

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Lucas Henrique Costa de Lima

Measurement and analysis of gab, an unmoderated social network system

Belo Horizonte
2019

Lucas Henrique Costa de Lima

Measurement and analysis of gab, an unmoderated social network system

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte
2019

© 2019, Lucas Henrique Costa de Lima.
. Todos os direitos reservados

Lima, Lucas Henrique Costa de.

L732m Measurement and analysis of gab, an unmoderated social network system [manuscrito] / Lucas Henrique Costa de Lima. — 2019.
xxiv, 65 f. il.

Orientador: Fabrício Benevenuto de Souza.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.

Referências: f. 57-65

1. Computação – Teses. 2. Gab– Teses. 3. Twitter – Teses. 4. Redes sociais – Teses. 5. Discurso de ódio – Teses. 6. Liberdade de expressão – Teses I. Souza, Fabrício Benevenuto de II. Título.

CDU 519.6*04(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

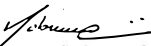
FOLHA DE APROVAÇÃO

Measurement and Analysis of Gab, an Unmoderated Social Network System

LUCAS HENRIQUE COSTA DE LIMA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:


PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG


PROF. FABRÍCIO MURAI FERREIRA
Departamento de Ciência da Computação - UFMG


PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA
Departamento de Ciência da Computação - UFMG


PROF. EMILIO FERRARA
Computer Science Department - University of Southern California

Belo Horizonte, 16 de Dezembro de 2019.

To my parents, brother, and loved ones.

Acknowledgments

First of all, I would like to thank my parents, Edison and Lídia, for all their love and dedication so I could achieve my goals. To my brother, Luiz, for all the support and encouragement. Without you, none of this would have been possible.

To João Vítor, for all the love and support since the beginning of this journey. Thanks for your patience and for always being by my side, even in the moments of struggle and anxiety. A very special thanks to my dearest friends Rangel and Natércia, for your friendship and for sharing the struggles of grad students. You were the best friends that I could have had during this journey. To my dear friends Elizabeth and Simone, Amanda, all the friends from LOCUS, Glumpor, Seicho-No-Ie, old friends from CEFET-MG, Waleska and Izabella, and to my special friends Alicia, Keenal, and Nikha for all your support and friendship. I am really thankful for having you all in my life.

I also would like to thank the professors and community of the DCC-UFMG. In particular, thanks to the professors Pedro Olmo, Ana Paula, and Mirella Moro who I had the opportunity to closely work with, still as an undergraduate student. A special thanks to Fabricio Murai who collaborated in this study. To Leandro Araújo and Juhi Kulshrestha for also collaborating with this work. Thanks to CAPES and MASWeb for the funding.

Last but not least, I would like to thank my academic advisor, professor Fabrício Benevenuto, who has shown to be an incredible mentor and leader throughout this journey. Thanks for the encouragement, support, opportunities, and for always believing in my potential. I have learned so much from you. I am really thankful for having you as my advisor.

“What is essential is invisible to the eye.”
(Antoine de Saint-Exupéry)

Resumo

Mídias sociais vêm mudando a forma como nossa sociedade se comunica, tornando-se locais populares para usuários consumirem, produzirem e disseminarem conteúdo. Apesar das valiosas interações sociais que esses ambientes promovem, cria-se também um espaço para discursos potencialmente prejudiciais a diferentes grupos de pessoas. Recentemente, há um longo debate entre a regulação de conteúdo e a liberdade de expressão nas redes sociais. A moderação de conteúdo em muitas redes sociais, como Twitter e Facebook, motivou o surgimento de uma nova rede social com o foco em liberdade de expressão, chamada Gab. Rapidamente, o aplicativo do Gab foi removido da Google Play Store por violar a política de discursos de ódio da empresa e foi rejeitado pela Apple por motivos semelhantes. Neste trabalho, apresentamos um estudo aprofundado sobre o Gab, com o objetivo de entender quem são os usuários que aderiram ao sistema e o tipo de conteúdo que eles compartilham nessa rede social. Nossas descobertas mostram que Gab é um sistema extremamente politicamente orientado que hospeda usuários banidos de outras redes sociais, alguns deles devido a possíveis casos de discurso de ódio e associação com extremismo. Nós fornecemos a primeira medição da disseminação de notícias dentro de uma câmara de eco politicamente conservadora onde os leitores raramente são expostos a conteúdo que atravessa linhas ideológicas, mas são alimentados com conteúdos que reforçam suas visões políticas ou sociais atuais. Por fim, apresentamos uma análise comparativa de discurso de ódio no Gab e no Twitter, uma rede social com uma rígida moderação de conteúdo. Mostramos que existem diferenças linguísticas significativas entre um conteúdo moderado e não moderado, além de mostrar que um ambiente não moderado como o Gab pode conter proporcionalmente mais discurso de ódio do que mídias tradicionais. Finalmente, mostramos os tipos mais comuns de ódio encontrados em cada uma das redes sociais. Esperamos que nossas análises possam contribuir para a discussão em torno da moderação de discurso e beneficiar abordagens de detecção de discurso de ódio.

Palavras-chave: Gab, Twitter, Redes Sociais, Notícias, Discurso de Ódio

Abstract

Social media systems have changed the way our society communicates, becoming popular places for users to consume, produce, and disseminate content. Despite the valuable social interactions that the online media promote, these systems also provide space for speech that would potentially be detrimental to different groups of people. Recently, there has been a long debate between content regulation and freedom of expression in social networks. The moderation of content in many social media systems, such as Twitter and Facebook, motivated the emergence of a new social network for free speech, named Gab. Soon after that, the Gab app has been removed from Google Play Store for violating the company's hate speech policy and it has been rejected by Apple for similar reasons. In this work, present a deep study about Gab, aiming at understanding who are the users who joined it and what kind of content they share in this system. Our findings show that Gab is a very politically oriented system that hosts banned users from other social networks, some of them due to possible cases of hate speech and association with extremism. We provide the first measurement of news dissemination inside a right-leaning echo chamber, investigating a social media where readers are rarely exposed to content that cuts across ideological lines, but rather are fed with content that reinforces their current beliefs. We present an analysis of posts from Gab, while comparing them with those from Twitter, a content-moderated social network. Our findings support that unmoderated environments have significant different linguistic features from moderated environments, and proportionally more hate speech. Finally, we show the most common type of hate in both social systems. We hope our analysis and findings may contribute to the discussion around moderation of speech and benefit hate speech detection approaches.

Keywords: Gab, Twitter, Social Media, News, Hate Speech

List of Figures

2.1	Gab logos as observed in (a) August 2017 and (b) November 2019.	19
3.1	Flowchart representing the methodology for identifying hate posts from Gab and Twitter posts.	30
4.1	Cumulative distribution function (CDF) for the number of followers and friends of Gab users.	34
4.2	Percentage of users who joined Gab per month since its creation.	36
4.3	Distribution of bias scores for 16,804 Gab users.	37
5.1	Word cloud for all Gab posts.	43
5.2	Number of times news sources were shared in Gab posts categorized as news, grouped by political leaning (Republican, Democrat, and Neutral) as inferred by [12] (a), [53] (b), [8] (c), and AllSides (d). Each box plot shows minimum, 25-percentile, median, 75-percentile and maximum.	46
6.1	Percentage of Gab and Twitter posts which contain at least one word or token per LIWC dimension.	52
6.2	Cumulative Distribution Function (CDF) for (a) Overall Sentiment Score and (b) Toxicity Score.	53
7.1	Percentage of Gab and Twitter posts which contain at least one word or token per LIWC dimension.	57
7.2	Intersection size of the sets of different types of hate for the 775 posts which are associated with more than one type of hate.	59

List of Tables

2.1	List of all 15 Gab categories as observed in August 2017.	19
2.2	Example of Gab post along with information about users and posts' attributes.	20
3.1	Lexicon of 34 hate terms from Hatebase presenting their categories. At the bottom, the number of terms in each category and its relative percentage within the parenthesis.	31
3.2	Total number of posts for Gab and Twitter, as well as the total number of labeled hate posts for each social media.	32
4.1	Top 10 most followed users as observed in August 2017.	34
4.2	Demographic distribution of nearly 36 thousand Gab users.	36
4.3	List of profiles considered extremist by the SPLC and ADL who were found in Gab.	39
4.4	Top 10 news spreaders.	40
5.1	Frequent bigrams and trigrams in Gab posts along with the percentage of posts in which they appear.	43
5.2	Most popular posts' categories in Gab.	43
5.3	Top 10 languages with more posts.	44
5.4	Number of domains found in Gab posts which are categorized as news and coexistence in each dataset.	45
5.5	Top 30 news sources in posts and their respective domain, percentage over all posts.	47
5.6	Top 15 most popular links shared in Gab, their number of shares in Gab, and popularity according to Bit.ly (larger means more popular).	48
6.1	LIWC dimensions, subdimensions and attributes used in the present study. . .	51
7.1	Manual evaluation. Each triple shows the number of posts agreed as hate, without agreement and agreed as non-hate, respectively.	56
7.2	Top 10 most frequent hate terms in Gab and Twitter hate posts.	58
7.3	Examples of hate posts with multiple types of hate. The associated types of hate for each post are in bold.	60

Contents

1	Introduction	13
1.1	Objectives and Goals	14
1.2	Main Contributions	15
1.3	Organization	16
2	Background and Related Work	18
2.1	The Gab Social System	18
2.2	Echo Chambers in Social Media Systems	20
2.3	News Sharing and Propagation in Social Media	21
2.4	Social Media Content Moderation	22
2.5	Online Manifestations of Hate Speech	23
3	Datasets and Methodology	25
3.1	Datasets	25
3.1.1	Common Users Dataset	26
3.2	Measurement Methodology	26
3.2.1	Measuring Demographic Factors of Users	26
3.2.2	Language Processing Methods	27
3.2.2.1	Linguistic Analysis	27
3.2.2.2	Sentiment Analysis	28
3.2.2.3	Toxicity Analysis	28
3.2.3	Assessing Online Hate Speech	29
3.3	Potential Limitations	30
3.4	Summary	32
4	RQ1 - Who are the Gab users?	33
4.1	Network Structure	33
4.2	Demographic Factors of Users	35
4.2.1	Location	35
4.2.2	Gender and Race	36
4.2.3	Political Leaning	37
4.3	Extremism in Gab	38
4.4	News Spreaders	38
4.5	Concluding Remarks	40

5	RQ2 - What do users share on Gab?	42
5.1	Popular Words, Topics, and Languages	42
5.2	News sharing	44
5.2.1	Most Shared Domains	45
5.2.2	Top Stories	48
5.3	Concluding Remarks	49
6	RQ3 - Distinguishing characteristics of moderated and unmoderated content	50
6.1	Linguistic Features	50
6.2	Sentiment Analysis and Toxicity	53
6.3	Concluding Remarks	54
7	RQ4 - Different types of hate across moderated and unmoderated environments	55
7.1	Manual Validation	55
7.2	Types of Hate	56
7.2.1	Frequent Types of Hate	56
7.2.2	Frequent Hate Terms	57
7.2.3	Multiple types of Hate	58
7.3	Concluding Remarks	59
8	Conclusions and Future Work	61
	References	63

Chapter 1

Introduction

The Web has changed the way our society communicates, giving rise to social platforms where users can share different types of content and freely express themselves through posts containing personal opinions. Unfortunately, with the popularization of this new flavor of communication, toxic behaviors enacted by some users have been gaining prominence through online harassment and hate speech. These platforms have become the stage for numerous cases of online hate speech, a type of discourse that aims at attacking a person or a group on the basis of race, religion, ethnic origin, sexual orientation, disability, or gender [40].

Hate speech, as well as extremism in ideological opinions and fake news, are examples of problems currently faced by our society, in particular in online social media [74, 8, 73, 66]. They have been recognized as serious problems by many societal segments and authorities across different countries [31], such as Germany, where the first steps towards regulating hate speech in social networks have already been taken. As many online platforms increasingly need to detect and counter the dissemination of online hate, hate speech has effectively become a Computer Science problem [22, 47, 54].

Recently, to prevent the proliferation of toxic content, most online social networks prohibited hate speech in their user policies and enforced this rule by deleting posts and banning users who violate it. Some news websites even turned their comments section off due to lack of resources for manual moderation¹, as human moderation of content can be a costly operation because of the large volume of data². Particularly, Twitter, Facebook, and Google (YouTube) have largely increased removals of hate speech content [48]. Reddit also deleted some communities related to fat-shaming and hate against immigrants [15].

This scenario has motivated the emergence of a new social network system, named Gab³, that has nearly 1 million users (as announced in June 2019). In essence, Gab is very similar to Twitter, but barely moderates any of the content shared by its users. According to Gab guidelines, the website promotes freedom of expression and states that “*the only*

¹<https://medium.economist.com/help-us-shape-the-future-of-comments-on-economist-com-fa86eeafb0ce>

²<https://www.nytco.com/press/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/>

³<https://gab.ai/>

valid form of censorship is an individual's own choice to opt-out". They, however, do not allow illegal activity, spam, or form of illegal pornography, promotion of violence and terrorism⁴. In spite of advocating liberty and freedom of speech, Gab has received several criticisms regarding the content shared there.

The lack of moderation in this system started to attract users banned due to hate speech from other social networks [84], and it has also been pointed out as a safe space for alt-right groups [83]. This environment favors the emergence of clusters of like-minded individuals and the polarization of opinions, creating groups of individuals who are often fed with information that reinforces their current beliefs. This phenomenon is known as "echo chambers" and it has been extensively studied recently [73, 62]. Despite the importance of all these efforts, little is still known about what happens inside an echo chamber. This lack of understanding is likely due to the difficulty of splitting apart what is a bubble in social networks such as Facebook and Twitter.

In this context, understanding what happens inside Gab can provide us valuable insights into the study of right-leaning echo chambers. More important, contrasting the content shared in Gab with the content shared in Twitter might give us important insights about the result of very different content policies. Thus, Gab is a valuable source of data for researchers. The next section summarizes the goals of this work.

1.1 Objectives and Goals

The main goal of this work is to analyze a right-leaning echo chamber, performing a deep characterization of the Gab social network, its users and posts. Moreover, this research identifies the textual characteristics of a set of unmoderated data and compare them with characteristics of moderated data. We analyze and investigate the existence of hate speech and its different types in both moderated and unmoderated environments. Overall, the main research questions investigated by this work are summarized as follows.

- **Research Question 1 (RQ1):** Who are the users who joined Gab in terms of extremist views, political leaning, and ability to spread news?
- **Research Question 2 (RQ2):** What kind of content is shared in Gab? What are the news shared within this system?
- **Research Question 3 (RQ3):** What are the distinguishing characteristics of unmoderated content in Gab and moderated content in Twitter in terms of linguistic

⁴<https://gab.ai/about/guidelines>

features, sentiment, and toxicity?

