INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Decoding Hate: Analyzing Twitter Social Networks for Hate Speech and User Behavior

Catarina da Silva Pontes

Master's degree in Integrated Business Intelligence Systems

Supervisor
PhD António Jorge Filipe da Fonseca, Assistant Professor,
ISCTE-University Institute of Lisbon

Co-Supervisor
PhD Sérgio Miguel Carneiro Moro, Full Professor,
ISCTE-University Institute of Lisbon

September, 2024

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Decoding Hate: Analyzing Twitter Social Networks for Hate Speech and User Behavior

Catarina da Silva Pontes

Master's degree in Integrated Business Intelligence Systems

Supervisor
PhD António Jorge Filipe da Fonseca, Assistant Professor,
ISCTE-University Institute of Lisbon

Co-Supervisor
PhD Sérgio Miguel Carneiro Moro, Full Professor,
ISCTE-University Institute of Lisbon

September 2024

To my biggest support for all the criticism and advice.
Thank you for helping me grow.

Acknowledgements

I would like to express my gratitude to Professor António Fonseca, whose expertise in networks was fundamental to the development of this work. His generosity in sharing knowledge and his constant availability were priceless. I would also like to thank Professor Sérgio Moro for his endless encouragement and guidance, which helped me to continually improve. His openness to support me at all times was crucial to the completion of this project. A special thanks to Professors Fernando Baptista and Ricardo Ribeiro, who, as part of the team, always motivated me and maintained a positive attitude, creating an inspiring working environment.

Finally, I would like to thank my family, who, even when I was completely wrapped up in work, reminded me of the importance of maintaining a balance between academic and personal life. To my father who always did everything he could to provide me with the best opportunities to grow, to my mother who was always my shoulder to lean on, to my sister who made me laugh even when I was sad and tired and to my grandmother who brought me up to be the person I am today. All of this is thanks to you. A special thank you to my boyfriend for his constant support and careful reviews of my work.

To everyone, my sincere thanks.

Funded by the European Union: CERV-2021-EQUAL (101049306)
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or kNOwHATE Project. Neither the European Union nor the kNOwHATE Project can be held responsible for them.

Resumo

Esta dissertação vem explorar a dissiminação do discurso de ódio num conjunto de tweets em português, provenientes de Portugal, selecionados com base em palavras-chave que possam indicar a presença de discurso de ódio.

Nesta investigação foi identificada uma lacuna na literatura existente relativamente à intersecção de dois temas dferentes, a análise de redes e o discurso de ódio. A literatura disponível centra-se predominantemente na classificação e deteção utilizando machine e deep learning e não na utilização de métodos de análise de redes sociais pelo que acaba por ser uma abordagem inovadora neste ramo.

Utilizando esses métodos, este estudo fornece uma análise das métricas mostrando como se comporta a rede combinando com representações gráficas da rede que ajudam também na análise. Tal análise pode ajudar a identificar grupos específicos influentes na disseminação de informação e particularmente do discurso de ódio.

A análise também inclui o estudo das mudanças de participação (p-shifts), que ajudam a compreender a dinâmica das interações e de que modo um utilizador consegue a sua vez de intervir, mostrando como estas podem evoluir para discursos de ódio ou contra-discursos.

Os resultados provenientes deste estudo têm como objetivo informar possíveis ações de intervenção futura correspondendo com os objetivos e esforços recentes que a União Europeia mostra no combate ao discurso de ódio.

Palavras-chave: Análise de Discurso de Ódio, Análise de Redes Sociais, Comunidades Online, Twitter, Redes Sociais Portuguesas, Discurso de Ódio Online.

Abstract

This dissertation explores the dissimination of hate speech in a set of tweets in Portuguese from Portugal, selected on the basis of keywords that might indicate the presence of hate speech.

This research identified a gap in the existing literature regarding the intersection of two different topics, network analysis and hate speech. The available literature focuses predominantly on classification and detection using machine and deep learning and not on the use of social network analysis methods, so it turns out to be an innovative approach in this field.

Using these methods, this study provides an analysis of the metrics showing how the network behaves combined with graphical representations of the network that also help in the analysis. Such analysis can help identify specific groups influential in the dissemination of information and particularly hate speech.

The analysis also includes the study of p-shifts, which help to understand the dynamics of interactions and how a user gets their turn to intervene, showing how these can evolve into hate speech or counter-discourse.

The results of this study aim to inform possible future intervention actions in line with the European Union's recent objectives and efforts to combat hate speech.

Keywords: Hate Speech Analysis, Social Network Analysis, Online Communities, Twitter, Portuguese Social Media, Online Hate Speech.

Contents

List of Figures

List of Tables

Acronyms

ADL: Anti-Defamation League.
API: Application Programming Interface.

CICDR: Commission for Equality and Against Racial Discrimination.
CIG: Roma.
COVID-19: Coronavirus Disease 2019.
CRISP-DM: Cross-Industry Standard Process for Data Mining.
CS: Counter Speech.

DB: Data Base.
DHS: Direct Hate Speech.
DSR: Design Science Research.

IHS: Indirect Hate Speech.

LGBT: LGBTIQA+: Lesbian, gay, bisexual, transgender, intersex, queer/questioning, asexual and many other terms.
LGBTIQA+: Lesbian, gay, bisexual, transgender, intersex, queer/questioning, asexual and many other terms.
LIWC: Linguistic Inquiry and Word Count.

NLP: Natural Language Processing.

OS: Offensive Speech.

P-Shifts: Participation Shifts.
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis.
PT-BR: Portuguese from Brazil.
PT-PT: Portuguese from Portugal.

RAC: Racism.

SNA: Social Network Analysis.

URL: Uniform Resource Locator.

VADER: Valence Aware Dictionary and Sentiment Reasoner.

XEN: Xenophobia.

CHAPTER 1

Introduction

Twenty-First Century society has the answers to all questions at the distance of a click as information has become accessible for everyone, everywhere at anytime. Social media more than a place to find information, is a place where we can communicate with each other, and with that comes some problems that only few could anticipate.

Nowadays there is a growing problem related to hate speech [3], access to the internet has become easier than ever, allowing people to communicate and interact more on an online context, and with a larger community. The greater the number of people involved, the greater the possibility of starting a disagreement that can lead to the use of hate speech and, due to the already established network, increase its dissemination which encourages its use. [4]

Compared to all the other popular social networks in use, Twitter may be the one that instigates more interactions between users that may or may not know each other since it is home to a lot of debates on a wide range of topics [5]. Besides that a lot of celebrities from the more diverse areas (politics, sports, etc.) use this platform to expose their thoughts which also drives its massive use. Twitter is made for users to broadcast short text posts than may or may not contain also videos, photos or links and because of that it becomes a great case of study.

This study seeks to analyse a set of tweets from Portugal and written in Portuguese that were selected based on certain keywords that may suggest the presence of hate speech. This Portuguese context was chosen because this work is part of a project funded by the European Union that aims to understand Portuguese online hate speech. This dataset contains tweets published between 2021 and 2022 and with those, we aim to identify patterns in the use of the language in those tweets. The main goal is to get important insights into how hate speech is disseminated online so that we can apply actions in key places to mitigate it in the best way possible. For example, this analysis can help identify specific groups that are the big influences in the dissemination of hate speech or even understand the behaviour of a conversation so that it can be signaled as *risk of hate speech*.

The biggest difficulty in this project is the fact that we have tweets in Portuguese as our object of study which is something that has been studied very little, so we have to recognize that not everything can be translated directly into one language and treated in the same way as other researchers may have done with tweets in English, for example. Each language has its own dialect and specific vocabulary so each one needs a different approach [6]. If it is already difficult to compare different dialects of the same language , the difficulty intensifies when trying to treat two completely different languages. On the positive side, we are contributing to the study of hate speech in Portuguese, something that has not been studied much yet.

1.1. Framework

Unfortunately, there are no official statistics from the justice system on hate speech in Portugal, as they are only released if there are more than three incidents in a year, which is rarely the case [7].

To give context to the problem in Portugal, we turned to other organizations that provide data based on complaints received. The Commission for Equality and Against Racial Discrimination (CICDR) indicates that in 2022 there were 291 complaints in which more than half were victims of discrimination because of their nationality [8]. Also between 2020 and 2022, ILGA's Observatory on Discrimination Against LGBTI+ People (a non-governmental organization) received a total of 469 complaints [9]. In both of these cases, the proportion of complaints about hate speech on social media is small, with 15% at CICDR and around 13% at ILGA, this does not mean that there are fewer cases, but that they are not being complained about. Often, even when upset, the victims do not realize that they are suffering from a crime that can be punished and so they end up not pressing charges.

1.2. What is Hate Speech?

To better study it, it is important to understand what is hate speech and how it differs from other types of speech. We shall start with Parekh's definition "expresses, encourages, stirs up, or incites hatred against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation"[10]. United Nations also came up with a definition that broadened by adding "or other identity factor" at the end of the definition [11] and saying that most of the time this type of speech is directed to minorities.

Due to the regulation of this type of speech, there has been a debate about whether is it against freedom of speech or not, but for the Portuguese Ministry of Justice, it is already treated as a crime and can be punished from 6 months to 5 years in prison.[12]

1.3. Thesis Motivation

Hate speech can have a significant impact on the people and communities that are surrounded by it. This type of speech can affect the mental and emotional health of its victims, causing anxiety, stress, and even depression. It can also lead to isolation, making people feel excluded from society.

According to the Anti-Defamation League (ADL), the most common side effect of being a victim of hate speech is having trouble sleeping, but what at first sight may seem bearable can be deadly because it can lead to suicidal thoughts. This way it is important to create a society prevented from hate speech, for that, we need to analyse how hate speech is disseminated throughout social media.

Reports showed an increase in Online Hate Speech (OHS) during 2020-2021, boosted by the pandemic, which fueled stigmatization of minority groups. Portugal, such as all European Union, has seen a rise in hate speech against immigrants, racial/ethnic groups, and LGBTIQA+ communities, but tools to detect, monitor, and prevent OHS are lacking. Due to this, our project

is motivated by filling this gap with a comprehensive, culturally sensitive approach to analyzing, detecting, and countering OHS in Portuguese.

The motivation behind this research is to use social network analysis, as used in other languages' environments, to respond to the need to comprehend the evolving landscape of social interactions and its consequent rise of usage of hate speech in Portugal's Portuguese language, so that we can target some actions to mitigate it in the social media environment and we can have a society free of hate speech.

This research is embedded in the broader effort of the European Union to combat OHS, by creating the kNOwHATE project. The kNOwHATE consortium, was funded by the Citizens, Equality, Rights and Values European Commission Programme.

1.4. Thesis Objectives

The goal of this dissertation is to analyse the conversation networks on the former Twitter, now X, relating it with the presence of hate speech and its targets. For that is necessary to first retrieve data from Twitter's API and create a cleaning method for the data. After that, we chose to do an exploratory analysis of the data to have a contextualisation of what we had in hand.

We aim to characterize our network based on network metrics and compare some subsections with others, for example, the subsection with hate speech versus the subsection without it. For that, we aim to create representations of the network using qualified software and auxiliary libraries that illustrate the reality of our network. In the end, we aspire to analyse user behaviour to eventually identify communities within our network.

In summary, this study's goal is to model the behaviour of information dissipation with hate speech on social networks, so that it is possible to find areas of attack on hate speech and to boost counter-speech.

1.5. Thesis Contribution

A gap was identified in the existing literature, the lack of comprehensive studies on the intersection between network analysis and hate speech since most of the times it focus on hate speech classification or detection.

We aim to fill the gap by understanding how hate speech spreads within social networks and identifying the types of conversations that are more likely to evolve into this harmful phenomenon or what type of conversation evolves from it. This approach offers valuable insights into the understanding and prevention of hate speech, providing a solid basis for targeted and effective interventions. Ultimately, this study contributes to the advancement of knowledge by offering a unique and comprehensive perspective on the phenomenon of hate speech on social media.

1.6. Research Methodology

The methodology adapted in this study was a Design Science Research (DSR) [13] approach as the primary framework. DSR provided a guide to develop and validate the prescriptive knowledge in information science. This approach allowed for a broader exploration of the phenomenon in study and of the problem, considering various perspectives and dimensions.

This problem-solving paradigm focus on the creation and evaluation of artefacts by combining the scientific precision with practical relevance.

Additionally, to ensure effective project management and implementation, the Cross-Industry Standard Process for Data Mining (CRISP-DM) [14] was used as a project plan. CRISP-DM enabled us to outline specific tasks, timelines, resources, and deliverables necessary for the exploratory analysis of the networks. Originally designed as a systematic framework for data analysis, CRISP-DM has key stages such as business and data understanding, data preparation, modeling, evaluation, and deployment. However, when applied to social network analysis, additional considerations are necessary. In addition to understanding the business context and the data, an essential step was added which involves the understanding of the network and its structure. This adapted methodology model also includes pre-processing network data, identifying patterns unique to social networks, and effectively deploying the analysis results.

Therefore we get an adapted CRISP-DM, shown in Figure 1, and described as following:

- Business Understanding - business context
- Data Understanding - initial exploration (structure and quality)
- Network Data Preparation - transformation of the data into a format suitable for network analysis
- Network Modelling - social network analysis methods and algorithms
- Network Evaluation - checking the quality of the results and any possible limitations or biases in the analysis
- Deployment - communicating the results of the network analysis

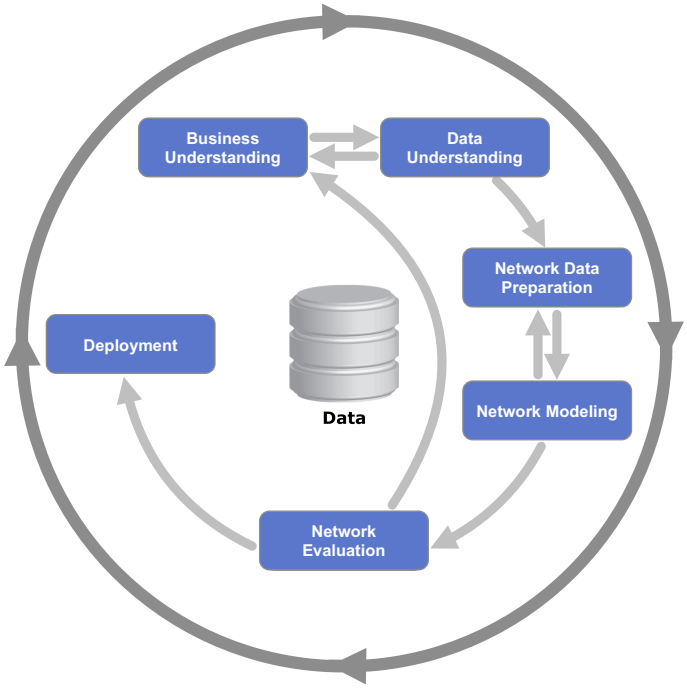

Figure 1. *Adapted CRISP-DM.*

By incorporating both DSR and CRISP-DM, as explained in Figure 2, we seamlessly integrated the wider research framework with a more detailed project planning, making sure that the study is well structured and ensuring its comprehensibility.

1.7. Research Questions

Within the scope of the examined subject, the following are the research questions that drive the analysis:

$[RQ1]$ How prevalent is hate speech within Twitter's social network connections?

$[RQ2]$ How to characterize Hate Speech and Counter Speech in the Portuguese community framework?

$[RQ3]$ How does hate speech dissipate, and what influences this process?

1.8. Thesis Organisation

This section provides an overview of this dissertation's structure summarizing its contents and objectives. This dissertation is organized in 5 main chapters which are: Introduction, Literature Review, Data Collection and Analysis, Social Network Analysis, and Conclusions. Each chapter is designed to build upon the previous one, creating a coherent study. It is important to acknowledge that some of the work in this thesis was also used in a publish paper [15].

By using DSR and CRISP-DM methodologies its structure reflects in the organization of this dissertation. The correspondence of the chapters to the phases of these cycles is as follows:

Chapter 1: The Introduction provides an overview of the problem defining its framework and all the context and background behind it. Additionally, it sets the stage for the research by formulating research question, defining its objectives and describing the methodology adopted. This chapter is aligned with the DSR phase of Problem Identification and Research Objective Definition and the CRISP-DM phase of Business Understanding.

