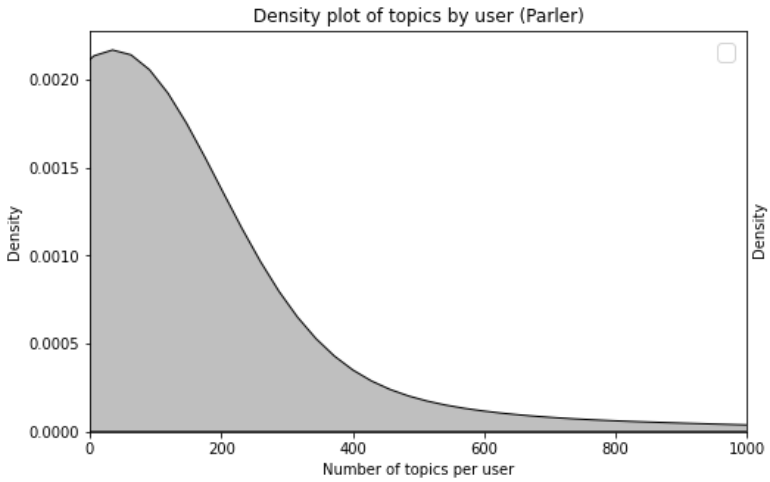


Plot 1a: (Parler)

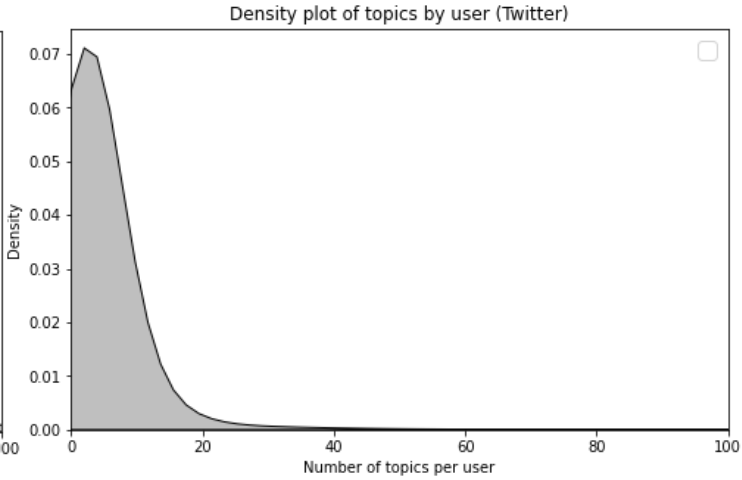


Plot 1b: (Twitter)

To get a better idea of each platform's posting behavior, density plots were created. Plot 1a and 1b above display the distribution of posts created by users on Parler and Twitter, respectively. Focusing on plot 1a, the peak of the curve is situated at around 12 parlays while the maximum number of parlays observed by a user was 535. Plot 1b similarly shows the density of posts (aka. Tweets) created on Twitter. The peak of the density curve for number of tweets per user is just 2 while maximum number of tweets posted by a user is 87. Nearly all of the users on Twitter made less than 5 posts. Broadly generalizing, Parler users posted more messages with nearly 6 times as much than the average Twitter user.



Plot 2a: (Parler)



Plot 2b: (Twitter)

The density plots 2a and 2b presents the distribution of topics by users on Parler and Twitter, respectively. Starting with plot 2a, the peak of the curve is somewhere around 50 topics while the maximum number observed was 4640. In contrast, the peak of the curve for Twitter users is around 5 topics and the majority of users share less than 20 topics. The maximum number of topics observed was just 358. Generally speaking, Parler users endorsed extensively more topics in their messages as opposed to Twitter users. This finding already signals abnormal commitment and engagement in topics of conversations by Parler users. Nevertheless, these interpretations should be taken with a grain of salt as they are small datasets of much larger population groups.

3.4 Community detection system

Latent communities of a social network population can be discovered in many ways, though typically exploiting the methods of network graphs or clustering. As stated previously in the methodology, the approach taken in this thesis makes use of clustering. Clustering is a form of unsupervised machine learning that involves the process of grouping similar items together based on certain common attributes. The algorithm does not figure out the correct output, but rather explores the data and draws inferences from datasets to describe hidden structures from unlabeled data (RSDharra 2021). In order to start clustering, a clustering algorithm typically requires a distance or similarity metric defining each observation in a n-dimensional space. The goal here is to cluster Parler users and categorize them into groups by computing a pairwise similarity score between users based on their interests, namely the hashtags that they use. The following work and computations were done on Python v3.8 involving a number of ML[11] and data manipulation libraries.

3.4.1 Hashtag similarity

The hashtag similarity metric developed by Zhang et al. (2012) for their work on community discovery on Twitter is used as the distance measure as input for a clustering algorithm. A pairwise similarity score is calculated between two users based on the number of their common hashtags and the importance of these hashtags. Thus, a given hashtag may have different weights in these two users based on the number of occurrences. The formula for hashtag similarity between users is defined as follows:

Hashtag Similarity between users U_i and U_j can be calculated as:

$$sim_{hashtag}(i,j) = \sum_{k=1}^{n}(1 - \left|\frac{N_{ik}}{|H_i|} - \frac{N_{jk}}{|H_j|}\right|)\frac{N_{ik} + N_{jk}}{|H_i| + |H_j|}$$

(Zhang et al. 2012)

Where $|H_i|$ is the total number of hashtags published by U_i, n is the number of hashtags that appear in both U_i and U_j. N_{ik} represents the number of the k^{th} hashtag in user U_i.

[11] Machine Learning

$\left| \frac{N_{ik}}{|H_i|} - \frac{N_{jk}}{|H_j|} \right|$ represents the difference of the k^{th} hashtag's weights in U_i and U_j.

$\frac{N_{ik} + N_{jk}}{|H_i| + |H_j|}$ represents the k^{th} hashtag's weight in the two users as a whole.

Generally, the bigger n is (the two users have more common hashtags), the smaller $\left| \frac{N_{ik}}{|H_i|} - \frac{N_{jk}}{|H_j|} \right|$ is and the importance of the kth hashtag in two users are more similar.

The bigger $\frac{N_{ik} + N_{jk}}{|H_i| + |H_j|}$ is, then the hashtag is more important between these two users and the larger the hashtag similarity score will be. Essentially, a hashtag similarity score of 1 indicates perfect similarity between two users while a score of 0 indicates no similarity.

The hashtag similary metric was computed for a random sample of 1000 out of the available 2327 users for faster computing time. This subsample boasted a total of 3226 unique topics. The output from this computational task is a user-user matrix with the pairwise similarity score as values which can now be fed into a clustering algorithm.

3.4.2 HDBSCAN clustering for community detection

There are many algorithms created in the field of machine learning for clustering, each have their own advantages and disadvantages. The approach taken here draws on HDBSCAN[12] clustering as it is proven to be more robust towards clustering data with high dimensionality (i.e., tons of features) (Healy, 2019). HDBSCAN is a density-based hierarchical clustering algorithm that finds clusters of point features within surrounding noise based on their spatial distribution (Campello, Moulavi & Sander 2013). Density-based algorithms are well suited at detecting clusters of high density separated from clusters of low density. HDBSCAN takes this concept a step further as it is self-adjusting meaning that it uses varying epsilon (or radius) distances, allowing for clusters with varying densities based on cluster stability and is more robust to parameter selection (McInnes, Healy & Astels 2016). This is beneficial when it is uncertain how many true clusters exist in the data, there are outliers or noise in data and when the clusters' shapes are arbitrary.

HDBSCAN takes two parameters as input: `min_cluster_size` and `min_samples`. The primary parameter to influence the resulting cluster is `min_cluster_size` which sets the smallest size grouping a cluster can be. The secondary parameter of `min_samples` provides a measure of how conservative you want your clustering to be. The larger the value given, the more

[12] Hierarchical Density-Based Spatial Clustering of Applications with Noise

conservative the clustering; more points will be declared as noise, and clusters are restricted to progressively denser regions (McInnes, Healy & Astels 2016). After testing out various configurations of parameter selection, `min_cluster_size` of 12 and `min_samples` of 4 were nominated as parameter inputs for the grouping of observations into four unique clusters surrounded by noise. While more clusters could have been detected by the algorithm by reducing either input parameter, the current output is of reasonable size to explore the research question within the respective timeframe. Clustering results may likely vary depending on the range of parameter selection and sub-sampling of the users.

3.4.3 Corpus similarity evaluation

Now that potentially radicalized communities have been identified on Parler, this makes us wonder how similar are their corpora? A statistical measure of corpus similarity would be very helpful in determining whether the rhetoric of users between clusters belong to the same homogenous population or not. If it is found that two corpora are indeed similar according to the significance test, then it is reasonable to combine them when conducting corpus linguistics. To check for corpora similarity within the discovered communities, the word frequency measures of each cluster were considered. Frequency lists are useful as they are a representation of the text, which is susceptible to automatic, objective manipulation whereas the full text is very rich in information that cannot be readily used to make similarity judgements (Kilgariff & Rose 1997). Moreover, word frequency measures are cheap and easy to generate which can serve as a quick heuristic to judge the corpora similarity when a more extensive evaluation of the two corpora is not viable.

Cluster 3 was selected as the candidate cluster to compare corpora similarity with other clusters as it is a cluster that clearly shows substantial interest in radical ideologies with noticeably greater engagement. This finding is elaborated further in section 4.4 of the Results. Previous works into investigating statistical metrics for corpora similarity by Kilgarriff & Rose (1997) found that they had achieved the best results with 320 or 640 words. For the experiments below, the most frequent 100 words in the union of the two corpora is considered as the linguistic data concerns short text documents. Since the measure does not directly permit comparison between corpora of different sizes and each corpus is indeed different in size, the word frequencies have been normalized to their occurrence per 10,000 words. The formula for normalizing word frequencies for comparison between corpora is stated as follows:

$$f_n = f_o \cdot n/c$$

Where f_n is the normalized frequency, f_o is the observed (or raw) frequency, c is the total number of words in the corpus, and n represents the number we want to normalize to.

The method for computing corpus similarity is as follows:

1. Divide corpus A and corpus B into ‘slices’;
2. Create two subcorpora by randomly allocating half the slices to each;
3. Iterate with different random allocations of slices;

4. Calculate mean and standard deviation over all iterations.

This method is repeated until all clusters as well as the twitter corpus have been compared against the candidate cluster. The output from this method is reported in section 4.5 under Results.

3.5 Corpus linguistics

Corpus linguistics is a field which focuses upon a set of procedures, or methods, for studying language and register (McEnery & Hardie 2012). Broadly, corpus linguistics seeks to uncover what patterns are associated with lexical and grammatical features in a corpus. A corpus is a large, principled collection of naturally occurring examples of language stored electronically. A corpus if nothing else, is evidence of language use (Tognini-Bonelli 2001). Corpora[13] are invariably explored using software linguistic tools as the set of texts is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe (McEnery & Hardie 2012). There are two main theoretical approaches to conducting corpus linguistics which is corpus-based versus corpus-driven. The current thesis focuses on the corpus-based type which typically use linguistic data in order to explore a research question or hypothesis with the aim to validate, refute or refine existing assumptions. Nevertheless, both approaches are predicated on four major characteristics when conducting language analysis (Biber, Conrad, & Reppen 1998):

1) It is empirical, analyzing the real patterns of language use in natural texts.
2) It utilizes a large, natural, and principled collection of texts as the basis for analysis.
3) It makes extensive use of computers for analysis.
4) It relies on both quantitative and qualitative analytical techniques.

3.5.1 Descriptive statistics

Descriptive statistics are statistics that simply describe the data in some way. The most widely used statistical measure to report on is a 'frequency count' referred to as a simple tallying of the number of instances of something that occurs in a corpus (McEnery & Hardie 2020). Which are the most common words? Where does a word lie in a continuum from very common to very uncommon? Such answers to these questions provide a first step towards understanding the patterns of use associated with a word (Biber, Conrad & Reppen 1998). Moreover, word frequency measures can be used to analyze the vocabulary contained in a corpus in order to find domain-specific vocabulary (Weisser 2013).

As for descriptive statistics of lexical components, N-gram features (up to n=3) of each corpus were extracted, and their frequency counts are described upon in the results section. An n-gram is a contiguous sequence of *n* items from a given sample of text (Broder 1997). They can be uni-gram (1 word), bi-gram (2 words) or tri-gram (3 words) depending on the value of *n*. Bi-grams and tri-grams are useful features for extraction as they are a prominent way of studying phrases, a technique known as collocation. Collocation is simply the statistical tendency of words to co-occur. From studying collocations, it is well known that there is a tendency for each collocate of a word to be associated with a single sense of that word. This is discernible when looking at

[13] Plural form of corpus.

the phrases using *trump – stop trump, president trump*: *stop trump* refers to a formative action against Trump whereas *president trump* is a referral to the entity Donald Trump. Studying collocations can also help us better understand particular words used in a certain phrase (Bennett 2010). In order to make direct comparisons in the word frequencies of different corpora, the frequency scores are normalized to a fixed number of total words which varies depending on the n-gram type.

3.5.2 Beyond descriptive statistics

Domain-specific vocabulary significance test

Domain-specific vocabulary refers to the language that is primarily used within one area of knowledge but not others (Drew 2019). The domain specificity of the language is also a marker of inclusion within a social community. Corpus linguists make use of frequency counts to test for a statistical difference in the frequency of a word between two corpora. Knowing that individual word forms in one corpus occur more or less often than in another corpus may help characterize some generic differences between those texts (Wordhoard 2021). Statistics on keyword occurrences between texts provide a framework for judging how likely or unlikely observed differences are to have occurred by chance, and so deserve further attention and interpretation. The log-likelihood statistical measure is used to test for this significance (Pojanapunya & Todd 2018). The domain-specific vocabulary applied in this research regards political far-right extremism. To date, there are no official or scientific resources of right-wing extremist specific vocabulary for research purposes. Therefore, selection of the terms was manually curated based on reporting of previous literature that examined relevant traits that are typical of extremist discourse as discussed by Torregrosa et al. (2021) and Hitkul et al (2021) while taking into consideration the context of election fraud and the Capitol riots. It is important to note as well that this is not an exhaustive list and was created by a non-expert on the matter, but the terms present in this lexicon[14] should suffice as a preliminary endeavor. There is a decent chance that other keywords of interest may have been overlooked. Nonetheless, the generated lexicon comprises a total of 17 items. The lexicon includes nouns (e.g., patriot, qanon, death etc.) and verbs (e.g., kill, hang, shoot etc.) that are typically used to share and discuss far-right ideals.

Concordance analysis

One crucial task in corpus linguistics is to search in a corpus of text for illustrative samples of lexical features and retrieve them, a technique known as concordance. Concordance analysis involves identifying instances of a word in a corpus, presented with the words surrounding it. Reading concordances means looking at targeted words in their context of occurrence in texts and allows the analyst to study the meaning of the word in the text, and to see how meaning is created in the particular case. Simply searching through a corpus and looking at examples individually is to treat the corpus like a text; it is through concordancing that the patterns of usage and the paradigms are revealed (Kytö, Lüdeling & de Gruyter 2007). They are essential for checking results derived by automatic procedures and to examine examples in text in more detail as a way to characterize and make substantive claims about the nature of the corpus. Typically, in such a task, the substantial quantitative results generated from the corpus are taken as keywords and analyzed upon qualitatively in a manner called ‘keyword in context’ (KWIC) to find

[14] The vocabulary of a person, language, or branch of knowledge.

significance in language. This means that concordance lines are arranged by a sorting criterion whereby the keyword is placed in the middle of the line with a variable or fixed length of context on either side of it (Tang 2021).