- **Research Question 4 (RQ4):** What are the most common types of hate in an unmoderated and moderated environment?

We hope our analyses contribute to the understanding of the behavior of users on a social network without the strict policies of moderation present in other media. We also believe that our analysis and findings may contribute to the debate about hate speech and free speech, and benefits systems aiming at deploying hate speech detection approaches.

1.2 Main Contributions

The main contributions of this study can be summarized as follows.

- We show that Gab is a very politically oriented system. Furthermore, we show that the majority of Gab users are conservative, male, and caucasian. We were able to identify many users listed as extremists by the main media who showed to be influential and very active in the Gab network. The most popular type of discussion in Gab posts is focused on politics and conservatism.
- We provide the first measurement of news dissemination inside a right-leaning echo chamber. We show that, although most of the (unique) news domains present in shared URLs are considered left-leaning, conservative news outlets comprise a larger fraction of the shared links. By quantifying the popularity of sites whose links are disseminated inside this ideology-biased community, we show that popular news domains shared on Gab are not popular on Facebook or the Internet.
- Our analysis shows that content in Gab and Twitter have different linguistics patterns, with higher toxicity and a more negative overall sentiment score in the unmoderated Gab content. Additionally, we find that, in general, Gab has more hate posts than Twitter. We show that Gender and Class types of hate are more frequent on Twitter, whereas Disability, Ethnicity, Sexual Orientation, Religion, and Nationality types tend to appear proportionality much more in Gab.
- Our findings not only highlight the importance of creating moderation policies as an effort to fight online hate speech in social systems, but also point out possible points for improvement in the design of content policies for social media systems. Additionally, our findings suggest that the unmoderated content found in Gab might

be an appropriate data source for the development of learning approaches to detect hate speech. Thus, as a final contribution, we make our hate-labeled Gab posts available for the research community. We hope the labeled messages exchanged in Gab, categorized into different types of hate, can foster the development of future hate speech detection systems.

The results presented in this work are part of the following publications:

- Lima, L., Reis, J. C., Melo, P., Murai, F., and Benevenuto, F. (2020). Characterizing (Un)moderated Textual Data in Social Systems. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).
- Lima, L., Reis, J. C., Melo, P., Murai, F., Araujo, L., Vikatos, P., and Benevenuto, F. (2018). Inside the right-leaning echo chambers: Characterizing Gab, an Unmoderated Social System. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

1.3 Organization

The rest of this work is organized as follows.

- **Chapter 2: Background and Related Work.** This Chapter provides information about the Gab social system and reviews related work along four dimensions: (i) echo chambers in social media systems, (ii) efforts related to news sharing and propagation in social media, (iii) social media content moderation, and (iv) recent efforts towards online hate speech understanding and detection.
- **Chapter 3: Datasets and Methodology.** In this Chapter, we describe our strategy for data collection, the methods, and frameworks used in our analyses, as well as the datasets used in this work.
- **Chapter 4: RQ1 - Who are the Gab users?** In this Chapter, we provide a series of analyses aiming at depicting who are the users that joined Gab, presenting demographic and political leaning related analyses for Gab users. We also investigate the presence of extremist users and news spreaders.
- **Chapter 5: RQ2 - What do users share on Gab?** This Chapter analyzes the content shared in Gab. We characterize typical language usage and investigate the dissemination of news in this social media.

-
- **Chapter 6: RQ3 - Distinguishing characteristics of moderated and unmoderated content.** This Chapter analyzes distinguishing characteristics of moderated and unmoderated content in terms of (i) linguistic features, (ii) sentiment, and (iii) toxicity scores.
 - **Chapter 7: RQ4 - Different types of hate across moderated and unmoderated environments.** In this Chapter, we evaluate hate speech and its different forms in both moderated and unmoderated environments.
 - **Chapter 8: Conclusions and Future Work.** Finally, this Chapter presents concluding remarks and offers directions for future work.

Chapter 2

Background and Related Work

In this Chapter, we review background information and studies related to this work. We start by providing an overview of the Gab social system. Then, we present related work along four dimensions directly correlated with the main research questions approached this work. First, we discuss echo chambers and how they manifest in social systems. Next, we investigate efforts related to news sharing and propagation in social media. Then, we discuss the impact of content moderation on social media. Finally, we present efforts concerning online hate speech understanding and detection.

2.1 The Gab Social System

Gab has been designed to be a social network that promotes online free speech. The main page of Gab describes itself as a “*social network that champions free speech, individual liberty and the free flow of information online. All are welcome*”. During the past years, Gab has changed many features of the network. It used to feature a green frog as its logo, which raised doubts regarding its relatedness with the hate symbol “Pepe the Frog”¹. Figure 2.1 shows one of the first logos (Fig. 2.1a) and the logo as observed in November 2019 (Fig. 2.1b).

Posts are limited to 3,000 characters and users can reply, repost, quote or favorite them. Each post has an ID, an associated user ID, and a timestamp. Additionally, posts could be categorized as News, Politics, Art, etc., topics that users could use to find popular content on specific subjects. The complete list of Gab categories, as observed in August 2017, is shown in Table 2.1. Users can follow and be followed by other users. Each user has a user ID, name, screen name, number of friends, number of followers, associated posts, account creation date, information of whether the profile is verified, PRO (subscribers), or Premium (users paid for their content), profile picture URL, and a profile bio. Table 2.2

¹<https://www.adl.org/news/press-releases/adl-adds-pepe-the-frog-meme-used-by-anti-semites-and-racists-to-online-hate#.V-rqlvkrJaQ>



Figure 2.1: Gab logos as observed in (a) August 2017 and (b) November 2019.

Table 2.1: List of all 15 Gab categories as observed in August 2017.

Categories
Art
Ask Me Anything (AMA)
Cuisine
Entertainment
Faith
Finance
Humor
Music
News
Politics
Philosophy
Photography
Science
Sports
Technology

shows an example of Gab post from its founder, Andrew Torba, along with the attributes we have in our dataset. This post has been published under the context of Gab being removed from the Google Play Store for hate speech². Andrew Torba has now (November 2019) over a thousand friends, more than 150 thousand followers, and over 12 thousand posts.

Since Gab is a very recent and controversial network, there are few other recent studies on this social network. [85] present concurrent work on Gab. [50] investigates the diffusion characteristics of the posts made by hateful and non-hateful users. Overall, our work is complementary as we provide a much deeper investigation about Gab users and posts, performing analysis on hateful posts on Gab while comparing with hateful posts on Twitter.

²<https://www.theverge.com/2017/8/18/16166240/gab-google-play-removed-hate-speech>

Table 2.2: Example of Gab post along with information about users and posts' attributes.

users' attributes	
<i>id</i>	31
<i>name</i>	Andrew Torba
<i>screen name</i>	a
<i>following count</i>	1, 485
<i>followers count</i>	32, 780
<i>posts count</i>	8, 601
<i>account creation</i>	August 2016
<i>verified profile (bool)</i>	True
<i>PRO profile (bool)</i>	True
<i>Profile Bio</i>	Patriot. CEO of @Gab I'm fighting for a better internet that puts people first and promotes free speech for all. "Your freedom to be you includes my freedom to be free from you." -Andrew Wilkow Exodus 8:2-7
posts' attributes	
<i>id</i>	10888485
<i>body</i>	Attention #GabFam, We are exploring all of our legal options in regards to Google's unfair, discriminatory and unjust treatment of our app. Stay tuned.
<i>published at</i>	2017-08-18 T19:56:03+00:00
<i>type</i>	Repost
<i>category title</i>	News

2.2 Echo Chambers in Social Media Systems

Facebook, in particular, with its *News Feed*, shows users stories based on social and content-based features [77], personalizing it as an effort to make it meaningful and informative³. The surrounding context and impact of the News Feed algorithm have been the object of several studies [37, 38, 63, 23, 11], though many Facebook users were not aware of the News Feed curation algorithm's existence [27]. According to [73], these news feeds are constructed in a way that consumers are selectively exposed to certain kinds of news, which, therefore, leads to the establishment of groups containing like-minded people where the polarization of opinions may happen, resulting in echo chambers in social media.

This phenomenon has been extensively studied in recent works. In the context of behavioral and psychological studies, [80] show through a comparative analysis of two distinct polarized communities on Facebook, that emotional behavior is affected by the users' involvement inside the echo chamber. [10] investigates whether users in echo chambers have similar personality traits, and highlights that specific personality traits might lead users to join polarized communities. Notice that both works focused on the un-

³<https://www.facebook.com/facebookmedia/solutions/news-feed>

derstanding of particular pages and communities on Facebook. Differently, we provide an in-depth study of a complete social network, highlighting features that characterize Gab as an entire right-leaning echo chamber. We also provide an investigation on the psycholinguistic features of Gab posts, contributing to the discussion presented in the context of behavioral and psychological studies.

In the political context, [32] examine the interaction between shared opinion and the social network as an enabling environment for dissemination. Based on metrics of users' production and consumption, the authors identify three types of user profiles on Twitter: (i) users who are exposed to content that reinforces their opinions, (ii) users who try to bridge the echo chambers by sharing content with diverse leaning, and (iii) the gatekeepers, which are users who consume diverse content but produce content that fuels the bubbles. Finally, these results are used as input for inferring user profiles. [24] shows that greater interest in politics and more media diversity reduces the likelihood of individuals being in an echo chamber. In this work, we perform a characterization of the news media and news spreaders presented in Gab so we can understand whether there is a diversity of content in this social network.

Overall, our work is complementary to such studies as we investigate a social network that has shown to be unique and an entire right-leaning echo chamber, differently from traditional media, such as Facebook and Twitter. Understanding what happens inside Gab can provide researchers valuable insights into the study of right-leaning echo chambers.

2.3 News Sharing and Propagation in Social Media

As the characterization of news plays an important role in this work, we next review related work concerning news sharing and propagation in social media.

Social media, such as Facebook and Twitter, have recently become popular places for users to obtain, share and discuss different news in the online environment. Whereas in the middle '90s only 12% of U.S. adults got news online [61], this number has grown to about 81%, with an increasing number of U.S. adults consuming news primarily from social media sites [52]. One characteristic of these social media is that they allow anyone to be registered as a news publisher, which results in the emergence of many news outlets, in addition to the traditional news media that are increasingly migrating to social media [69, 46]. Users have now a large number of options when it comes to deciding what and where to get news in social media.

The many factors that influence the news sharing intention in social media, in

particular information seeking, socializing, entertainment, status seeking and prior social media sharing experience, are explored by [44]. [65] investigate the demographics of users on Twitter and how it affects news sharing. They show that white and male users tend to share more news on Twitter, biasing the news audience towards the interests of these demographic groups. Our work is complementary to these studies as we provide the first measurement of news dissemination inside a right-leaning echo chambers, showing that news spreaders in Gab can reach a larger number of users of the network within just one hop.

Recently, many studies have focused on the understanding of misinformation and fake news dissemination in social media. [67] study political-oriented groups as an effort to understand information and misinformation dissemination on WhatsApp. The study of social bots and its impact on the spread of misinformation campaigns have been reviewed by [28]. Many other works are also concerned by fake news publishers posting and disseminating fake stories using followers which might be fake as well [3, 81, 43].

Complementary, our work provides an in-depth investigation of what kind of news is shared in Gab. We contribute to the understanding of an environment which is also prone to the dissemination of fake news and online activism.

2.4 Social Media Content Moderation

Back in 2010, the ongoing debate on content moderation in an online environment led Internet companies to reflect on the undesirable attention their sites can attract and the consequences of it⁴. As we have already discussed, the first steps towards regulating and moderating content have already been taken. Our work contributes to this discussion since we quantify the differences between moderated and unmoderated text data as an effort to shed light on the importance of creating different methodologies and policies to outline the boundaries of hate in social media.

There is a lot of debate on how social media platforms moderate their content, and how their moderation policies are shaped. Twitter and YouTube, for instance, make available their hate policies^{5,6} so users can actively report content that might violate their policies. [6] analyses how Muslims are being viewed and targeted by perpetrators of online abuse via the Twitter search engine and argues that new cyberhate policy is needed. Our work also brings the discussion of the impact these policies and the moderation of content

⁴<https://www.nytimes.com/2010/07/19/technology/19screen.html>

⁵<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

⁶<https://support.google.com/youtube/answer/2801939?hl=en>

might inflict on social systems content.

Besides, some other studies have focused on the understanding of systems that lacks moderation of content. [29] focus on making extensive analysis of antisemitism in 4chan's and Gab. Their results provide a quantitative data-driven framework for understanding this form of offensive content. Our effort is complementary to these prior works, as we compare the textual data shared from an unmoderated system like Gab with a moderated one as Twitter, highlighting some psycholinguistic differences between these two different environments.

The data within Gab now allow researchers to empirically measure what really happens in an unmoderated free speech environment. Thus, by characterizing this system, we hope to bring insights that can contribute to this discussion.

2.5 Online Manifestations of Hate Speech

As we understand that the polarized groups found in Gab might turn the place into an environment prone to the dissemination of hate speech, we next provide an extensive review about hate and efforts related to detecting it in social media.

Even before social networks became popular, the problem of racism and hate detection was already a Computer Science research theme. Back in 2004, there have been efforts that attempted to identify hateful web pages, containing racism or extremism [35]. Nowadays, comment features implemented by different systems raise the need to deal with hateful or aggressive messages. Particularly, [19] identify offensive language on YouTube comments. Also, comments' sections from major newspapers are becoming home for hostility, trolls, and negativism as well [64].

A vast number of studies were conducted to provide a better understanding of hate speech on the Internet. [74, 54] provide a deeper understanding of the hateful messages exchanged in social networks, studying who are the most common targets of hate speech in these systems. [71] create a taxonomy and use it to investigate how different features and algorithms can influence the results of the hatefulness classification of text using several machine learning methods. Some works also use a grading of how much hate there is in each message. [45] analyze data from Youtube and Facebook related to terrorist groups, asking whether hate speech was used to justify or promote the ideologies of the organization, or their tactics, besides denigrating their targets. [16] characterize two banned communities from *Reddit*, one about fat shaming and the other related to hate against immigrants in the US, and proposed a lexicon for hate speech detection. Our effort is complementary to these studies, as it quantifies the amount of hate shared in a

moderated and an unmoderated environment and highlights the different types of hate exchanged in these social systems, elucidating the importance of content moderation to fight hate on the Internet.