Chapter 2: Describes the research techniques, critically analyses key studies, identifies gaps in the current knowledge, and demonstrates the relevance of our research. This second chapter corresponds to the DSR phase of Theoretical Foundations and State of the Art, and the CRISP-DM phase of Data Understanding in terms of prior knowledge and research gaps.

Chapter 3: Details the pipeline used to retrieve data from Twitter API and also explains the analytical techniques used to process and understand the data in a early stage. This chapter aligns with the DSR phase of Artifact Design and Development and the CRISP-DM phase of Data Preparation.

Chapter 4: Focuses on specific methods and results of analyzing social networks. Provides a deep analysis on network metrics to better understand the dissemination of hate speech. This chapter can be matched with the DSR phase of Artifact Evaluation and the CRISP-DM phases of Modelling and Evaluation, since it involves applying and assessing analytical techniques to understand the research problem. This chapter was accepted and presented at the 20th IPMU Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, in which it was title "Unveiling Patterns of Hate Speech in the Portuguese Sphere: A Social Network Analysis Approach" [16] and during this year it will be published.

Chapter 5: Details the conclusion reached in the study and proposes future approaches to carry on with this work. It corresponds to the DSR phase of Communication and Research Outcomes and the CRISP-DM phases of Deployment and Final Evaluation.

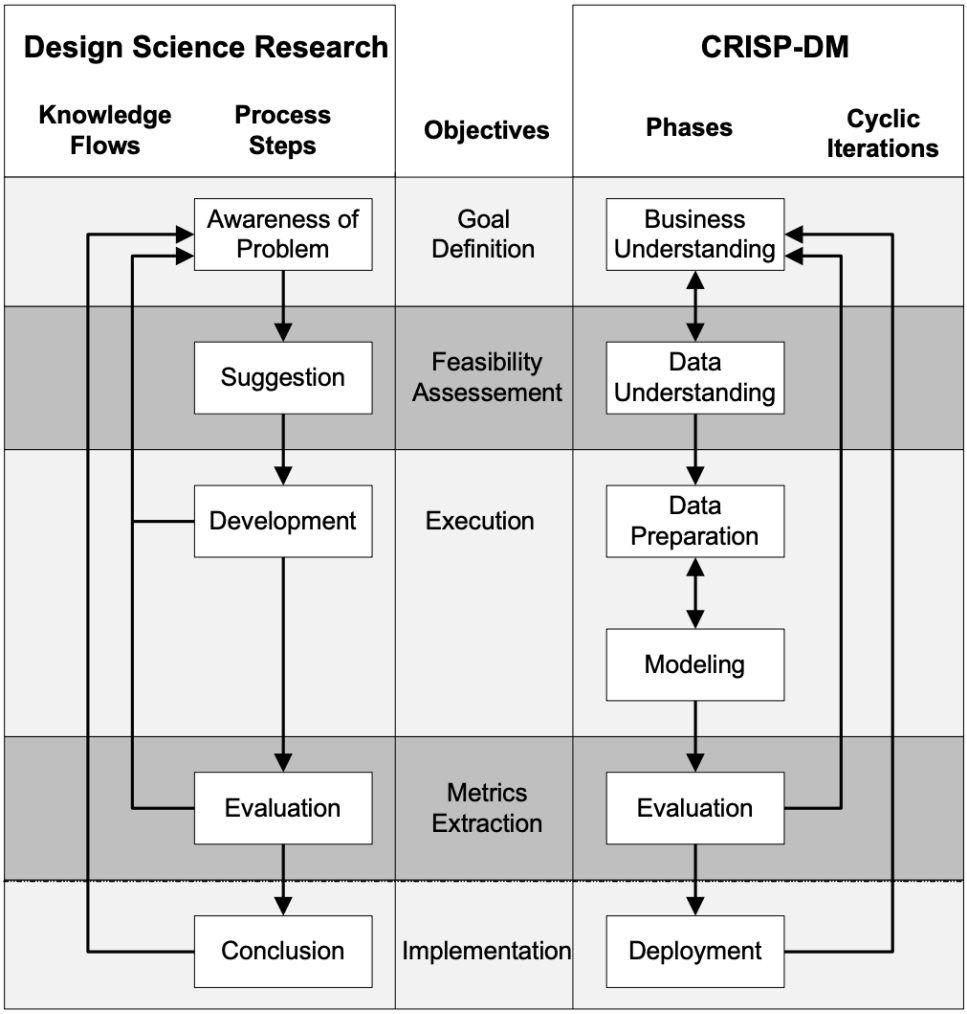

Figure 2. *Design Science Research and CRISP-DM side by side. [1]*

CHAPTER 2

Literature Review

2.1. Introduction

This project is focused on two main areas that normally do not intercept: hate speech and social network analysis. Hate speech research has already a substantial bulk literature and has recently gained much attention as seen in Figure 3. There is already a lot of work done when it comes to hate speech, either about its detection or related to text mining but there is a gap on the social network analysis topic.

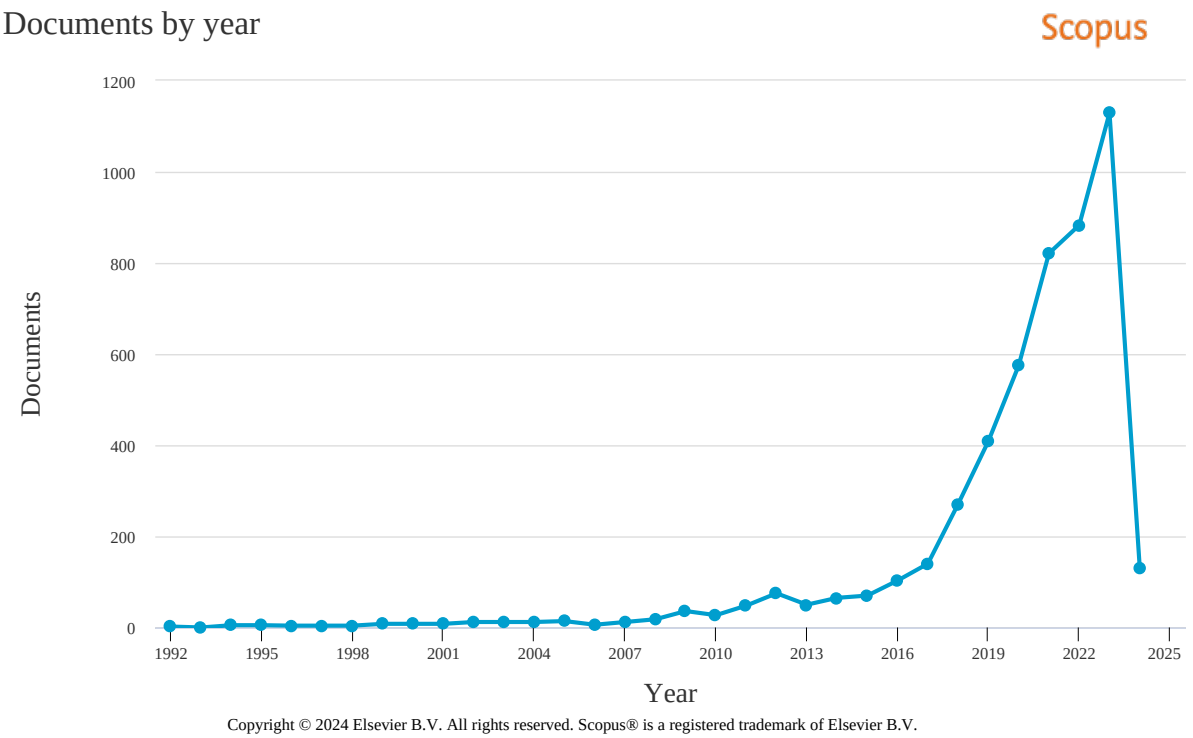

Figure 3. *Hate speech Documents Related with Detection OR Machine Learning by Year.*

This chapter exhibits the state of the art developed related to social network analysis and with particular focus on the presence of hate speech but to have more content we decided to broaden this search to a wider topic which also includes abusive and toxic speech.

Our main purpose with this state of the art is to identify if this type of studies was already done with the PT-PT Portuguese language and even with any other language and after that to identify some of the methods and approaches used by the researches to analyse social networks in the context of offensive speech.

2.2. Social network Analysis

Social Network Analysis (SNA) is a powerful methodology for understanding the complex fabric of social structures and relationships, offering a systematic approach to unraveling the dynamics that drive interactions within networks.

With the increasing usage of social media, it started to be a huge asset to organizations so they can better understand their targets. Because of that SNA is becoming increasingly important in understanding and analyzing these networks [17]. This technique has been applied in various fields, including marketing [18] and risk analysis [19].

This branch of network science is a synonym for analyzing interaction structures, such as graphs, by indicating objects that interact and how they do it. Which can often be representing people and their connections as a graph, this graph can be afterwords analysed using graph theory. [20] Social network analysis is a "fit method for studying dynamic and transient social contexts" because it uses a geometrical approach where individuals (users) are represented by a node and their connections or relationships by an edge that connects two individuals [21].

We can use two different techniques when analyzing social media networks, using a macro level or micro level analysis. The macro level studies the way people behave as a group, when unwritten laws rule it, which most of the time is very different from the micro level where people communicate as individuals.[22]

The number of Social Network Analysis documents has been steadily increasing, as we can see in Figure 4 and has reached its peak today, probably because organizations and researchers want to understand complex social interactions and networks and the widespread use of social media has also generated a vast amount of data for analysis. Additionally, we can say that the increase was steeper after the emergence of the Web 2.0, which enabled the proliferation of social networking platforms using forms of cooperative production and online information sharing.[23]

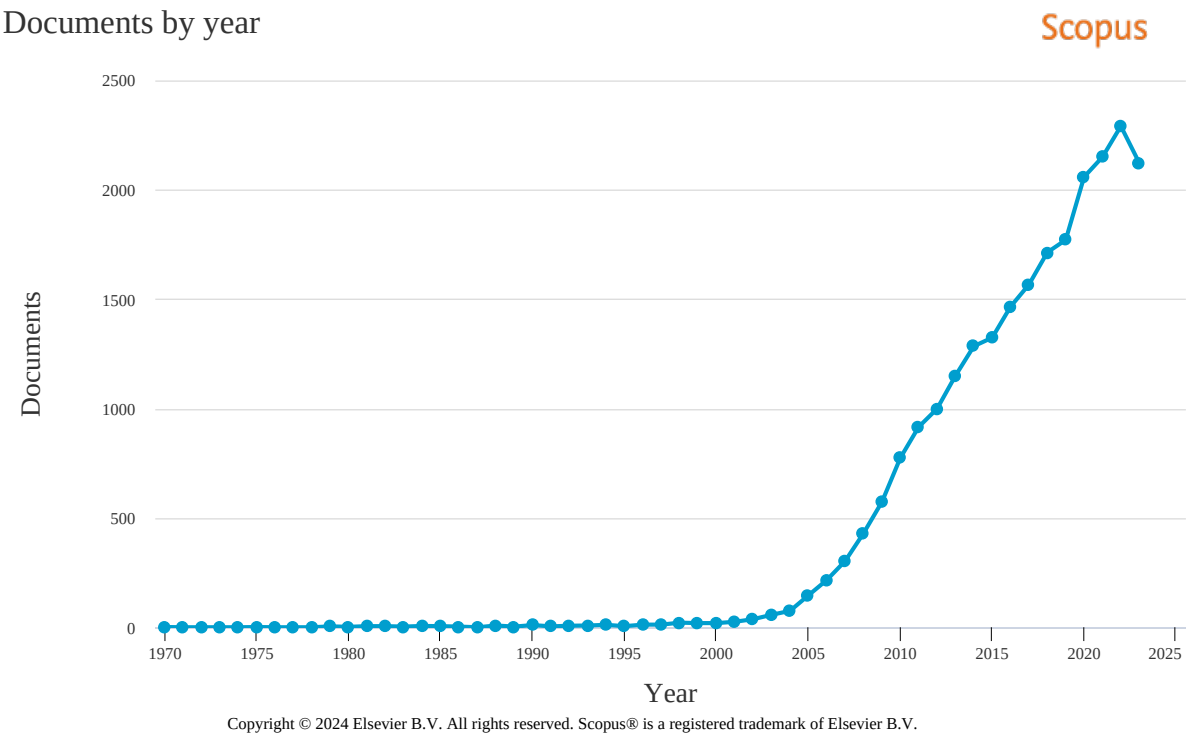

Figure 4. *Social Network Analysis Related Documents by Year.*

2.3. Hate Speech

Hate speech hardly has a definition accepted by every organization or researcher but for the sake of this study we will define it as a type of communication that disparages a group based on race, ethnicity, gender, or other characteristics and its presence in our society is a growing concern in both traditional and online media [24]. It is often targeted at vulnerable or minority groups [25], and its prevalence has led to increased research in the field, particularly in the areas of regulation, computational linguistics, and discourse analysis [25] [26].

The study of hate speech has undergone a remarkable turnaround, migrating from a predominantly social and humanistic approach to a more computer science-centered perspective. The computer science community has identified an opportunity to apply its skills in data analysis, machine learning and natural language processing to understand and combat hate speech in never-thought-before ways. As Figure 5 reflects, until 2015 the main subject area studying hate speech was Social Sciences followed by Arts and Humanities, representing together more than 80% of all the studies, at this time Computer Science was only present in 4%. But when we switch eras, more specifically when we analyse from 2016 until 2024 as in Figure 6, we identify that Computer Science took place having now a little more than 30% and other subjects also raised its number, such as Engineering, Mathematics and Decision Science, which all of them together had around 18%.

There are lots of algorithms being developed to identify patterns of hate speech, to more efficiently detect it on a large scale using different types of classifiers such as Deep Learning [27], Random Forest [28], other forms of Machine Learning [29] and many more.

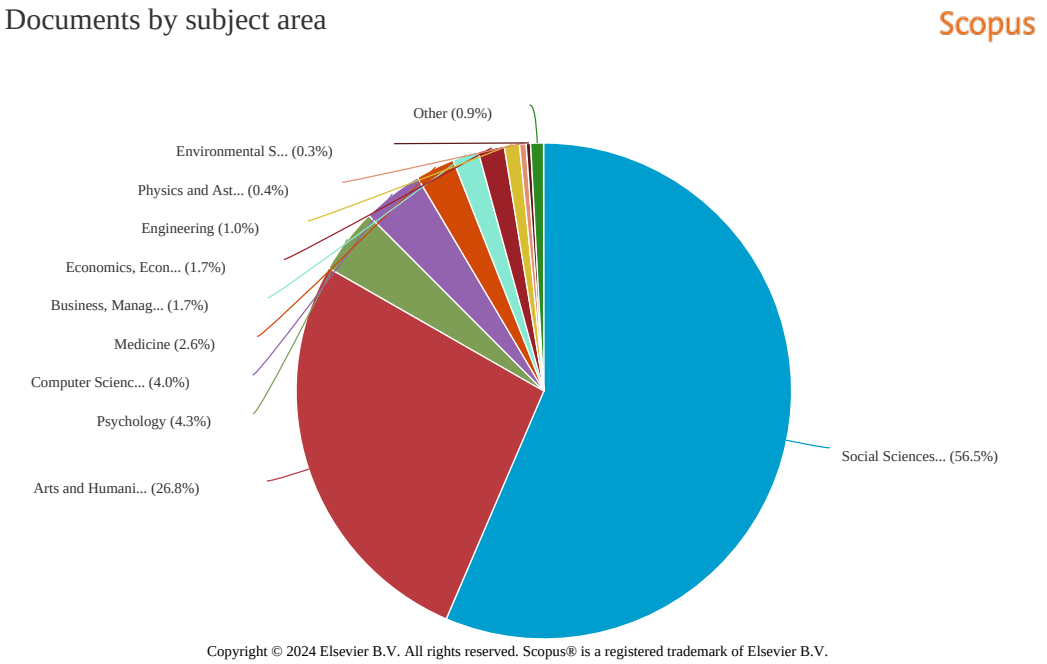

Figure 5. *Distribution of Scientifically Documents per subject areas from 1992 until 2015.*