Concordance analysis was carried out using the NLTK[15] (Natural Language Toolkit) library on Python. NLTK is a library that offers many useful functions important for human natural language processing. Before the concordance output can be generated, the text undergoes preprocessing steps that include removal of noisy text data (e.g., mentions, hashtags, URLs etc.), punctuations, white space, lemmatization, and tokenization. Tokenization and lemmatization thereafter are performed here as it allows for generating concordance lines of the keyword in context in their standardized dictionary form so that more examples can be systematically and invariably viewed together.

Frequency statistics and concordance exemplify respectively the two forms of analysis, namely quantitative and qualitative, that are equally important to corpus linguistics. Despite that, the frequency count and significance tests of linguistic terms are useful only to a certain extent. It is only when concordance occurs to see how salient keywords are used to convey meaning in discourse that meaningful claims about language arises out of it. Thus, concordance serves as a window into better understanding the nature of the corpus.

4 Results

The following sections provide an overview of the obtained results from this study. Section 4.1 reports the results from HDBSCAN clustering. Section 4.2 gives a tabular overview of the final datasets used. Section 4.3 depicts a t-SNE visualization of the detected sub-communities. Section 4.4 presents a tabular overview of the top hashtags in each cluster. Section 4.5 reports on the corpora similarity outcomes. Section 4.6 displays results gathered from n-gram frequency measures. Section 4.7 reports on the results for significance tests of domain-specific vocabulary comparing all Parler corpora against the Twitter corpus. Finally, Section 4.8 features the concordance analysis of specifically curated keywords.

4.1 HDBSCAN clustering

Label	Counts
-1 (noise)	548
0	23
1	60
2	289
3	80

[15] https://www.nltk.org/

Table 1: Number of users per cluster

Table 1 describes the number of users that belong to a particular cluster label found by HDBSCAN clustering using the hashtag similarity as the distance metric. HDBSCAN identified four clusters of users while the rest is labelled as noise. The largest cluster was label 2 with 289 users while the smallest cluster was label 0 with 23 users. Nearly 46% of the entire sample space belonged to an identified cluster while the rest was noise.

4.2 Overview of final datasets

Overview Data

	Source	Terms	Unique terms	Messages	Users	Lexical diversity
	Parler	168,347	13,905	10,420	452	0.083016
	Twitter	121,058	10,894	11,599	8141	0.090061
Difference		47,289	3011	-1179	-7689	-0.007045

Table 2: Summary of Parler and Twitter datasets

Table 2 describes the final details of the two datasets that will be used to acquire the results hereafter. The Parler dataset now comprises four discovered sub-communities. The table shows a larger number of unique terms on Parler over Twitter, but this finding is offset by the difference in total terms favored towards Parler. The difference in messages between both datasets is only 1179 which makes them sufficiently comparable, although it is important to keep in mind that information from the Parler dataset is analyzed as sub-communities of the social network. The table also show that there is little to no difference in lexical diversity and is therefore not a meaningful differentiating factor.

4.3 t-SNE visualization

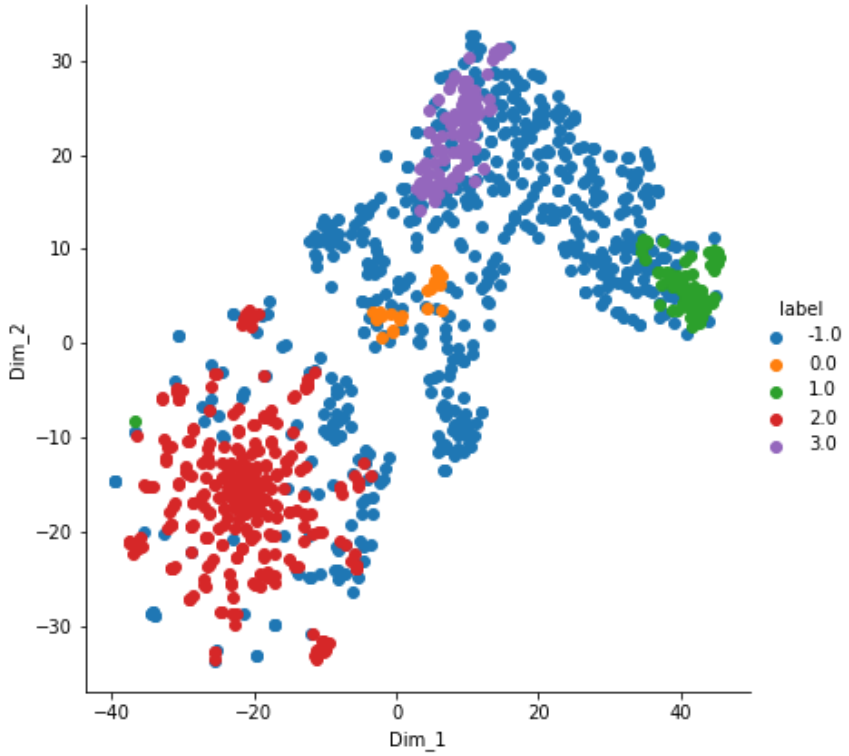


Figure 3: t-SNE visualization of the discovered clusters by HDBSCAN clustering

Figure 3 depicts a t-SNE plot of the four identified clusters projected onto a 2D-plane. t-SNE is a non-linear dimensionality reduction technique that is particularly well suited for visualization of high dimensional datasets (Vidiyala 2020). Essentially, t-SNE captures local structure in the data, meaning that neighboring points in the input space often tend to be neighbors in the low-dimensional space. Reducing high dimensionality of the data makes it more interpretable for human analysts to make sense of it visually. The t-SNE plot shows a clear distinction of clusters that have their neighbor points densely crowded together.

4.4 Top hashtags in each cluster

Cluster 0	Cluster 1	Cluster 2	Cluster 3
trump2020 - 439	stopthesteal - 288	italydidit - 35	stopthesteal - 1120
fyp - 434	italydidit - 60	stopthesteal - 34	maga - 1016
duet - 239	wethepeople - 46	georgia - 29	trump - 990
trump - 178	cd9tc - 41	antifa - 24	trump2020 - 781
trumpsquad2020 - 160	maga - 31	jonosoff - 22	voterfraud - 713
greenscreen - 113	stopthecoup - 31	covid19 - 21	electionfraud - 688
patriots - 95	notag - 31	fightback - 19	wwg1wga - 610

Cluster 0	Cluster 1	Cluster 2	Cluster 3
america - 95	7mdk7 - 28	raphaelwarnock - 17	draintheswamp - 601
foryoupage - 91	electionfraud - 28	billgates - 17	qanon - 555
maga - 86	contecomeclean - 26	kamalaharris - 15	dominion - 548
republican - 82	libertyfirst - 25	joebiden - 15	patriots - 523
georgia - 73	trumpwon2020crazyhorse - 23	mikepence - 14	fakenews - 521
usa - 69	fightback - 20	parler - 14	freedom - 492
patriotparty - 60	italygate - 18	stormthecapitol - 14	donaldtrump - 468
proudboys - 56	stopthedeepstatecriminalpoliticians - 15	contecomeclean - 13	trumptrain - 463
election2020 - 56	voetsekanc - 15	briankemp - 13	covid19 - 461
foryou - 56	wakeywakey - 15	oprahwinfrey - 13	deepstate - 437
iii - 56	millionmagamarch - 15	blm - 12	americafirst - 437
trumptrain - 52	electionfraud2020com - 13	memes - 12	fightback - 434
trump2021 - 52	voterfraud - 13	mitchmcconnell - 11	parler - 408
womenfortrump - 47	fuckyoucyril - 11	wwg1wga - 11	usa - 405
projectcar - 47	antifa - 11	hillaryclinton - 11	truth - 393
americaproud - 43	thiswewilldefend - 11	chuckschumer - 10	presidenttrump - 375
stitch - 43	holdtheline - 11	maga - 10	bidencrimefamily - 371
freedom - 43	libertyordeath - 11	trump - 10	kag - 368
biden2020 - 43	wearethecure - 11	presidenttrump - 10	maga2020 - 346
draintheswamp - 43	trumpwon2020 - 11	treason - 9	gop - 343
dc - 43	1776 - 11	chinazi - 9	election - 342
biden - 43	tyrannyisthevirus - 11	spygate - 9	conservative - 342
conservative - 39	bideniscorrupt - 11	sundarpichai - 9	scotus - 337

Table 3: Overview top 30 hashtags of Parler communities (normalized to per 100 users)

Table 3 provides an overview of the top 30 hashtags used by communities on Parler. The counts of the hashtags have been normalized accordingly to a group of 100 users in order to draw meaningful comparisons between groups. The majority of topics across all communities largely concentrate around elements of political grievances, calls for retaliation, conspiracy theories and far-right groups. Across all results, communities show a strong affinity towards topics that promote Trump ideology (e.g., trump2020, maga2020, trumptrain etc.) with many references also alluding

to voter fraud and far-right group memberships such as *patriotparty, proudboys* and *qanon*. The community that shows the highest engagement with extremist-related topics is cluster 3 as all their popular hashtags, on average, are used multiple times whereas other communities do not share this distinctive quality. The top hashtags *stopthesteal, maga,* and *trump* were used on average 10 times by the users in this group. This observation demonstrates a high motivational imbalance in individuals of cluster 3 towards extremist topics of conversations which is indicative of online radicalization. Cluster 2 appears to be the least coherent and radicalized community in terms of common hashtags used since the highest value observed is only 35. Nevertheless, a lot of the hashtags in this cluster still bear signs of far-right ideology such as *stormthecapitol* and *fightback* which hint towards possible instances of radicalized speech. Though, all communities exhibit evidence of online radicalization to a notable degree as observed by their collective use of hashtags.

4.5 Corpora similarity

Corpus A	**Corpus B**	**Mean χ2 value**	**Standard deviation of χ2**	**Exceed the test statistic of** 148.230 **at p=0.01?**
Cluster 3	Cluster 0	1056.41	637.74	✓
Cluster 3	Cluster 1	1765.87	697.64	✓
Cluster 3	Cluster 2	1552.78	527.57	✓
Cluster 3	Cluster –1 (noise)	1240.86	254.47	✓
Cluster 3	Twitter	1357.71	239.88	✓

Table 4: Chi-squared tests for corpora similarity between HDBSCAN clusters

Table 4 provides results for corpora similarity tests. According to the chi-squared distribution table[16] published online at MedCalc (2021), the critical value for significance at p=0.01 is given as 148.230 when the degrees of freedom equals 99. In statistics, degrees of freedom act as variables in the final calculation of a statistical measure that are used to determine the outcome of different scenarios in a system (Frost 2021). Degrees of freedom can be calculated as (m – 1) * (n – 1) where m = number of columns and n = number of rows of a contingency table. In this scenario, the number of columns is two representing corpus A and corpus B while the number of rows represent the number of word frequency measures given as 100.

The chi-squared distribution table can be used to compare the observed chi-squared value against the critical value in testing the null hypothesis. Corpus similarity tests a null hypothesis asserting that various populations are equal with respect to some characteristics of interest (i.e., word frequency measures). This means that if the calculated χ2 value exceeds the given critical value, then the null hypothesis can be rejected in favor of the alternative hypothesis (H_1) which states that the word frequency measures are significantly different in their distributions between the two corpora. The higher the χ2 value is, the less likely it is that the differences are due to random chance. It can be seen from table 4 that all corpora comparisons exceeded the specified

[16] https://www.medcalc.org/manual/chi-square-table.php

critical value suggesting that none of the corpora are drawn from the same population. As such, the corpora are allowed to remain in their current state for subsequent analyses.

4.6 N-gram frequency measures

By comparing the frequency measures of n-grams between corpora, it was possible to establish similarities and differences across the results. These results provide an indication of what lexemes[17] are commonly being used in the communities that define the general scope of the conversations.

4.6.1 Uni-gram

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Twitter	
counts	word	counts	word	counts	word	counts	word	counts	word
328	tiktok	124	trump	151	trump	206	trump	173	trump
293	music	100	election	82	president	111	follow	99	amp
284	create	86	vote	71	election	97	echo	97	capitol
282	short	82	patriot	59	people	88	patriot	90	people
208	sound	58	people	57	get	82	president	68	yesterday
203	original	57	president	56	vote	73	election	65	not
160	trump	57	not	52	patriot	69	news	61	election
72	president	54	say	52	say	68	vote	57	get
64	election	54	get	50	capitol	63	comment	57	call
56	vote	51	fraud	45	news	57	people	54	say
54	biden	51	one	44	not	54	capitol	53	like
53	say	50	antifa	41	one	49	supporter	51	good
51	viva	48	know	40	supporter	46	see	49	see
50	antifa	47	take	39	know	46	say	49	police
45	live	45	georgia	37	state	45	us	49	know
45	watch	45	us	36	police	45	get	48	need
45	patriot	43	make	36	america	41	break	47	medium
42	state	42	capitol	36	time	41	watch	47	stop
42	georgia	42	like	36	us	40	police	46	violence
37	house	40	call	36	like	40	like	44	one

Table 5: Top 20 uni-gram features of Parler and Twitter corpora (normalized to per 10,000 counts)

Table 5 shows an overview of the most frequent uni-gram features from each platform. The prominent uni-grams (also referred to as 'terms') among Parler communities show strong ties to far-right political views and discussions about the Capitol riots and election fraud. For example, the terms *trump*, *election* and *patriot* commonly appear across the Parler communities illustrating that most conversations are focused on a political dimension that is more right-leaning.

[17] A basic lexical unit of a language, consisting of one word or several words, considered as an abstract unit.

Interestingly, the uni-gram features of cluster 0 is notably different from the rest. The terms *tiktok* and *music* appear most frequently but other less frequent terms like *trump, election*, and *vote* indicate that conversations here may still be relevant to the research question. Twitter uni-grams also feature many of the terms commonly found in Parler communities, but also feature context relevant distinct terms not common on Parler like *violence* and *stop.*

4.6.2 Bi-gram

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Twitter	
count	bigram	counts	bigram	counts	bigram	counts	bigram	counts	bigram
597	short tiktok	61	trump supporter	72	president trump	80	trump supporter	70	social medium
597	tiktok music	50	president trump	68	trump supporter	72	president trump	59	incite violence
597	create short	42	patriot listecho	34	electoral college	71	follow echo	53	sign petition
429	original sound	42	commentfollow seat	31	mike pence	60	caboose loose	53	medium ban
422	music original	42	listecho commentfollow	29	donald trump	47	donald trump	53	dangerous rhetoric
98	president trump	30	seat hide	26	vice president	41	comment follow	53	spread dangerous
61	trump supporter	27	echo echo	22	storm capitol	40	follow patriot	53	ban good
38	electoral college	25	look like	21	stop steal	40	add handle	53	stop trump
34	joint session	23	high level	20	capitol building	36	mike pence	53	violence social
31	trump rally	23	level part	19	movie review	36	loose follow	53	add name
31	sidney powell	22	election fraud	17	lin wood	36	handle comment	53	rhetoric incite
27	antifa thug	18	online voice	16	electoral vote	36	echo add	52	call stop
24	bless jan	18	voter fraud	16	joe biden	35	list comment	52	petition call
24	viva president	18	hide hide	15	election fraud	35	patriot list	52	trump spread
24	donald trump	18	vocaroo online	15	white house	34	electoral college	52	good add
24	freedom plaza	18	georgia senate	15	save america	33	vice president	41	trump supporter
24	brad raffensperger	18	white house	14	election result	30	good news	26	capitol police
24	armed soldier	17	antifa blm	14	college vote	29	president donald	25	storm capitol
24	soldier viva	17	electoral vote	14	president mike	29	election fraud	24	donald trump
24	worry armed	17	steal election	14	president donald	27	app offer	23	domestic terrorist

Table 6: Bi-gram features of Parler and Twitter corpora (normalized to per 20,000 counts)

Table 6 provides an overview of the most frequent bi-gram features on Parler and Twitter. Looking at bi-grams gives a clearer picture on the collocates that produce a distinct sense of the lexeme. Across all Parler communities, the bi-grams *trump supporter* or *president trump* appear most frequently with the exception of cluster 0. Interestingly, *president trump* does not appear at all in the Twitter results. Cluster 0 exhibits a noticeable difference as compared to the other clusters that is characterized by an abnormal proportion of TikTok related terms when compared to other bi-grams in that cluster. TikTok is a video-sharing focused social networking service. This distinction shows that cluster 0 revolves around engagement of sharing TikTok related content on Parler. Other bi-gram features among Parler clusters such as *echo echo*, *comment follow, and* echo *add* convey motivations by Parler users to further spread their messages so that it can gain more traction within the entire platform, an observable effort of media manipulation. Other bi-grams that echo the rhetoric of the attacking mob worthy of mentioning include *antifa thug, stop steal, hide hide*, *save america*, and *storm capitol*. Meanwhile, the top bi-gram features on Twitter paint a stark difference in opinion. Bi-grams such as *stop trump,* spread dangerous, *dangerous rhetoric,* and *sign petition* advocate the immediate removal of Trump from office in an effort to punish his incitement to violence from his followers. Also, some users on Twitter make references to the attacking mob as *domestic terrorist* whereas this is not present on Parler.