Some other studies focus on the instigators of hate content. [26] have a peer-to-peer approach to hate speech, in which they investigate the personality of instigators of hate speech and their targets. [17] dissect the #Gamergate controversy, where some blog posts turned out to generate a polarizing issue involving social justice and some related topics such as sexism and feminism in the gaming industry. The authors' analysis shows that new and popular accounts are generally more engaged, posted greater negative sentiment, and displayed more hate than general users. In contrast, we focus on the analysis of textual hate data and its characteristics in different environments.

Finally, several other efforts have attempted to provide detection approaches for hate speech [9, 34, 2, 82]. [18] review three recent studies that aim to detect the presence of racism or offensive words on Twitter. They show that although simple text searches for hate words in tweets represent a good strategy to collect hate speech data, it also has a major shortcoming: the context of the tweets is lost (the word "crow" or "squinty", for instance, is a racial slur in the United Kingdom, but it can also be used in multiple non-hate related contexts).

Our work makes an important step towards the development of automated hate speech detectors. We believe that an unmoderated hate dataset, as the one analyzed in this work, can help the development of better hate speech detection approaches in future works.

Chapter 3

Datasets and Methodology

This Chapter describes our datasets, the methods, and frameworks used in our analyses. Given the importance of data acquisition in the analyses presented in this work, we start by describing our dataset building process.

3.1 Datasets

Our Gab dataset comprises posts from users crawled following a Breadth-First Search (BFS) scheme on the graph of followers and friends. We used as seeds users who authored posts listed by categories in the Gab main page. We implemented a distributed and parallel crawler that ran in August 2017 which took three days to gather all users reachable from the seeds. In total, our dataset comprises **171,920 users** (the estimated number of users in August 2017 was 225 thousand [30]) and 12,829,976 posts.

Part of our analyses compare Gab and Twitter posts. The Twitter dataset contains English posts randomly selected from the Twitter 1% Streaming API. For consolidating our dataset and keep data consistency, we consider only random tweets published in the same period as Gab posts, which gives us also 12,829,976 tweets. Many works present efforts to mitigate bias and limitations of Twitter samples [33, 41, 56, 55]. Even though the 1% random sample from Twitter may not be completely representative of all Twitter, this is the best available option at our disposal to analyze online moderated content.

After preprocessing and removing duplicated posts in both datasets, we have a total of **7,794,990 Gab posts** and **9,118,006 tweets**. These are the final sets of posts for each media that are going to be further analyzed in this work.

3.1.1 Common Users Dataset

Part of our analysis is centered on users who have accounts in both Gab and Twitter and their posts. For discovering such users, we design a two-step methodology described as follow.

First, we searched for the screen name of all Gab accounts on Twitter by using Twitter REST API, identifying 62,291 (36.23% of Gab users) accounts with exactly the same screen name in both systems. Next, we compare the profile name of these users on Gab with their respective profile names on Twitter, keeping only those for which we found an exact match, and then collecting their Twitter timelines. This results in 23,030 users, which corresponds to about 13% of the total of Gab users from our data, in which 3,983 have at least one valid publication in both social media, i.e. a publication which is not only a hyperlink and has more than one character. These users are going to be further investigated in the political leaning analysis performed in this work.

3.2 Measurement Methodology

To provide a deep understanding of Gab data, we use several methods and frameworks recently explored in the scientific literature. In this section, we describe the methods used for the major analyses provided in this study.

3.2.1 Measuring Demographic Factors of Users

To infer the gender and race of Gab users, we use a methodology recently explored in previous efforts [51, 14]. The strategy consists of gathering the profile picture Web link of Gab users and submit these links into the Face++ API.

Face++¹ is a face recognition platform based on deep learning which can identify gender (i.e. male and female) and race (limited to Asian, Black, and White) from recognized faces in images. We have discarded users whose profile does not have a face picture or does not have a recognizable face according to Face++. From the total number of

¹<https://www.faceplusplus.com/>

Gab users we crawled, less than a half (47.22%) have a profile picture Web link. Given these 82,215 users, only 35,493 have valid profile pictures that we were able to collect demographic information.

For measuring the political bias of Gab users, we use a recently introduced framework [42], kindly shared by the authors. Their approach measures the political leaning of Twitter users, given a set of information from their friends. In particular, they measure the bias of a Twitter user after examining how close are the interests of this specific user with the interests of representative sets of users who are known to have a democratic or republican bias.

Thus, to identify the political leaning of Gab users, we use the set of users with accounts in both Gab and Twitter, as described by the aforementioned dataset section. Next, for each user identified as being the same on Twitter and Gab, we further gathered her lists of Twitter friends from the Twitter API. Then, the framework used to measure these users' political leaning was able to identify the leaning of 16,804 users out of the total of 23,030, nearly 73% of the matched pairs.

3.2.2 Language Processing Methods

Next, we describe the methods used in this work to perform language processing characterization. We rely on established methodologies to measure linguistic differences among textual content.

3.2.2.1 Linguistic Analysis

One of our goals is to understand the distinguishing linguistic characteristics of posts on Gab and Twitter and contrast them. Thus, we use the 2015 version of the Linguistic Inquiry and Word Count (LIWC) [78] to extract and analyze the distribution of psycholinguistic elements posts of both media. Since it has been proposed, LIWC has been widely used for several different tasks, including sentiment analysis [68] and discourse characterization in social media platforms [21].

LIWC is a psycholinguistic lexicon system that categorizes words into psychologically meaningful groups, which is organized as a hierarchy of categories and subcategories, all of which form the set of LIWC attributes. Examples of categories include *Linguistic*

Style, Affective Processes, and Cognitive Processes. Examples of attributes (or subcategories) of the Affective category include *Positive emotions, Negative emotions, Anxiety* and *Anger*. Each attribute is characterized by a set of words from LIWC’s dictionary. Examples of words representing *Anger* in the dictionary are *hate, kill, pissed*. Then, in a given text, the LIWC software counts the occurrence of the words in a sentence for each attribute. The output is the proportion of the words in each attribute to the total words of the text. In sum, LIWC classifies more than 6,400 words into nearly 90 attributes [60].

3.2.2.2 Sentiment Analysis

Methods that aim to extract and analyze subjective information from people’s emotions and attitudes towards something else have been widely used for many tasks and researches. In this context, numerous advances emerged in the field of sentiment analysis in recent years. We perform sentiment analysis on Gab and Twitter posts as a complementary effort to characterize the differences between moderated and unmoderated accounts.

We use an established opinion mining method to measure sentiment score on our messages: the SentiStrength [79], which has shown to be an effective tool for sentiment analysis in social media posts [1]. This method implements a combination of supervised learning techniques with a set of rules that impact the polarity of the feeling expressed by the algorithm. We apply the standard English version of SentiStrength to quantify positive $P \in [+1, +5]$ and negative $N \in [-5, -1]$ sentiments in each post, as well as their overall sentiment score, which is given by the difference between P and absolute N values for a post.

3.2.2.3 Toxicity Analysis

Finally, also as a complementary analysis to elucidate the difference between moderated and unmoderated content, we measure the toxicity of posts with the Perspective API². This API, created by Jigsaw and Google’s Counter Abuse Technology team, uses a Convolutional Neural Network (CNN) trained with word-vector inputs to create a classifier that scores the toxicity a comment might have on a conversation. This score measures

²<https://www.perspectiveapi.com>

how “*toxic*” a message can be perceived by a user.

The API creators define a toxic message as a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion. The value does not represent a degree of “toxic severity” of a particular message, but instead the probability that someone perceives that message as toxic. Scores range from 0 to 1, where scores closer to 1 indicate that posts are likely to be perceived as toxic. Also, they have related trained models for other dimensions that comments can be scored on, for example, “obscene”, “thoughtful”, “off-topic”, “spam”, etc.

3.2.3 Assessing Online Hate Speech

Many recent research efforts have attempted to operationalize the concept of hate speech, that is, to define hate speech in terms of measurable factors to be able to identify and counter it. One of the key challenges in doing so is that, even in our society, there is not an accepted definition of hate speech. For this reason, in this work, we reproduce a recent methodology to classify hate speech on social media posts with minimal noise.

[26] present a semi-automated classification approach for the analysis of directed explicit hate speech which relies on keyword-based methods and the Perspective API. The authors validate their methodology by incorporating human judgment using Crowdfunder, concluding that their final hate speech dataset is reliable and has minimal noise. The methodology to detect hate implemented in this work is inspired by the referred work and it is similar to it with minor changes.

Figure 3.1 illustrates through a flowchart the methodology for identifying hate posts implemented in this work. First, both datasets go through a pipeline where the initial step is to query the Perspective API using each post as input. Besides the toxicity score, we gather the *attack on commenter* score of posts, which measures direct and personal offense or injury to another user participating in the discussion. Next, we filter posts which have toxicity score higher than 0.8 and attack on commenter higher than 0.5 (these thresholds were defined by [26] so as to yield a high-quality dataset). Finally, we check whether these filtered posts contain at least one hate word, and, if so, we assume these are hate posts.

The list of hate words is also obtained from the study of [26]. The authors curated a compressed lexicon of Hatebase³ terms which are likely to indicate hate speech content across different hate classes for characterizing the types of hate in social media. Hatebase is described as “*the world’s largest structured repository of regionalized, multilingual hate*

³<https://hatebase.org/>

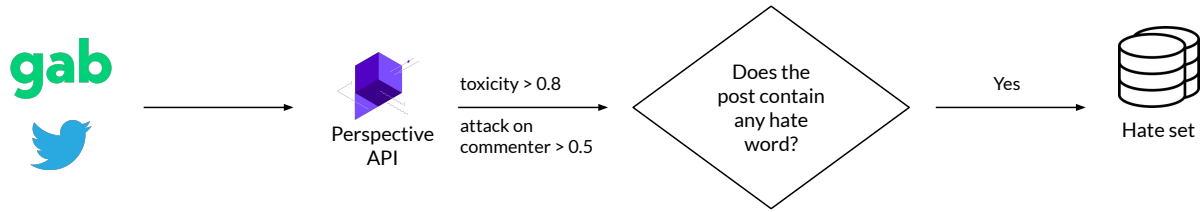


Figure 3.1: Flowchart representing the methodology for identifying hate posts from Gab and Twitter posts.

speech”. Table 3.1 shows the 34 terms obtained from the lexicon categorized according to different types of hate (ethnicity, class, disability, nationality, religion, gender, and sexual orientation). We removed terms that are not currently in Hatebase or which are not associated with any type of hate. Wherever present, the ‘*’ has been inserted by us, in order to lessen the impact that the offensive terms may inflict on some people, and was not part of the original word or text.

Following the aforementioned methodology, **9,554** (0.12%) Gab posts and **2,392** (0.03%) tweets are labeled as **hate** and are going to be further explored on the hate speech analysis presented in this work. This methodology expands the hate speech assessment for Gab explored by [85], as it provides finer granularity for the investigation of hate speech in this social media, and also presents an analysis on the different types of hate.

3.3 Potential Limitations

There are few limitations with our datasets and methods, that we describe next.

Accuracy of the inference by Face++. Gender and race inferences are challenging tasks, and we are limited by the accuracy of Face++ in the inference. A recent work [14] evaluated the effectiveness of the inference made by Face++, using human annotators to label randomly selected profile images from Twitter. They measured the inter-annotator agreement in terms of the Fleiss’ K score which was 1.0 and 0.865 for gender and race, respectively. Even though, we acknowledge that some inferences made in this work might be mistaken. Furthermore, we were able to collect gender and race for only 20.65% of Gab users, which also represents another limitation of this work.

Bias inference. One limitation of this analysis is that it requires us to match Gab and Twitter accounts as a first step, thus reducing the set of suitable users to 13.4%. This is due to the fact the framework provided by [42] is a Twitter-based approach. Furthermore, under the assumptions of the framework, one user may not be classified as both pro-republican and pro-democrat. However, as we observe through the other

Table 3.1: Lexicon of 34 hate terms from Hatebase presenting their categories. At the bottom, the number of terms in each category and its relative percentage within the parenthesis.

	Ethnicity	Class	Disability	Nationality	Religion	Gender	Sexual Orientation
bamboo coon	×			×			
Gray bint						×	
bitter clinger		×					
Gray boojie	×	×					
camel fu**er	×			×	×		
Gray chinaman				×			
cu*t						×	
Gray d*ke						×	×
fa**ot							×
Gray house ni**er	×						
limey				×			
Gray moon cricket	×						
muzzie					×		
Gray ni**er	×						
plastic paddy	×			×			
Gray raghead	×						
redneck		×					
Gray retard			×				
retarded			×				
Gray rube		×					
sideways pu**y	×			×			
Gray soup taker				×	×		
spic	×						
Gray surrender monkey				×			
trailer park trash	×	×					
Gray trailer trash	×	×					
t*at						×	
Gray wetback	×						
whi**er	×			×			
Gray white ni**er	×			×			
white trash	×	×					
Gray wi**er	×			×			
yo**o		×					
Gray zionazi	×			×	×		
Number of terms (%)	18 (52.9)	8 (23.5)	2 (5.9)	12 (35.3)	4 (11.8)	4 (11.8)	2 (5.9)

analyses provided in this work a strong conservative bias, we believe the direction of the estimations performed by the framework would remain consistent.

Hate speech assessment. Finally, our Twitter dataset is shaped by the limitations of getting a sample from all Twitter with the Streaming API. Moreover, hate speech classification is inherently difficult. There is no universal definition for it and many important variables, such as context, are not easily measured with common hate detection approaches. Our methodology to detect hate relies, as a first step, on an external API which uses Wikipedia data as training set which might also lead to inaccurate toxicity scores for social media posts. Furthermore, [39] have shown that subtle changes on highly toxic sentences may assign significantly lower scores to them, which may indicate that many posts could not be classified as hateful in our work. Also, the limitation of number

Table 3.2: Total number of posts for Gab and Twitter, as well as the total number of labeled hate posts for each social media.

	Gab	Twitter
Total number of posts	7,794,990	9,118,006
Number of hate posts	9,554	2,392

of characters for tweets might impact on the way users write on this social network, therefore our methodology may not be able to get all the forms of hate in this social network. In spite of that, we have shown that our approach builds on previous work to accurately identify many forms of hate.

3.4 Summary

In this Chapter, we describe the data collection process, datasets, and main methodologies we followed to perform the analyses further described in this work. We perform a BFS on the graph of followers and friends from initial seeds and crawled information and posts of over 170,000 Gab users. Throughout this work, different sets of data are used for specific analysis, in particular, we define the news dataset, hate evaluation datasets, common users dataset, and hate datasets. Table 3.2 summarizes our main datasets and the respective number of hate posts found in each of them.