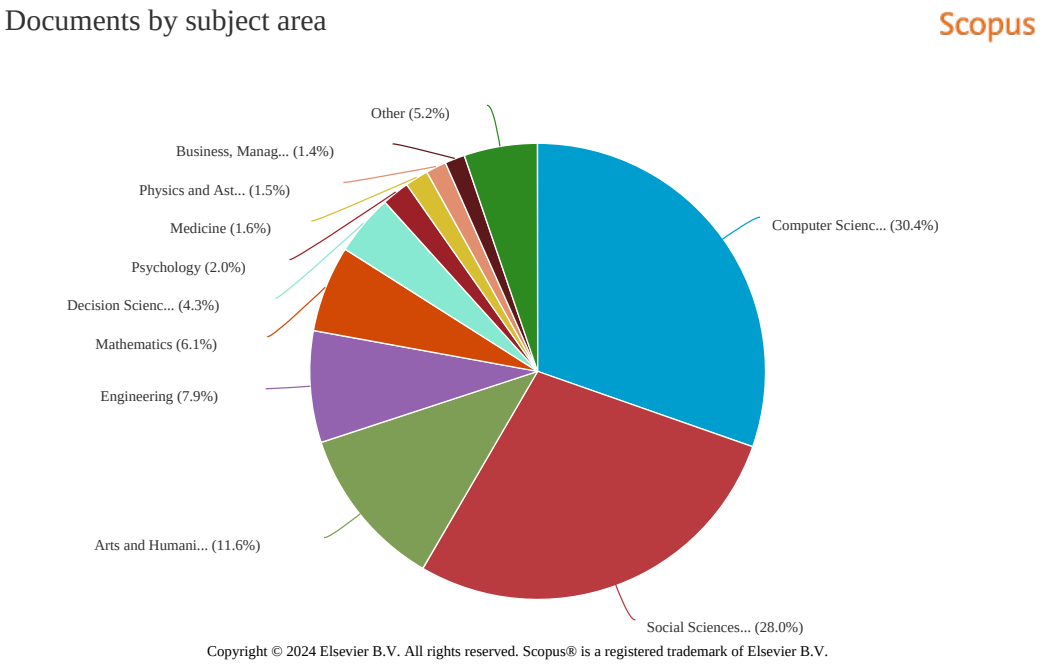

Figure 6. *Distribution of Scientifically Documents per subject areas from 2016 until 2024.*

2.4. Participation Shifts

To analyse interaction sequences in conversations we use a framework called participation shifts (P-shifts) [2]. P-shifts refer to how people switch roles between speaking, listening, and being an unaddressed recipient during conversations. Using cluster analysis helps identify the multiple ways individuals move between these roles, providing a clearer understanding of their conversational dynamics. The group function has a more significant impact on these roles than factors like gender, and it is closely linked to the nature of exchanges facilitated by each group function. There are sixteen types of P-shifts which are classified based on how the second speaker gets their turn from the first one. The shifts are divided into some groups: addressing the group, addressing a third person, or speaking after being addressed.

2.5. Search Techniques

We decided to use the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) [30] methodology in the creation of our literature review. This methodology refers to a set of guidelines that help to keep a common base for the systematic reviews. When searching for the papers we used IEEE Xplorer and Scopus databases, and we filtered it to only the ones written in English. After that we applied two other filters so the results became more aligned with our needs: Only articles, conference papers, or reviews and only in the branch of Computer Science. The last filter was applied because there was a large number of papers related to social sciences that focused on more theoretical studies than we wanted.

The research technique was based on a query with some keywords that specify the studies to what we are looking for. This query searched for the following words into articles' titles, abstracts and keywords: "Hate Speech" and "Social Network" AND "Analysis" and not "Deep Learning" and not "detection". It was very important that the query had the entries with "and not" because most of the work done around hate speech is with its detection and we wanted to take the noise out of our search.

Using the graph on Figure 7 to analyse the yearly evolution of documents published related to our query we can see that there is been a big increase of interest in combining those subjects, but in terms of quantity is still not a lot when compared, for example, with the numbers on the graph of Figure 4.

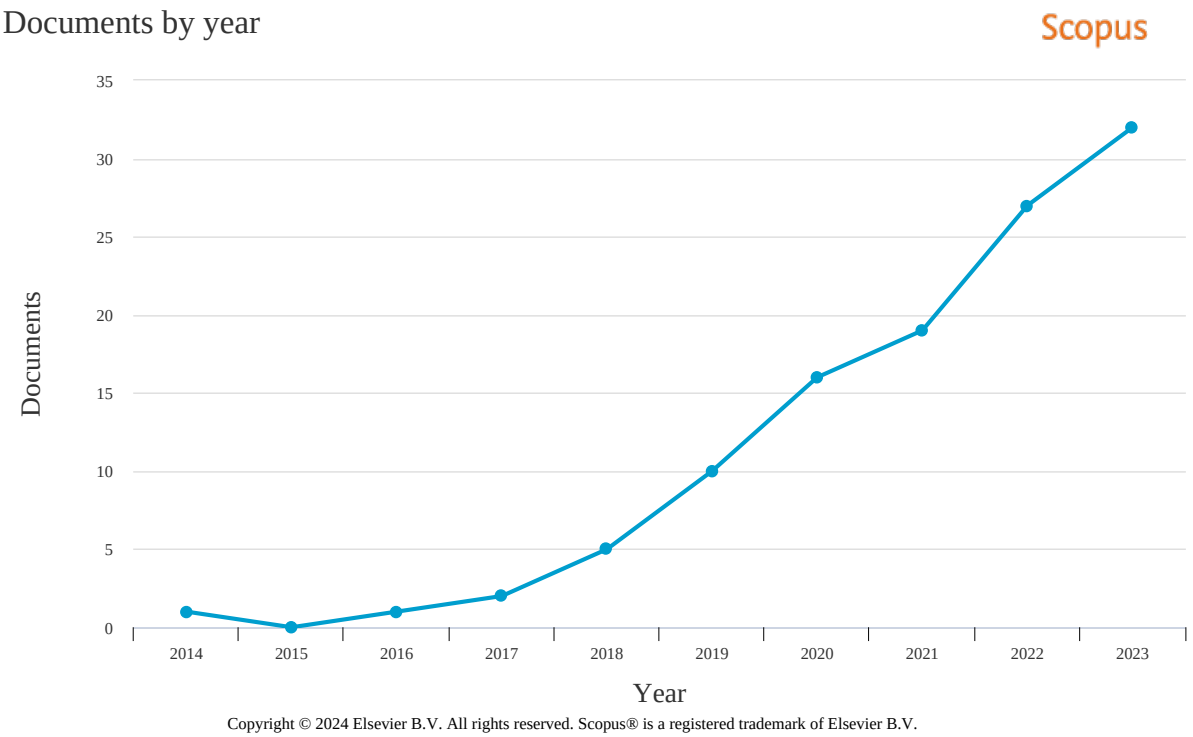

Figure 7. *Query Resulted Documents by Year.*

After applying all the filters in the Databases, as seen in Table 1, we were left with 34 documents that needed careful analysis one by one to see if they match our study's purpose.

<table>
<tr><th>Concept</th><th>Context</th><th>Exclusions</th><th>Limitations</th></tr>
<tr><td>Social Network

AND

Analysis</td><td>Hate Speech</td><td rowspan="3">AND NOT
Detection

AND NOT
Machine Learning</td><td rowspan="3">Only articles or conference papers or reviews

In English

In Computer Science</td></tr>
<tr><td>72 244 documents</td><td>4772 documents</td></tr>
<tr><td colspan="2">193 documents</td></tr>
<tr><td colspan="4">34 documents</td></tr>
</table>

Table 1. *Query and Its Results.*

To identify the eligible documents we performed a PRISMA [30] analysis as described in Figure 8. Even after the definition of the keywords we still end up with some documents that did not match our study, most of them because they emphasized some kind of machine learning to classify, detect or predict hate speech, which was not our purpose.

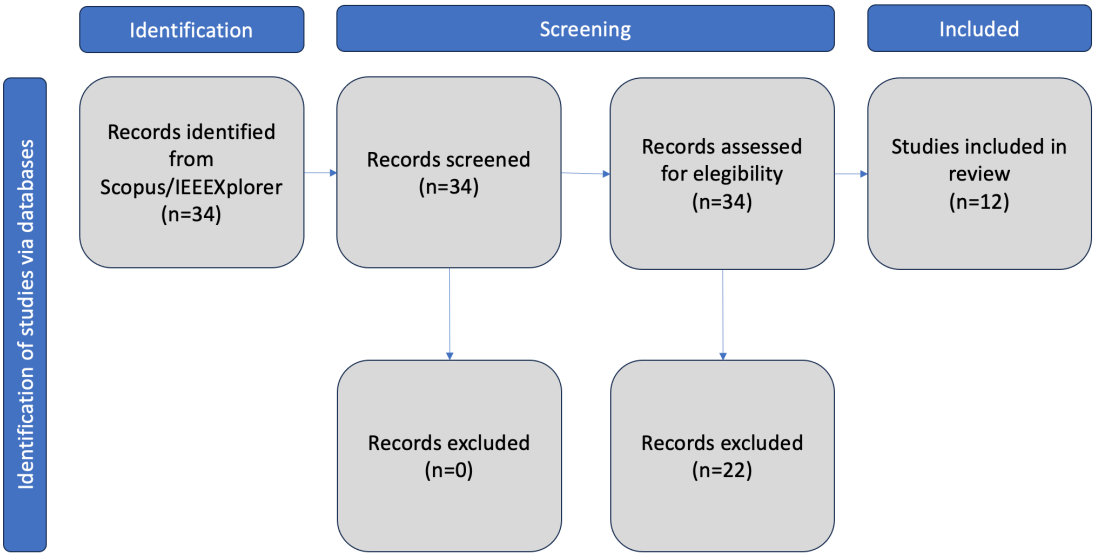

Figure 8. *PRISMA Analysis.*

2.6. Related Work

Because of the growing problem, which is the propagation of hate speech on the Internet, Le Nguyen [31] aimed to understand how hate speech spreads on dark-web forums and at which speed can it influence people for that they combined the following techniques: sentiment analysis, social network analysis, and graph theory. In this article the sentiment is treated as a disease so they can study its spread and to see how effective their approaches are in mitigating it.

During the peak of the COVID-19 pandemic, we have witnessed a significant increase in hate speech (specially anti-asian) on social media, which has become a problem worthy of Kim and Kesari's study [32]. This phenomenon has been exacerbated by social isolation, which has left us with little more to do than engage on social media. This context has generated a greater need for online interaction, thus fueling the spread of hate speech.

In their article DiCicco et al. [33] focus on COVID-19 vaccine discussions on Twitter and Parler. A Detoxify model was used to create an index that translates the presence of toxicity in a text with a score. By using NetworkX and creating co-hashtag network graphs both Twitter and Parler dynamics were compared but to taper off the study they filtered the dataset to only tweets or posts that had more than 0.5 on the toxicity score. After their toxicity analysis Twitter was defined as more toxic than Parler in almost all the cases. The use of social network analysis came to define clusters of users also known as communities, and even a misinformation echo chamber was founded.

Arce-García and Menéndez-Menéndez [34] used other type of analysis, it was investigated how conversations on Twitter about gender and sexual identities had origins and which characteristics they have, for that they collected a sample of over 1 million tweets (referring to one year) related to women's rights, the LGBTIQ+ collective and trans people. They applied network theories to be able to carry out the study, and using the Louvain algorithm they could analyse the presence of groups highly interconnected and without clear references, they also could find the presence of coordinated networks that propose to cause damage and provoke confrontation, but also other groups such as queer, trans, feminists and LGBT groups.

Gephi is a great tool to calculate various network measures, to visualize the networks and even to apply filters to it [35], Weng and Lin [36] analyse the information propagation path using Gephi's tools. The retweet relationship was directly related to topic diffusion behaviour and therefore it translates better the willingness of the users to spread it, because of that retweets are studied here over "likes" or "comments". They choose an approach of analyzing centrality metrics to better understand which accounts are the main influencers on the spread of information, so in the end, they could understand how social bots take part in that.

Still on the topic of social bots and their roles in hate speech dissemination, Riquelme et al. [37] suggest an interdisciplinary study combining computation with philosophy and sociology to better understand and model their behaviours. The conduct of 5 opinion leaders was analysed around key events such as the start of a massive protest in Chile at the end of 2019. Using different techniques including descriptive, quantitative (data aggregation, centrality measures,

and statistical analysis), and qualitative (text analysis) techniques they analysed user profiles, their activity, and their content which led to the identification of hundreds of social bots that were specifically created to spread ideological ideas. This analysis found out that Chile's right wing may have made up bot accounts, acting as amplifiers of the speeches spread by specific political leaders, presumably created to function as echo chambers in political campaigns.

Online speech analysis from a political perspective can be fundamental to define a more efficient approach to the people. In this paper, Volkovskii and Bodrunova [38] examine the relationship between civility and incivility in Russian online political discussions, focusing on how they are influenced by the political parallelism of media platforms. They analytically characterize speech using some metrics and the analysis of perceptual data. The study concluded that while incivility is common in Russian online discussions, it is partly offset by explicit civility, but there is still over 10% of posts containing intolerant, offensive, and rude expressions. Besides that there was not significant diversity in incivility between oppositional and pro-state media platforms, with both exhibiting levels higher than 2%.

In this next study Park and Kim [39] also used a key and real event to focus the analysis on, which in this case was the suicide of a famous artist that is thought to have been caused by abusive discourse on social media. They were aimed to detect the structural characteristics of the information flows and the patterns present in the conversations about abusive comments that lead to the artist's death. They created a semantic network analysis where each node is a word or expression present on the data corpus and an edge is formed if both words are in the same comment, this way they could analyse the relations and associations between words, reveal important topics, metaphors, and key themes from this large-scale textual data.

Torregrosa et al. [40] opted to use some network metrics such as various types of centrality metrics (n-degree, eigenvector, k-shells, betweenness, and closeness) to measure the relevance of each selected user. Besides that it was analysed which linguistic indicators of the extremist discourse are the most used and if the use of this type of discourse increases the relevance of the actor in the network. Therefore, the tweets content was analysed by looking into the linguistic indicators used and tone of the text using LIWC and VADER. To validate their hypothesis of the existence of a relationship between user relevance and the use of abusive discourse they test it on texts about other topics. They came up with the conclusion that the retweets received by high relevant users had more aggressive, racist, supremacist and group-directed type of language.

In Blanco-Herrero and Calderón's [41] study they explore the impact of fake news about minorities on the existence and rise of hate speech directed at those same minorities. They apply a three-step routine to analyse that, which consists of one survey to analyse the society's opinion, a social network analysis to understand the dissipation of the content, and lastly, an experimental survey to recognize how people interact with these types of contents and if they actually believe in that or not. Their social network analysis is not only based on graph and cluster analysis, besides that they analysed the cascades to understand depth, size and breadth of a specific content.

In the following article, Kargar and Rauchfleisch [42] investigate the dimension and the authors of online hate, harassment, and abusive speech in the particular case of opposing to Iranian emigrants. This study is based on two pillars: qualitative interviews and a quantitative analysis of related individuals' Instagram accounts. The quantitative approach explores how fast speech spreads, who are the main responsible users for that, and which patterns of information dissemination are found. All this is done as an attempt to mitigate the voices of users who tend to use Instagram as a place for practicing and instigate hate speech, with less violent and hateful content people are less encouraged to hate and their minds can actually change when it comes to Iranian emigrants.

The main aim of Pereira-Kohatsu et al. [43] is to create an intelligent system that can identify and monitor hate speech on Twitter. Although the final objective is detection, this study has the particularity of having an automatic network analysis stage which is called Social Network analyser, this uses graph theories to identify social structures. This analyser creates visualizations such as word clouds and users' mentions graphs that help understand more visually the dissipation of hate in this social network. This tool can be an excellent addition to state organizations that aim to prevent hate speech because with this they can monitor hate without having to have technicians who understand network analysis.