4.6.3 Tri-gram

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Twitter	
counts	trigram	counts	trigram	counts	trigram	counts	trigram	counts	trigram
957	create short tiktok	68	listecho commentfollow seat	23	electoral college vote	58	loose follow echo	89	social medium ban
957	short tiktok music	68	patriot listecho commentfollow	22	vice president mike	56	caboose loose follow	88	spread dangerous rhetoric
676	music original sound	47	commentfollow seat hide	22	president donald trump	56	follow echo add	88	medium ban good
660	tiktok music original	41	echo echo echo	21	matt movie review	56	echo add handle	88	incite violence social
38	god bless jan	37	high level part	20	president mike pence	56	comment follow patriot	88	violence social medium
38	bless jan 2021	29	vocaroo online voice	14	front page political	56	add handle comment	87	dangerous rhetoric incite
38	viva jenna ellis	21	recorder vocaroo quick	14	america front page	56	handle comment follow	87	rhetoric incite violence
38	viva president trump	21	vocaroo quick easy	14	page political news	55	follow patriot list	87	trump spread dangerous
38	soldier viva rudy	21	voice recorder vocaroo	14	citizen free press	55	patriot list comment	87	stop trump spread
38	president trump worry	21	easy way share	13	high level part	43	app offer good	87	sign petition call
38	rudy giuliani viva	21	way share voice	12	joint session congress	43	news app offer	87	call stop trump
38	trump worry armed	21	quick easy way	12	storm capitol building	43	offer good news	87	petition call stop
38	armed soldier viva	21	online voice recorder	12	trump supporter storm	43	president donald trump	86	good add name
38	worry armed soldier	21	seat hide hide	11	short tiktok music	43	dml news app	86	ban good add
38	powell viva jenna	21	voice message interwebs	11	create short tiktok	43	good news reporting	18	police involve capitol
38	viva rudy giuliani	21	share voice message	11	music original sound	41	follow echo without	18	capitol attack yesterday
38	viva sidney powell	19	brand fastener patch	9	tiktok music original	41	please like share	17	involve capitol attack
38	sidney powell viva	19	velcro brand fastener	9	save america rally	41	like share follow	17	confirm police involve
33	join rsbn crew	19	georgia senate runoff	9	movie review matt	41	conversation comment4 follow	14	need hold accountable
33	freedom plaza ahead	17	patch proudly make	9	review matt movie	41	echo without comment	11	wait mondaydo today

Table 7: Tri-gram features of Parler communities and Twitter (normalized to per 30,000 counts)

Table 7 presents an overview of the most frequent tri-grams on respective platforms. Trig-rams may be indicative of reshared posts or of people speaking on a common subject. Again, results of cluster 0 show disproportionately high counts for TikTok-related language. Nevertheless, due to the other far less occurring tri-grams in the cluster such as *viva president trump, viva rudy giuliani*, and *god bless jan*, cluster 0 does indeed show a deep affiliation towards Trump ideology to an extent. Prominent tri-grams among other Parler clusters can fall under the categories of signal diffusion and news reporting. Signal diffusion refers to the acts of spreading the message throughout the platform's network as evidenced by many tri-gram features like *follow echo add, please like share,* and *like share follow* revealing a collective drive to normalize the rhetoric displayed on Parler. In contrast, Twitter results shown here build on the findings arrived at from the bi-gram results and the conclusions gathered by Hitkul et al. (2021). Indeed, a lot of the tri-grams on Twitter voice a strong opposition towards Trump's endorsing violence from his followers as well as the actions of the rioters with urgent calls to remove Trump on both social media and from office (e.g. *stop trump spread, petition call stop*, *social medium ban*, *need hold accountable* etc.)

4.7 Domain-specific vocabulary significance test

This section provides an overview of the results obtained from domain-specific vocabulary significance tests comparing all clusters against the twitter corpus. The column headers 'O1' and 'O2' represent the observed frequencies from each corpus, respectively. '1%' and '2%' are the observed frequencies normalized to their percentage form. The LL score is the log-likelihood, which indicates whether the result can be treated as significant. The LL must be above 3.84 for the difference to be significant with 95% confidence level or when $p < 0.05$. Log-likelihood measures highlighted in yellow signal that the item on average, appears more often in the first corpus whereas LL scores highlighted in light blue means the item is more prominent on Twitter. This is also indicated by the positive (i.e. Parler) or negative (i.e. Twitter) sign respectively that comes after the percentage value on column '2%'. The higher the LL is, the less likely it is that the result is due to random chance (McEnery & Hardie 2012). Therefore, if more terms are salient

for a Parler corpus than for Twitter, then this is indicative of radicalized rhetoric being used and these terms may be nominated as keywords that warrant a closer look in their language patterns when conducting concordance analysis in order to validate or refute such assumptions.

Cluster 0 vs. Twitter					
Item	**O1**	**1%**	**O2**	**2%**	**LL**
antifa	31	0.49	342	0.4	0.94
death	4	0.06	28	0.03	0.71
fight	5	0.08	309	0.36 -	18.06
fuck	7	0.11	59	0.07	0.81
gun	0	0	31	0.04	1.9
hang	1	0.02	9	0.01	0.06
kill	2	0.03	31	0.04	0.02
march	3	0.05	164	0.19 -	8.07
patriot	28	0.44	396	0.46	0.02
protest	9	0.14	227	0.27	3.5
qanon	0	0	8	0.01	0.01
riot	1	0.02	86	0.1 -	5.06
shoot	3	0.05	88	0.1	1.54
storm	9	0.14	158	0.18	0.39
trump	101	1.6	1319	1.54	0.09
war	8	0.13	69	0.08	0.88
white	16	0.25	105	0.12 +	5.4

^Table 8: Cluster 0 vs. Twitter

For lexicon[18] significance tests between Cluster 0 and Twitter, only one word, *white* was significantly different in favor of Parler. In contrast, three words (e.g., march, riot, and fight) were more salient on Twitter. This adds further evidence that the content in cluster 0 may be more benign in nature regarding far-right language use.

Cluster 1 vs. Twitter					
Item	**O1**	**1%**	**O2**	**2%**	**LL**
antifa	84	0.5	342	0.4	2.8
death	4	0.02	28	0.03	0.15
fight	34	0.2	309	0.36 -	11.82
fuck	23	0.14	59	0.07 +	6.15
gun	8	0.05	31	0.04	0.2
hang	3	0.02	9	0.01	0.15
kill	13	0.08	31	0.04 +	3.87
march	16	0.09	164	0.19 -	8.19

[18] the vocabulary of a person, language, or branch of knowledge.

patriot	138	0.81	396	0.46 +	29.28
protest	22	0.13	227	0.27 -	11.78
qanon	3	0.02	8	0.01	0.28
riot	6	0.04	86	0.1 -	7.32
shoot	24	0.14	88	0.1	1.51
storm	25	0.15	158	0.18	0.93
trump	209	1.23	1319	1.54 -	9.34
war	29	0.17	69	0.08 +	9.57
white	18	0.11	105	0.12	0.2

^Table 9: Cluster 1 vs. Twitter

Table 9 highlights the significance tests between cluster 1 and Twitter. Keywords that include *patriot*, *fuck*, *kill* and *trump* are more salient on Parler whereas 5 out of the remaining 13 terms are more significant on Twitter. A large LL score for *patriot* in cluster 1 signals an unusually high usage of the term which warrants a closer examination in its pattern of use during concordance. Other keywords that draw on themes of anger and violence like *fuck* and *kill* are also of interest.

Cluster 2 vs. Twitter					
Item	**O1**	**1%**	**O2**	**2%**	**LL**
antifa	381	0.32	342	0.4 -	8.65
death	94	0.08	28	0.03 +	18.35
fight	181	0.15	309	0.36 -	88.73
fuck	170	0.14	59	0.07 +	25.01
gun	71	0.06	31	0.04 +	5.19
hang	33	0.03	9	0.01 +	6.86
kill	120	0.1	31	0.04 +	29.74
march	111	0.09	164	0.19 -	34.57
patriot	616	0.52	396	0.46	2.93
protest	132	0.11	227	0.27 -	65.79
qanon	9	0.01	8	0.01	0.04
riot	31	0.03	86	0.1 -	46.92
shoot	186	0.16	88	0.1 +	10.52
storm	223	0.19	158	0.18	0.01
trump	1796	1.51	1319	1.54	0.33
war	116	0.1	69	0.08	1.39
white	123	0.1	105	0.12	1.5

^Table 10: Cluster 2 vs Twitter

Table 10 describes the significance tests between cluster 2 and Twitter. Among the many terms that was significant in cluster 2, the terms *kill, fuck,* and *death* had relatively high LL scores. All of these terms relate to possible hate speech which is a strong indicator of extremist discourse taking place. In spite of this, there were many terms that was significant in favor of Twitter too, however,

these terms when considered together appear generalizable to regular discourse in covering the event of the Capitol riots.

Cluster 3 vs. Twitter					
Item	**O1**	**1%**	**O2**	**2%**	**LL**
antifa	99	0.38	342	0.4	0.11
death	13	0.05	28	0.03	1.14
fight	54	0.21	309	0.36 -	15.17
fuck	13	0.05	59	0.07	0.85
gun	7	0.03	31	0.04	0.27
hang	2	0.01	9	0.01	0
kill	26	0.1	31	0.04 +	12.73
march	12	0.05	164	0.19 -	32.74
patriot	227	0.88	396	0.46 +	54.63
protest	53	0.2	227	0.27	2.79
qanon	4	0.02	8	0.01	0.23
riot	9	0.03	86	0.1 -	11.17
shoot	65	0.25	88	0.1 +	26.84
storm	56	0.22	158	0.18	0.86
trump	531	2.05	1319	1.54 +	29.91
war	25	0.1	69	0.08	0.42
white	30	0.12	105	0.12	0.03

^Table 11: Cluster 3 vs Twitter

Table 11 provides results for significance tests between cluster 3 and Twitter. Cluster 3 showed significant differences in terms including *patriot, trump, shoot* and *kill* with relatively high LL scores. Such keywords, when used in language typically portray a violent rhetoric and calls for further inspection when applying concordance analysis.

4.8 Concordance analysis

This final section of the results presents a comprehensive overview of the concordance lines of carefully chosen keywords for each sub-community on Parler and conversations on Twitter. Each corpus features 2-3 concordance searches accompanied by brief analytical interpretations of the patterns and typicalities found in the generated concordance lines. Each concordance presents 20 random samples of the keyword in context with a surrounding margin of 8 words on either side. Most of the keywords used in this analysis are salient terms derived from the domain-specific vocabulary tests of section 4.4 while others have been derived from keyword frequencies.

4.8.1 Cluster 0

1 'shinn mommashinn have create a short video on tiktok with music original soundwho else be feed upcheatersneverwin ',
2 'capitalbuildingprotuser6911328812911 bsengineer have create a short video on tiktok with music original soundthis guy want the cameraman ',
3 'crawford fedupwithpoliticians2 have create a short video on tiktok with music original soundduet with blackbeard03x try it ',
4 'reynolds blakereynolds22 have create a short video on tiktok with music original soundhttpsvmtiktokcomzmjw52bgp bastilleday2021jan6 ww3 january6 inc
5 'walker kmwalk31 have create a short video on tiktok with music original soundmemehsnana trumpwon trumpwontconcede democratscheate draintheswa
6 'badactorjustmenow pink3515 have create a short video on tiktok with music original soundduet with yahoonew plan patriot ',
7 'gustafson fam_man84 have create a short video on tiktok with music original soundfyp boom trump curruption patriot ',
8 'viralhttpsparlercompostfc9bb1834a64401cb1ecae371b30d04a vaccine bewareyo thatcrazyforyou fyp fypシ foryoupage tik tiktok xyzbca public makemefam
9 'mcelroy jesslanamac have create a short video on tiktok with music original soundyou can hear the evil ',
10 'tiktoksunshine countrygirlcansurvive5 have create a short video on tiktok with music original soundhttpsvmtiktokcomzmjw56yne why do the news ',
11 'sanchez hotmessmom_3 have create a short video on tiktok with music original soundrepost draintheswamp trumptrain2020holdthelinehttpsvmtiktokcom
12 'hemp corypunk have create a short video on tiktok with music song of the viking my mother ',
13 'gallegos rednexican1776 have create a short video on tiktok with music original soundstitch with rosie fypシ fyp ',
14 'nave tomernave0 have create a short video on tiktok with music original soundduet with desi_simm67 trumpsquad2020 trumptrain2020trump2020 ',
15 'hernandez alexhernandez9513 have create a short video on tiktok with music star spangled bannerone of the proud ',
16 'aoc aoc14ny have create a short video on tiktok with music the aftermath sad dramaduet with havenothadmycoffee ',
17 'barlow craig_13arlow have create a short video on tiktok with music mr red white and blueamericaproud trump2020httpsvmtiktokcomzmjwtw61vamerica
18 'flagevanl18 evanl188 have create a short video on tiktok with music original soundi think the vet do ',
19 'serious canadian_canuck1867 have create a short video on tiktok with music original soundtrudeau have to go biden ',
20 'love youjesus_is_king1239 have create a short video on tiktok with music beethoven moonlight sonata high sound qualityduet '

^Image 1: Cluster 0 - 'tiktok'

Image 1 depicts the concordances of 'tiktok'. The results for cluster 0 reaffirm the notion gathered from previous results that this sub-community is largely communicating in the same manner - through the sharing of TikTok videos. Although, additional empirical evidence is required to make inferences about whether the TikTok videos themselves display far-right characteristics. Nonetheless, this is an inquiry that is outside the scope of the research as it considers a separate modality from language. In light of the previous evidence gathered, some of the language used in cluster 0 do show glimpses of radical rhetoric. For example, towards the end of lines: 1, 5 and 11. The descriptive texts at these parts hint towards a perceived grievance (line 1: 'who else be feed up cheatersneverwin') or fringe beliefs that the election was stolen (line 5: 'democratscheated') and a recognition that drastic measures should occur (lines 5 & 11: 'draintheswamp').