We also describe the main methodologies and frameworks that we use in this work. Many of the analyses performed in this work use frameworks and techniques already explored by other works in the literature. We describe the methodologies for gathering political leaning and demographic characteristics of users, measuring psycholinguistic features, sentiment, and toxicity of posts, and detecting hate in social media. In the next Chapters, we present and discuss the results of the use of these methods.

Chapter 4

RQ1 - Who are the Gab users?

In this Chapter, we provide a series of analyses aiming at depicting who are the users that joined Gab and what are their characteristics. We start by investigating its complete network structure and analyzing some network metrics.

4.1 Network Structure

The network structure can be investigated by characterizing its graph representation. We represented Gab as a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. G is a directed and unweighted graph, where the nodes represent Gab users and the edges indicate that users have a relation of following or being followed by other users.

The complete Gab network has a total of 171,920 users (nodes) and 11,162,492 connections (edges). Figure 4.1 shows the cumulative distribution function (CDF) for the outdegree (number of friends) and indegree (followers) values of nodes. Notice that the x axis is in logarithmic scale, therefore, the 10,153 (5.9%) users who have 0 followers and the 47,926 (27.9%) ones who have 0 friends were not considered. We notice that an extremely high number of users, nearly 90% of the total, have less than 100 followers/friends, whereas only 1% of users have over 1,000 followers. The indegree and outdegree distribution of the network follows a power-law like curve, i.e., there are few users with high indegree or outdegree, but a large number of users with low degree values. This is a common feature of many complex networks and other social systems.

Next, we indicate the most popular users of the network. There are many ways of measuring the popularity of users which are not related to the network's topological features, for instance applying PageRank [57] or analyzing the number of users' publications. For Twitter, users may also be ranked by the number of retweets for a certain tweet. As the average number of publications for all users is low (55.69), we use the easiest way to estimate the popularity of users in social media, which is investigating their indegree

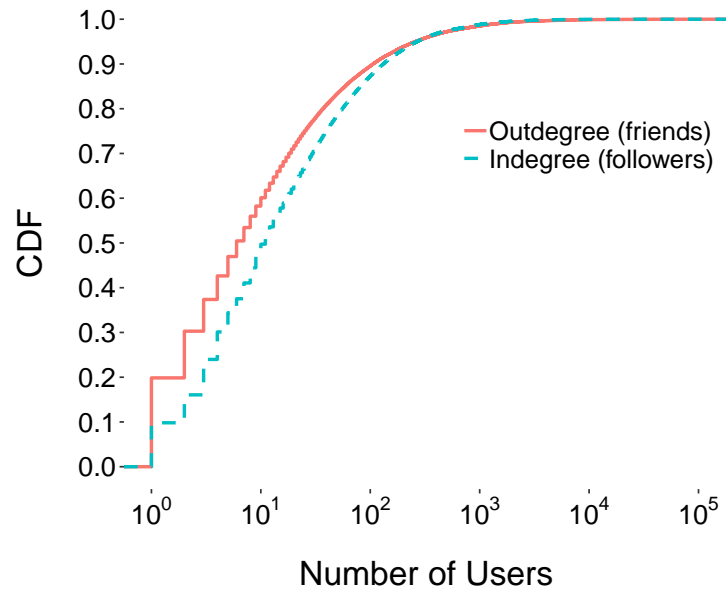


Figure 4.1: Cumulative distribution function (CDF) for the number of followers and friends of Gab users.

Table 4.1: Top 10 most followed users as observed in August 2017.

Screen name	Profile name	Number of followers
PrisonPlanet	PrisonPlanet	40,821
m	Milo Yiannopoulos	39,891
a	Andrew Torba	32,780
Ricky_Vaughn99	Ricky Vaughn	28,826
Cernovich	Mike Cernovich	27,462
stefanmolyneux	Stefan Molyneux	23,629
BrittPettibone	Brittany Pettibone	23,335
DeadNotSleeping	Jebs	22,202
RightSmarts	J. Allen - Right Smarts	18,615
voxday	Vox Day	18,551

values, i.e., the number of followers.

Table 4.1 shows the top 10 users with the highest number of followers as observed in August 2017. Out of these 10 users, only two have not had their profiles verified (*DeadNotSleeping* and *RightSmarts*). Despite the high number of followers (mean equals to 27,611.2 and median equals to 25,545.5), these users do not follow many other users back (the median number of following of the top 10 users is only 183.5, one order of magnitude lower than the number of followers). Also, differently from other social networks, such as Twitter and Facebook, in which popular users often include either celebrities or mass media, through manual inspection, we notice that the top 10 popular users of Gab do not directly fit in any of these categories. They are supporters of free speech, American nationalists, political activists, authors, writers, and editors.

Finally, we decompose the network into groups (sets of highly connected nodes). Most users (99.2%) are part of the largest connected component (LCC) (the largest subgraph in which any pair of nodes is connected by paths) corresponds to of the entire network. The rest of the users (0.8%) are singletons or part of smaller components. To compare, one of the pioneer studies on Twitter found that the largest connected component of the network contained about 94.8% of users of their dataset [13]. These results show that Gab is a social network where users tend to be as connected as the ones in traditional medias like Twitter. In the next subsections, we dive deeper on the analysis of these Gab users to understand their characteristics and peculiarities.

4.2 Demographic Factors of Users

The profile of users can be shaped by looking at some demographic factors, such as race, gender, and political leaning. In this section, we provide a series of analyses aiming at depicting the characteristics of Gab users in terms of demographic aspects. We start by investigating the location of Gab users.

4.2.1 Location

The Gab API does not provide any information regarding users' location. As an effort to mitigate that, we analyze user activity on the social network. Since the release of the service in August 2016, Gab has experienced a constant and slow increase in its number of users, except for a large peak around November 2016, during the US presidential election period.

Figure 4.2 shows the percentage of new users who joined Gab per month. From all users we crawled, 51.4% joined Gab from October to December, in 2016, 28.7% of them just in November 2016. This suggests that the US presidential election represented an important motivation for users to join Gab. Because of that, we conjecture that most Gab users are from the United States. Furthermore, Gab has become popular among users in the occurrence of political events such as the 2017 Charlottesville Unite the Right rally [85]. This is another evidence that users might be predominantly from USA. Next, we investigate gender and race of users.

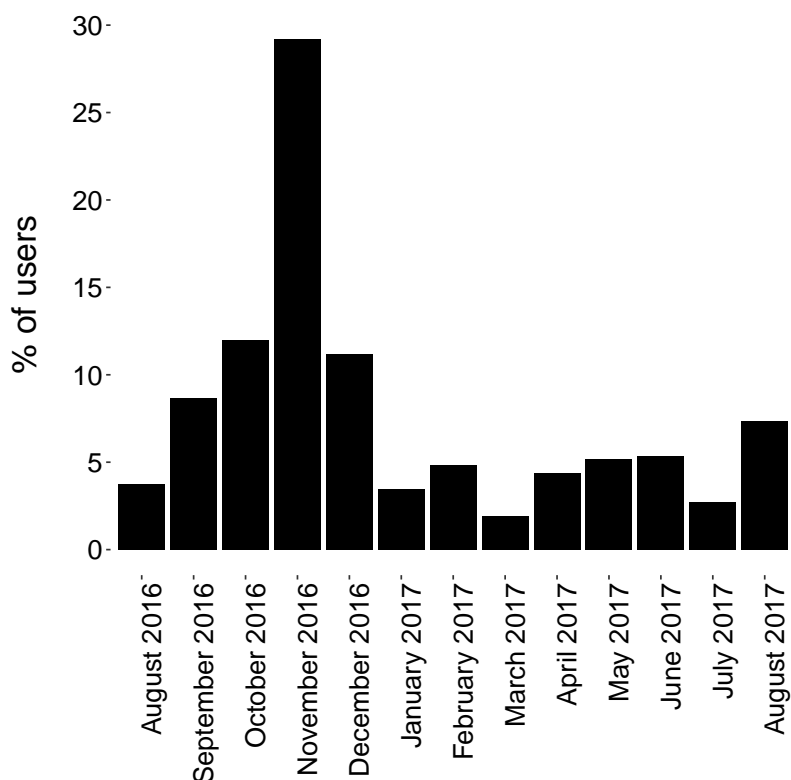


Figure 4.2: Percentage of users who joined Gab per month since its creation.

Table 4.2: Demographic distribution of nearly 36 thousand Gab users.

2*Race	Gender		2*Total
	Male	Female	
Asian	3,676 (10.4%)	1,920 (5.4%)	5,596 (15.8%)
Black	2,106 (5.9%)	787 (2.2%)	2,893 (8.2%)
White	18,078 (50.9%)	8,926 (25.1%)	27,004 (76.1%)
Total	23,860 (67.2%)	11,633 (32.7%)	35,493 (100%)

4.2.2 Gender and Race

Table 4.2 reports the demographic distribution of the 35,493 users in our dataset that Face++ was able to infer race and gender values. We observe a prevalence of men (67.2%) in comparison to women (32.7%) and a high predominance of Whites (76.1%) in comparison to Asians (15.8%) and Blacks (8.2%). This means that if we pick users randomly in our dataset, we would expect demographic groups with these proportions. These proportions, in particular the ones regarding race, differ from Facebook, where White men correspond to 29.35% of the total population [76].

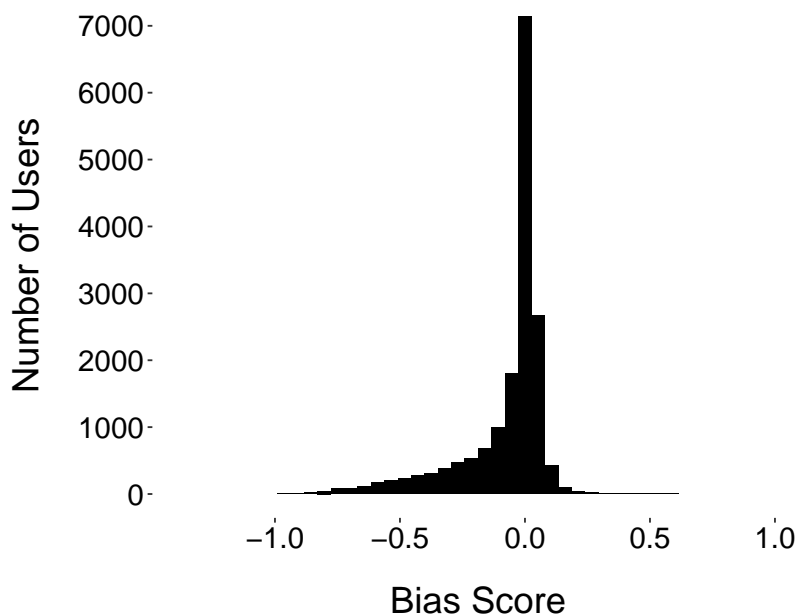


Figure 4.3: Distribution of bias scores for 16,804 Gab users.

4.2.3 Political Leaning

Next, we present an analysis on the political leaning of users. Figure 4.3 shows the distribution of bias scores for the 16,804 users who have accounts on Gab and Twitter and the framework was able to infer bias scores. The closer the bias score is to +1, the more liberal a user is. Users with bias score between -0.03 and +0.03 are considered neutral or moderate. We note a large number of extremely conservative users (i.e. those with scores close to -1) in Gab and most of the liberal ones have ideological leaning scores quite close to 0 (i.e. moderate). Out of the 16,804 users we sampled, 6,237 (37.1%) were inferred to be conservative, 2,925 (17.4%) as liberal, and 7,642 (45.5%) as moderate. Hence, the ratio between conservatives and liberals is about 2.13.

To contrast these values with those from other social networks, we use the Facebook audience API, following the same methodology of recent efforts [76]. Facebook provides this API for advertisers to estimate the number of users who are likely to match advertising criteria¹. We then used the Audience API to gather the political leaning of all Facebook users in the US. Overall, the inferred fraction of conservatives on Facebook is 33.6%, 42.6% of liberals and 23.8% of moderates. The ratio between conservatives and liberals is about 0.79. Therefore, in comparison with Facebook, Gab has a much more conservative than liberal crowd. For this reason, in the next section, we explore deeply these political concerns as an effort to identify potential extremist users in Gab.

¹developers.facebook.com/docs/marketing-api/audiences-api

4.3 Extremism in Gab

We want to investigate whether extremely conservative users are present in Gab. According to the SPLC², which is an American nonprofit legal advocacy organization specializing in civil rights dedicated to fighting hate, extremists in the United States come in different forms and follow a wide range of ideologies. As we look through the individual profiles of extremists provided by the SPLC [75] and the Anti-Defamation League (ADL) [4], we notice the presence of a few members of these movements in the platform. Table 4.3 presents the list of 29 listed extremist users by the two sources with Gab accounts and their number of Gab followers as observed in August 2017. In particular, most people listed by the ADL (61.11%) have a Gab account, reinforcing that this social network has been attracting a large number of extremely conservative users.

A key question that arises about these listed extremist users in Gab is whether they are widely followed and whether they are active users in the system. Among the top 10 most followed users, four are extremist users according to the SPLC and ADL, and all of them have verified accounts (that is a form of verifying identities). These considered extremist users, namely Milo Yiannopoulos (@m), Mike Cernovich (@Cernovich), Brittany Pettibone (@BrittPettibone) and Vox Day (@voxday), have on average 27,309.75 Gab followers. The average number of followers over the list of 29 users is 6,160.2, two orders of magnitude higher than the average number of followers for all Gab users we crawled (72.23). This means that posts of listed extremist users can reach a large number of different users within one hop. Approximately 35% of Gab users we crawled follow at least one of these 29 extremist users.

In regards to the number of posts, we also note that these users share in average more posts than the average Gab user. The average number of Gab posts overall users is 55.69, whereas the average for listed extremist users is 571.66. This indicates that these users are not only more followed, but they are also more active in the system.

4.4 News Spreaders

Finally, we analyze the profile of news spreaders in our dataset, i.e. people who share news sources within the news category in Gab posts. It is interesting to note that the majority of the active users within the news category (60.14%) have posted at least

²Southern Poverty Law Center <https://www.splcenter.org/>

Table 4.3: List of profiles considered extremist by the SPLC and ADL who were found in Gab.