Authors	Title	Year
Blanco-Herrero, D. and Calderón, C. A.	Spread and reception of fake news promoting hate speech against migrants and refugees in social media: Research Plan for the Doctoral Programme Education in the Knowledge Society	2019
Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M.	Detecting and Monitoring Hate Speech in Twitter	2019
Kargar, S. and Rauchfleisch, A.	State-aligned trolling in Iran and the double-edged affordances of Instagram	2019
Torregrosa, J., Panizo-Lledot, Bello-Orgaz, G., and Camacho, D.	Analyzing the relationship between relevance and extremist discourse in an alt-right network on Twitter	2020
Park, S. and Kim, J.	Tweeting about abusive comments and misogyny in South Korea following the suicide of Sulli, a female K-pop star: Social and semantic network analyses	2021
Kim, J. Y. and Kesari, A.	Misinformation and hate speech: The case of anti-Asian hate speech during the COVID-19 pandemic	2021
Weng, Z. and Lin, A.	Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic	2022
Arce-García, S. and Menéndez-Menéndez, M.-I.	Inflaming public debate: a methodology to determine origin and characteristics of hate speech about sexual and gender diversity on Twitter	2022
Riquelme, F., Rivera, D., and Serrano, B.	Analyzing the far-right political action on Twitter: the Chilean constituent process	2022
Volkovskii, D. and Bodrunova, S.	Incivility Balanced? Civil vs. Uncivil Speech in Online Political Discussions as Dependent on Political Parallelism	2023
Nguyen, L. and Rastogi, N.	Graph-based Approach for Studying Spread of Radical Online Sentiment	2023

DiCicco, K., Noor, N., Yousefi, N., Maleki, M., Spann, B., and Agarwal, N.	Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Social Media Platforms	2023

Table 2. *Related Work Overview.*

2.7. Research Gap

There is a significant gap in the literature on the study of hate speech using network science, since most researchers use machine or deep learning for its detection or prediction. The lack of exploratory research using network science methods limits our understanding of the complex dynamics of social interactions that may fuel hate speech online.

The use of social network analysis is an innovation in this hate speech area of research, contrasting with the predominant use of machine learning. While most previous studies have focused on predicting or detecting hate speech, SNA offers a different and with added value approach, studying the structure of the social interactions. This provides a more global understanding of this phenomenon, identifying user communities and patterns of information propagation.

With this type of approach we can make more effective interventions, targeting specific users that have a more relevant role in the propagation of hate speech to mitigate the impact of hate speech and that s why we decided to use Social Network Analysis as the basis for this study.

CHAPTER 3

Data Collection and Analysis

3.1. Data Collection

During the initial phase of our research, we carried out a data retrieval process when Twitter still allowed free access through its APIs, which made it easier to obtain the necessary information. However, given the sensitive nature of our goal of analyzing hate speech, it was essential to annotate the dataset. This involved a meticulous process of identifying and tagging the relevant content, ensuring that we could carry out an accurate and meaningful analysis.

In this chapter, the extraction and annotation phase will be thoroughly described, providing detailed insights into the methods used to obtain and prepare the data for analysis.

3.1.1. Retrieval

The way of retrieval of data was based on the usage of twitter api and other dataset already retrieved for Hate-COVID-19 Project [44]. The approach is specified in Figure 9. To access tweets it was necessary to use the Twitter API that provides various data accesses, with the limits being much lower for approved users who use it for Academic Researches. This project focuses on 4 themes that were predefined by its coordinator: Roma (CIG), Racism (RAC), Xenophobia (XEN), and LGBTIQA+ (LGBT). Therefore, tweets that had certain keywords determined beforehand (related to the 4 themes above) were extracted. In the set of keywords were identified some that would have ambiguous meanings, and so it was decided to extract tweets with these words only when associated with insults that are on a list also previously defined. For these extractions, the following Queries were used:

The query for non-ambiguous words/expressions: “place country:pt lang:pt (keyword OR ... OR keyword)”

The query for ambiguous words combined with insults: ”place country:pt lang:pt (keyword OR ... OR keyword) (insult OR ... OR insult)”

The collected data was also added to the data already collected in the previous project Hate-COVID19 that followed the same selection criteria. Time span We restricted the data collection to a time span of 2 years from January 1, 2021, to December 31, 2022.

Language We have filtered the collection to only tweets written in Portuguese but this way we had a huge percentage of PT-BR (Portuguese from Brazil) instead of the desired PT-PT (Portuguese from Portugal).

Geography To make sure we only retrieve tweets that were posted in Portugal, this way most of them will be pt-pt.

Keywords Within the kNOwHATE project, a list of 259 words and expressions, that are related to the targets in study (Roma, Racism, Xenophobia, LGBTIQA+), was created. In this

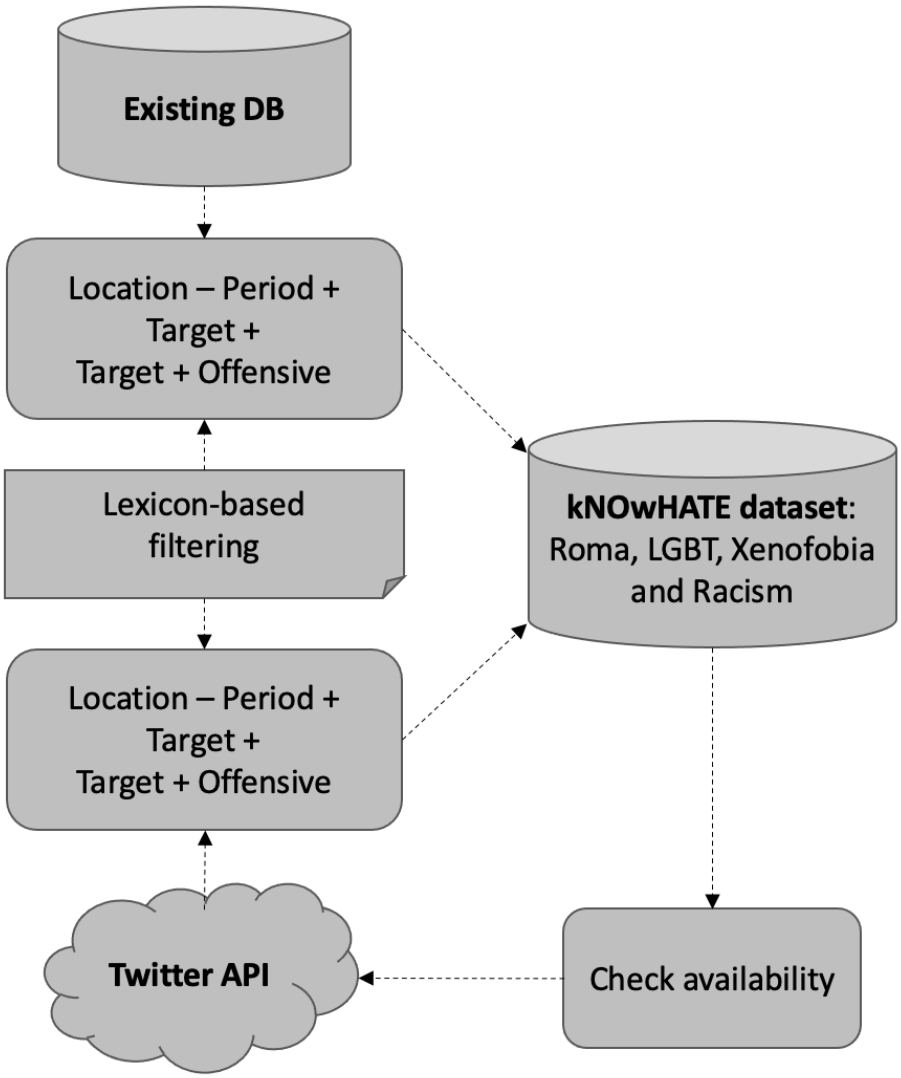

Figure 9. *Data Collection Pipeline Approach*

lexicon with have both ambiguous and non-ambiguous words/expressions, and because of that we had the necessity to create a different approach for each one of them. For the ambiguous words we decided to retrieve all the tweets that contain them, but for the non-ambiguous, we had to retrieve them only when combined with an insult. To be able to make the non-ambiguous words' retrieve we had to create a lexicon with approx. 800 mostly adjectives or expressions that are often used as a form of insult for the already mentioned targets.

As we can see from Table 3, there were already more than 27 000 tweets retrieved in the Existing DB but, for the enrichment of the study we decided to retrieve more using the Twitter API, there were of course overlaps of tweets that were both retrieved from the API and the Existing BD but we also had approx. 10 000 new tweets for our final data set. The total number was calculated as follows: Existing DB + Twitter API - Both

Data Source	Total
Existing DB	27 133
Twitter API	18 082
Both	7 297
Total	36 918

Table 3. *Distribution of all tweets collected by its source*

In addition to the collected tweets we also decided to retrieve the conversations to which they belong but only if the parent-tweet had been published in Portugal in order to mitigate the collection of Brazilian tweets that are not the focus of the study. Those conversations' tweets were collected using the Twitter API which used the following filter: "conversation id: ID" Because the parent-tweet does not have a conversation ID associated, a third extraction needed to be done, which was made by creating a list of the unique conversation id's present on the data set and retrieve tweets with those same ID's on the tweet id feature.

Parent-Tweet Tweet that originates the conversation, its conversation ID is created based on its own ID.

3.1.2. Annotation

For this study, we specifically focused on tweets that were part of the conversations identified during the second retrieval since, in order to study networks and the dissipation of hate speech, it is necessary to analyse the dynamics that are part of a conversation. Consequently, only 29,531 tweets underwent the following described process of annotation.

The data underwent annotation by a team of linguists, who meticulously identified various linguistic elements within the content. These annotations encompassed the discernment of direct hate speech, indirect hate speech, counter speech, and offensive speech. Additionally, the annotators specified the target of the speech, shedding light on the intended recipient or subject. Furthermore, the presence of specific sentiments such as Anger, Disgust, Fear, and Hope was systematically recorded, offering a comprehensive analysis of the emotional undertones embedded in the text. The linguists also scrutinized various linguistic features within the content, such as Lexical creativity, Metonymy, Metaphor, and many others.

All the 29,531 tweets were meticulously annotated by an interdisciplinary team of researchers with backgrounds in language sciences and social psychology, who meticulously identified various linguistic elements within its content. These annotations, crucial for our analysis aiming to understand the dissemination of hate speech, included spotting instances of direct hate speech, indirect hate speech, counter-speech, and offensive speech.

Additionally, the annotators identified the target mentioned in messages, shedding light on the intended recipient or subject. Each tweet can have more than one type of target and speech, depending on the context and content.

After the annotation, we did a preliminary analysis that showed the prevalence of tweets with no toxic or toxic-related speech, representing almost 83% of the dataset. Regarding the distribution of speech types in the dataset, Direct Hate Speech and Offensive Speech are present

in a smaller portion of the dataset. However, the values rise when looking into Indirect Hate Speech and Counter Speech, as we can see in Figure 10.

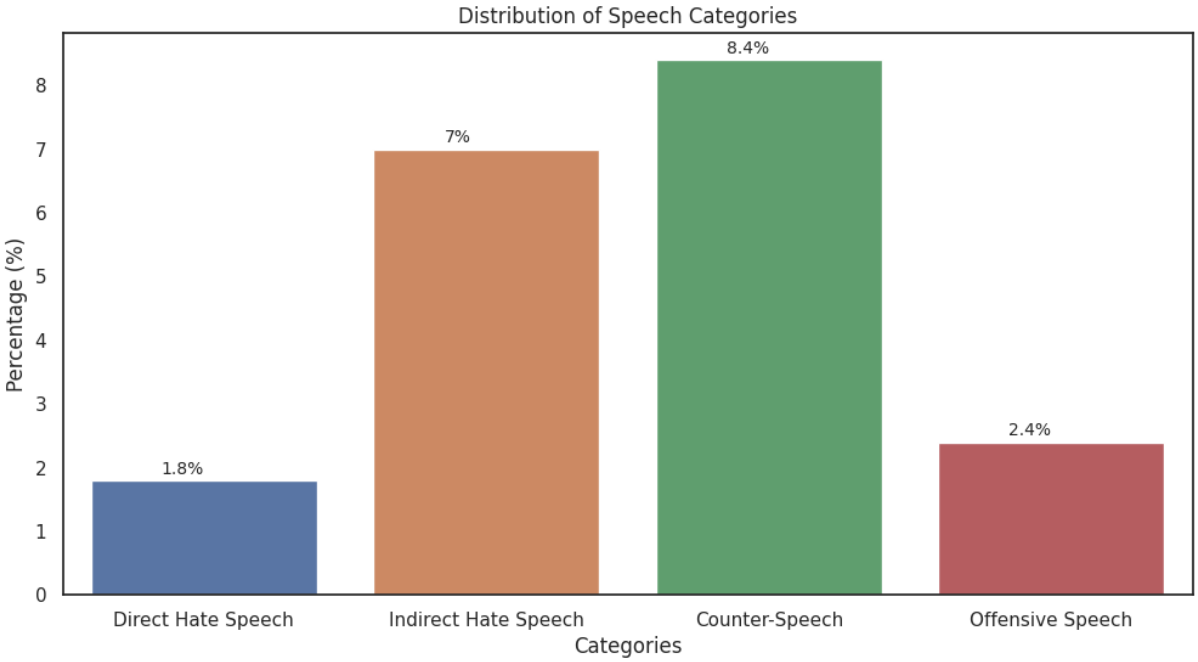

Figure 10. *Distribution of the type of discourse identified. Offensive speech, Direct Hate speech, Indirect Hate speech or Counter hate Speech.*

3.2. Data Understanding

In the Data Understanding chapter, we will thoroughly explore the fundamental characteristics of our data set. This will include a detailed analysis of the types of data present, identifying missing values, determining the total amount of data available and other relevant descriptive statistics.

By examining these characteristics, we will be better prepared to understand the structure and quality of our data, allowing us to make informed decisions about the most appropriate analysis methods to apply. This chapter will serve as a solid foundation for our study, providing a comprehensive overview of the context in which our analysis will be conducted.

Regarding the columns, there are 62 columns actively in use as seen in the table. The data collected consists of a total of 29531 instances, representing 2796 different conversations. Analyzing these conversations, we observed an average of approximately 10.5 tweets per conversation. Notably, the smallest conversation contains just two instances, while the largest extends impressively to 1070 tweets.

Column Name	**Description**	**Data Type**
c_id	Conversation ID	Numeric
tweet_id	Tweet ID	Numeric

in_reply_to_tweet_id	In-reply-to Tweet ID	Numeric
user_id	User ID	Numeric
time	Tweet Timestamp	Date & Time
username	Username of the tweet's author	Text
text	Tweet Text	Text
c_len	Conversation Length	Numeric
alvo_c	Conversation Target	Text
alvo	Target Categorical	
D. O. Direto	Direct Hate Speech	Binary
D.O. Indireto	Indirect Hate Speech	Binary
Contra-discurso	Counter-speech	Binary
D. Ofensivo	Offensive Discourse	Binary
Racializadas	Racialized	Binary
Ciganas	Anti-Gypsy	Binary
LGBTQA+	Anti-LGBTQA+	Binary
Migrantes	Anti-Migrant	Binary
Interseccional	Intersectional Discourse	Binary
Outra	Other	Binary
Fav. Endogrupo	Presence of In-group Favoritism	Binary
Der. Exogrupo	Presence of Out-group Derogation	Binary
Zero-sum	Presence of Zero-sum Logic	Binary
Criatividade lexical	Presence of Lexical Creativity	Binary
Metonímia	Presence of Metonymy	Binary
Metáfora	Presence of Metaphor	Binary
Comparação	Presence of Comparison	Binary
Hipérbole	Presence of Hyperbole	Binary
Apelo ao medo	Presence of Fear Appeal	Binary
Apelo à ação	Presence of Call to Action	Binary
Perg. retórica	Presence of Rhetorical Question	Binary

Ironia / humor	Presence of Irony / Humor	Binary
Intertextualidade	Presence of Intertextuality	Binary
Outras Falácias	Presence of Other Fallacies	Binary
Outro	Presence of Others	Binary
Inversão de papéis	Presence of Role Reversal	Binary
Estereótipo	Presence of Stereotype	Binary
Negação do ódio	Presence of Denial of Hate	Binary
Desumanização	Presence of Dehumanization	Binary
Silenciamento	Presence of Silencing	Binary
Ameaça Realística	Presence of Realistic Threat	Binary
Ameaça Simbólica	Presence of Symbolic Threat	Binary
Recategorização	Presence of Recategorization	Binary
Into contra-estereotipica	Presence of Counter-stereotypical Information	Binary
Empatia	Presence of Empathy	Binary
Normas legais	Presence of Legal Norms	Binary
Normas sociais	Presence of Social Norms	Binary
Positivo	Presence of Positive Emotions	Binary
Negativo	Presence of Negative Emotions	Binary
Ódio	Presence of Hatred	Binary
Raiva	Presence of Anger	Binary
Nojo	Presence of Disgust	Binary
Medo	Presence of Fear	Binary
Culpa	Presence of Guilt	Binary
Vergonha	Presence of Shame	Binary
Esperança	Presence of Hope	Binary
Obs:	Additional Observations	Text
in_reply_to_user_id	ID of the user to whom it responds	Numeric
HateSpeech	Presence of Hate Speech (Any type)	Binary
conversation_id	Conversation ID	Numeric

HS_conversation	Tweets with Hate speech in the Conversation	Numeric
Ratio_HateSpeech	Proportion of hate speech in the conversation	Numeric
pshift	Participation Shift	Categorical

Table 4. *Description of columns in the dataset.*

3.3. Tweets Analysis

Tweet analysis plays an important role in understanding the data we are investigating. By exploring and examining tweets in depth, we can gain valuable insights about the characteristics of the discourse present on X/Twitter. This analysis also helps us to better understand the nuances and dynamics of online social interactions. Therefore, in this chapter, we will focus our attention on the analysis on all the conversation tweets.