4.8.2 Cluster 1

```
1  'hour what be they up to eye open patriot they start right away i could nt get ',
2  'i hope flynn be right i support the patriot war sadly a lot of these be not ',
3  'elliebofficial elliebofficial thing that make we go hmmmm patriot in dcthing that make we go hmmmm patriot ',
4  'in one place just know i love you patriot we ve have a rough day but take ',
5  'fuck you 100 antifa the democrat street soldier patriot be tell to keep an eye out for ',
6  'taghttpsvideoparlercomhf1shf1sya0409xd_smallmp4 great news to start 2021echo so other patriot can know too7mdk7 7mdk7
7  'from trump train to maga boat parades to patriot planes we the people takin overfight fight fight ',
8  'off it s head proclaim this with i patriot warriors antifa disguised hat back share for patriots ',
9  'heart warming this be the great country ever patriot god bless you well there you have it ',
10 'good to serve the groundswell that be the patriot movement in the unitedhttpswwwneonrevoltcom20210106atthehighestlevel
11 'wall or we have a new rule at patriot shit outfitters all of our patch be proudly ',
12 'like an easter egg hunt but bloodier fellow patriot take heed when the bullet start fly do ',
13 'over lin wood trump and all of our patriot hero s and leader include whistle blower omg ',
14 'loop patch we have a new rule at patriot shit outfitters all of our patch be proudly ',
15 'blm that say they would infiltrate dress like patriot and try to make we look bad david ',
16 'i need your help blah blah blah play patriot like a fiddle down in ga asshatwethepeoplestopthestealtrumpwon2020crazyhorsetr
17 'patriot in dcthing that make we go hmmmm patriot in dc if you see an antifa look ',
18 'we be take it all backx my people patriot pay attentionx patriot as far as the eye ',
19 'qanon say her name ababbitt you be a patriot we will defend your name maga wikileak just ',
20 'wall or we have a new rule at patriot shit outfitters all of our patch be proudly '
```

^Image 2: concordance lines for ‘patriot’

Image 2 outlines the concordances for ‘patriot’. ‘Patriot’ is a commonly used noun by far-right groups to call themselves as and is a major part of their collective identity. Indeed, a lot of the messages express solidarity towards the group membership (lines: 2, 4, 7, 8, 9, 12, 14, 19). In other cases, messages are accusing out-groups for disguising as patriots and committing violence to sabotage the reputation of their social group contributing to the disinformation narrative (lines: 8, 15). These results show clear evidence that authors of this community, using ‘patriot’ in their rhetoric, share common beliefs and a self-identity that is supportive of the far-right movement.

1	'guess as it be obvious to all but trump right again okay let s count the damage ',
2	'1 bus load of antifa thug infiltrate peaceful trump demonstrator as part of a false trump flag ',
3	'be cause the violence it be not like trump supporter to be violent at all however i ',
4	'do nt believe you dc lawyer secretly advise trump quit after leak georgia callstopthestealtop dc lawyer who ',
5	'biden ten of thousand of vote cast for trump be adjudicate and switch tohttpwwwelectionfraud2020comsharephpidvideo_s
6	'peaceful trump demonstrator as part of a false trump flag ophttpstwittercompaulsperry_status1346940301307310082 disclo
7	'to state that want to change their elector trump will win the president speak to the american ',
8	'car rally in toronto to show support for trump ita fight between good and evil right now ',
9	'a board member of leonardohttpsrumblecomvch2gwexplosiveobamaandrenziformerpmofitalyorchstratedthetheftofushtmlr
10	'bad actor be really antifa everyone that be trump patriot be peaceful protester any rino that do ',
11	'godblesstheusax we think it be bad when president trump be cheat out of the presidency we think ',
12	'to respect the police which i suspect the trump supporter have support the police i suspect antifa ',
13	'be cause the violence it be not like trump supporter to be violent at all however i ',
14	'frame smhand the msm create the narrative that trump supporter be the violent extremist when several have ',
15	'be all connect former italian pm link to trump spying and obama be also a cohort of ',
16	'on time but democrat one need all nightx trump addressyour browser do not support the video taghttpsvideoparlercomjkp1
17	'so happytrump deploy national guard for capitol protestpresident trump have direct the national guard to be deploy ',
18	'be do please listen to president donald j trump he be ask the same thing that we ',
19	'be an orchestrated attempt to shut down the trump election fraud protest at the capitol today we ',
20	'prayer and pray psalms 91 over lin wood trump and all of our patriot hero s and '

^Image 3: concordance lines for 'trump'

Image 3 depicts the concordances for 'trump'. Some of the messages express their support for Trump and his ill-founded campaign to contest the presidential election (lines: 4, 8, 18, 19, 20). Some of these messages even frame Trump as their savior to lead them in the 'fight against evil' which could be interpreted as an us versus them distinction (e.g. democratic values, the institution, capitol police etc.) of this radical group identity thereby legitimizing a need for violence. In other instances, 'trump' is a collocate to refer to the demonstrators or 'patriots' where the speaker appears to direct blame on antifa (lines: 2,10) for causing the violence, whereas trump demonstrators are viewed as 'peaceful' (line 6).

1	'gon big up the action of those motherfucker fuck that man yall get short ass memory patriot8404 ',
2	'100 fraud and total bs electionfraud these two fuck be now do i agree mcconnell hateful move ',
3	'treason trumpwon wwg1wga fucking mcconnel give his final fuck you to trump stopthesteal leadermcconnell patriot holy fucking ',
4	'and your guilty ass have condone it so fuck you and when you get whatcome there will ',
5	'a serious discussion about this atrocity what the fuck be this a snl skit also roger stone ',
6	'out same hat same guy 100 antifa deepstatefalseflagx fuck blmfuck antifa contecomecleanitalydiditarrestobamastopthestealleonardopatriot a
7	'you fucker be guilty of treason now so fuck you 100 antifa the democrat street soldier patriot ',
8	'fuck joe biden anyone who vote for he fuck that traitor mike judas pence all these politician ',
9	'tell that pos to sit his ass down fuck joe biden anyone who vote for he fuck ',
10	'help we today when they vote name hide fuck you i ve never trust you and i ',
11	'mean keep that orange fucking retard in office fuck off cindy azbarnpros bye bitchazbarnpros he certainly can ',
12	'political suicide she must have be watch fox fuck ga the people of that state do nt ',
13	'on gab anyone can be hear this guy fuck stopthesteal mvpx i know i say i would ',
14	'jackdorseyisapunkassbitch markfuckface killaryclinton thefirstgaypresident aka the one who fuck bigmikeobama all in one picture traitorsshou
15	'society let this happen the same way again fuck the medium judge dc police doj fbi and ',
16	'by block the 2000 stimulus the gop be fuck it up enough without help from anyone else ',
17	'that unacceptable loeffler and a purdue can both fuck off for be gutless and ignore the people ',
18	'i be much far along the learning curve fuck the 6 pm curfew teamtrump sidneypowellstopthestealbideniscorrupttrumpwon2020x last night ',
19	'result i would say i tell you so fuck it i tell you sorogerstone stopthesteal 2020 stopthestealrally ',
20	'dick out of your ass and shut the fuck up azbarnpros you and just you can get '

^Image 4: concordance lines for ‘fuck’

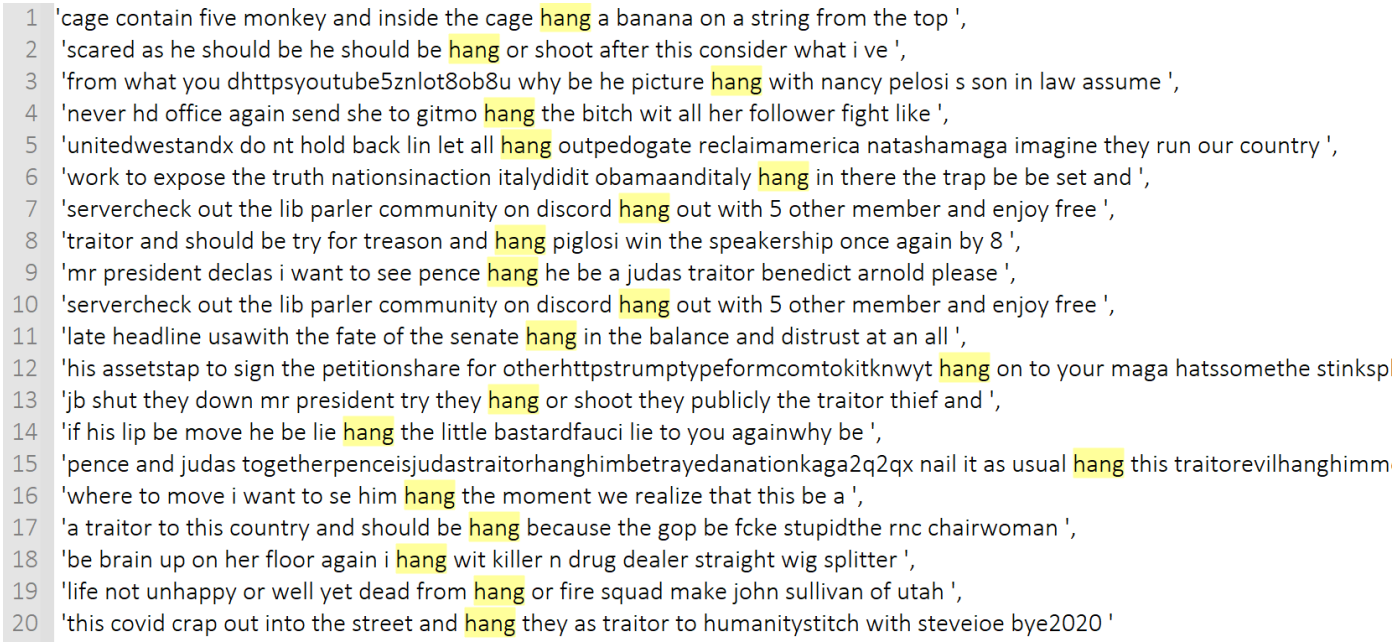
Image 4 presents concordances for ‘fuck’. The term ‘fuck’ is particularly salient when used to express a grievance towards an out-group. This allows the analyst to quickly identify whom the far-right radicals label as their enemies or view as a perceived threat to their social group. Evidence shows that the keyword is being used as profanity where the authors are verbally attacking certain political figures (lines: 3, 8, 9, 11, 17, 20) and other entities (lines 6, 20) like antifa.

4.8.3 Cluster 2

```
1  'the clinton s have have many good americans kill we the people be not have it anymore ',
2  'table today they kill jesus for it they kill that woman today christianity turn the world upside ',
3  'border report the woman who be shoot and kill as demonstrator storm the capitol in washington dc ',
4  'border report the woman who be shoot and kill as demonstrator storm the capitol in washington dc ',
5  'individual liberty you be literally go to be kill should you not listen to good sense if ',
6  'fake medicine it be a sick system that kill people we need to go back to our ',
7  'marxist order correction to my last post woman kill inside the capital building be report by fox ',
8  'ripashleykusi news confirm identity of woman shoot and kill inside us capitolhttpsyoutubemwxq84olsv4 pray for our countryx time ',
9  'lin wood they have the video of they kill kid wowabsolutely shocking text from lin wood they ',
10 'in his right ear special op coup who kill that young innocent trump supporter or be that ',
11 'and looter patriot what should we call theyx kill your neighbor your family your boy in blue ',
12 'assholes the woman that be shot and apparently kill murdered by the scumbag capital cop will now ',
13 're fist bump after an unarmed vet be kill nice blue tie too you traitorx follow i ',
14 'fight cus of his family until the redcoat kill family member just wait because it will get ',
15 'conservative censoring social mediause the obama create internet kill switch selectivelyhow trump can neutralize the conservative censoring ',
16 'so i can wear un buttoned bowling shirt kill wild pig and party at the ebach all ',
17 'be in massive support over fucking us then kill us because theu got caught drink that nectar ',
18 'heavy hand with trump supporter tohttpswwwthegatewaypunditcom202101protrumpmarinegoesoffdcpolicemacestopstealprotestersfingbackair
19 'of hatethank you anthony brian loganu200dx 4threich now kill each other dumb americans make it easy to ',
20 'undermine president now it show traitorrootdescriptionhttpswwwnewsmaxcomtnewsmaxarticle1004406ns_mail_uid016aea09ab854116a0a58b
```

^Image 5: Cluster 2 - ‘kill’

Image 5 shows concordance for ‘kill’. Some of the messages here use the keyword to discuss an event where a woman was shot by Capitol police while participating in the riots (lines 3, 4¸10, 12, 17). In other cases, the keyword is framed as an action committed towards the in-group. For example, line 1 claims that the Clintons have many good Americans ‘killed’ and that the radical individual views the current circumstances as unacceptable – a sign of perceived grievance.



^Image 6: Cluster 2 - 'hang'

Image 6 portrays the concordance for 'hang'. The keyterm has many definitions (Merriam-Webster[19]) but can take on violent connotations depending on its use in context. Evidence shows that indeed, 'hang' is being repeatedly used an intransitive verb to refer to the act of dying by suspension of the neck. The violent sense of this word is being directed towards political figures (lines: 2, 4, 8, 9, 10, 15, 19) and other entities (lines: 2, 13, 17). This illustrates that many of the users are calling for violence echoing negative sentiments similar to the attacking mob.

[19] Hang | Definition of Hang by Merriam-Webster

4.8.4 Cluster 3

1 'support the video taghttpsvideoparlercomraf3raf3pqes02rlmp4 an unarmed woman be shoot by law enforcement in cold blood and no ',
2 'and woman shoot for jog in their neighborhood shoot for fit a description i m piss the ',
3 'capitol building have die the unarmed woman be shoot through a window on a closed door by ',
4 '16 year old trump supporter deserve to be shoot videoif you enjoy my post please like share ',
5 'support the video taghttpsvideoparlercomuskauskaee1ajruzmp4updatethe female trumpsupporter that be shoot by capitolpolice have passwarne gr
6 'for their equality to white man and woman shoot for jog in their neighborhood shoot for fit ',
7 'the unarmed 16 year old girl who be shoot by capitol police during wednesday s protest may ',
8 'constitution wethepeople draintheswamp maga stopthesteal wait before you shoot i let i get a pic hey look ',
9 'old trump supporter may have deserved to be shoot dead by capitol police videoin a shocking statement ',
10 'dc and be probably the person responsible for shoot and murder ashlibabbitt maga rally murder murderer jakeangeli ',
11 'the unarmed 16 year old girl who be shoot by capitol police during wednesday s protest may ',
12 'have die executethe pro trump protestor who be shoot by police inside the capitol building have die ',
13 'parler contenttommyrobinson tommyrobinson this angle show the woman shoot inside thethis angle show the woman shoot inside ',
14 'neverconcede deplorable break videos reportedly show dc police shoot unarmed peaceful female trump supporter in neck national ',
15 'have an amazing 4 year black citizen get shoot while sleep shoot in the back knee to ',
16 'dod usairforcewho be ashli babbit pro trump protester shoot at the capitol buildingher husband describe ashli babbit ',
17 '4 year black citizen get shoot while sleep shoot in the back knee to their throat shoot ',
18 'supporter paul sperry behttpswwwocconservativecompostformerfbiagentconfirmsantifathugsinstigatedfalseflagcausinghavocatuscapitoltodaybreak
19 'they drew first blood innocent unarmed patriotic woman shoot in the neck inside capital building by police ',
20 'have die executethe pro trump protestor who be shoot by police inside the capitol building have die '

^Image 7: Cluster 3 - 'shoot'

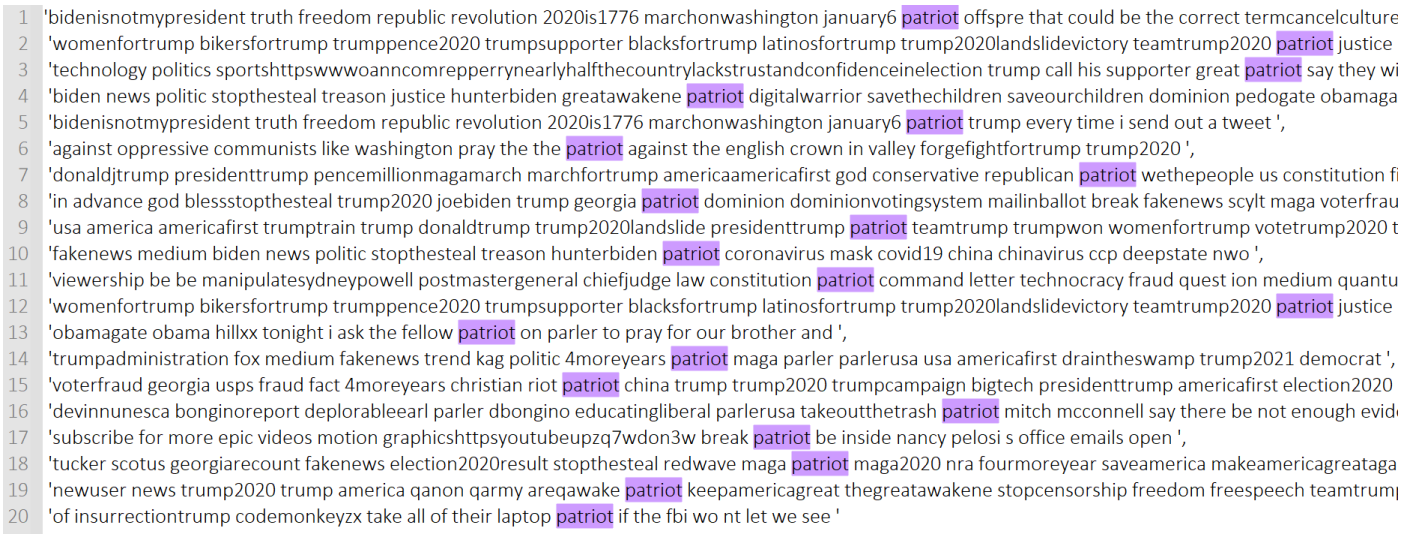
Image 7 depicts the concordance for 'shoot'. Similar to the concordance of 'kill' within cluster 2, a lot of the messages are reporting on the murder of a woman who participated in the Capitol riots (lines: 1, 3, 5, 7, 10, 13, etc.) while voicing their solidarity with the deceased.