Screen name	Username	# of followers
RealAlexJones	Alex Jones**	14,962
Alex_Linder	Alex Linder	963
AndrewAnglin	Andrew Anglin**	4,853
AndyNowicki	Andy Nowicki	75
thewizardofthorntonpark	Augustus Invictus	128
BillyRoper	Billy Roper	407
occdissent	Brad Griffin	534
BrittPettibone	Brittany Pettibone*	23,335
Cantwell	Christopher Cantwell***	2,728
DanielFriberg	Daniel Friberg	299
RealDavidDuke	David Duke****	982
GavinMcInnes	Gavin McInnes	2,681
Posobiec	Jack Posobiec*	2,944
jartaylor	Jared Taylor*	147
TheMadDimension	Jason Kessler*	454
mattforney	Matt Forney**	4,322
matthewheimbach	Matthew Heimbach	43
mattparrott	Matthew Parrott	56
Cernovich	Mike Cernovich*	27,462
mikeenoch	Mike Peinovich	703
m	Milo Yiannopoulos*	39,891
Pamela	Pamela Geller*	3,238
ramzpaul	Paul Ray Ramsey	2,951
pax	Pax Dickinson	12,218
Richardbspencer	Richard B Spencer	6,833
RobertSpencer	Robert Spencer	1,317
TaraMcCarthy	Tara McCarthy***	5,565
voxday	Theodore Beale*	18,551
Microchimp	Tim Gionet	4

* verified; ** verified and PRO; *** verified, PRO and premium; **** PRO

one URL and, from these, 62.71% have posted more than one URL.

Table 4.4 shows the top 10 news spreaders and their total number of posts as of August 2017. These users have shared on average 10,838.5 posts. The user *Constitutional Drunk*, associated with the right-biased USSA News website, had the largest number of posts, 59,378. In fact, most publications from this user share the content of that website, which happens to be the most shared domain. Only 37.9% of the users listed as extremists have posts categorized as news.

Aiming at understanding the influence of these users in Gab, we also analyze their number of followers. While the average number of followers among all crawled users is low (72.23), the top 10 news spreaders have on average 4,128.8 followers, two orders

Table 4.4: Top 10 news spreaders.

Screen name	Username	# of posts
USSANews	Constitutional Drunk	59,378
Zlatford	Zak	13,388
wrath0fkhan	wrath 0fkhan	8,189
histanvan	Harry2	6,459
Kek_Magician	Kek_Magician	4,248
Lakeem	Lakeem Khodra***	4,016
Arwen777	Dani	3,913
weeklyflyer	Jerry	3,213
rabite	Stankpipe	2,925
OpenQuotes	OpenQuotes	2,656

*** *verified, PRO and premium.*

of magnitude higher than the overall mean. In total, 20,176 out of 171,920 Gab users follow at least one of these news spreaders, which implies that posts from the top 10 news spreaders can reach 11.74% of the users of the social network. These results also show that news spreaders tend to be, overall, more followed and more active in the Gab social system.

4.5 Concluding Remarks

This Chapter answers the first research question of this work, providing an understanding of who the Gab users are. We start by characterizing users' engagement, analyzing common social media metrics, such as friends and followers. We notice that most Gab users joined the network around the occurrence of political events, mainly the presidential election in the US (November 2016).

Our analyses on political leaning and demographics showed that the majority of Gab users are conservative, male, and Caucasian. We were able to identify many known banned users from other social networks for cases of hate speech and association with extremism, who showed to be influential and very active in the Gab network. Finally, we show that news spreaders have on average more followers and posts than all Gab users, reaching a large number of users of the network within just one hop. The top news spreader is associated with a right-biased website.

As we notice that Gab is a very politically oriented system, in the next Chapter we investigate the content shared in this network. We expect to understand the behavior of such users in a system without the rigorous moderation of policies from other media

(i.e. Facebook and Twitter), providing a meticulous overview of Gab posts.

Chapter 5

RQ2 - What do users share on Gab?

In this Chapter, we analyze the content shared on Gab. We start by characterizing typical language usage, detailing popular words, topics, and languages. Then, we characterize the news shared in the Gab social system.

5.1 Popular Words, Topics, and Languages

Figure 5.1 illustrates through a word cloud¹ frequent terms of Gab posts, and Table 5.1 shows frequent bigrams and trigrams in Gab posts, i.e. frequent two or three adjacent elements from strings of tokens. Overall, the word cloud shows a strong political emphasis in Gab posts, with the existence of several terms that have been used by conservative political campaigns and their supporters. In particular, the hashtags *MAGA*, which stands for *Make America Great Again*, and *DRAINTHESWAMP* were popularly mentioned as part of the Donald Trump’s 2016 presidential campaign. Notice that frequent bigrams and trigrams usually contain hashtags which associates the free speech with conservative terms. Other frequent hashtags include *TRUMP*, *GABFAM*, *SPEAKFREELY*, *NEWS*, *POLITICS*. This is another evidence that many conservative users joined Gab as an alternative to traditional media with strong moderation policies.

The analysis of the posts categories reinforces the strong political tendency of the Gab social network. Table 5.2 presents the percentage of posts for the top 5 most popular categories. The category with the highest number of posts is News, usually associated with politics, followed by the categories Politics, Humor, AMA (Ask Me Anything) and Entertainment. The other categories, i.e. Music, Technology, Art, Sports, Faith, Philosophy, Photography, Science, Finance and Cuisine, sum up to 225,611 (17.39%) posts. As news and politics are the most popular categories in gab, our next section dives into the analyses of news sharing on Gab.

We characterize the presence of other languages in Gab by running a popular

¹<http://www.wordle.net/>

Table 5.3: Top 10 languages with more posts.

Language	% of posts
English	72.68
German	5.44
Hungarian	0.80
French	0.38
Indonesian	0.31
Afrikaans	0.28
Dutch	0.28
Danish	0.24
Italian	0.24
Catalan	0.23

increased in the occasion of the Brazilian election period in 2018 and that Portuguese has become the second most frequent language in the social platform [70].

5.2 News sharing

This section analyzes the 463,663 posts shared in the news category in Gab. Inferring bias of news sources often relies on strategies that combine analyses of the audience of news outlets and inspection of the published content. Recent research works such as [12], [53], [8], and AllSides³ infer the political leaning of over 600 news outlets using these techniques. Next, we briefly discuss each of these studies and, based on them, we analyze the presence of biased news sources on Gab.

Following a survey-based approach, [53] classify the audience of popular news media outlets based on a ten-question survey covering a range of issues like homosexuality, immigration, economic policy, and the role of government, classifying the political leaning of the audience in five categories: consistently liberal, mostly liberal, mixed, mostly conservative, and consistently conservative. On the other hand, [8] follow a news-based approach in which the authors derive the alignment score of media outlets by first identifying the political leaning of over 10 million Facebook users based on self-declarations and then considering how users with different political leanings shared the stories published by these outlets.

Other studies include approaches based on (a) content and (b) crowdsourcing. The former refers to the research of [12] where the authors use a content-based approach to identify the slant of the top 13 U.S. news outlets and two popular political blogs.

³<https://allsides.com/media-bias/media-bias-ratings>

Table 5.4: Number of domains found in Gab posts which are categorized as news and coexistence in each dataset.

Dataset	Republican (%)	Democrat (%)	Neutral (%)
[12]	4 of 4 (100.00)	11 of 11 (100.00)	-
[53]	6 of 7 (85.71)	19 of 23 (82.61)	1 of 3 (33.33)
[8]	165 of 205 (80.49)	203 of 260 (78.08)	-
AllSides	29 of 40 (72.50)	26 of 40 (65.00)	24 of 32 (75.00)

Regarding the latter, the website `AllSides.com` infers bias by encouraging its users to rate different news outlets in one of the five categories: left, lean left, center, lean right, and right.

Table 5.4 presents, for each of the aforementioned datasets, the number of news outlets inferred as Republican, Democrat, and Neutral, which were also shared on Gab posts categorized as news. We notice that regardless of the bias of news outlets, most domains are in fact found in Gab posts (e.g., 100% of the news outlets with political leaning inferred by [12] have been shared in Gab). However, we show next that news sources inferred as Republican are shared on average more often than those inferred as Democrat and Neutral.

Figure 5.2 shows box plots of the number of times news sources inferred as Republican, Democrat, and Neutral have been shared on Gab posts categorized as news. Clearly, news outlets biased towards conservative audience are consistently shared on average more than the others across all four methodologies. In particular, considering only news outlets found in the study of [12] (Figure 5.2a), Gab posts comprise on average 5,894 right-leaning news sources (median 2,116) and only 645.6 leftist news sources (median 538).

5.2.1 Most Shared Domains

We present an analysis on the most shared domains in Gab. First, we extract the domains of all links we collected from 463,663 Gab posts categorized as news. Table 5.5 shows the proportion of the top 30 most frequently shared domains in this social media⁴. We observe that `ussanews.com` is the most frequently shared news domain, accounting for more than 16% of the news URLs shared on Gab, followed by `youtube.com`, which accounts for more than 15%. Although YouTube is not specifically designed for news,

⁴We exclude from this ranking domains not belonging to news outlets (i.e. `twitter.com`, `i.imgur.com` and `pbs.twimg.com`)

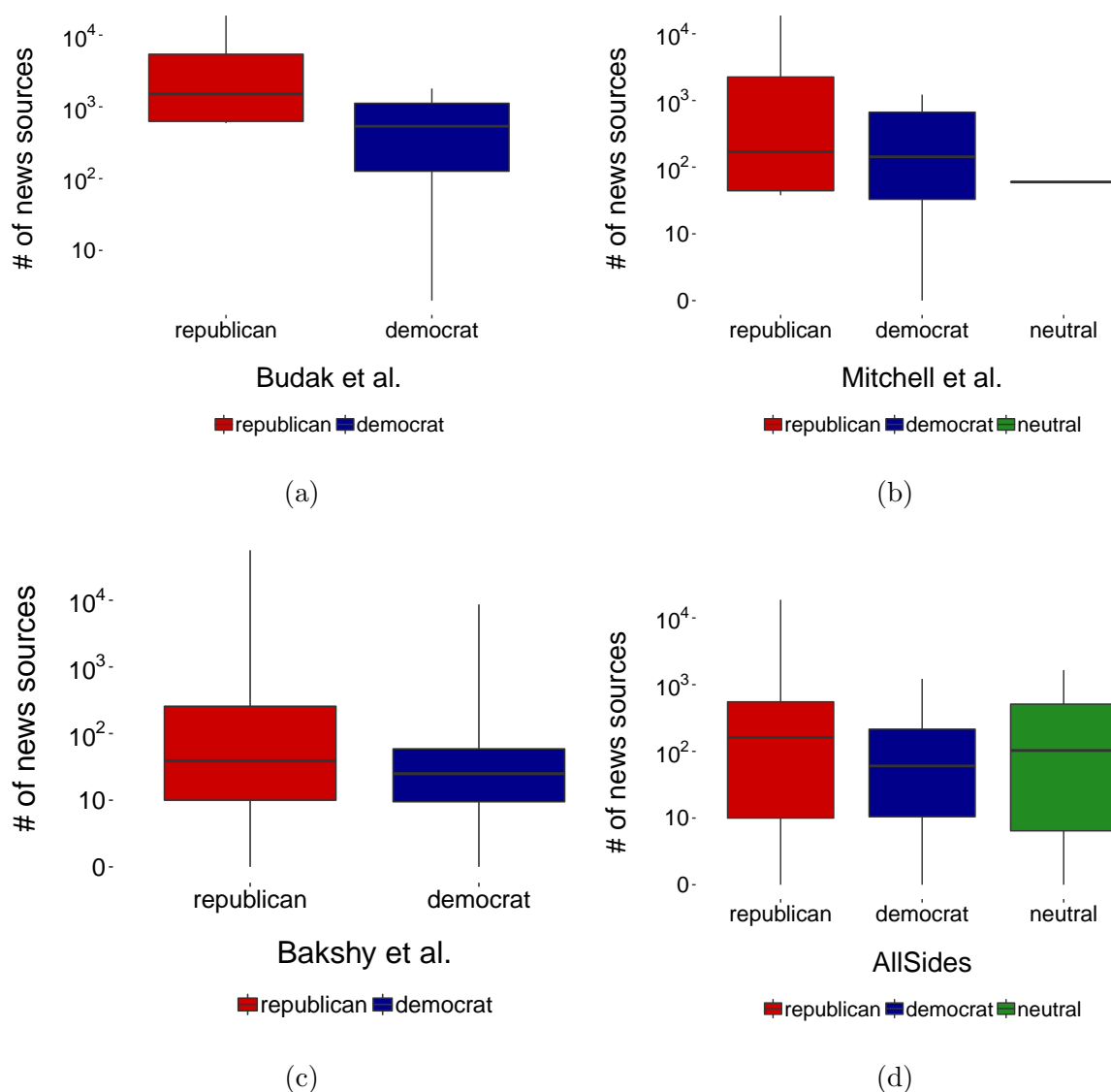


Figure 5.2: Number of times news sources were shared in Gab posts categorized as news, grouped by political leaning (Republican, Democrat, and Neutral) as inferred by [12] (a), [53] (b), [8] (c), and AllSides (d). Each box plot shows minimum, 25-percentile, median, 75-percentile and maximum.

this social media is widely used for publishing news videos [59, 36]. Moreover, we note a substantial presence of alternative news outlets like breitbart.com, zerohedge.com, infowars.com and thegatewaypundit.com, which correspond to more than 5%, 2%, 1.9% and 1.5%, respectively. Finally, the newspaper dailymail.co.uk (1.5%) figures among the most shared domains, followed by rt.com (1.4%), dailycaller.com (1.2%) and fightagainst-tyranny.com (1.2%).

Some news domains widely shared in Gab are not very popular according to Alexa.com. For example, the most shared domain in Gab (ussanews.com) appears in a very low position in the Alexa rank (409,260). The same is true for others domains such as fightagainst-tyranny.com, truthfeed.com, behoerdenstress.de, worldwideweirdnews.com.

Table 5.5: Top 30 news sources in posts and their respective domain, percentage over all posts.

Source	Domain	(%)
USSA News	ussanews.com	16.40
Youtube	youtube.com	15.99
Breitbart	breitbart.com	5.17
Zero Hedge	zerohedge.com	2.08
Infowars	infowars.com	1.91
The Gateway Pundit	thegatewaypundit.com	1.59
Daily Mail Online	dailymail.co.uk	1.52
RT News	rt.com	1.48
The Daily Caller	dailycaller.com	1.25
Tyranny News	fightagainst-tyranny.com	1.21
EXPRESS	express.co.uk	1.20
Fox News	foxnews.com	0.99
Truth Feed News	truthfeed.com	0.63
Sputnik News	sputniknews.com	0.63
ABC News	abcnews.go.com	0.57
Reuters	reuters.com	0.50
Washington Examiner	washingtonexaminer.com	0.49
Behoerdenstress Whistleblower	behoerdenstress.de	0.46
The Hill	thehill.com	0.46
Worldwide Weird News	worldwideweirdnews.com	0.45
The Daily Informer	thedailyinformer.net	0.44
WND	wnd.com	0.44
New York Post	nypost.com	0.41
The Washington Post	washingtonpost.com	0.34
The Last Refuge	theconservativetreehouse.com	0.34
The Telegraph	telegraph.co.uk	0.33
BBC News	bbc.com	0.33
The Washington Free Beacon	freebeacon.com	0.32
INDEPENDENT	independent.co.uk	0.30
The Sun	thesun.co.uk	0.30

com and `thedailyinformer.net`.