3.3.1. Top Words

The frequency of the top 20 words present in the tweets was done after putting all letters in lowercase and after deleting all stop words. In most cases, as seen in Table 5 the most frequent words in the tweets are directly related to the target as is the case of gay/gays to LGBT, however, there are words that by themselves have nothing to do with it such as ”transportes” and ”transmissão” to LGBT and therefore can be studied in more detail as they may indicate some specific event that may have happened.

LGBT	RAC	XEN	CIG
3107 gay	1495 racismo	733 zuca	1049 lelo
1404 pra	1081 racista	564 brasileiro	546 ciganos
1057 gays	979 privilégio	558 xenofobia	282 rt
1004 transportes	746 pessoas	400 zucas	259 cigano
988 vai	744 cor	366 brasileiros	213 cigana
964 pessoas	553 pra	332 pra	122 lelos
930 boiola	521 ter	289 portugal	89 wee_lelo
900 ter	510 racistas	263 brasileira	77 pra
888 portugal	489 pessoa	232 zucacritica	74 todos
866 dia	478 todos	222 brasil	73 porque
809 lgbt	458 portugal	214 vai	72 tudo
762 ainda	437 vai	201 xenófobo	70 pessoas
743 bem	421 preto	194 aqui	69 bem
724 trans	416 africanos	191 bem	67 comunidade
722 tudo	408 catinga	182 merda	66 ciganada
713 transmissão	385 ainda	182 rt	64 aqui
708 porque	370 tudo	155 vergonha	64 etnia
687 ver	365 bem	152 país	62 casa
675 todos	358 ver	146 ver	62 ventura
669 aqui	356 sobre	143 racista	60 agora

Table 5. *List of the 20 most frequent content words, considering our four targets.*

3.4. Tweets Distribution over the time

The final data set consists of 29,531 tweets where 2 372 are referring to the LGBT (LGBTIQA+) target, 1 177 to RAC (Racism), 737 to XEN (Xenophobia), and finally 257 to CIG (Roma). There was a greater trend for tweets with alleged hate speech in the year 2022, with the most significant increase being in the LGBT target.

Figure 11 shows how the tweets are distributed over time (by month-year) containing one line for each target, this way we can analyse how the different targets behave throughout the time span. Tweets referring to the LGBT target are the ones that are more common but are the most irregular class with a lot of ups and downs throughout the time span. Besides that there was a huge peak in the number of tweets in June and July of 2022 this was probably because June is the International Pride Month and there was also a controversy over a professor from the University of Aveiro who was suspended for allegedly making homophobic comments [45].

The other classes are relatively constant with the exception of August 2022 in the Racism and Xenophobia targets, which saw an increase. This increase can be related with the controversy surrounding Giovanna Ewbank's children [46], which was highly engaged with on social media.

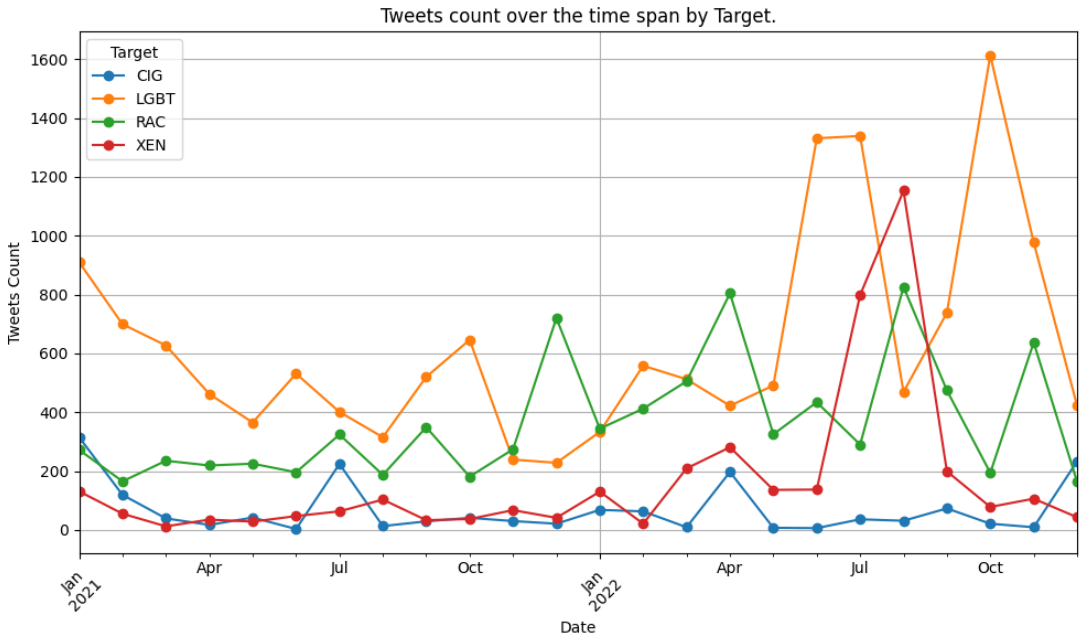

Figure 11. *Tweets count over the time span by Target.*

Figure 12 shows an increase in tweets in 2022 compared to 2021 by approximately 70% of the total can provide valuable insights into trends or changes in social media activity over time.

In 2021, because of the permanence of COVID-19 restrictions, many people could not leave their home. This meant less face-to-face interaction, which might have led to fewer situations where people felt the need to go to online spaces and use hate speech when reporting about events during their day.

	Number of Conversations	Number of Tweets
Roma	143	1 646
LGBT	1 674	15 149
Racism	693	8 756
Xenophobia	256	3 943
Other/Intersecctional	30	37

Table 6. *Number of Conversations and Tweets by Target.*

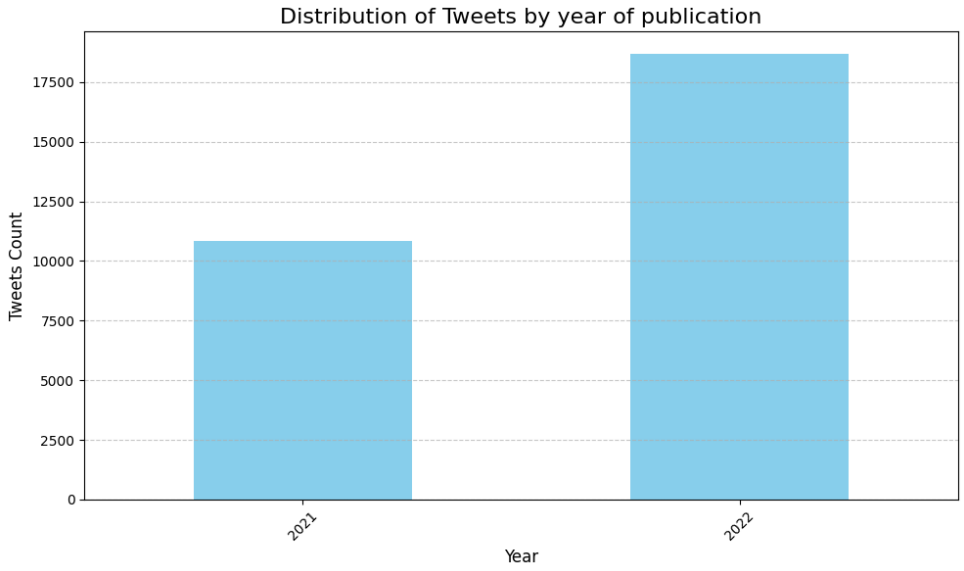

Figure 12. *Tweets Count by Year of Publication*

3.4.1. Conversation Tweets

As already mentioned, it was felt that collecting the conversations from our already collected tweets could be a way to enrich the study. That said, it was necessary to impose a rule, to only collect conversations in which the parent tweet is of Portuguese origin because in other attempts to collect without this rule, we ended up collecting conversations with more than 10 000 tweets but these belonged to the Brazilian community, which is not the object of our study. This way we get a data set with 29 531 occurrences referring to 2 796 different conversations for all the targets, being LGBT the target with the most associated conversations as we can see in Table 6.

There was a drastic decrease in the number of conversations now being studied compared to the first retrived data but this is because there are a lot of tweets without an associated location and therefore these are not retrieved and also because some of them don't have any replies so there is no conversation to be extracted. Regarding the characterization of the conversations, the average number of tweets in the conversation is 10, the minimum is 1, and the maximum is 1070 tweets. Most of the conversations are in the group of conversations with between 1 and 49 tweets, as shown in Figure 13.

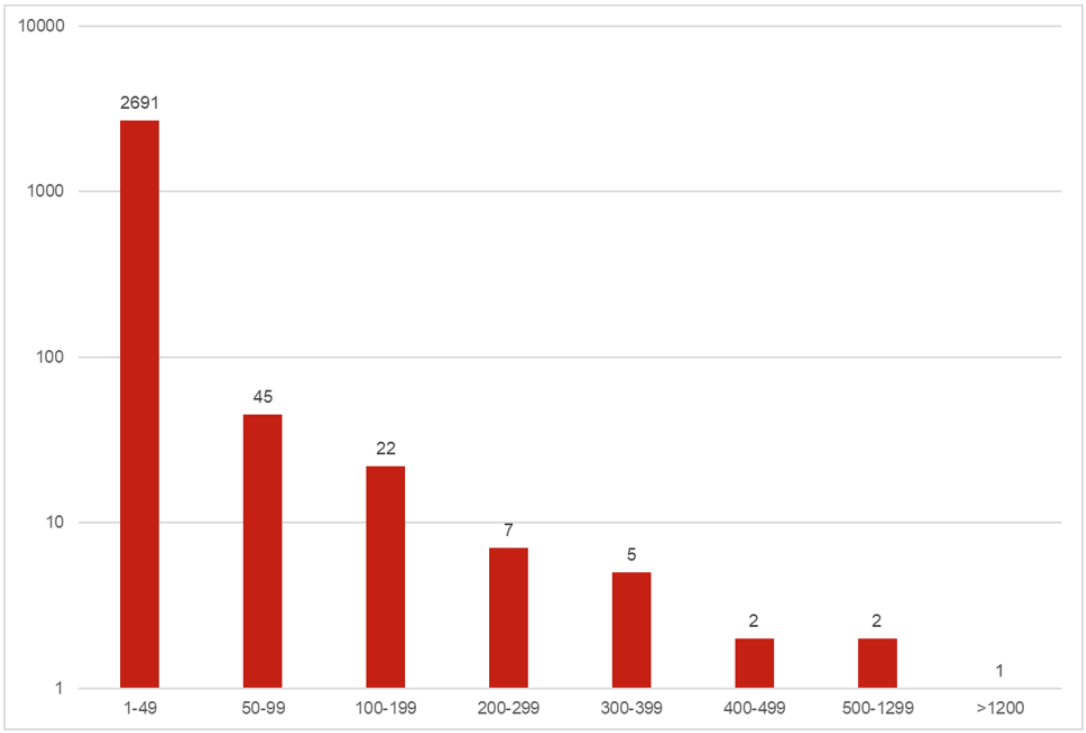

Figure 13. *Distribution of Conversations by the Number of Tweets in it.*

3.4.2. Unavailable Tweets

As there were two collections in different time slots it was possible for us to verify tweets that are no longer available for different reasons, which can be: Deleted Tweets, Unauthorised, Suspended User, or Blocked from User, its distribuition is presented in Figure 14. Tweets can get deleted whether after getting many reports from other users or because the author wants to delete them.

To extract this data, we used the Twitter API. The API allowed us to retrieve all the tweets that were already on the existing DB. However, there were instances where the API was unable to retrieve certain tweets. In such cases, an error was pointed out, which we collected and saved in a different column of our final data set. The fact that we were able to analyse tweets that have been removed from the platform provides a unique advantage for our study. It allowed us to gain insights from the content of tweets that may have otherwise gone unnoticed or unaddressed.

In the Existing DB there were 5 054 tweets that were not available on March 7th of 2023, from those tweets it was analysed in more detail to which target they belonged, and it was concluded that the LGBT target is the one with the most removed tweets as seen in Figure 15. While we cannot definitively say whether the removed tweets contained hate speech without analyzing them one by one, it is likely that they did. The fact that the LGBT community was the most targeted group suggests that there may be a larger issue of discrimination and hate speech on social media platforms. As we can identify in Table 7 the top 1 most frequent word in those tweets is also related to this target: ‘gay’ with 514 occurrences but aggregating that with the word ‘gay/s’ makes 761 occurrences. The words ‘racismo’ and ‘racista’ were not aggregated because they can lead to different kinds of discourse.

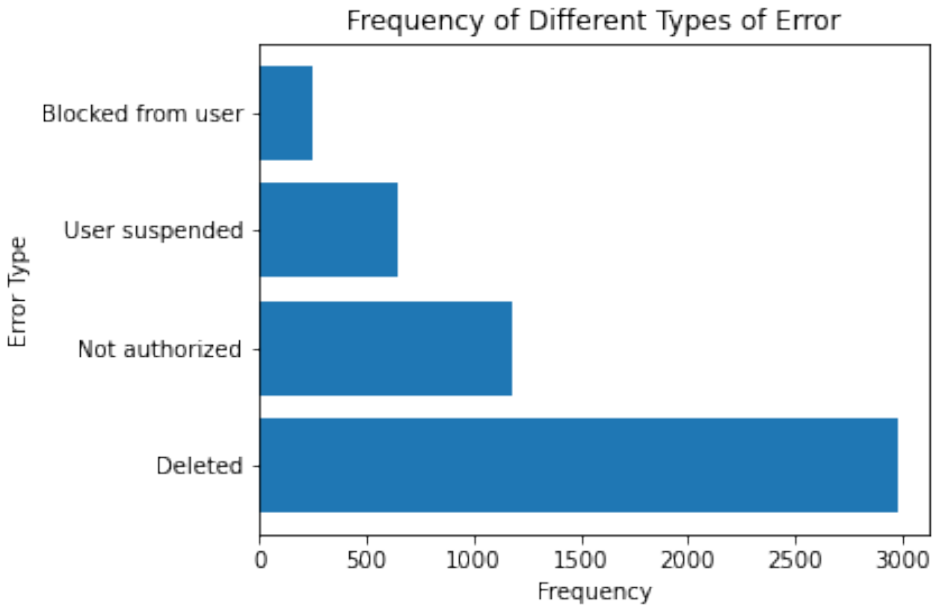

Figure 14. *Number of occurrences by each motive why tweets are no longer available*

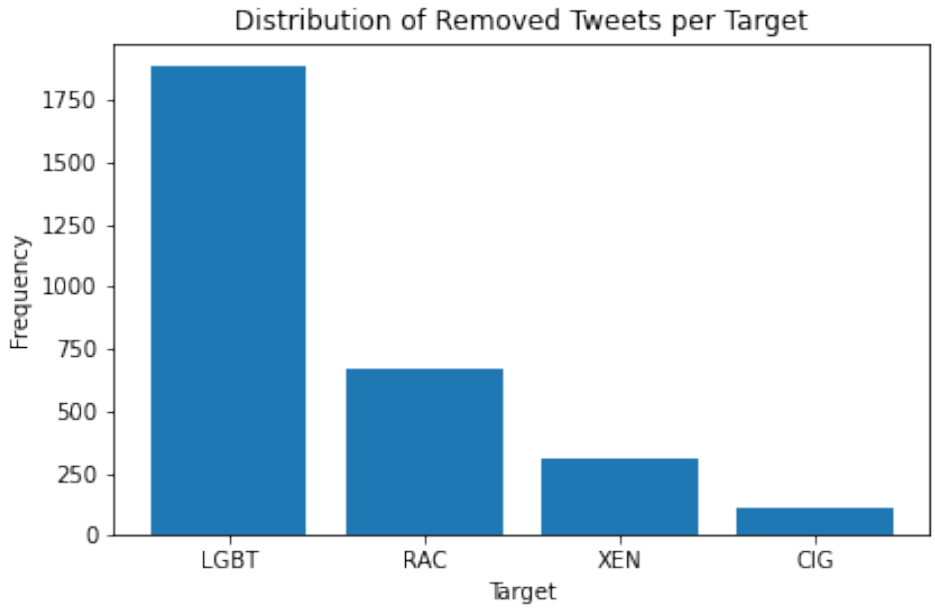

Figure 15. *Number of unavailable tweets by each target.*

3.4.3. Preliminary Corpus Analysis

The first procedure is the removal of links and URLs from the text. This step is vital because links and URLs don't give any significant information for the analysis of the text, and they can in some cases alter the results. Deleting them helps to guarantee that the analysis is centered exclusively on the content substance.