1 bidencrimefamilynomandatoryvaccination billgate vaccinedamage maga2020 bigpharma newyorklockdown californialockdownnothingcanstopwhatiscome digitals
2 'freedom justice stopthesteal cancelthecoup news infowarsgop rep wears trump won mask to swearing in ceremonymask signal support ',
3 'linwood americafirst obamagate teamtrumpgreatestpresidentever molonlabeweloveyou pedogate maga2020 kag2020 trump poyb usa wethepeople wwg1wga a
4 'and criminal scum revolutionstopthesteal fightback saveamerica americafirst maga trump veteransfortrump election electionfraud voterfraud fraud briankemp tr
5 'isaackappy lizardsquad ccp mossad fbi m16 andhttpswwwyoutubecomwatchv66w6mvvzqfkfeatureyoutube president trump files two lawsuit against dirty georgia
6 'least one bus load of antifa thugs infiltrated trump demonstrationvoterfraud stopthesteal electionfraud michigan fraudtestimony trump q trump2020 ',
7 'mask covid19 china chinavirus ccp deepstate nwo meme trump trump2020 2020 2020election maga savethechildren pizzagate pedogate obamagate ',
8 'the truthplease follow for more content and updatefamily trump 2amendment god freedomusspaceforce deepstate usconstitution trump2020 trump draintheswa
9 'help engineer the entire coup attempt against president trump in an attempt to oust trump and promote ',
10 'for war i suggest you do the same trump trump2020 2020 2020election boom maga republicans democrats dominion ',
11 'stopthesteal voterfraud electionfraud uspolitic maga trump2020just in donald trump tweets out big news in pennsylvania everyone needs ',
12 'stopthefraud protest rally trumprally fight fightback 1776 istandwithtrump trump trump2020 digitalsoldiers infowars deepstate swamp draintheswamp politic sav
13 'empirestate chrysler army usafirst military freedom freedomofspeech bestpresident trump trumpforever goverment covid badtime fraudelection fraudvote stays
14 'mask covid19 china chinavirus ccp deepstate nwo meme trump trump2020 2020 2020election maga savethechildren pizzagate pedogate obamagate ',
15 'wsjentertainmentnews stopthesteal objectjanuary6army usafirst military freedom freedomofspeech bestpresident trump england picadilly dominion gb wakeupa
16 'capitol police possibly fatal warning graphic videosa female trump supporter have be shoot by capitol police and ',
17 'treasonisdeath execution executionscontinue wethepeople maga kraken q takebackourcountry trump djt holdtheline nevergiveup deepstate corruption drainthe
18 'trump2020 2a 2ndamendment maga2020 veteran racism fakeracism dominion trump dominionsoftware xa0 deepstatecorruption corruptdemocrat soro corrupt
19 'truth freedom republic revolution 2020is1776 marchonwashington january6 patriot trump over 2000 felon vote in ga electionfraud fraud ',
20 'parlerusa takeoutthetrash patriot mainefortrump savethechildrentrump2020 redliner wwg1wga qanon trump teamtrump marklevinshow dbongino devinnunesc

^Image 8: Cluster 3 - 'trump'

Image 8 displays the concordance for 'trump'. At a glance, the lines presented in the window appear to be quite gibberish as nearly all of the messages here take on an unconventional grammatical structure not typically seen in proper written language. The words that make up the lines have no discernible semantics when reading them normally. Instead, the messages appear to be a continuous stream of hashtag-esque words that refer strongly to far-right themes such as election fraud (line 2: 'stopthesteal', line 4: 'voterfraud'), support for Trump (line 1: 'maga2020',

line 12: ‘istandwithtrump’), conspiracy theories (line 7: ‘obamagate’, line 20: ‘qanon’), and the capitol riots (line 4: ‘fightback’, line 19: ‘marchonwashington’). The results observed here are quite intriguing as hashtags were already removed during the preprocessing steps before generating concordance lines. Posts of this nature allude to the notion that they may be echo posts or propaganda that serve to escalate the far-right rhetoric on Parler by widespread dissemination while leveraging organic traffic manipulation for their objectives.



^Image 9: Cluster 3 - ‘patriot’

Image 9 provides the concordance for ‘patriot’. The results here show very similar language patterns to the messages seen previously in cluster 3, the messages are unconventionally structured and is more or less a continuous flurry of topics that are directly associated with the far-right narrative.

4.8.5 Twitter

The two keywords that have been selected for closer examination in the Twitter corpus are ‘fight’ and ‘protest’. Both of these terms had very high LL scores when compared to corpus 2 with 88.73 and 65.79 respectively. As the scores were unusually high in relation to far-right specific vocabulary, they qualified for further analysis in concordance.

1 'a train wreck right now but we still fight onn capitolriots merica happynewyear2021 httpstcoq7utoejyt0 fbi amp police ',
2 'have the courage to storm the capital to fight against the electionfraud that cheat bernie in 2016 ',
3 'number of trump support dickhead say they be fight for democracy any irony completely lose on these ',
4 'n stopthesteal repleezeldin kanekoathegreat thank you sir keep fight for truth and justice we demand the fair ',
5 'cawthornfornc open with wow this crowd have some fight in it and i be thankful for each ',
6 'the left s reaction stopthesteal httpstcoslk436veqp heckyessica jackposobiec fight after can be later stopthesteal nicholecordovas nmsair
7 'be great evidence when you say people must fight markmeadow btw you all be gross do trump ',
8 'senategop fakenews witchhunt spygate obamagate stopthesteal nn httpstcozeebdlstja fight for democracy in america trump2020 trumpı
9 'in presidential history itonly the beginning of our fight to make america great again n ok realdonaldtrump ',
10 'in front we and always choose the wrong fight this be the gleeful response to the capitolriots ',
11 'uscapitol insurrection dontrumpjr kimberlyguilfoyle markmeadow dontrumpsr trumpcrimefamily gloria fight stopthesteal q qanon horse
12 'america do nt let they silence your voice fight for america stopthesteal httpstcoiktjm37kcy capitol siege stop the ',
13 'i will be donate to youkeep up the fight stopthesteal dominionvotingsystem i be move soon to parler ',
14 'on opposite side n in 2021 these flags fight on the same side nn republicanlivesmatter nn there ',
15 'the last word not man represent us amp fight for justice amp protectrepvernonjone democrats social media and ',
16 'stand with you mr president nn i will fight for you n i will win for you ',
17 'bring it if congress and scotus wo nt fight to protect our constitution then we the people ',
18 'what he call an army for trump to fight for president trump still not sure why people ',
19 'bring the mountains of evidence and help we fight backstopthesteal enough be enough random insomniac think nn ',
20 'i i be now a conservative independent the fight be on and i be ready draintheswamp stopthesteal '
21

^Image 10: Twitter - ‘fight’

Image 10 outlines the concordance for ‘fight’ on Twitter. Nearly all of the messages in this concordance sample illustrate a strong affiliation towards Trump and right-wing extremist views. Lines like ‘have the courage to storm the capitol to fight’ (line 2) and ‘stand with you mr president.. I will fight for you’ (line 16) convey a positive in-group response to Trump’s endorsement of violent behavior. This provides evidence that there is still a presence of far-right communities on Twitter although they appear to be part of the minority in this platform.

1 'nn stopthedeepstate ptmarigan dgpurser there be a stopthesteal protest in carson city nv also i know because ',
2 'the need to explore theconnection between capitolriots and protest on several state capital that be stop because ',
3 'scotus you be squarely at fault for the protest in wa dc have spurn potus s attempt ',
4 'fact stopthesteal the difference between yesterday and the protest in the 60s nn trump supporter be lose ',
5 'trip to washington dc for today s stopthesteal protest some of those protest turn into a violent ',
6 'rip who be kill for attempt to peacefully protest us election fraudstopthesteal httpstcomy870n2ilp get ready for another ',
7 'on a joint session of congress w known protest outsideour ic do nt think there be more ',
8 'session to certify the electoral vote after the protest read pre written speeches hmmmalmost like they know ',
9 'local store do when blm activist be peacefully protest because these maga terrorists be a real threat ',
10 'recognize the truth at all nn capitolriot capitolriotschaotic protest at the capitol in washington dc httpstcomgxfsg6jxe via ',
11 'bidencheated2020 they be try to turn the stopthesteal protest into a terrorist attack yet blm amp antifa ',
12 'httpstcow62rzngz8c 250 year of frustration cause the blacklivesmatter protest fck you john catanzara you need to be ',
13 'capitol they be be refer to as peaceful protest capitolriots bias twitter joe biden call i a ',
14 'by capitol hill police inside the capitol during protest innocent unarmed woman shoot in the neck and ',
15 'of federal democracy constitution and freedom of peaceful protest nmsm trump whiteprivilege msmistheenemyofthepeople capitolbuilding
16 'when demonstrator storm the united states capitol to protest electionfraud fightback stopthesteal maga httpstcoq35sulcz6e david bowie be
17 'be make n electionfraud methink the lady doth protest too much n electionfraud httpstcod51jld8xax lindseygrahamsc your career ',
18 'lesson from blm black people know how to protest amp if provoke riot they have have a ',
19 'elect they be not the same racistamerica protest protest blacklivesstillmatter cruz s effort to see the objection ',
20 'send antifa and blm to infiltrate our peaceful protest nn we need to fight back the time '

^Image 11: Twitter - ‘protest’

Lastly, image 11 presents the concordances for ‘protest’. The examples given show mixed results in how the keyword is used in language by certain authors. In some cases, language is used to merely discuss the protest that is transpiring in Washington D.C. from different sides (lines: 2, 3, 8, 16 etc.). While other instances lean more towards a far-right agenda in disseminating misinformation such as lines 11 and 20 where the authors frame the attack of the Capitol on Antifa and supporters of BLM.

5 Discussion and Limitations

5.1 Discussion

In light of the unprecedented 2021 Capitol riots and the 2020 U.S. presidential election, there is no doubt that social media technologies like Parler facilitated online radicalization of individuals towards violence. Parler’s role as a hyper-conservative platform and the composition of its far-right userbase gives it strong connections to the storming of the Capitol (Munn 2021). As a result, the content on Parler, its lax positioning on content regulation, and heavily one-sided political leanings rendered it an attractive environment for empirically studying normative behaviors in a social network that is predominantly right-wing radical. The purpose of this thesis set out to answer the research question: *“How does the language of corpora from radicalized communities discovered on Parler compare to online conservations on Twitter regarding the Capitol riots and election fraud?”.* The problems posed within this question can be viewed and answered as a two-fold structure: the first concerns implementing a computational approach for community detection, and the other subsequently drawing on computational linguistic tools to infer knowledge about online extremism.

In terms of community detection within a social network, the implementation of HDBSCAN clustering of Parler users into sub-communities based on the similarity of their hashtag patterns made it possible to identify virtual social groups that subscribed to a wide assortment of far-right views. Drawing on the theoretical framework of radicalism by Kruglanski et al. (2014) and literature insights on the hashtags and terms used by political far-right groups (Torregrosa et al. 2020, Hitkul et al 2021), clear descriptive measures of radicalism within a sub-community including the discrepancies across sub-communities could be established. The sub-communities detected on Parler exhibited telltale signs of a high degree of radicalism towards far-right beliefs as indicated by a collective and motivationally imbalanced use of far-right hashtags. In light of Zhang et al. (2012), hashtags can serve as the symbol of a community by linking users with similar ideas. The results show that the most popular topics of conversations among Parler communities were centered around Trumpism, conspiracy theories, far-right extremist groups, voter fraud and conservative in-group favoritism.

Subjecting the texts of radicalized communities to corpus linguistic tools revealed key insights into the radical rhetoric of the groups. As expected, the general findings of the content on Parler and Twitter in this thesis corroborate with Hitkul et al.’s (2021) work - Parler exhibited an extreme right-wing narrative echoing the propaganda of disinformation, calls for violence, and conspiracy theories comparable to the attacking mob. By contrast, content on Twitter showed

strong opposition and disdain towards the actions of the rioters and Trump, calling for his removal on both social media and from office.

In terms of findings gathered from n-gram frequencies, the prominent unigrams among sub-communities were using common political types of terms to construct their discourse which display strong ties to far-right views when discussing the Capitol riots and election fraud. Moreover, from bi-gram and tri-gram feature extractions, it was discovered that radicalized communities used persuasive language (see sections 4.6.2 and 4.6.3) in promoting and endorsing far-right beliefs to others among their community via manipulation of social media engagement features. Interestingly, this behavior was not immediately present in the discourse on Twitter. Moreover, radicalized communities showed clear in-group favoritism towards Trump ideology as evidenced by the prominence of 'president trump' and 'patriot' among the sub-communities whereas this is not present on Twitter.

From the results gathered by domain-specific vocabulary significance tests, it was possible to derive which far-right terms were more salient in Parler sub-communities when compared to a more neutral corpus like Twitter. Doing so provided key insights of the important linguistic markers far-right actors would use within a radicalized community. It was found that salient terms drawing on emergent themes of anger and violence were typically common among the radicalized communities. Although significance tests of more terms along these psychological constructs should be explored to establish firmer conclusions. On the other hand, the set of terms that were more significant on Twitter when compared to Parler reflect a more benign nature when considered together. Evidence suggests that the terms served merely as descriptors for the violent attack.