Overall, these results show that there is a large diversity of news domains shared in Gab. While most domains are shared very few times, a few domains comprise most of the shared URLs related to news. Interestingly, there is *no intersection* between the top 30 most shared domains in Gab and the top 10 most shared news domains in traditional media like Facebook⁵. On Twitter, previous works showed that news URLs from online newspapers like “The New York Times” are shared very often [65]. However, this domain does not appear frequently in Gab shares, highlighting important differences between these social networks.

⁵<https://www.statista.com/statistics/265830/facebook-daily-newspapers-top-ten/>

Table 5.6: Top 15 most popular links shared in Gab, their number of shares in Gab, and popularity according to Bit.ly (larger means more popular).

URL News Link	Domain	Shares	Bit.ly
http://bit.ly/2qNboDV	petitions.whitehouse.gov	99	949
https://youtu.be/4Emd7urcWbc	youtube.com	60	-
https://wikileaks.org/ciav7p1/	wikileaks.org	44	4,025
https://youtu.be/MHZSfhd1X_8	youtube.com	39	-
http://bit.ly/2EJ8WpZ	express.co.uk	35	-
http://washex.am/2ivMBhv	washingtonexaminer.com	35	772
http://dailym.ai/2kubL0X	dailymail.co.uk	32	0
http://bit.ly/2nZdrD7	infowars.com	27	-
http://dailym.ai/2qSZ5mi	dailymail.co.uk	25	-
http://bit.ly/2nB285x	infowars.com	24	4,492
http://bit.ly/2C0QdW1	infowars.com	24	1
http://bit.ly/2Es9CAT	ussanews.com	24	-
http://bit.ly/2kSgBZC	ktla.com	24	7,356
http://bit.ly/2iH0mLy	thesun.co.uk	24	274
http://bit.ly/2nZ1X1K	ussanews.com	24	-

5.2.2 Top Stories

Next, we shift our focus to understand what are the top stories shared in the category news of the platform. Table 5.6 shows the 15 most popular links shared in Gab, i.e. the most frequently posted links. We also show their popularity according to Bit.ly⁶, which we use as a proxy for the number of clicks made to each of these links. Bit.ly is a well-known URL shortening service that shortens millions of URLs daily. The service API provides the possibility of checking the total number of clicks that a shortened link has received [5]. In general, these links presented in Table 5.6 have been posted on average 36 times, where the first link in the rank appears in 99 posts and the last one appears in 24 posts. Some popular links shared in Gab have Bit.ly popularity higher than links to popular news sources such as BBC News, DailyMail, New York Times, and Reuters [64].

The most popular link is a petition created on May 19, 2017, requesting action of the Congress for the following issue: “Appoint a Special Prosecutor to investigate the murder of Seth Rich, the alleged Wikileaks email leaker”. Online petitions are often cheaper ways of demanding acts to address perceived problems by some groups of people [25]. Given that the most popular link shared in Gab within the news category is a petition, we conjecture that Gab is not only a place for political discussion but also an environment prone to online activism.

Other popular links include YouTube videos (the second most popular link talks

⁶<http://bit.ly/>

about the alternative media), WikiLeaks news, Donald Trump-related news, and also news which might not be completely accurate. For instance, the sixth most popular link has been rated by the fact-checking website Snopes as mostly false⁷. This suggests that Gab is prone to the dissemination of fake news, similarly to other social networks. However, the fake news that spread in Gab are usually those that reinforce right-leaning beliefs.

5.3 Concluding Remarks

In this Chapter, our analysis about what users share in Gab unveiled a variety of political statements. On the characterization of popular words and topics, we show that frequent terms present in Gab posts have strong political emphasis, with the presence of many hashtags and terms used by conservative politicians and their supporters. The analysis of the most frequent categories in Gab also highlights the strong political tendency.

Our analysis also shows the existence of great diversity in the news domains shared within Gab. Most of these domains are not popular on other social media or the Internet as a whole, and part of them are known for spreading news with very biased content, rumors, as well as fake news. We conjecture that Gab is prone to the dissemination of fake news that reinforces right-leaning beliefs.

These analyses lead us to our third research question, which dives deeper into the characterization of Gab posts and compares with textual data from moderated social media, represented by Twitter. As there is any increasing debate between free speech and regulation of content, we hope our efforts contribute to this discussion.

⁷<https://www.snopes.com/child-prostitution-legalized-in-california/>

Chapter 6

RQ3 - Distinguishing characteristics of moderated and unmoderated content

In this Chapter, we show the differences between moderated and unmoderated content shared in Twitter and Gab, respectively. We evaluate the distinguishing characteristics of such discourses on the analysis of (i) linguistic features, and (ii) sentiment and toxicity scores.

6.1 Linguistic Features

We analyze linguistic differences of moderated and unmoderated content by computing the distributions values for each LIWC attribute in both sets of posts. We aggregate these attributes into four distinct dimensions as shown in Table 6.1, following [58], who made this arrangement available¹. The table shows some examples of words contained in each dimension and the number of words per dimension in the LIWC dictionary, grouped into four categories. One word or token may be in more than one category. For example, *happy*, which is part of “Psychological Processes”, “Positive Emotions” and “Affective Processes”. The numbers of unique tokens for the Standard Linguistic Dimensions, Personal Concerns, Spoken Categories, and Psychological Processes are respectively 1, 233, 1, 410, 68 and 4, 562.

We start by investigating the volume of posts from both social media containing words in each LIWC dimension, as shown in Figure 6.1. More than 80% of Gab and Twitter posts contain at least one term of either the Standard Linguistic Dimensions or the Psychological Processes dimensions. Nearly 50% of posts from both social media contain words of Personal Concerns. Interestingly, for the Spoken Categories, 43.9% of

¹http://lit.eecs.umich.edu/~geoliwc/LIWC_Dictionary.htm

Table 6.1: LIWC dimensions, subdimensions and attributes used in the present study.

(a) Standard Linguistic Dimensions			(d) Psychological Processes		
Dimension	Example	# Words	Dimension	Example	# Words
Pronouns	I, them, itself	153	Social Processes	talk, us, friend	756
Articles	a, an, the	3	Friends	pal, buddy, coworker	95
Past tense	walked, were, had	341	Family	mom, brother, cousin	118
Present tense	is, does, hear	428	Affective Processes	happy, ugly, bitter	1413
Future tense	will, gonna	97	Positive Emotions	happy, pretty, good	640
Prepositions	with, above	74	Negative Emotions	hate, worthless, enemy	744
Negations	no, never, not	62	Anxiety	nervous, afraid, tense	116
Numbers	one, thirty, million	36	Anger	hate, kill, pissed	230
Swear words	af, a**hole	131	Sadness	grief, cry, sad	136
(b) Personal Concerns			Cognitive Processes	cause, know, ought	797
			Insight	think, know, consider	259
			Causation	because, effect, hence	135
Dimension	Example	# Words	Discrepancy	should, would, could	83
Work	work, class, boss	444	Tentative	maybe, perhaps, guess	178
Achievement	try, goal, win	213	Certainty	always, never	113
Leisure	house, TV, music	295	Perceptual Processes	see, touch, listen	436
Home	house, kitchen, lawn	100	Seeing	view, saw, look	126
Money	audit, cash, owe	226	Hearing	heard, listen, sound	93
Religion	altar, church, mosque	174	Feeling	touch, hold, felt	128
Death	bury, coffin, kill	74	Biological Processes	eat, blood, pain	748
(c) Spoken Categories			Body	ache, heart, cough	215
			Sexuality	horny, love, f*ck	131
Dimension	Example	# Words	Relativity	area, bend, exit, stop	974
Assent	agree, OK, yes	36	Motion	walk, move, go	325
Nonfluencies	uh, ri*	19	Space	Down, in, thin	360
Fillers	blah, you know, I mean	14	Time	hour, day, o'clock	310

Twitter posts contain at least one word of this dimension, whereas only 9.1% of Gab posts contain at least one of the referred words. We believe this difference might be due to the characteristics of the audience and posts of the Gab network, which has a strong political bias where users tend to share a larger number of news and politics related posts, as observed in the previous Chapter, whereas Twitter has traits of behavior that may enhance informal communication [86].

Next, we compare the distributions of each LIWC attribute for both Gab and Twitter by running the Kolmogorov-Smirnov (KS) test [49], which is a non-parametric test of equality of continuous distributions, in which the null hypothesis posits that the two input samples have the same distribution. We find a significant statistical difference ($p - value < 0.05$) for all the distributions, indicating that moderated and unmoderated posts have different psycholinguistic features. To get some insight on how different these distributions are, we measure the similarity of each LIWC dimension between the two systems according to the following steps.

First, we calculate the mean values of each LIWC attribute for posts from Gab and Twitter. Then, we calculate the Euclidean distance according to the formula $d(p, q) = \sqrt{\sum_i^n (p_i - q_i)^2}$, where p and q are the mean values for each social network for the attribute

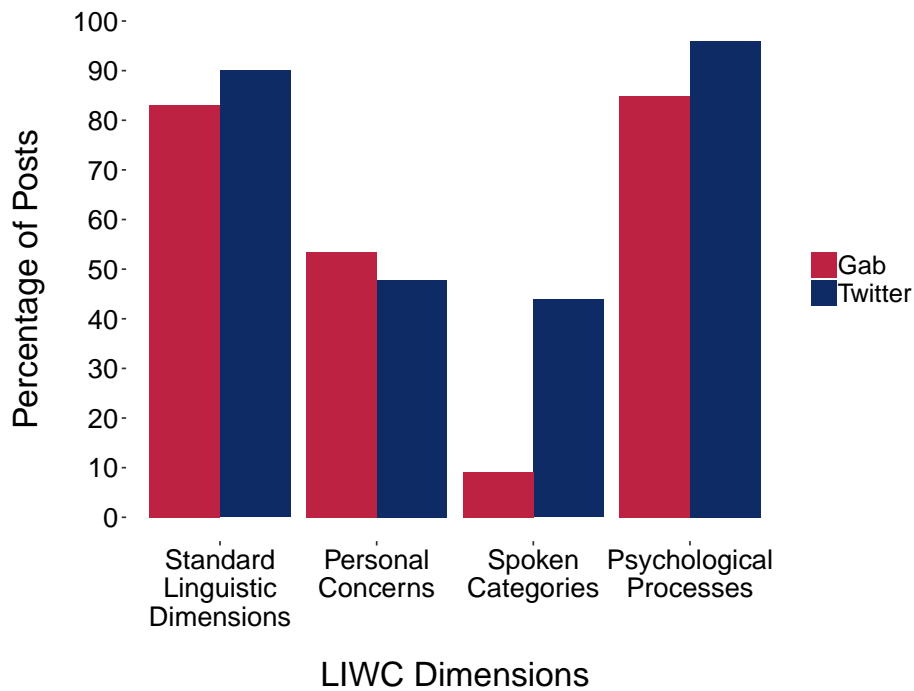


Figure 6.1: Percentage of Gab and Twitter posts which contain at least one word or token per LIWC dimension.

i of the LIWC dimension containing n other attributes. Then, we calculate the similarity between the social networks for each LIWC dimension according to the formula $\frac{1}{1+d(p,q)}$. Similarity scores closer to 1 indicate that attributes of a particular LIWC dimension from both Gab and Twitter have equal mean values.

The similarity scores for Gab and Twitter LIWC dimensions are 0.022, 0.019, 0.046, and 0.005, for the Standard Linguistic Dimensions, Personal Concerns, Spoken Categories, and Psychological Processes, respectively. We notice a very low similarity between Gab and Twitter content for all dimensions, indicating that these contents do not share much linguistic similarity. We dive deeper into the dimension which has a lower similarity between the content of Gab and Twitter, namely Psychological Processes, to understand which sub dimensions and attributes have a higher mean difference.

We apply the same aforementioned methodology to each psychological processes dimension (i.e. social, affective, cognitive, perceptual, biological processes and relativity) for measuring similarity, and we find that biological processes have the lowest similarity score among all the others. Posts containing sexuality related terms on Gab have on average 6.95% sexual terms, whereas the Twitter counterparts have about 13.19%, i.e., almost twice as much. We analyze both Gab and Twitter posts containing sexual terms to understand this high discrepancy, and we observe that Twitter has a very large number of explicitly pornographic posts, which increases mean and median values of posts for the sexuality attribute in Twitter. These findings highlight the differences between Gab and Twitter in terms of linguistic characteristics.

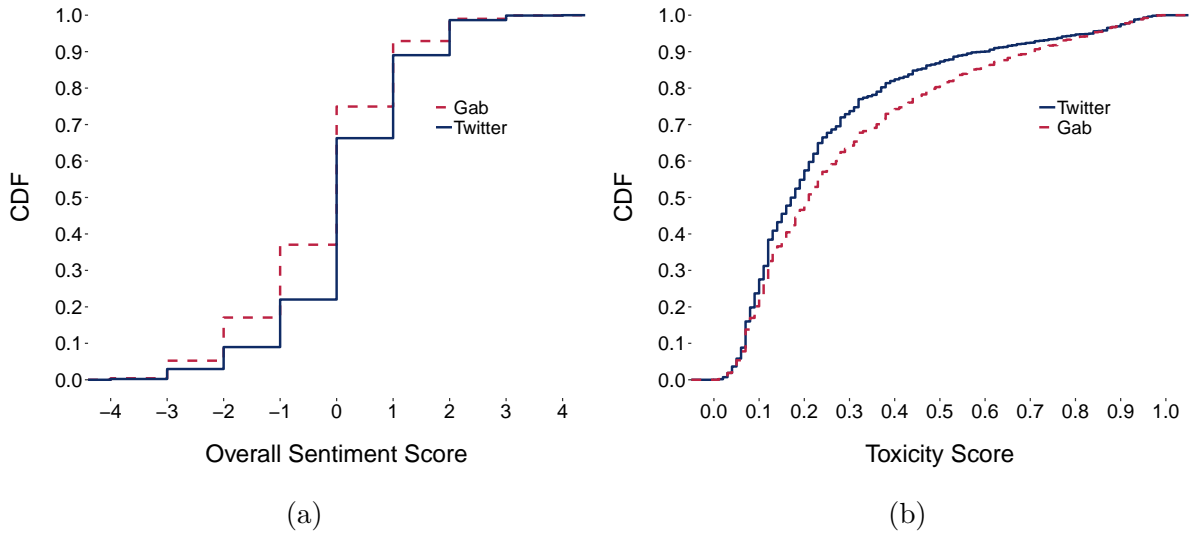


Figure 6.2: Cumulative Distribution Function (CDF) for (a) Overall Sentiment Score and (b) Toxicity Score.