Top 10 Most Frequent Words on Unavailable Tweets
761 gay/s
349 pra
294 racismo
255 racista
231 vai
216 pessoas
214 portugal
204 dia
192 ter
186 tudo

Table 7. *Top 10 most frequent words on Unavailable Tweets.*

The second procedure applied was tokenization, which includes breaking the content into individual words or tokens. This step is one of the principal assignments in NLP and is fundamental for numerous other text-processing techniques.

Stop word removal is another important NLP strategy that needs to be done. Stop words are common words that are regularly expelled from content information since they don't carry much meaning on their own. By removing stop words, the examination can focus on the more significant words within the content instead of being distracted by words that do not add anything.

Then it was collected the frequency of each remaining word within the text data. This method distinguishes the most frequent words, which can give insights into the most common topics or themes in the text. Word clouds are a visualization strategy commonly utilized in NLP (Natural Language Processing) to speak to the most frequent words in a text corpus.

Figure 16. *WordCloud for Conversation Tweets.*

Finally, it was performed a sentiment analysis of the tweets to identify and extract subjective data from the text, such as opinions or emotions. We used TextBlob to calculate the extremity of each text in the text column, which represents the sentiment of the content as positive, negative, or impartial.

In addition, it is also leverage to know the average word count per tweet, as well as the minimum and maximum values because by analyzing the distribution of word counts across the tweets, we can better understand the typical length of messages in our dataset and potentially identify any outliers or patterns in the data. That said, on the originally collected tweets the average length was 21 words whereas on the conversation tweets it was 19. About the maximum and minimum, there were no huge differences, the minimum was 1 for both and the maximum was 66 for the originals and, with only less 2 words, 64 for the conversation tweets.

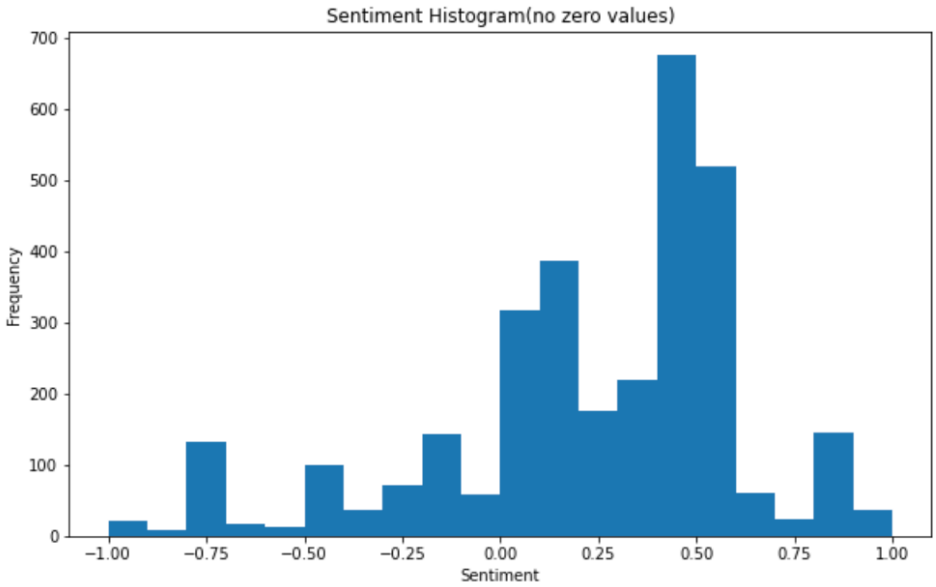

Figure 17. *Sentiment Analysis of the Conversation Tweets without zero values.*

CHAPTER 4

Social Network Analysis

4.1. Whole Graph

We aim to find if there are any communities by analyzing the interactions between users. In this network, each node represents a different user that is present in our dataset and each edge between nodes is a response from one user to another, these edges are directed, starting with the user who is and ending with the user which the other was replying to (source: *user_id*; target: *in_reply_to_user_id*). Because each interaction between users can have a different approach we decided not to merge parallel edges, this way every edge has the same weight of 1.

Our network has 9952 nodes and 24532 edges and most of the graph is connected, but around it, we can see some isolated groups.

The figure 18 shows the directed graph of all the tweets collected, except for those that either replied to a tweet that subsequently became unavailable or is just a parent tweet but with no reply available.

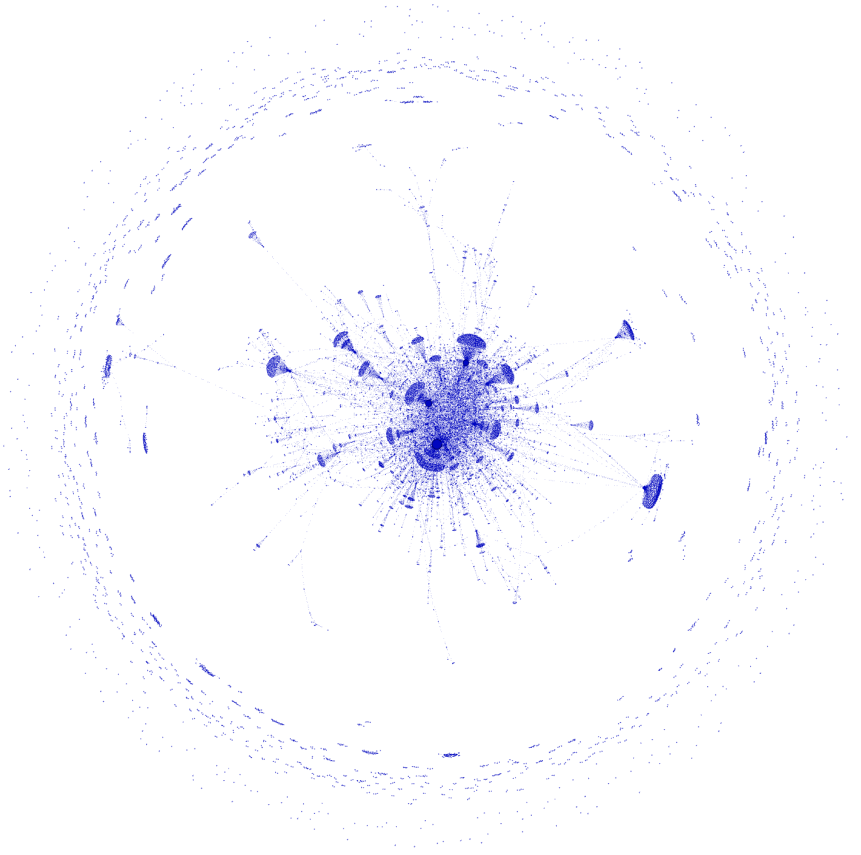

Figure 18. *Graph Representation of all the Conversations Data.*

Table 8. *Whole Graph Metrics.*

Number of Nodes	9 952
Number of Edges	24 532
Average Degree	2.465
Network Diameter	20
Average Path Length	5.53
Power Law Alpha	2.2966

Just by looking at it, we can see that there are three groups that behave in completely different ways. The central core is practically all interconnected, which means that it ends up functioning as kind of a community. Around it, we can see small clusters that represent small groups of people who interact only within that circle. And finally, even further away from the center, the users that only had responses from themselves and were not connected to any other node.

4.1.1. Density

This network has a very low-density value, 0.000159, so it is categorized as a sparse network. This means that the number of links in the network is much lower than the maximum possible in this network. In general, real networks tend to be sparse [47], meaning that they cover a large area but they are not well connected.

4.1.2. Diameter

The Diameter of this network is 20, which means that the maximum number of edges you have to traverse to get from one node to another in the shortest way possible is 20. This number might indicate that there are certain nodes that are not directly or closely connected to most other nodes which can affect the efficiency of information flow or communication in the network.

4.1.3. Degree

With the help of Figure 19 and the data in Table 8, we can see that most of the nodes have a low degree, while the nodes with a higher degree are less frequent. Apart from these outliers which have a high degree, we can say that the network is homogeneous. The average degree of this network is 2.465, which means if every node had the same degree it would be 2.465, but the problem with averages is that most of the time is not representative of reality, in this case, there are even nodes with a degree greater than 100. Additionally, it is crucial to mention that the degree distribution follows a power-law, indicating that the network is scale-free, as expected. This property is highly significant in Network Theory as it determines the behaviour of the complex system, offering insights into its structure and dynamics. By applying power law analysis to visualize the distribution of node degrees in a network, we can better capture the its structure. This provides more accurate modelling and deeper insights into the behaviour and dynamics of this network.

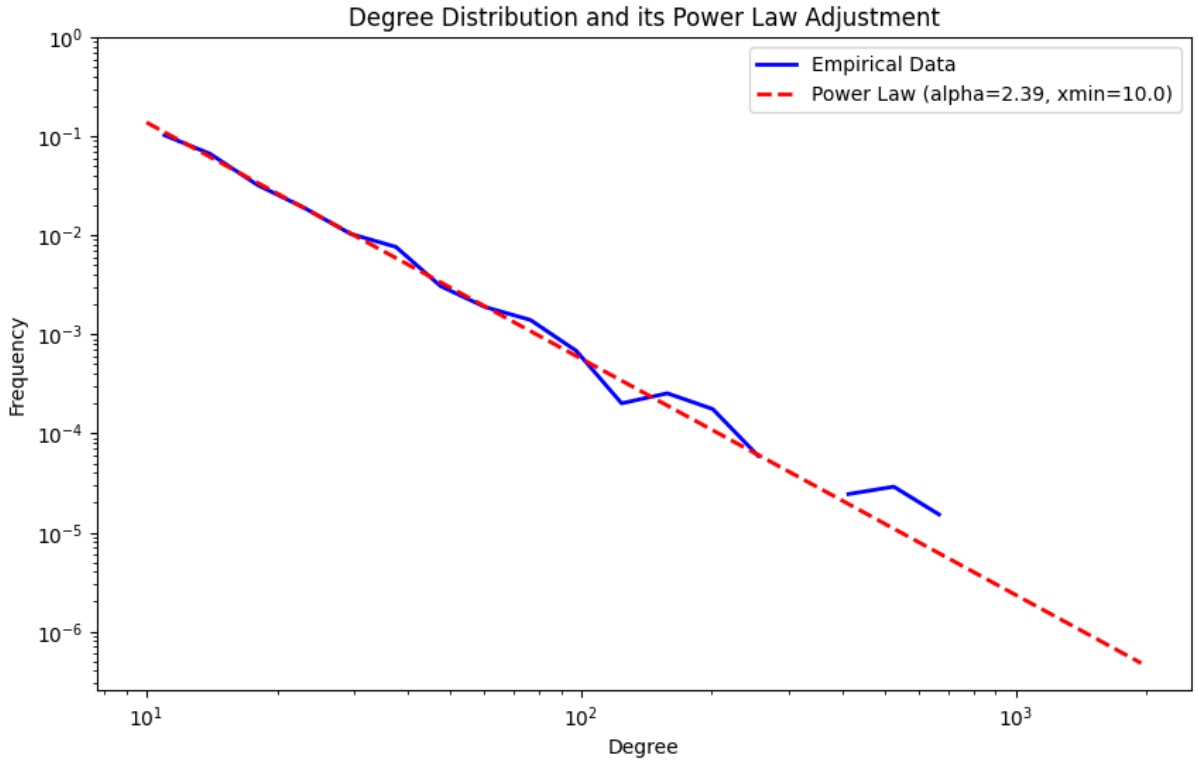

Figure 19. *Degree Distribution and its Power Law Adjustment.*

By applying power law analysis to visualize the distribution of node degrees in a network, we can better capture its structure. This provides more accurate modelling and deeper insights into the behaviour and dynamics of this network.

4.2. Central Core Analysis

The central core, as represented in Figure 20, contains approximately 69% of all the nodes in the original graph and 78% of the edges. These numbers mean that the central core is more connected than the peripheries, as it has more edges for fewer nodes. In this network, all the nodes are connected, so there is no isolated node. This graph ensures the flow of information because there are no isolated nodes.

This central core has a high significance on the graph, this means that any alteration in that may have a big impact on the network's dynamics. For example, if we remove the highest degree node there is going to be the need to reconfigure the communication pathways, which might affect the efficiency and speed of information flow dissemination.

4.2.1. Betweenness Centrality Analysis

In network analysis the Betweenness Centrality measure is highly used to identify the importance of a given node within the network. The importance of the node translates to how many times this node acts like a bridge or intermediary in the path between the other two nodes. The nodes with higher Betweenness centrality are normally used as a way to get a wider audience and as a connector between different clusters within the network. Frequently these users are

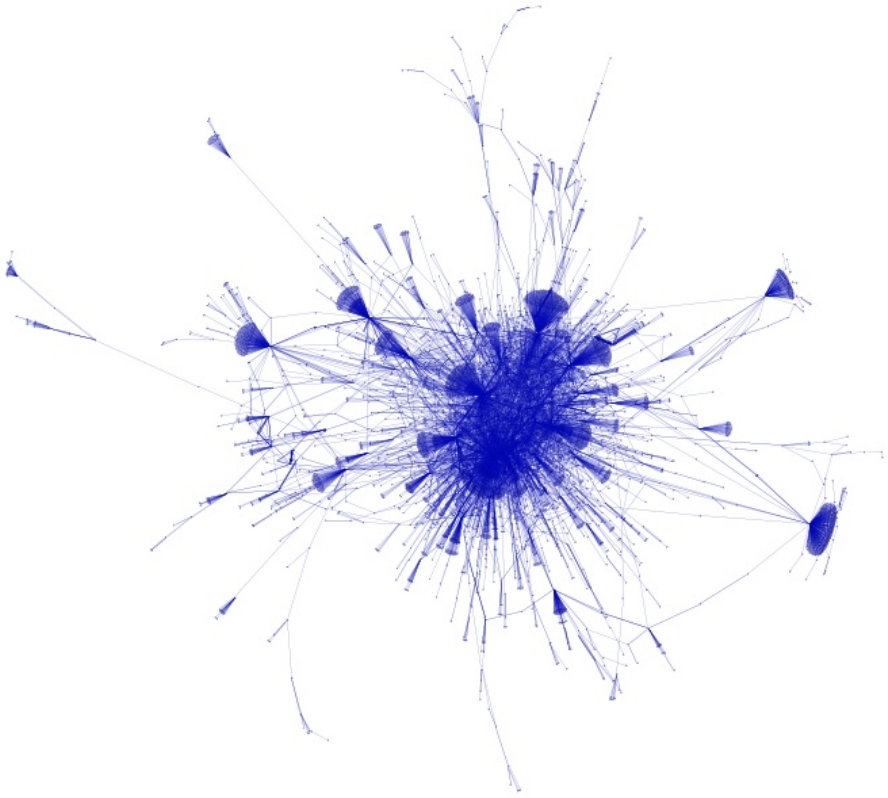

Figure 20. *Graph Representation of the Central Core.*

opinion leaders. In our network, we can highlight a few users that fit in that description, as shown in Figure 21.

Looking at the relative magnitude of the Degree Centrality (normalized in the dataset) of the Top 10 users with higher betweenness centrality, shown in Figure 22, are well involved in the discussions comparing to the data set, since their Normalized Degree seems to be correlated with the normalized betweenness centrality. To support this analysis we calculated the betweenness centrality correlation with degree centrality which is 0.89.