The salient far-right terms were then submitted to the analysis of concordances. By interpreting the concordance searches, certain sub-communities displayed clear evidence of the elements that make up a radical group as indicated by their collective discourse. These findings corroborate strongly to previous literature on the extremist discourse (Sakki & Pettersson 2016; Fortuna & Nunez 2018; Furlow & Goodall Jr. 2011). These were expressions of political grievances, hate speech and discursive speech touching on war terminology and 'otherness'. The generated samples of concordances supports the revealed pattern of polarized thinking and opposition that according to Meloy (2012), is considered to be a type of warning behavior that indicates radicalization. Moreover, concordance analysis identified the largest radicalized community to be expressing extremist discourse in a novel way not previously reported by existing literature. Radical users of this sub-community (see sections 4.8.4) were observed to be aggressively promoting radical ideas in an attempt to reify them as mainstream views. This normative behavior could be attributed to the 'group extremity shift' or 'group polarization' mechanism of political radicalization proposed by McCauley & Moskalenko (2008) which states that groups of strangers brought together to discuss issues of political opinion show consistently two kinds of change: increased agreement about the opinion at issue, and a shift in the average opinion of group members. The shift is toward increased extremity on whichever side of the opinion is favored by most individuals before discussion (Brown 1985). The intent may be to propagandize the far-right beliefs and fringe opinions on mainstream channels so that they may be gradually normalized among the population, a tactic not unheard of by far-right actors on social media as evidenced by Winter (2019).

5.2 Limitations

Although the computational approaches and linguistic analyses taken in this thesis strongly elucidates a range of behaviors distinctive of far-right communities, they are not without their shortcomings. Nevertheless, various rooms for improvement and implications in the approaches and nature of the datasets can be highlighted.

Firstly, Parler data had extremely limited metadata associated with its textual content. For instance, there is no way to check whether a given parlay is an echo parlay meaning it has been taken from another individual and reposted in the same manner. Studying this would be interesting as this information reveals that the implications behind the message has been supported by another individual. Including this feature in the analysis would offer a more comprehensive overview of how language and normative behaviors intersect in propagating extremist views. Another important remark is that this thesis examined the phenomenon of radicalization in a snapshot timeframe. In other words, the temporal factor was not considered in the analysis. From existing literature (Rehman et al. 2020, COPS 2021), it is well known that radicalization is a gradual process and typically necessitates a temporal component to rigorously examine the full scope of how language plays a role in perpetuating and escalating radicalization.

Secondly, the novel Twitter dataset acquired for this thesis research was not balanced in their number of candidate hashtags according to the expectation of API requests made for retrieving historical tweets (see section 3.3.1). This could have affected the quality of the analysis gathered from the linguistic data. In addition to this, there is an important distinction to be made between the Parler and Twitter datasets. The Parler dataset has complete information on all the textual content made by a user, however, the Twitter dataset only returns a fragment of a user's posting history of that day. With the current data acquisition method used, it was not possible to gather complete information of textual content by a user for Twitter. A more robust approach towards acquiring a comprehensive and fair Twitter dataset would be to scrape text data from a random sample of users who posted on the day of the riots as this permits no preconceived notions on what the content might entail and the conclusions thereafter.

Thirdly, from studying social media with NLP, it is well known that authors on microblogging sites are notorious for their inconsistencies in syntax. Parler is no exception to this which can influence the veracity of results obtained if there are too many errors or too much noise present in the data. This can especially have an impact on interpretations made during concordance analysis. In some cases, judgements made from language are subjective and open to interpretation as the real intent of the speaker can never be known without asking them directly. Nonetheless, identifying patterns in the language by finding supporting evidence from multiple distinct sources should alleviate this concern on its own.

A fourth observation is that this thesis did not make use of a ground-truth radical corpus to compare for similarity against the community detected corpora. The conclusions from this research could be strengthened by comparing corpora of discovered communities to a ground-truth radical corpus of known extremist actors belonging to a social network. Doing so will allow researchers to make more accurate judgements on the nature of the corpus and can be used as an indicator for radical discourse.

6 Conclusion and Future work

6.1 Conclusion

The work of this thesis has shown that language is a prevailing force in driving the far-right narrative on social media. By combining the powerful analytical tools of unsupervised clustering and corpus linguistics, the current thesis was able to identify distinct patterns in language that are indicative of online radicalization within the framework of right-wing extremism. Users in radicalized communities deployed a number of strategies in their discourse that is characteristic of perpetuating the far-right narrative thereby exposing more and more people to opportunities for radicalization. This entailed constructing a violent and aggressive discourse with warmongering qualities effectively creating an 'us vs. them' dichotomy that legitimizes their need for violence. In another special instance, it was found that radical users as a collective group would aggressively promote and disseminate a flurry of radical ideas into the mainstream limelight as a means to recommend fringe opinions to the entire community thereby painting those views as favorable; a tactic that is reminiscent of the down-the-rabbit hole effect. The techniques used in thesis are proven to work in detecting radicalized sub-communities within a known extremist social network. The work here can also be viewed as an explorative tool in building comprehensive labelled datasets for extremism detection which help to prevent radicalization from reaching their ultimate stages like terrorism and violent attacks. In addition to, removing radical content more efficiently from mainstream sources.

6.2 Future work

Online radicalization is a complex and burgeoning societal issue that can take on many distinctive qualities which necessitates studying it from a wide array of analytical perspectives. This thesis leveraged only a small fraction of the analytical tools that can be performed on linguistic data. Recent advances in NLP opens up a large avenue of computational techniques applicable to the study of online extremism. Future work would do well to incorporate such techniques towards better understanding of normative behaviors associated with radicalization and extremism. Some notable suggestions for meaningful progress and contributions related to this work are outlined below:

- Apply significance tests of far-right extremist vocabulary between key actors of a known extremist group and regular users. This would elucidate which terms are more salient to a radicalized individual which could be a linguistic marker for radicalization.
- Compare the within-differences of language use in radicalized users across the two social media platforms.
- Conduct a comparative sentiment analysis between discovered radicalized communities and Twitter following a key event to see if there any interesting differences.

- Analyze the temporal component of certain keywords over time such as before the Capitol riots and afterwards for gathering key insights in changes to propaganda.

- Explore motivational theories of radicalization with LIWC (Linguistic Inquiry and Word Count) software on detected communities to automate the process of extracting

psychological constructs from textual content. This approach can be powerful in describing normative behaviors rooted in social science theories of radicalization.

- Make use of POS (parts-of-speech) tagging as to critically analyze pronoun use by radical users vs. non-radical users among other parts-of-speech features.
- In this thesis, user clustering was performed on one textual feature (i.e. hashtag similarity), however other textual features such as URL similarity, text similarity, and retweeting similarity can be exploited and combined for more nuanced and fine-grained community discovery results as previously done by Zhang et al.(2012) and Kelly (2017).

7 Acknowledgements

This project was done in collaboration with the Dutch national police force. I would like to sincerely thank Paul Verhaar and Max Kemman from Utrecht University and Mijke van den Hurk from The National Police Lab AI for their insights, feedback, and support in this work.

I would also like to show a tremendous and heartfelt appreciation towards my parents for their never-ending and endearing support. You guys are always pushing me harder so that I may one day create meaningful success in my life. I hope that this work also contributes as a step towards that direction. Thank you and much love.

8 References

Abd-Elaal, A. I. A., Badr, A. Z., & Mahdi, H. M. K. (2020). Detecting Violent Radical Accounts on Twitter. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 8). www.ijacsa.thesai.org

ADL. (2020). *Parler: Where the Mainstream Mingles with the Extreme | Anti-Defamation League*. https://www.adl.org/blog/parler-where-the-mainstream-mingles-with-the-extreme

Aldera, S., Emam, A., Al-qurishi, M., Alrubaian, M., & Alothaim, A. (2021). *Online Extremism Detection in Textual Content:A Systematic Literature Review*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9371676

Aliapoulios, M., Bevensee, E., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., & Zannettou, S. (2021). *An Early Look at the Parler Online Social Network*. https://doi.org/10.5281/zen

Alkarkhi, A. F. M., & Alqaraghuli, W. A. A. (2020). *Cluster Analysis - an overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/cluster-analysis

Ashour, & Omar. (2011). Online De-Radicalization? Countering Violent Extremist Narratives: Message, Messenger and Media Strategy. *Perspectives on Terrorism*, *4*(6). http://www.terrorismanalysts.com/pt/index.php/pot/article/view/128/html

Backlink. (n.d.). *How Many People Use Social Media in 2021? (65+ Statistics)*. 2021. Retrieved May 14, 2021, from https://backlinko.com/social-media-users

BBC News. (2021). *QAnon: What is it and where did it come from? - BBC News*. https://www.bbc.com/news/53498434

Ben-David, A. (2016). *Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain | Ben-David | International Journal of Communication*. https://ijoc.org/index.php/ijoc/article/view/3697

Benigni, M. C., Joseph, K., & Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLoS ONE*, *12*(12), e0181405. https://doi.org/10.1371/journal.pone.0181405

Benigni, M. C., Joseph, K., & Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLoS ONE*, *12*(12), e0181405. https://doi.org/10.1371/journal.pone.0181405

Bennett, G. R. (2010). *An IntroductIon to corpus LInguIstIcs Part 1 Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. http://www.press.umich.edu/titleDetailDesc.do?id=371534

Berger, J. M., Strathearn, B., & Al Momani, H. (2013). *Who Matters Online: Measuring influence, evaluating content and countering violent extremism in online social networks*. www.icsr.

Berger, J. M. (2018). *Extremism* . https://mitpress.mit.edu/books/extremism

Bermingham, A., Conway, M., Mclnerney, L., O'Hare, N., & Smeaton, A. F. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009*, 231–236. https://doi.org/10.1109/ASONAM.2009.31

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use - Douglas Biber, Biber Douglas, Professor Douglas Biber, Susan Conrad, Randi Reppen - Google Boeken*. Cambridge University Press. https://books.google.nl/books/about/Corpus_Linguistics.html?id=2h5F7TXa6psC&redir_esc=y

Binder, M. (2020, September 11). *What is Parler? Everything you need to know about the conservative social network.* https://www.msn.com/en-us/news/technology/what-is-parler-everything-you-need-to-know-about-the-conservative-social-network/ar-BB1aQTqG

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. In *Physics Reports* (Vol. 424, Issues 4–5, pp. 175–308). North-Holland. https://doi.org/10.1016/j.physrep.2005.10.009

Broder, A. Z. (1997). Syntactic clustering of the Web. *Computer Networks*, *29*(8–13), 1157–1166. https://doi.org/10.1016/s0169-7552(97)00031-7

Brown, R. (1981). *Group polarization*. Social Psychology: The Second Edition.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7819 LNAI(PART 2), 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7819 LNAI*(PART 2), 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Castro, A. (2020). *Parler, a conservative Twitter clone, has seen nearly 1 million downloads since Election Day - The Verge*. https://www.theverge.com/2020/11/9/21557219/parler-conservative-app-download-new-users-moderation-bias

Clary, G. (2021). *Parler Wasn't Hacked, and Scraping Is Not a Crime - Lawfare*. https://www.lawfareblog.com/parler-wasnt-hacked-and-scraping-not-crime

COPS. (2014). *Defining Online Radicalization*.

Correa, D., & Sureka, A. (2013). *Solutions to Detect and Analyze Online Radicalization : A Survey*. http://www.isiconference.org/

Crys, E. (2021). *FSI - Parler's First 13 Million Users*. https://fsi.stanford.edu/news/sio-parler-contours

Curran Benigni, M. (2016). *Thesis Proposal: Online Extremist Community Detection, Analysis, and Intervention*.

Data Visualisation Catalogue. (2021). *Density Plot - Learn about this chart and tools to create it*. https://datavizcatalogue.com/methods/density_plot.html

Dharra, S. (n.d.). *Intro To ML*. Retrieved June 22, 2021, from https://rsdharra.com/blog/lesson/32.html

Doosje, B., Loseman, A., & van den Bos, K. (2013). Determinants of Radicalization of Islamic Youth in the Netherlands: Personal Uncertainty, Perceived Injustice, and Perceived Group Threat. *Journal of Social Issues, 69*(3), 586–604. https://doi.org/10.1111/josi.12030

Doosje, B., Moghaddam, F. M., Kruglanski, A. W., De Wolf, A., Mann, L., & Feddes, A. R. (2016). Terrorism, radicalization and de-radicalization. *Current Opinion in Psychology*, *11*, 79–84. https://doi.org/10.1016/j.copsyc.2016.06.008

Doosje, B., van den Bos, K., Loseman, A., Feddes, A. R., & Mann, L. (2012). "My in-group is superior!": Susceptibility for radical right-wing attitudes and behaviors in dutch youth.

Negotiation and Conflict Management Research, *5*(3), 253–268. https://doi.org/10.1111/j.1750-4716.2012.00099.x

Drew, C. (n.d.). *What is Domain-Specific Vocabulary? - 121 Examples (2021)*. Retrieved June 25, 2021, from https://helpfulprofessor.com/domain-specific-vocabulary/

El-Said, H., & Barrett, R. (2011). Radicalisation and Extremism that Lead to Terrorism. In *Globalisation, Democratisation and Radicalisation in the Arab World* (pp. 199–235). Palgrave Macmillan UK. https://doi.org/10.1057/9780230307001_11

Farivar Masood. (2020). *Researchers: More Than a Dozen Extremist Groups Took Part in Capitol Riots | Voice of America - English*. VOA. https://www.voanews.com/2020-usa-votes/researchers-more-dozen-extremist-groups-took-part-capitol-riots

Farivar, M. (2021). *Researchers: More Than a Dozen Extremist Groups Took Part in Capitol Riots | Voice of America - English*. https://www.voanews.com/2020-usa-votes/researchers-more-dozen-extremist-groups-took-part-capitol-riots

Floridi, L. (2021). Trump, Parler, and Regulating the Infosphere as Our Commons. In *Philosophy and Technology* (Vol. 34, Issue 1, pp. 1–5). Springer Science and Business Media B.V. https://doi.org/10.1007/s13347-021-00446-7

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. In *ACM Computing Surveys* (Vol. 51, Issue 4). Association for Computing Machinery. https://doi.org/10.1145/3232676

Fortunato, S. (n.d.). *Community detection in graphs*. Retrieved June 30, 2021, from www.amazon.com

Fothergill, R., Cook, P., & Baldwin, T. (n.d.). *Evaluating a Topic Modelling Approach to Measuring Corpus Similarity*.

Frenkel, S. (2021). *How The Storming of Capitol Hill Was Organized on Social Media - The New York Times*. https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html

Frost, J. (2021). *Degrees of Freedom in Statistics - Statistics By Jim*. https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/

Fuchs, C. (2016). *Racism, nationalism and right-wing extremism online: The Austrian Presidential Election 2016 on Facebook* (Vol. 5, Issue 3). www.momentum-quarterly.org

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021). *Online Extremism Detection: A SystematicLiterature Review With Emphasis onDatasets, Classification Techniques,Validation Methods, and Tools*. IEEE Acess. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9383220

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Papadopoulos, S., Vrochidis, S., & Kompatsiaris, I. (2017). Detection of terrorism-related twitter communities using centrality scores. *MFSec 2017 - Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, Co-Located with ICMR 2017*, 21–25. https://doi.org/10.1145/3078897.3080534

Goyvaerts, J. (2019). *Regular Expression Tutorial - Learn How to Use Regular Expressions*. https://www.regular-expressions.info/tutorial.html

Graff, G. (2021). *Behind the Strategic Failure of the Capitol Police - POLITICO*. https://www.politico.com/news/magazine/2021/01/08/capitol-police-failure-456237

Graham, R. (2016). Inter-ideological mingling: White extremist ideology entering the mainstream on Twitter. *Sociological Spectrum*, *36*(1), 24–36. https://doi.org/10.1080/02732173.2015.1075927

Hapal, D. K. (2015). *Twitter-terror: How ISIS is using hashtags for propaganda*. https://www.rappler.com/technology/social-media/twitter-isis-hashtag-propaganda

Hartung, M., Klinger, R., Schmidtke, F., & Vogel, L. (n.d.). *Identifying Right-Wing Extremism in German Twitter Profiles: a Classification Approach*. https://doi.org/10.1007/978-3-319-59569-6_40

Hartung, M., Klinger, R., Schmidtke, F., & Vogel, L. (n.d.). *Identifying Right-Wing Extremism in German Twitter Profiles: a Classification Approach*. https://doi.org/10.1007/978-3-319-59569-6_40

Healy, J. (2019). *HDBSCAN, Fast Density Based Clustering, the How and the Why - John Healy - YouTube*. YouTube. https://www.youtube.com/watch?v=dGsxd67IFiU&t=1624s

https://developer.twitter.com/en/developer-terms/agreement-and-policy/source. (2020, April). *Developer Agreement*.