6.2 Sentiment Analysis and Toxicity

Next, we analyze the differences of sentiment and toxicity for moderated and unmoderated content. The Perspective API is still not able to infer the toxicity of posts which are not in English and thus, we were unable to retrieve this score for many posts. In practice, 13.72% of Gab posts and 3.44% tweets do not have a toxicity score. As we showed, English is the most popular language in Gab, which explains the larger number of posts that do not have toxicity information. In the sentiment analysis, posts which are either (i) not in English or (ii) that do not have any text that impacts the polarity of the post, the SentiStrength framework returns neutral ($P = 1$ and $N = 1$) values.

Figure 6.2 shows the Cumulative Distribution Function (CDF) for the (a) overall sentiment score, calculated as the difference between the positive and the absolute value of negative scores given by SentiStrength ($P - |N|$), and (b) toxicity score for both Gab and Twitter. First, we compare these distributions using the KS test. For each metric, we find a significant statistical difference ($p - value < 0.05$) between the distributions, i.e., posts from moderated and unmoderated social media are statistically different in terms of sentiment and toxicity scores.

We notice that unmoderated publications tend to be more negative and to have higher toxicity than moderated ones. For the overall sentiment scores, Figure 6.2a shows that nearly 37% of unmoderated posts have negative overall sentiment, i.e., have an absolute negative score greater than a positive score, whereas only 22% of moderated posts have negative overall sentiment. Furthermore, Figure 6.2b shows that 19.4% of Gab posts have toxicity scores higher than 0.5, whereas this percentage is 13% for Twitter posts,

indicating that unmoderated posts tend to be perceived as more toxic than moderated posts. We observe that the fraction of posts that have toxicity above 0.8 is higher in Gab than on Twitter, as there are 6.5% of such posts on Gab and 5.5% of tweets in all posts. These results suggest that Gab is slightly more toxic than Twitter. One possible explanation for this is the lack of moderation on Gab, allowing harmful speech in this network to fester unchecked.

6.3 Concluding Remarks

In this Chapter, we evaluate the differences between moderated and unmoderated content in terms of linguistic features, sentiment, and toxicity analyses. Our results show that unmoderated content has more negative sentiment overall and higher toxicity scores than moderated content. We show that the distribution of linguistic features for moderated and unmoderated content is also different and that Twitter exhibits proportionally more sexuality related terms than Gab in its posts. This is an interesting result as the appearance of the body and sexual attributes often appear in hateful speech with the use of words related to genitals and sexual appeal regularly used as offense (e.g., *cu*nt*, *d*ck*).

As we notice these differences between moderated and unmoderated posts, this leads us to our final research question. In the next Chapter, we approach the different types of hate found in each of these social networks.

Chapter 7

RQ4 - Different types of hate across moderated and unmoderated environments

This Chapter presents our efforts to measure the differences between hate displayed in a free speech social network, represented by Gab, and that in a moderated social media, represented by Twitter. We analyze 9,554 Gab posts and 2,392 tweets labeled as hate in this work.

7.1 Manual Validation

Before diving into the analysis, we first evaluate the quality of our hate set by manually annotating a random sample of Gab and Twitter posts. We take 100 random posts from Twitter/Gab that were marked as hate/non-hate by the framework described in Section 3.2.3. Because people have distinct opinions about the definition of hate speech, two of the authors of this work independently annotated these posts by hand as hate or non-hate according to their perspectives, but following the definitions of Hate Speech by Twitter¹ and Facebook². We hide the previously assigned labels from the annotators. Table 7.1 shows the results of our manual evaluation for the given samples. Each triple shows the number of posts agreed as hate, without agreement and agreed as non-hate, respectively. We notice low numbers of false positive and false negative considering the agreement of the annotators.

The annotators have agreed on the label of 383 posts, resulting in a Cohen's kappa coefficient (κ) equals 0.92. Assuming the labels created by the annotators of these 383 posts as the correct ones, we evaluate the quality of the predictions made by the methodol-

¹<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

²https://www.facebook.com/communitystandards/hate_speech

Table 7.1: Manual evaluation. Each triple shows the number of posts agreed as hate, without agreement and agreed as non-hate, respectively.

	Gab	Twitter
Labeled by framework as Hate	(93, 4, 3)	(90, 6, 4)
Labeled by framework as Non-Hate	(5, 4, 91)	(0, 3, 97)

ogy described in this work to detect hate. To assess the effectiveness of this methodology, we adopted metrics commonly used in Machine Learning and Information Retrieval [7]. The methodology to identify hate presented in this work was able to correctly classify 371 labels: accuracy 96.87%, precision 96.32%, and recall 97.34%. These results reinforce the quality of the methodology to identify hate posts on social media with minimal noise.

7.2 Types of Hate

We associate one hate post with different types of hate, according to the types shown in Table 3.1: *ethnicity*, *class*, *disability*, *nationality*, *religion*, *gender*, and *sexual orientation*. One hate post can be associated with up to seven types of hate (or none). For instance, for the following hateful post “*Ni**er retard,ni**er retard,ni**er retard,ni**er retard with aids,ni**ers carry the most aids and have the lowest IQ’s in the world. You Dumb worthless Ni**er.*”, the term *ni**er*, which is related to ethnicity, appears six times, whereas the disability term *retard* appears four times. Therefore, this post is associated with both ethnicity and disability types of hate. Notice that being associated with one type of hate does not necessarily imply that one post is being hateful towards that particular group, but rather that it makes use of terms associated with that type of hate to write a comment that is perceived as rude, disrespectful, or unreasonable by different groups.

7.2.1 Frequent Types of Hate

Considering our hate set, Figure 7.1 shows the percentage of Gab and Twitter posts associated with each type of hate. Our findings show that Gab and Twitter posts are predominantly associated with disability and gender types of hate. It is interesting

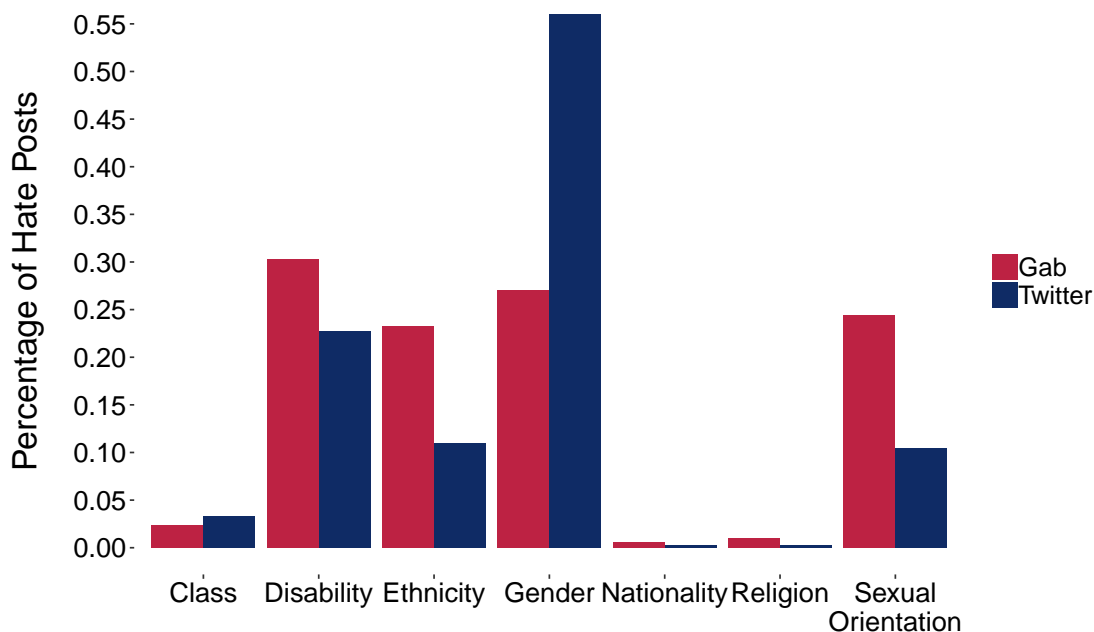


Figure 7.1: Percentage of Gab and Twitter posts which contain at least one word or token per LIWC dimension.

to notice that 56.06% of hate tweets are associated with gender, followed by 22.74% of hate tweets being related to disability, a difference of over 30%. For Gab, this difference is smaller, as disability is associated with 30.25% and is followed by gender, with 27.04%. Gab has also a large number of hate posts associated with sexual orientation and ethnicity (over 20% each). These results suggest that an environment which lacks moderation of content like Gab is more prone to the dissemination of hate speech of many types than moderated one, which still needs to improve their hate speech policies and methods in order to avoid hate speech towards specific groups of people.

7.2.2 Frequent Hate Terms

Table 7.2 shows the 10 most frequent hate terms in Gab and Twitter hate sets. The complete rankings have significant ($p - value < 0.05$) Kendall rank correlation coefficient (0.76). On this table, we observe that all gender related hate terms appear proportionally more on Twitter than Gab, which helps explaining the larger number of gender related hate posts observed in Twitter in comparison with Gab on Figure 7.1. The term c^{*nt} appears in more than 45% of Twitter hate posts whereas this number for Gab drops to near 22%. T^{*at} and d^{*ke} are the others gender related terms on the top 10 which appear

Table 7.2: Top 10 most frequent hate terms in Gab and Twitter hate posts.

Gab			Twitter		
term	category	%	term	category	%
retarded	disability	30.25	c*nt	gender	45.31
fa**ot	sexual orientation	23.96	retarded	disability	22.74
c*nt	gender	22.10	t*at	gender	9.65
ni**er	ethnicity	21.31	fa**ot	sexual orientation	9.53
t*at	gender	4.68	ni**er	ethnicity	8.52
redneck	class	0.99	white trash	ethnicity; class	1.88
muzzie	religion	0.96	redneck	class	1.17
white trash	ethnicity; class	0.84	d*ke	gender; sexual orientation	0.87
d*ke	gender; sexual orientation	0.49	bint	gender	0.41
spic	ethnicity	0.41	muzzie	religion	0.25

proportionality more on Twitter hate posts than Gab’s, corroborating our findings on the analysis of linguistic differences between these networks, which shows that Twitter has almost two times more sexual related terms than Gab.

Class related hate terms also appear more on Twitter than Gab. The other terms, *retarded*, *fa**ot*, *ni**er*, *muzzie*, and *spic* appear more in Gab posts than in tweets. Interestingly, even though there are 12 nationality related terms, none of them are in the top 10 most frequent terms for either social media. The most frequent nationality term on Gab is *wi**er*, which appears in 0.23% of hate posts, and for Twitter is *limey*, which is in 0.13% of hate tweets.

7.2.3 Multiple types of Hate

We investigate posts which are associated with more than one type of hate, i.e., hate posts containing hate terms associated with different categories. Out of the 9,554 and 2,392 Gab and Twitter hate posts, 775 (8.11%) and 94 (3.93%) are associated with more than one type of hate, respectively. Figure 7.2 depicts the number of Gab and Twitter posts for different combinations of hate for posts associated with more than one type of hate using a matrix design created by Conway et al. [20]. The black dots right below each bar represent the combinations of hate types. For instance, there are 7 Gab hate posts simultaneously associated with Gender, Sexual Orientation, and Ethnicity types of hate. Table 7.3 shows the top 10 posts with highest toxicity scores associated with different types of hate. The associated types of hate for each post are in bold.

Overall, hate posts associated with multiple types of hate are more often related to Class and Ethnicity (there are 121 and 48 Gab and Twitter posts, respectively). However,

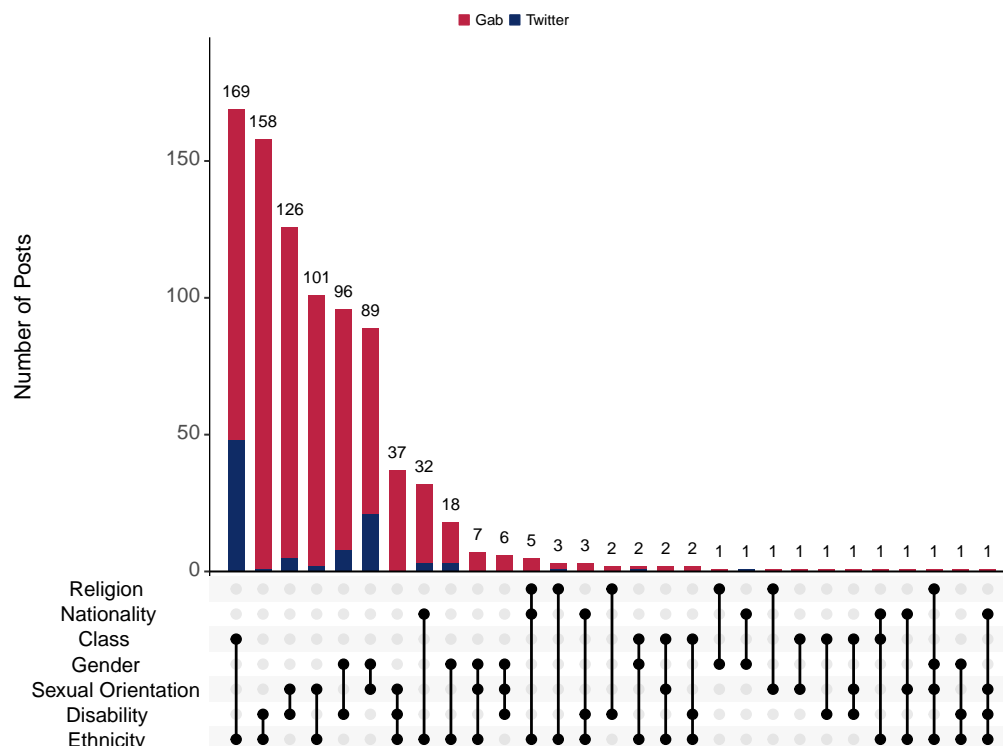


Figure 7.2: Intersection size of the sets of different types of hate for the 775 posts which are associated with more than one type of hate.

Gab posts which have multiple types of hate are mainly associated with Disability and Ethnicity (157 posts). Other common combinations of hate for Twitter are Gender and Sexual Orientation, followed by Gender and Disability. For Gab, Disability and Sexual Orientation, Ethnicity and Sexual Orientation are common combinations of hate. These results indicate that Gab posts tend to have more distinct types of hate at the same time when compared with Twitter.