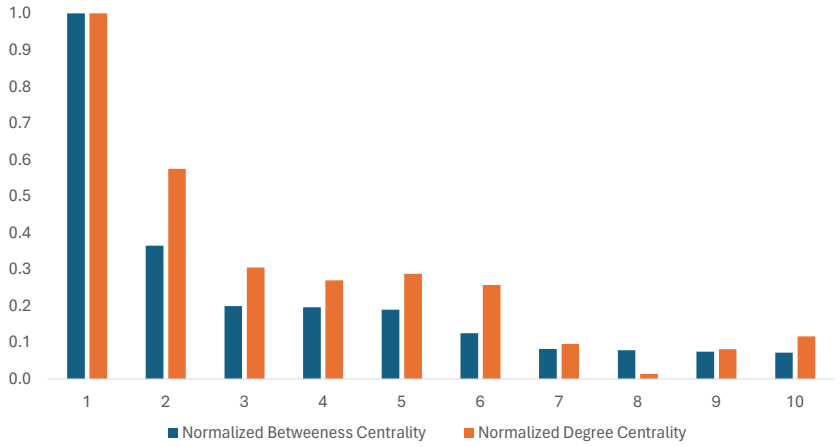

Figure 22. *Relative Magnitude of the Degree Centrality and Betweenness Centrality of the Top 10 users.*

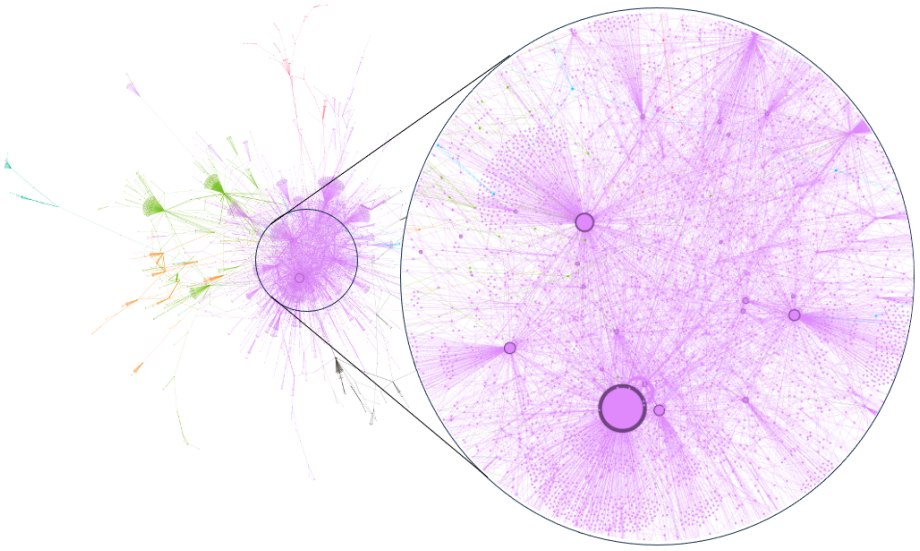

Figure 21. *Graph Representation of the Betweenness Centrality on Central Core.*

4.2.2. Closeness Centrality Analysis

Unlike Betweenness Centrality, closeness centrality measures inhow many steps does a given node need to reach the other nodes in the network. This means that when looking for high closeness centrality nodes we are facing a user that has a wide direct audience.

4.3. Peripheries Analysis

This part of the network is highly disconnected with just a few clusters of users that seem to create a small community. The peripheries of the graph, represented in Figure 23 may seem smaller than it looks because its nodes represent around 30% of the whole graph. In this part of the graph users do not commonly participate in more than one conversation and most of the conversations are rather small.

From now on we will proceed with the study with only the data from the core of the graph, we decided to use this approach so we can focus our study on a group of people that is more present in the community and that is more connected in-between.

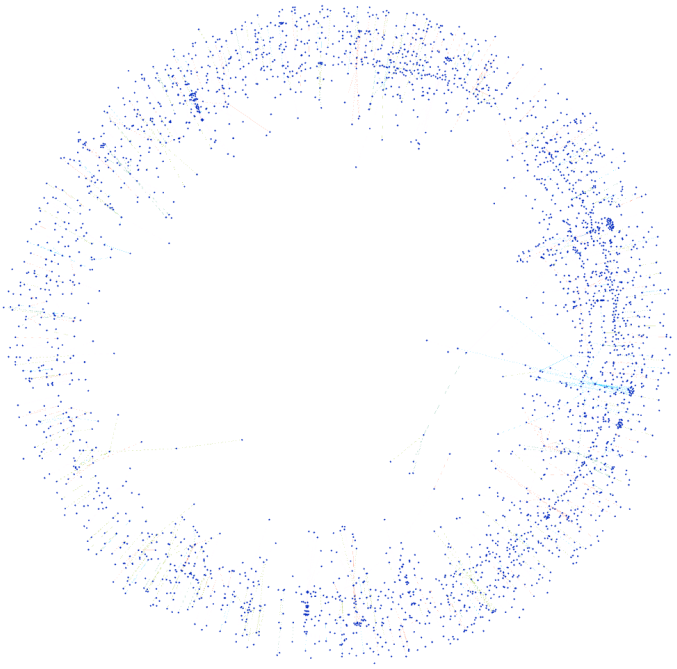

Figure 23. *Graph Representation of the Peripheries.*

4.4. Target Analysis

In this study we are analyzing, as shown previously, 4 target groups: LGBT, Racism, Xenophobia, and Roma. Because there were also tweets that did not particularly fit in any of these targets we created a new category: 'other'.

It is important to reinforce that only the edges have a target associated because they represent the tweet itself whereas the node is the user that may discuss every topic. After using the target feature as a colour partition for the edges we can see that there are groups of users that are extremely focused on that topic, but for example, this does not mean that they are literally being racist instead they are talking about race-related content.

The distribution of the used topic may differ when analyzing the hate, because of that with Figure 24 we can only assume that LGBT is the topic with more tweets about it.

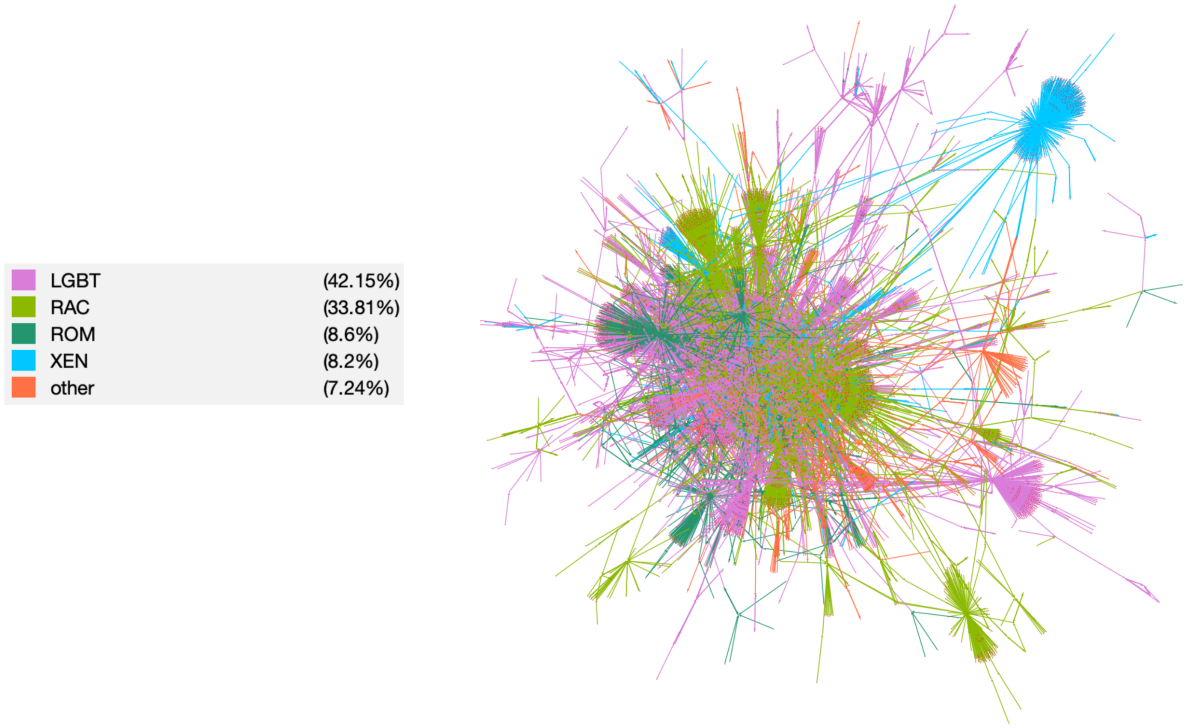

Figure 24. *Graph Representation of Edges Target in Central Core.*

4.5. Hate Speech Analysis

Around 12% of the users were involved in an exchange of tweets that indicated the presence of hate speech either direct or indirect. In this hateful subsection of the graph, we have 1181 users who are involved and 1575 edges that represent their interactions.

Furthermore, the previously analysed users with the highest betweenness centrality are present in this graph, represented in Figure 25, which may indicate a huge problem for the dissipation of hate in social media.

4.5.1. Density

This sub-network has a still low-density value, 0.001, but it is way higher when compared with the whole graph, nevertheless it is still relatively sparse, as the proportion of edges is quite low compared to the total number of possible edges in the graph.

4.5.2. Diameter

The Diameter of this subsection of the network is 10, which is less than the diameter of the whole graph. This decrease indicates that the nodes in this subsection are relatively more closely connected compared to the rest of the network.

4.5.3. Degree

With this filter to only study the users and edges related to hate speech, the graph became better connected because a lot of the not-so-many active users were deleted. While better connected it does not mean that it will have a higher average degree, in this case, the number decreases

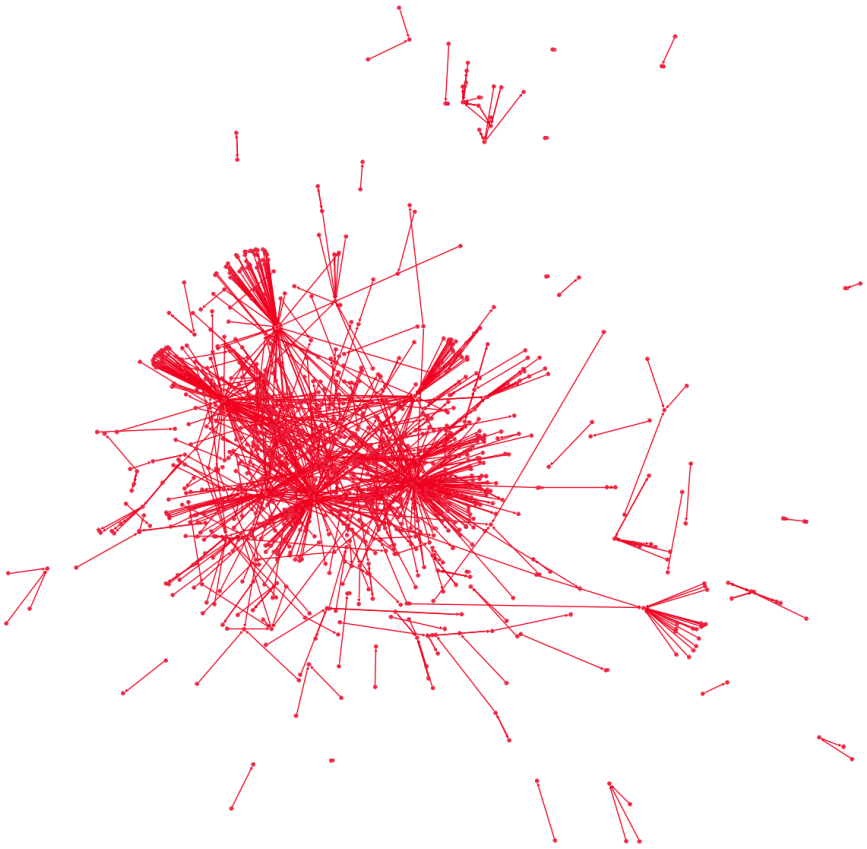

Figure 25. *Graph Representation of Hate Speech in Central Core.*

to 1.334 which means that if every user had the same number of edges attached, this number would be 1.334.

As seen in Figure 26 we can understand most nodes have low degrees, which means few connections, and the distribution decreases rapidly to around 25 degrees and then stabilizes. This is common in many real-world networks [48], where most nodes have a small number of connections, and only a few have a very large number.

There is a smaller proportion of nodes in the network that have a higher number of connections, but these nodes are still present and contribute to the overall connectivity of the network.

4.5.4. Hate Speech Types

The hate speech can either be direct or indirect, before analyzing which are the cases in our study, it is important to better understand both of these definitions.

Both forms of hate speech are extremely harmful and can lead to violence and other forms of discrimination, but there are some differences between direct and indirect hate speech.

Direct Speech is more explicit where the bully uses abusive, toxic, and derogatory language to put people down. On the other hand, indirect hate speech is in a more subtle form, for example, it can be in the form of a joke, metaphor, euphemism, or rhetorical question.

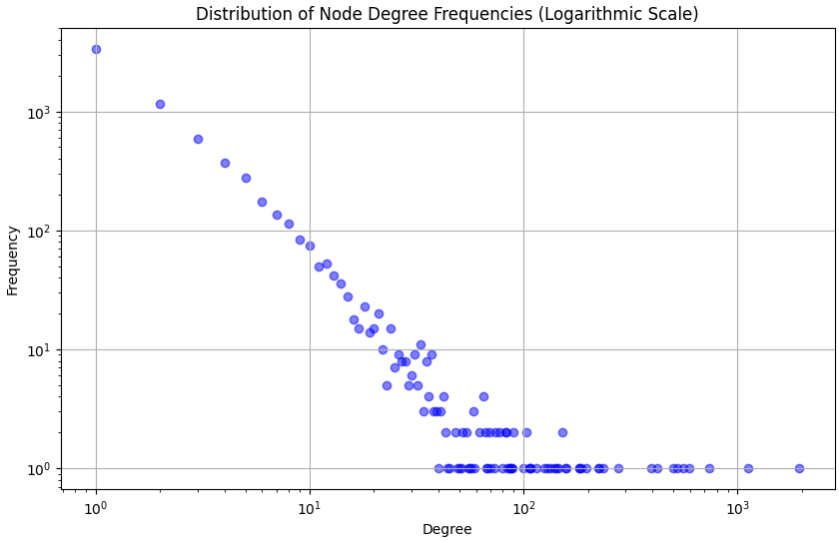

Figure 26. *Distribution of Node Frequencies (Logarithmic Scale).*

In our data, about 6% of the interactions use indirect hate speech, and almost 2% use direct. This may look like a small number, but in the case of hate speech even one would be a problem.

4.5.5. Targets of Hate Speech

In the Twitter user interaction graph, almost half of the hate speech incidents are targeted towards the LGBT community, while approximately 30% of the hate speech is directed at race. If we look closely enough at the graph represented in Figure 27 we can see that each user is related, almost every time, to only one target, this shows that there are already communities that focus on only one target instead of just wandering around or sharing opinions about other target groups.