ICO. (2025). *What is a DPIA? | ICO*. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/what-is-a-dpia/

Ines von Behr, Anaïs Reding, Charlie Edwards, & Luke Gribbon. (2021). Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism. In *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*. https://doi.org/10.7249/rr453

Jupskås R, A., & Segers B, I. (2020). *What is right-wing extremism? - C-REX - Center for Research on Extremism*. https://www.sv.uio.no/c-rex/english/groups/compendium/what-is-right-wing-extremism.html

Karell, D., Linke, A., & Holland, E. C. (n.d.). *Right-Wing Social Media Use Increases Political Unrest Across the United States*. https://doi.org/10.31235/OSF.IO/PNA5U

Karrel, D., Linke, A., & Holland, E. (2021). *Right-Wing Social Media Use IncreasesPoliticalUnrestAcross the United States*.

Kelly, P. (2017). Community Detection for Counter-Terrorism. *Senior Projects Spring 2017*. https://digitalcommons.bard.edu/senproj_s2017/361

Kilgarriff, A., & Rose, T. (1998). *Measures for corpus similarity and homogeneity*. http://www.itri.brighton.ac.uk

Koehler, D. (2014). *The Radical Online: Individual Radicalization Processes and the Role of the Internet | Koehler | Journal for Deradicalization*. https://journals.sfu.ca/jd/index.php/jd/article/view/8

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, *35*(SUPPL.1), 69–93. https://doi.org/10.1111/pops.12163

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 27, Issue 1). http://tempest.wellesley.edu/~ywang5/aaai/paper.html.

Kyto, M., Ludeling, A., & de Gruyter, M. (2007). *Handbook of Corpus Linguistics*. https://www.pala.ac.uk/searching-and-concordancing.html

Lifewire. (2021). *What Is Social Media?* https://www.lifewire.com/what-is-social-media-explaining-the-big-trend-3486616

Litvinova, T., Litvinova, O., Panicheva, P., & Biryukova, E. (2018). Using corpus linguistics tools to analyze a Russian-language islamic extremist forum. *Lecture Notes in Computer*

Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *11193 LNCS*, 54–65. https://doi.org/10.1007/978-3-030-01437-7_5

López Sánchez, D., Revuelta, J., De La Prieta, F., Gil-González, A. B., & Dang, C. (2016). *Twitter User Clustering Based on Their Preferences and the Louvain Algorithm*. https://doi.org/10.1007/978-3-319-40159-1_29

Lopez, C., & Wharton, J. (2021). *What Role Did Social Media Play In The Capitol Riots? | Connecticut Public Radio*. https://www.wnpr.org/post/what-role-did-social-media-play-capitol-riots

Lytvynenko, J., & Hensley-Clancy Molly. (2021). *Rioters Taking Over The Capitol Planned It Online*. https://www.buzzfeednews.com/article/janelytvynenko/trump-rioters-planned-online

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

McEnery, T., & Hardie, A. (n.d.). *Corpus Linguistics: Method, theory and practice*. Retrieved June 23, 2021, from http://corpora.lancs.ac.uk/clmtp/main-1.php

McInnes, L., Healy, J., & Astels, S. (2016). *How HDBSCAN Works — hdbscan 0.8.1 documentation*. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

Medcalc. (n.d.). *Values of the Chi-squared distribution table*. Retrieved June 9, 2021, from https://www.medcalc.org/manual/chi-square-table.php

Merriam-Webster. (n.d.). *Hang | Definition of Hang by Merriam-Webster*. Retrieved July 1, 2021, from https://www.merriam-webster.com/dictionary/hang

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (n.d.). *Distributed Representations of Words and Phrases and their Compositionality*.

Moghaddam, F. M. (2005). The staircase to terrorism a psychological exploration. In *American Psychologist* (Vol. 60, Issue 2, pp. 161–169). Am Psychol. https://doi.org/10.1037/0003-066X.60.2.161

Munn, L. (2021). *View of More than a mob: Parler as preparatory media for the U.S. Capitol storming | First Monday*. https://firstmonday.org/ojs/index.php/fm/article/view/11574/10077

Newhouse, A. (2020). *Right-wing users flock to Parler as social media giants rein in misinformation | PBS NewsHour*. https://www.pbs.org/newshour/nation/right-wing-users-flock-to-parler-as-social-media-giants-rein-in-misinformation

United Nations Office On Drugs and Crime. (2012). *UNITED NATIONS OFFICE ON DRUGS AND CRIME Vienna THE USE OF THE INTERNET FOR TERRORIST PURPOSES*.

Olteanu, A., Castillo, C., Boy, J., & Varshney R., K. (n.d.). *View of The Effect of Extremist Violence on Hateful Speech Online*. Retrieved June 22, 2021, from https://ojs.aaai.org/index.php/ICWSM/article/view/15040/14890

Omer, E. (2015). *Using machine learning to identify jihadist messages on Twitter*. http://www.teknat.uu.se/student

Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira, W., Almeida, V., & Meira, W.-N. (2018). *Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination*. *10*(18). https://doi.org/10.1145/3201064.3201081

PALA. (n.d.). *Searching and Concordancing - PALA*. Retrieved June 18, 2021, from https://www.pala.ac.uk/searching-and-concordancing.html

Panizo-LLedot, A., Torregrosa, J., Bello-Orgaz, G., Thorburn, J., & Camacho, D. (2020). Describing Alt-Right Communities and Their Discourse on Twitter During the 2018 US Mid-term Elections. *Studies in Computational Intelligence*, *882 SCI*, 427–439. https://doi.org/10.1007/978-3-030-36683-4_35

Papadopoulos, S., Vakali, A., & Kompatsiaris, I. (2011). *Community detection in Social Media*. https://www.researchgate.net/publication/233790771_Community_detection_in_Social_Media

Petrovskiy, M., & Chikunov, M. (2019). Online extremism discovering through social network structure analysis. *2019 IEEE 2nd International Conference on Information and Computer Technologies, ICICT 2019*, 243–249. https://doi.org/10.1109/INFOCT.2019.8711254

Pojanapunya, P., & Todd, R. W. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 133–167. https://doi.org/10.1515/cllt-2015-0030

Prentice, S., Rayson, P., & Taylor, P. J. (2012). The language of Islamic extremism. *International Journal of Corpus Linguistics*, *17*(2), 259–286. https://doi.org/10.1075/ijcl.17.2.05pre

Reeves, J., Mascaro, L., & Woodward, C. (2021). *Capitol assault a more sinister attack than first appeared*. https://apnews.com/article/us-capitol-attack-14c73ee280c256ab4ec193ac0f49ad54

Reeves, J., Mascaro, L., & Woodwards, C. (2021). *Capitol assault a more sinister attack than first appeared*. https://apnews.com/article/us-capitol-attack-14c73ee280c256ab4ec193ac0f49ad54

Reicher, S. D., & Haslam, S. A. (2016). Fueling Extremes. *Scientific American Mind*, *27*(3), 34–39. https://doi.org/10.1038/scientificamericanmind0516-34

Reid Meloy, J., Hoffmann, J., Guldimann, A., & James, D. (2012). The Role of Warning Behaviors in Threat Assessment: An Exploration and Suggested Typology. *Behavioral Sciences and the Law*, *30*(3), 256–279. https://doi.org/10.1002/bsl.999

Ríos, S. A., & Muñoz, R. (2012). Dark web portal overlapping community detection based on topic models. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2331791.2331793

Rowe, M., & Saif, H. (n.d.). *Mining Pro-ISIS Radicalisation Signals from Social Media Users Conference or Workshop Item Mining Pro-ISIS Radicalisation Signals from Social Media Users*. Retrieved May 5, 2021, from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13023/12752

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., & Alani, H. (2017). A Semantic Graph-Based Approach for Radicalisation Detection on Social Media. *Lecture Notes in Computer Science*, 571–587. https://doi.org/10.1007/978-3-319-58068-5

Sakki, I., & Pettersson, K. (2016). Discursive constructions of otherness in populist radical right political blogs. *European Journal of Social Psychology*, *46*(2), 156–170. https://doi.org/10.1002/ejsp.2142

Sakki, I., & Pettersson, K. (2016). Discursive constructions of otherness in populist radical right political blogs. *European Journal of Social Psychology*, *46*(2), 156–170. https://doi.org/10.1002/ejsp.2142

Saul, I. (2019). *White Supremacists Love Twitter Alternative Parler News – The Forward*. https://forward.com/news/427705/parler-news-white-supremacist-islamophobia-laura-loomer/

Schackmuth, A. (2018). *Extremism, fake news and hate: effects of social media in the Extremism, fake news and hate: effects of social media in the post-truth era post-truth era*. https://via.library.depaul.edu/etd

Schneider, G., & Lauber, M. (n.d.). *The Chi-Square Test*.

Schuurman, B., & Taylor, M. (2018). *Articles Reconsidering Radicalization: Fanaticism and the Link Between Ideas and Violence*.

Sharif, W., Mumtaz, S., Shafiq, Z., Riaz, O., Ali, T., Husnain, M., & Choi, G. S. (2019). An empirical approach for extreme behavior identification through tweets using machine learning. *Applied Sciences (Switzerland)*, *9*(18). https://doi.org/10.3390/app9183723

Smart Insights. (n.d.). *Global social media statistics research summary [updated 2021]*. 2021. Retrieved May 14, 2021, from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Smith, B. (2031). *polynomial's gists · GitHub*. https://gist.github.com/polynomial

Statista. (n.d.). • *Daily social media usage worldwide | Statista*. Retrieved June 15, 2021, from https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Stepin. (2015). D.S.: Features of the implementation of terrorist agitation and propaganda using Internet resources (on the example of the "KavkazChat" forum). *In: Problems of Theory and Practice of Combating Extremism and Terrorism*.

Stern, J. (2016). *Radicalization to Extremism and Mobilization to Violence: What Have We Learned and What Can We Do about It? on JSTOR*. https://www.jstor.org/stable/26361939

Tang, W. M. (2021). *Corpus Linguistics – A Short Introduction – in other words*. https://wmtang.org/corpus-linguistics/corpus-linguistics/

Teja, S. (2020). *What are Stop Words.How to remove stop words. | Medium*. https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company. https://b-ok.cc/book/2609794/782a83

Torok, R. (2011). The online institution: Psychiatric power as an explanatory model for the normalisation of radicalisation and terrorism. *Proceedings - 2011 European Intelligence and Security Informatics Conference, EISIC 2011*, 78–85. https://doi.org/10.1109/EISIC.2011.43

Torregrosa, J., Bello-Orgaz, G., Martinez-Camara, E., Ser, J. Del, & Camacho, D. (2021). *A SURVEY ON EXTREMISM ANALYSIS USING NATURAL LANGUAGE PROCESSING A PREPRINT*.

Torregrosa, J., Panizo-Lledot, Á., Bello-Orgaz, G., & Camacho, D. (2020). Analyzing the relationship between relevance and extremist discourse in an alt-right network on Twitter. *Social Network Analysis and Mining*, *10*(1). https://doi.org/10.1007/s13278-020-00676-1

Torregrosa, J., Thorburn, J., Lara-Cabrera, R., Camacho, D., & Trujillo, H. M. (2020). Linguistic analysis of pro-ISIS users on Twitter. *Behavioral Sciences of Terrorism and Political Aggression*, *12*(3), 171–185. https://doi.org/10.1080/19434472.2019.1651751

Trip, S., Bora, C. H., Marian, M., Halmajan, A., & Drugas, M. I. (2019). Psychological mechanisms involved in radicalization and extremism. A rational emotive behavioral conceptualization. *Frontiers in Psychology*, *10*(MAR), 6. https://doi.org/10.3389/fpsyg.2019.00437

Trip, S., Bora, C. H., Marian, M., Halmajan, A., & Drugas, M. I. (2019). Psychological mechanisms involved in radicalization and extremism. A rational emotive behavioral conceptualization. *Frontiers in Psychology*, *10*(MAR), 6. https://doi.org/10.3389/fpsyg.2019.00437

Twitter. (2018). *Expanding and building #TwitterTransparency*. https://blog.twitter.com/en_us/topics/company/2018/twitter-transparency-report-12

UK Home Office. (2015). *HM Government Counter-Extremism Strategy Cm 9148*. www.gov.uk/government/publications

Ul Rehman, Z., Abbas, S., Khan, M. A., Mustafa, G., Fayyaz, H., Hanif, M., & Saeed, M. A. (2020). Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning. *Computers, Materials and Continua*, *66*(2), 1075–1090. https://doi.org/10.32604/cmc.2020.012770

Van Bergen, D. D., Feddes, A. F., Doosje, B., & Pels, T. V. M. (2015). Collective identity factors and the attitude toward violence in defense of ethnicity or religion among Muslim youth of Turkish and Moroccan Descent. *International Journal of Intercultural Relations*, *47*, 89–100. https://doi.org/10.1016/j.ijintrel.2015.03.026

van de Weert, A., & Eijkman, Q. A. M. (2019). Subjectivity in detection of radicalisation and violent extremism: a youth worker's perspective. *Behavioral Sciences of Terrorism and Political Aggression*, *11*(3), 191–214. https://doi.org/10.1080/19434472.2018.1457069

Van Den Hurk, M., & Dignum, F. (2021). *Towards fundamental models of radicalization*.