7.3 Concluding Remarks

This Chapter presented an analysis of the 9,554 Gab and 2,392 Twitter hate posts along with their different types of hate. Starting with a lexicon of 34 hate terms from the Hatebase, we associate posts to different types of hate and then quantify the number of posts for each type and for combinations of types. We show that Gender and Class types of hate are more frequent on Twitter, whereas Disability, Ethnicity, Sexual Orientation, Religion, and Nationality types tend to appear proportionally much more in Gab. With the analysis of the multiple types of hate, we also conclude that Gab posts tend to have more distinct types of hate at the same time when compared with Twitter, likely due to

Table 7.3: Examples of hate posts with multiple types of hate. The associated types of hate for each post are in bold.

<i>Gab</i>		
[HTML]EFEFEF	Ethnicity, Disability	are you fuc*ing retarded? ARCHIVE THAT SHIT YOU STUPID NI**ER.
	Ethnicity, Gender, Sexual Orientation	@(anonymized) my office was in 1 WTC so statistically speaking you are a f*cking t*at. Go back to sucking muslim, ni**er c*ck you f*cking fa**ot. NoRefunds
[HTML]EFEFEF	Ethnicity, Sexual Orientation	stupid ni**er and fa**ot shit
	Disability, Gender	You f*cking Hindus are stupid c*nts. I bet your head is doing that retarded Indian wobble as you type your retarded responses you stupid c*nt.
[HTML]EFEFEF	Ethnicity, Sexual Orientation	f*ck you ni**er fa**ot
	Ethnicity, Disability	FULL RETARD That's your reasoning f*ck you are f*cking stupid are you part ni**er?
[HTML]EFEFEF	Ethnicity, Disability	you are just mad because you are a f*cking retarded ni**er
	Ethnicity, Gender	F*ck you you b*tch ass ni**er c*nt I hate black ni**ers
[HTML]EFEFEF	Disability, Gender	Amy Schumer is a fat gelatinous mass of self loathing lard. No one cares about you retard! Stay the f*ck outta politics with your dumb stank c*nt havin ass!
<i>Twitter</i>		
[HTML]EFEFEF	Ethnicity, Gender	@(anonymized) your an old fat c*nt stop making cod you sh*t dog fuc*ing ni**er
	Gender, Sexual Orientation	@(anonymized) you mad ignorant and weak for talking about her kids like that you bald headed ugly d*ke b*tch.
[HTML]EFEFEF	Ethnicity, Nationality	@(anonymized) @(anonymized) @(anonymized) Stupid f*cking wi**er doesn't understand black people hate his f*cking guts. Pathetic.
	Ethnicity, Class	RT @(anonymized): To the idiot who killed a famous bear during the NJ bear hunt, rot in hell. I know who you are, white trash with shit
[HTML]EFEFEF	Ethnicity, Class	@(anonymized) You are such a trailer trash, bimbo b*tch. You were the worse SECSTATE to ever hold office. TRAITOR!
	Gender, Sexual Orientation	You ni**as be p*ssy like d*kes I handle bars like bikes
[HTML]EFEFEF	Ethnicity, Nationality	@(anonymized) F*ck you wi**er
	Ethnicity, Gender	@(anonymized) @(anonymized) Should be out rounding up f*cking ragheads and sending the c*nts home to where they came from and their families..!
[HTML]EFEFEF	Gender, Sexual Orientation	RT @(anonymized): hate when a d*ke say she hate fake ni**as.. u is a fake ni**a.

the moderation policies put in place by the latter.

Chapter 8

Conclusions and Future Work

In this study we characterize Gab, a social network that emerged advocating liberty and freedom of expression, but received several criticisms about the content shared in it. As the debate between hate speech and free speech is an open and very controversial issue, Gab is an evident source of data for many researchers to explore deeper all those concerns. In this direction, our study provides a characterization of users and content shared on Gab and also contributes to understanding the behavior of such users in a system without the strict policies of moderation from other medias.

Our findings show that Gab is a very politically oriented system that hosts known banned users from other social networks. We also show that the majority of Gab users are conservative, male, and caucasian. Gab is also crowded by extremist users. Our analysis of what users share in Gab unveiled a lot of political statements, showing that Gab is extremely politically oriented, sharing many terms frequently used by conservative politicians and their supporters. We show the existence of a great variety of news domains shared within Gab. Most of these domains are not popular in other social media or on the Internet as a whole, and part of them are known for spreading politics-related news. These results indicate that an unmoderated social media such as Gab has become an echo chamber for right-leaning content dissemination.

The controversial discussion between hate speech and free speech has opened a long debate about speech regulation, especially in the online space. Our analysis on Gab, put into perspective with Twitter, showed that the unmoderated posts on Gab present more negative sentiment, higher toxicity, and different psycholinguistic features. Furthermore, one of the goals of this work is to provide a diagnostic of hate in the unmoderated content from Gab, by categorizing the different forms of hate speech in that system and comparing it with Twitter as a proxy for a moderated system. Our findings support that the unmoderated environment has proportionally more hate speech, and that posts often target diverse groups simultaneously. These findings give us insights on how content policies can be better designed to deal with the different shades of hateful discourse in social media systems. Additionally, our measurement study can be helpful to improve the operationalization of hate speech by online platforms to prevent the spread of online hateful content.

As a final contribution, we make our hate-labeled Gab data available for the research community. Most of the existing efforts on hate speech rely on labeled data from moderated systems, such as Twitter. This means that existing trained models for hate speech might not rely on all the hateful messages exchanged in these system [72]. As we show that Gab is an environment where different types of hate can be found, we believe that our unmoderated hate-labeled dataset can help the development of automated hate speech detectors. We hope that our dataset provides a valuable resource for those interested in detecting online hate speech.

Besides developing better automated hate speech detectors, we leave as future work to exploit hate speech in Brazil. Gab has received an increasing number of conservative Brazilian users, mainly on the occasion of the Brazilian election in 2018. Thus, analyzing Portuguese content on Gab might be helpful for identifying hate in this social media and in others, such as Facebook and YouTube.

References

- [1] Ahmed Abbasi, Ammar Hassan, and Milan Dhar. Benchmarking twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 823–829, 2014.
- [2] Swati Agarwal and Ashish Sureka. Using KNN and SVM based one-class classifier for detecting online radicalization on twitter. In *Distributed Computing and Internet Technology - 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings*, pages 431–442, 2015.
- [3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.
- [4] Anti-Defamation League. From alt right to alt lite: Naming the hate. <https://www.adl.org/education/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate>, 2017.
- [5] Demetres Antoniadis, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, Sotiris Ioannidis, Evangelos P. Markatos, and Thomas Karagiannis. we.b: the web of short urls. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 715–724, 2011.
- [6] Imran Awan. Islamophobia and twitter: A typology of online hate against muslims on social media. *Policy & Internet*, 6(2):133–150, 2014.
- [7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- [8] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- [9] Jamie Bartlett, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. Anti-social media. *Demos*, pages 1–51, 2014.
- [10] Alessandro Bessi. Personality traits and echo chambers on facebook. *Computers in Human Behavior*, 65:319–324, 2016.
- [11] Leticia Bode. Political news in the news feed: Learning politics from social media. *Mass Communication and Society*, 19(1):24–48, 2016.

- [12] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.
- [13] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*, 2010.
- [14] Abhijnan Chakraborty, Johnnatan Messias, Fabrício Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 22–31, 2017.
- [15] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *ACM on Human-Computer Interaction*, 1(CSCW):31:1–31:22, December 2017.
- [16] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *PACMHCI*, 1(CSCW):31:1–31:22, 2017.
- [17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring #gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1285–1290, 2017.
- [18] Irfan Chaudhry. #hashtagging hate: Using twitter to track racism online. *First Monday*, 20(2), 2015.
- [19] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Conference on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*, pages 71–80, 2012.
- [20] Jake R. Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.
- [21] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P. Gummadi. The many shades of anonymity: Characterizing anonymous

- social media content. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 71–80, 2015.
- [22] Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.
- [23] Michael A DeVito. From editors to algorithms: A values-based approach to understanding story selection in the facebook news feed. *Digital Journalism*, 5(6):753–773, 2017.
- [24] Elizabeth Dubois and Grant Blank. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5):729–745, 2018.
- [25] Jennifer Earl and Alan Schussman. Contesting cultural control: Youth culture and online petitioning. 2008.
- [26] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth M. Belding. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 52–61, 2018.
- [27] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. “i always assumed that I wasn’t really that close to [her]”: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 153–162, 2015.
- [28] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8), 2017.
- [29] Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. A quantitative approach to understanding online antisemitism. *arXiv preprint arXiv:1809.01644*, 2018.
- [30] Kerry Flynn. Tech is cracking down on hate speech—but it’s still thriving on Gab and 4chan. <https://mashable.com/2017/08/17/alt-right-free-speech-online-network-gab/#E10Lb40BQmqx>, August 2017.
- [31] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

- [32] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 913–922, 2018.
- [33] Carolin Gerlitz and Bernhard Rieder. Mining one percent of twitter: Collections, baselines, sampling. *Media and Culture Journal*, 16(2), 2013.
- [34] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.
- [35] Edel Greevy and Alan F. Smeaton. Classifying racist texts using a support vector machine. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 468–469, 2004.
- [36] Gary Hanson and Paul Haridakis. Youtube users watching and sharing the news: A uses and gratifications approach. *Journal of Electronic Publishing*, 11(3), 2008.
- [37] Eduardo M. Hargreaves, Claudio Agosti, Daniel Sadoc Menasché, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. Fairness in online social network timelines: Measurements, models and mechanism design. *Perform. Eval.*, 129:15–39, 2019.
- [38] Christopher M Hoadley, Heng Xu, Joey J Lee, and Mary Beth Rosson. Privacy as information access and illusory control: The case of the facebook news feed privacy outcry. *Electronic commerce research and applications*, 9(1):50–60, 2010.
- [39] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.
- [40] NF Johnson, R Leahy, N Johnson Restrepo, N Velasquez, M Zheng, P Manrique, P Devkota, and Stefan Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, 2019.
- [41] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s don’t make a whole: Comparing simultaneous samples from twitter’s streaming api. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 75–83. Springer, 2014.

- [42] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 417–432, 2017.
- [43] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- [44] Chei Sian Lee and Long Ma. News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, 28(2):331–339, 2012.
- [45] Ioanna K. Lekea and Panagiotis Karampelas. Detecting hate speech within the terrorist argument: A greek case. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 1084–1091, 2018.
- [46] A Lella. Traditional news publishers take non-traditional path to digital growth. <https://www.comscore.com/Insights/Blog/Traditional-News-Publishers-Take-Non-Traditional-Path-to-Digital-Growth>, March 2016.
- [47] Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L Williams. Fuzzy multi-task learning for hate speech type identification. In *The World Wide Web Conference*, pages 3006–3012. ACM, 2019.
- [48] Alastair Macdonald and Julia Fioretti. Social media firms have increased removals of online hate speech: EU. <https://www.reuters.com/article/us-eu-hatespeech/social-media-firms-have-increased-removals-of-online-hate-speech-eu-idUSKBN18S3FO>, May 2017.
- [49] F. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- [50] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182. ACM, 2019.
- [51] Johnnatan Messias, Pantelis Vikatos, and Fabrício Benevenuto. White, man, and highly followed: gender and race inequalities in twitter. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 266–274, 2017.

- [52] Amy Mitchell. Key findings on the traits and habits of the modern news consumer. <https://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/>, 2016.
- [53] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. <https://www.journalism.org/2014/10/21/political-polarization-media-habits/>, 2014.
- [54] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM, 2017.
- [55] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. When is it biased?: assessing the representativeness of twitter’s streaming api. In *Proceedings of the 23rd international conference on world wide web*, pages 555–556. ACM, 2014.
- [56] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. In *Seventh international AAAI conference on weblogs and social media*, 2013.
- [57] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- [58] Konstantinos Pappas, Steven Wilson, and Rada Mihalcea. Stateology: State-level interactive charting of language, feelings, and values. *arXiv preprint arXiv:1612.06685*, 2016.
- [59] Limor Peer and Thomas B Ksiazek. Youtube and the challenge to journalism: new standards for news videos online. *Journalism Studies*, 12(1):45–63, 2011.
- [60] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
- [61] Pew Research Center. Online news consumption. <https://www.people-press.org/1996/12/16/online-news-consumption/>, 1996.
- [62] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.
- [63] Emilee J. Rader and Rebecca Gray. Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 173–182, 2015.

- [64] Julio Reis, Fabrício Benevenuto de Souza, Pedro Olmo Stancioli Vaz de Melo, Raquel Oliveira Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 357–366, 2015.
- [65] Julio Reis, Haewoon Kwak, Jisun An, Johnatan Messias, and Fabricio Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 195–204. ACM, 2017.
- [66] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, Fabrício Benevenuto, and Erik Cambria. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.
- [67] Gustavo Resende, Philippe F. Melo, Hugo Sousa, Johnatan Messias, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The Web Conference, TheWebConf 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 818–828, 2019.
- [68] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.*, 5(1):23, 2016.
- [69] Filipe Nunes Ribeiro, Lucas Henrique, Fabrício Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P. Gummadi. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 290–299, 2018.
- [70] Ethel Rudnitzki and Felipe Sakamoto. Rede social de ultradireita chega ao Brasil com acenos a Bolsonaro. <https://apublica.org/2018/12/rede-social-de-ultradireita-chega-ao-brasil-com-acenos-a-bolsonaro/>, December 2018.
- [71] Joni Salminen, Hind Almerakhi, Milica Milenkovic, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 330–339, 2018.
- [72] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *SocialNLP*, pages 1–10, 2017.

- [73] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36, 2017.
- [74] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 687–690, 2016.
- [75] Southern Poverty Law Center. Extremists. <https://www.splcenter.org/fighting-hate/extremist-files/individual>, 2018.
- [76] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummedi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 5–19, 2018.
- [77] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*, 2009.
- [78] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.
- [79] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, pages 1–14, 2013.
- [80] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *CoRR*, abs/1607.01032, 2016.
- [81] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [82] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.
- [83] Mike Wendling. Gab: Free speech haven or alt-right safe space? By BBC Trending. <http://www.bbc.com/news/blogs-trending-38305402>, December 2016.

-
- [84] Jason Wilson. Gab: alt-right's social media alternative attracts users banned from Twitter. <https://www.theguardian.com/media/2016/nov/17/gab-alt-right-social-media-twitter>, November 2016.
- [85] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1007–1014, 2018.
- [86] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.