Vidiyala, R. (2019). *What, Why and How of t-SNE. Dimensionality Reduction using t-SNE in… | by Ramya Vidiyala | Towards Data Science*. https://towardsdatascience.com/what-why-and-how-of-t-sne-1f78d13e224d

Vosoughi, S., Vijayaraghavan, P., & Roy, D. (n.d.). *Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder*. https://doi.org/10.1145/2911451.2914762

Wattenberg, M., Viégas, F., & Johnson, I. (2017). How to Use t-SNE Effectively. *Distill*, *1*(10), e2. https://doi.org/10.23915/distill.00002

Weisser, M. (2013). *Corpus Linguistics*. http://martinweisser.org/courses/intro/corpusLing.html

Winter, A. (2019). Online Hate: From the Far-Right to the 'Alt-Right' and from the Margins to the Mainstream. In *Online Othering* (pp. 39–63). Springer International Publishing. https://doi.org/10.1007/978-3-030-12633-9_2

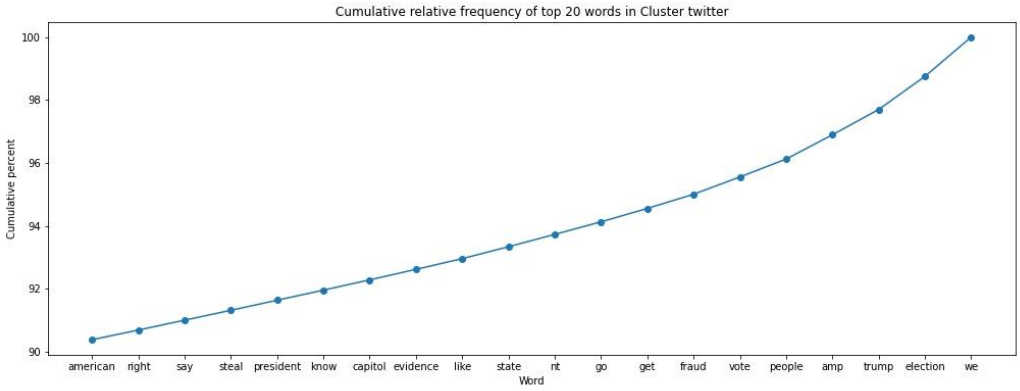
WordHoard. (2021). *Comparing Word Form Counts*. https://wordhoard.northwestern.edu/userman/analysis-comparewords.html

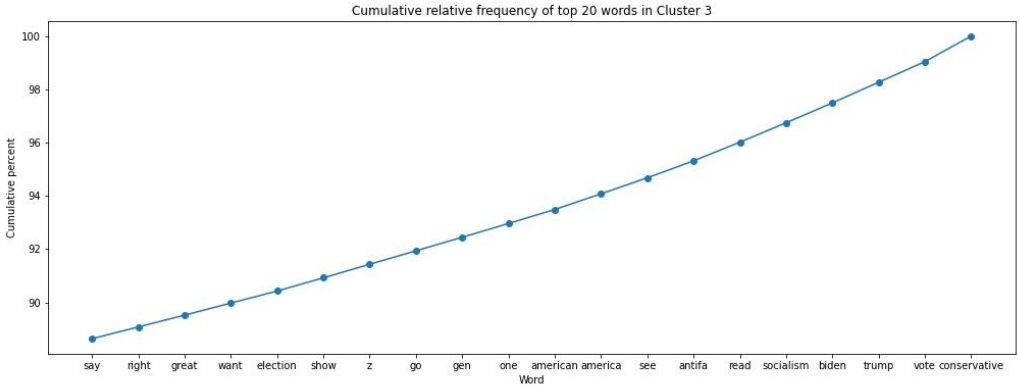
Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what @you #tag: Does the dual role affect hashtag adoption? *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, 261–270. https://doi.org/10.1145/2187836.2187872

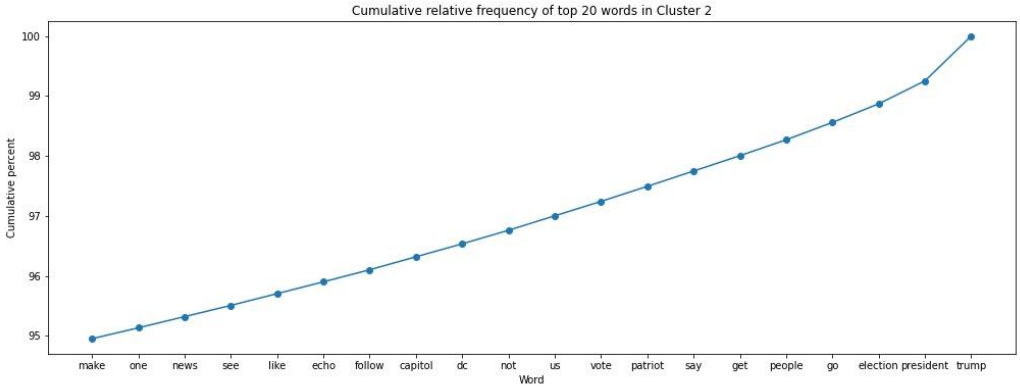
Zhang, Y., Wu, Y., & Yang, Q. (2012). Community Discovery in Twitter Based on User Interests. In *Journal of Computational Information Systems* (Vol. 8). http://www.jofcis.com

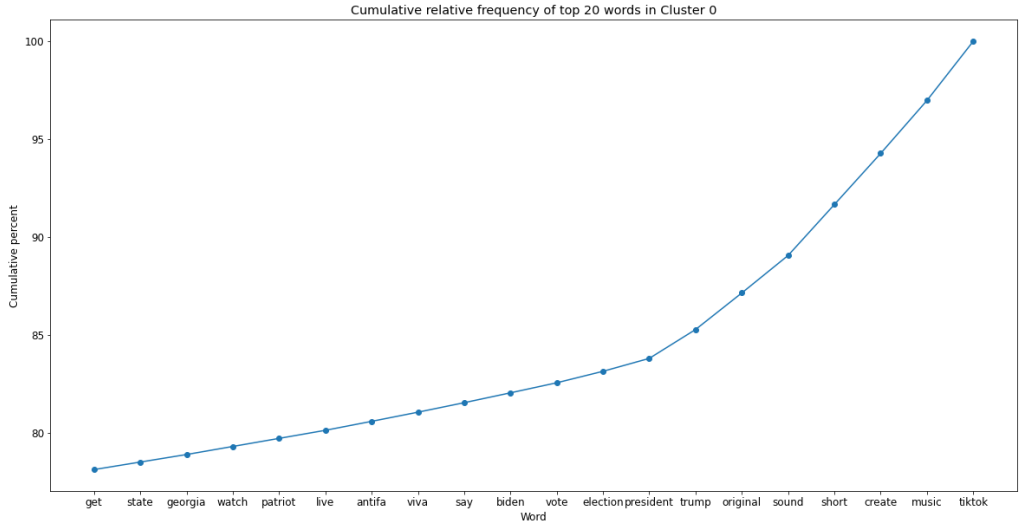
9 Appendix

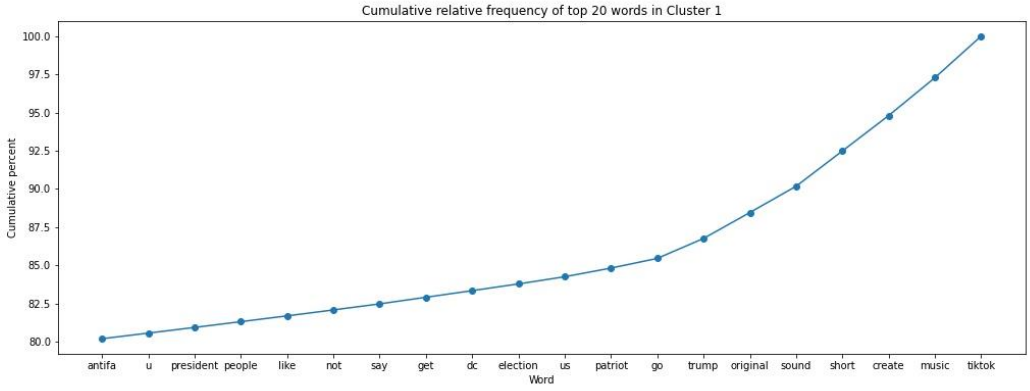
9.1 Parler EDA

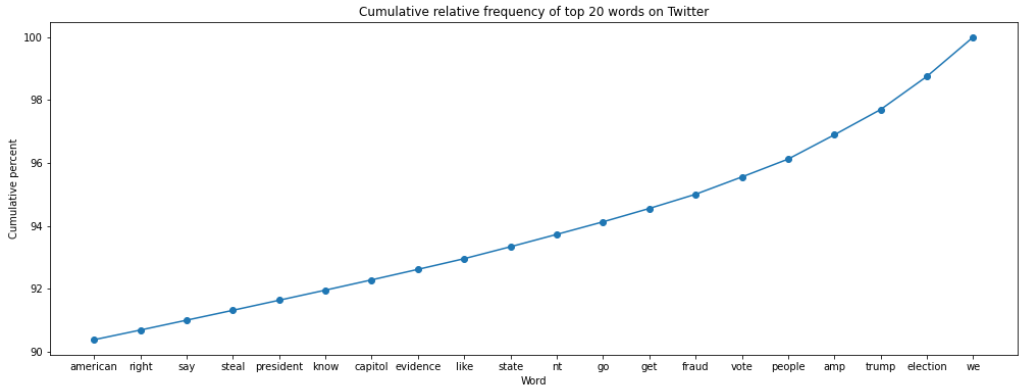




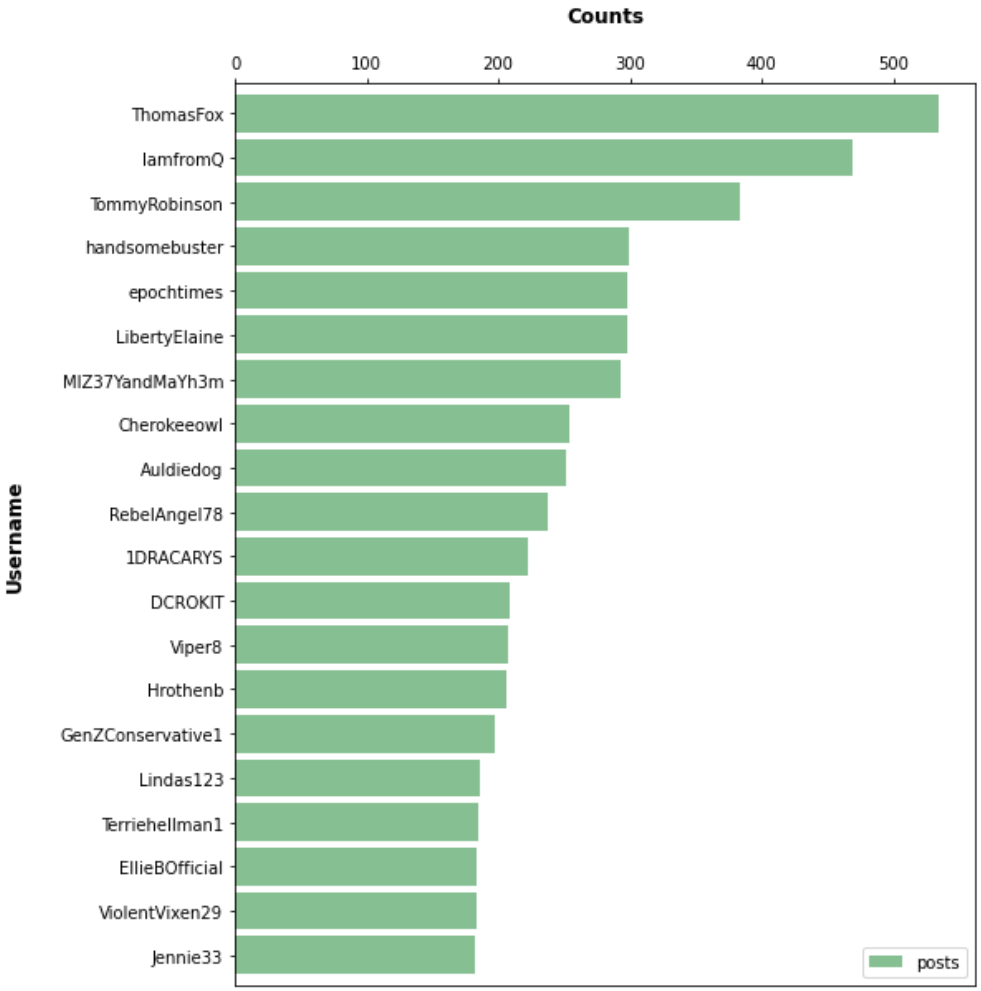



Cumulative relative frequency of top 20 words in Cluster 0
Cumulative percent
100
95
90
85
80
get
state
georgia
watch
patriot
live
antifa
viva
say
biden
vote
election
president
trump
original
sound
short
create
music
tiktok
Word

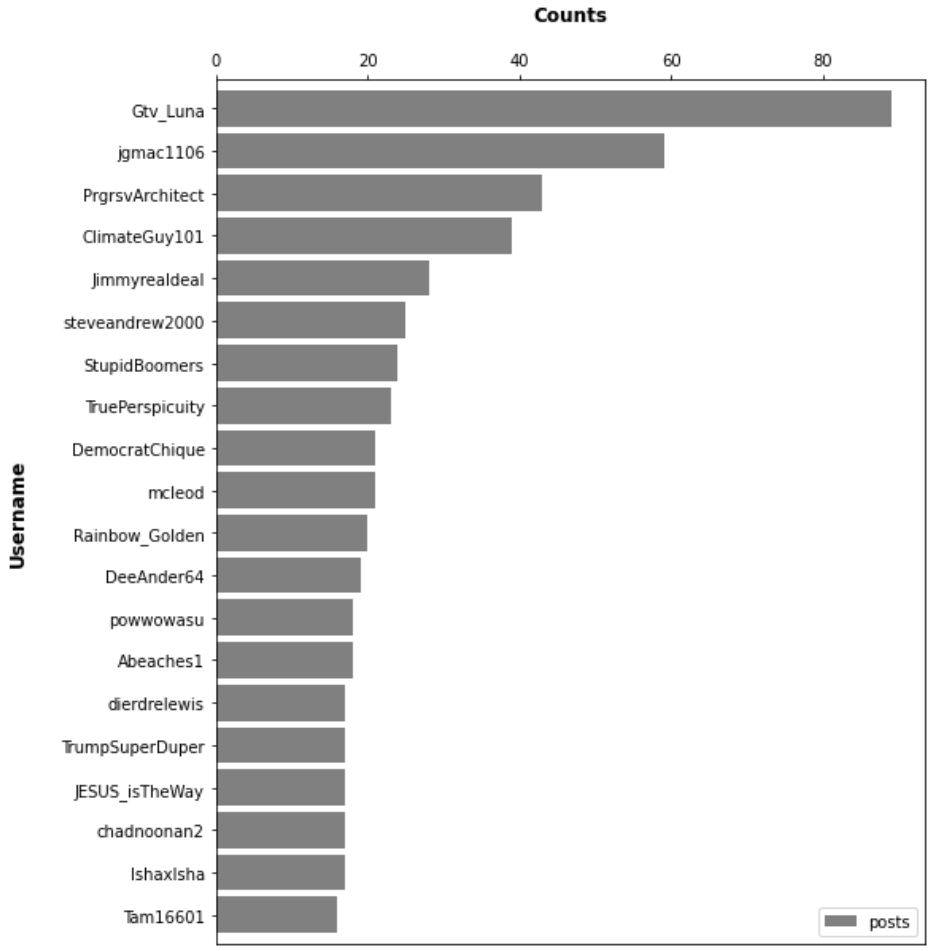

Cumulative relative frequency of top 20 words in Cluster 1
Cumulative percent
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
antifa
u
president
people
like
not
say
get
dc
election
us
patriot
go
trump
original
sound
short
create
music
tiktok
Word


Cumulative relative frequency of top 20 words on Twitter
Cumulative percent
100
98
96
94
92
90
american
right
say
steal
president
know
capitol
evidence
like
state
nt
go
get
fraud
vote
people
amp
trump
election
we
Word

	Top hashtags (Parler)		
	topics	**counts**	**percentage**
0	stopthesteal	7172	29
1	maga	5286	21
2	trump2020	4472	18
3	trump	4424	18
4	wwg1wga	2934	12
5	electionfraud	2884	11
6	fightback	2466	10
7	patriots	2202	9
8	voterfraud	2148	8
9	qanon	2044	8
10	wethepeople	1956	8
11	usa	1915	7
12	trumptrain	1911	7
13	draintheswamp	1799	7
14	americafirst	1793	7
15	kag	1700	6
16	election2020	1692	6
17	parler	1650	6
18	georgia	1631	6
19	Fakenews	1565	6


Counts
0
100
200
300
400
500
Username
ThomasFox
IamfromQ
TommyRobinson
handsomebuster
epochtimes
LibertyElaine
MIZ37YandMaYh3m
Cherokeeowl
Auldiedog
RebelAngel78
1DRACARYS
DCROKIT
Viper8
Hrothenb
GenZConservative1
Lindas123
Terriehellman1
EllieBOfficial
ViolentVixen29
Jennie33
posts

9.2 Twitter EDA



	username	posts	percentage
0	Gtv_Luna	89	0.77
1	jgmac1106	59	0.51
2	PrgrsvArchitect	43	0.37
3	ClimateGuy101	39	0.34
4	Jimmyrealdeal	28	0.24
5	steveandrew2000	25	0.22
6	StupidBoomers	24	0.21
7	TruePerspicuity	23	0.20
8	DemocratChique	21	0.18
9	mcleod	21	0.18