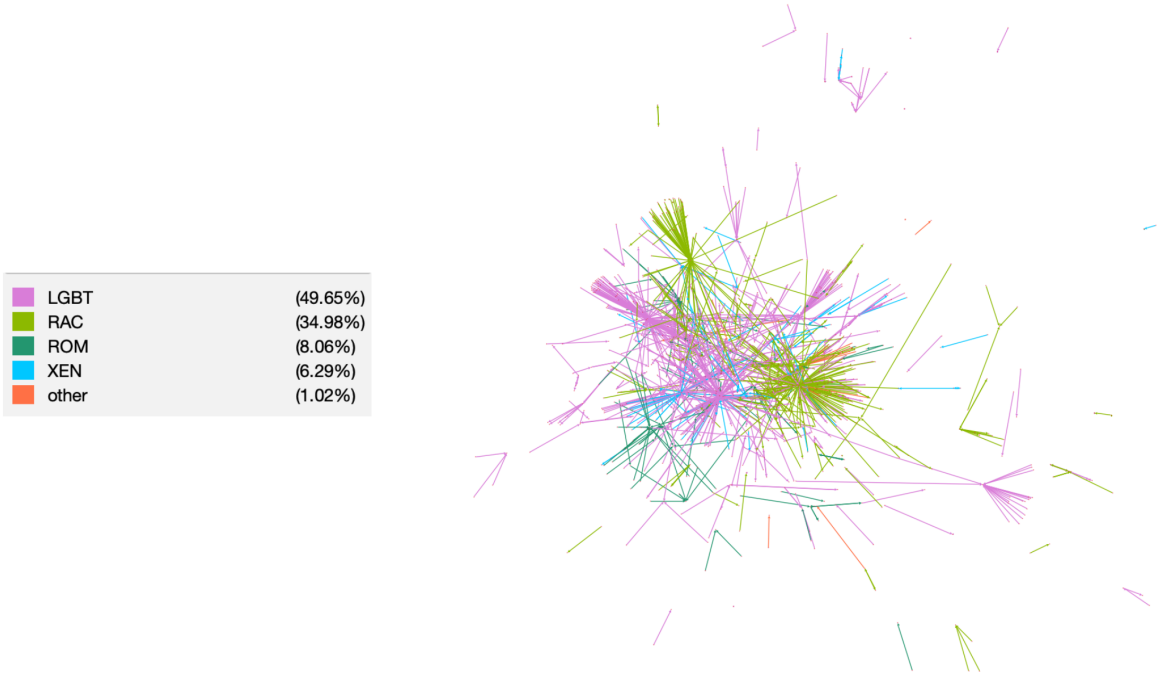

Figure 27. *Graph Representation of Target in Tweets with Hate Speech.*

4.6. Conversation Example Analysis

Although all the conversations are represented in the general network, we also chose to take a more micro approach where we analysed some of the conversations individually. For the most part, all the conversations have the appearance shown in figure 28, where one user remains in the center (being the author of the parent-tweet) and most of the other users respond directly to them. In this example we are analyzing a conversation with 199 tweets and 169 users. It is important to acknowledge that some tweets were replies to a tweet that was unavailable at the time of retrieval. As a result, they appear disconnected from the network. However, if the unavailable tweet were accessible, it would serve as the connector, linking these tweets to the rest of the conversation. The graphic representation of Figure 28 show some scattered connections, where some of them are more dense than others. This reflects the presence of 9 components in the network which shows that this network is not all connected. The average degree in this network is 1.93 which is highly bias since the majority of the nodes reply to the same node, which is the user from the parent-tweet.

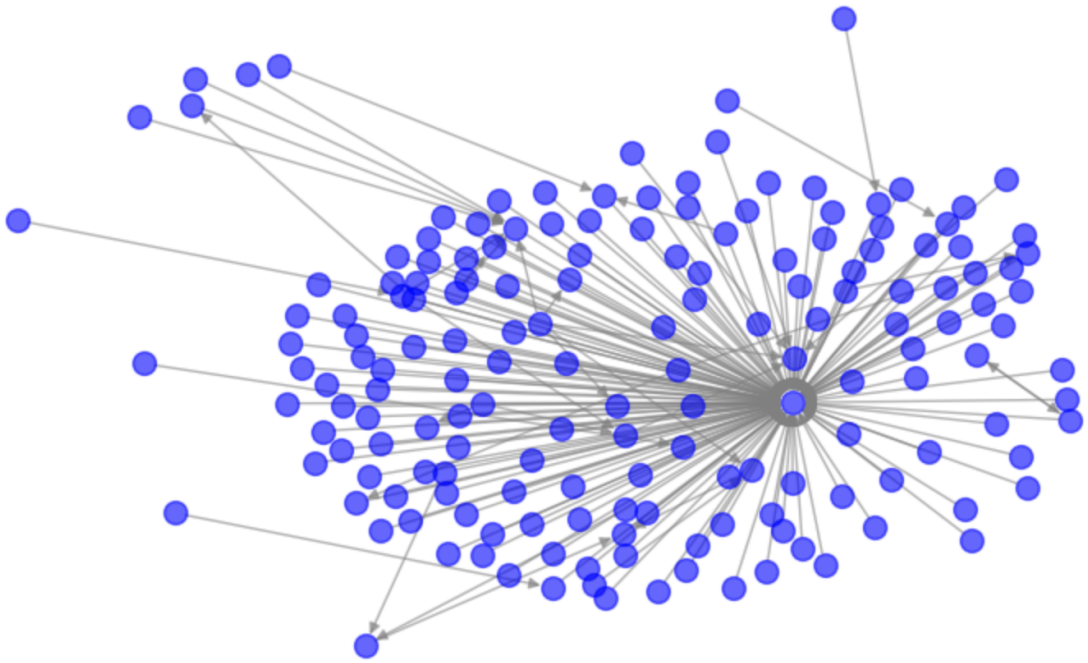

Figure 28. *Example of a Conversation with 199 tweets.*

As we can see in Figure 29, and as expected, the author of the parent-tweet has the highest degree centrality among the participants in the network analysed. This phenomenon occurs because the author of the original tweet receives the largest number of direct interactions. In the structure of this specific network, most users choose to reply directly to the initial tweet's author, rather than engage with other users in discussions that would develop like a tree with multiple branches and different users entering the conversation. As a result, interaction tends to concentrate around the original tweet, and in this case around its author, reinforcing the centrality of the parent-tweet author in the context of this conversation network of interactions.

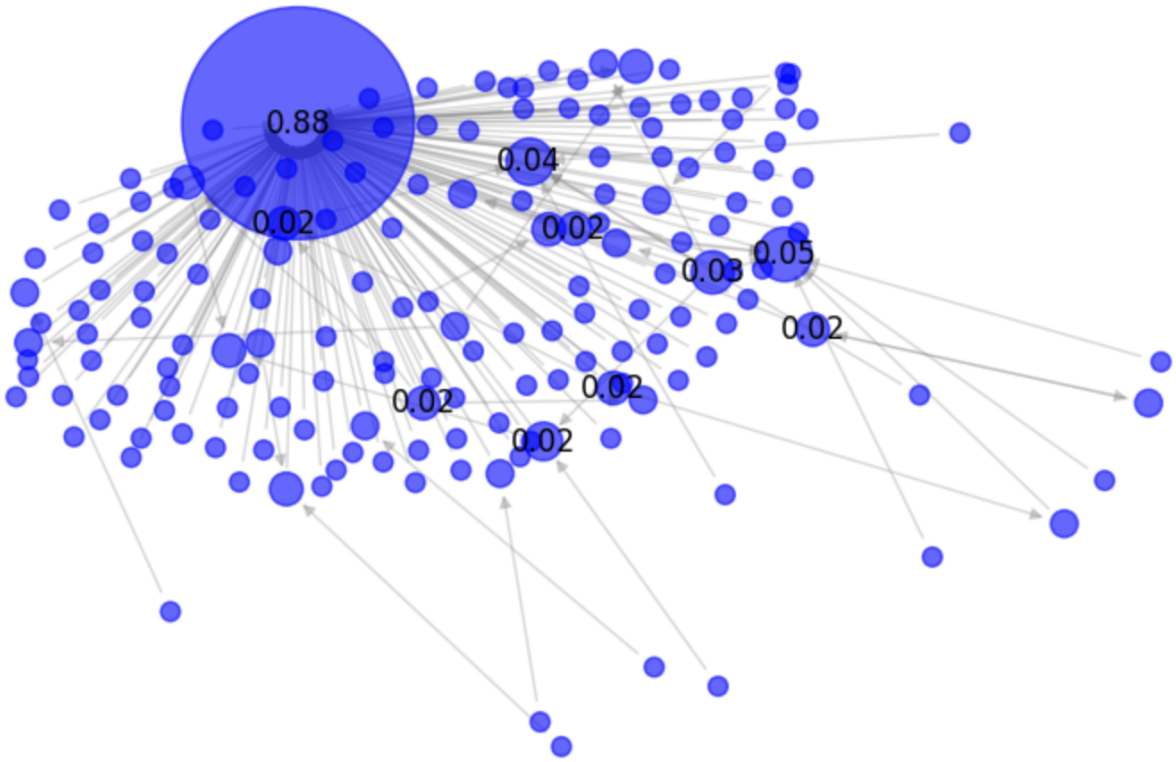

Figure 29. *Conversation Network with Degree Centrality.*

4.7. Conversation Dynamics

As proposed by Gibson [2] we used the framework called Participation Shifts (P-shifts) to classify individual tweets within a conversational context.

This framework delineates several distinct types of P-shifts, which can be further organized into four categories: Turn-Receiving, Turn-Claiming, Turn-Usurping, and Turn-Continuing.

Turn-Receiving is the action of an individual assuming their turn after to being directly addressed. Turn-Claiming occurs when an individual interjects the conversation after someone else addressing the group as a whole rather than a specific individual. Turn-Usurping refers when one participant gets control over another's turn. Finally, Turn-Continuing is when a participant who, while already engaged in the discourse, redirects its targets.

To better understand the categories we added some examples in Table 9.

P-shift [1]	
Turn receiving	
AB-BA	John talks to Mary, then Mary replies.
AB-BB	John talks to Mary, then Mary talks to herself.
AB-B0	John talks to Mary, then Mary addresses the group.
AB-BY	John talks to Mary, then Mary talks to Irene.
Turn claiming	
A0-X0	John talks to the group, then Frank talks to the group.
A0-XA	John talks to the group, then Frank talks to John.
A0-XY	John talks to the group, then Frank talks to Mary.
Turn usurping	
AB-X0	John talks to Mary, then Frank talks to the group.
AB-XA	John talks to Mary, then Frank talks to John.
AB-XB	John talks to Mary, then Frank addresses Mary.
AB-XY	John talks to Mary, then Frank addresses Irene.
Turn continuing	
A0-AY	John talks to the group, then addresses Mary.
A0-AA	John talks to the group, then talks more.
AB-A0	John talks to Mary, then makes a remark to the group.
AB-AA	John talks to Mary, then talks again.
AB-AY	John talks to Mary then to Irene.

Table 9. *Listing of the different types of participation shifts (Parshift) according to Gibson 2003 [2].*

In our dataset we got all the sixteen categories but were distributed heterogeneously as seen in Figure 30. The most common type is AB-XB which means that user A is addressing to user B and then someone unaddressed, called user X, also addressed user B.

User X's interjection in the AB-XB participation shift may be common from various motives. They might contribute insights or object to User A's statement without direct input from User B. Alternatively, they may interject to address a potential point or implication from User B, offering agreement or disagreement. In each scenario, User X's input enriches the conversation, fostering engagement and deeper exploration of the topic at hand.

[1]The initial speaker is always labeled A, and the initial target B, unless the group is addressed (or the target was ambiguous), in which case the target is O. Then the shift is summarized in the form [speaker.] [target.] -[speaker.] [target.], with A or B appearing after the hyphen only if the initial speaker or target serves in one of these two positions after the shift. When the speaker after the shift is someone other than A or B, X is used, and when the target after the shift is someone other than A, B, or the group, Y is used.

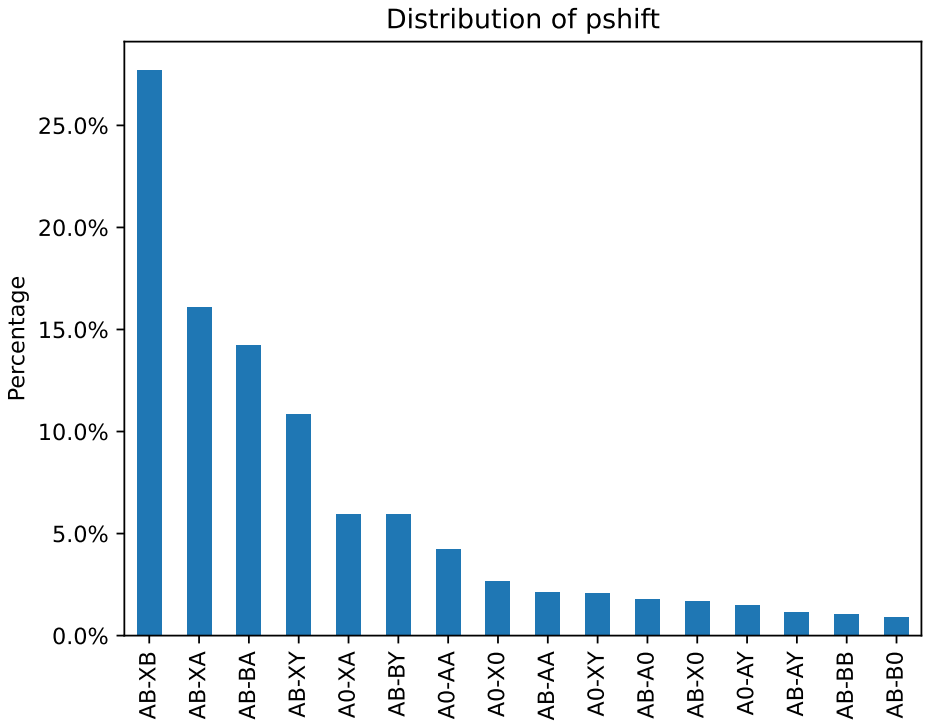

Figure 30. *P-Shift Distribution*

Figure 31 illustrates the distribution of speech type by participation shifts (Pshift), excluding tweets where no type of speech was detected. The distribution of speech types within a Pshift type is similar for all the Pshifts, meaning that the percentages do not vary much, therefore we applied a logistic regression for each type of speech against PShift in order to find any statistically significant association.

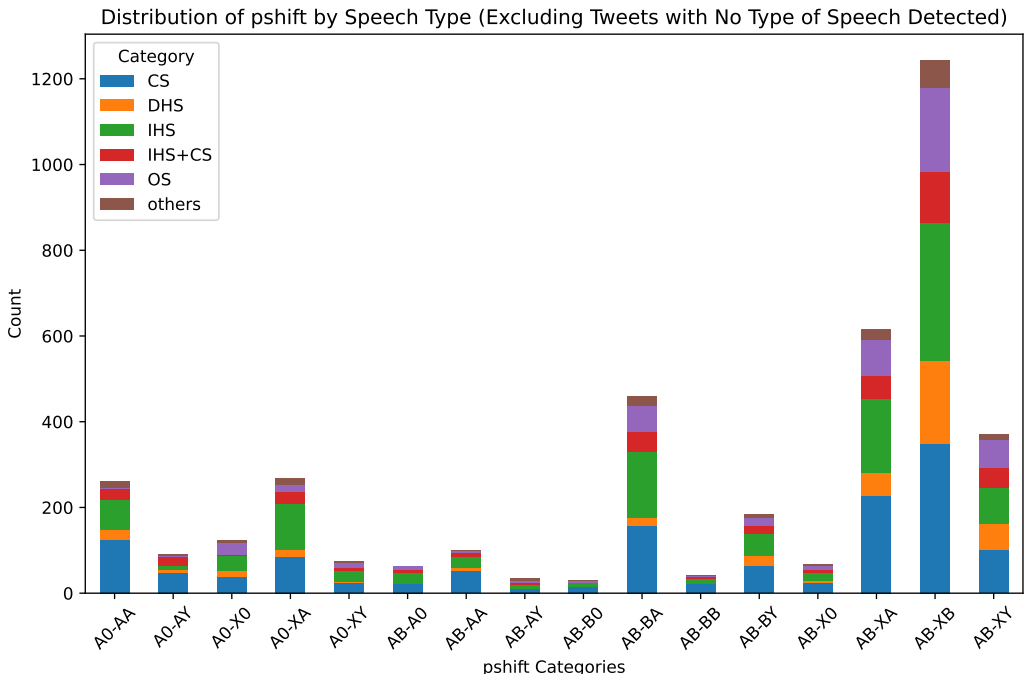

Figure 31. *Distribution of Speech Type by Pshift (Excluding Tweets with no Type of Speech Detected)*

To gain deeper insights into the relationship between P-shifts and all the speech types in study, we conducted a logistic regression analysis, examining each discourse type in relation to P-shift mode. This allowed us to determine whether any statistically significant associations exist between the two variables and if there is a specific way that these types of speech appear in the conversation.

The metrics in Table 10 support the statement that Direct Hate Speech is mostly used by external individuals in the conversation. We can also say, that this type of hate speech may occur outside the context of the conversation between two external speakers (which is supported by second row of Table 10 since the second half of the P-shift is XY). In both of these P-sifts there is the interjection of a third party on the conversation which in the X.

P-shift	**log(OR)**[2]	**95% CI**[3]	**p-value**
AB-XB	0.89	0.61, 1.2	<0.001
AB-XY	0.93	0.57, 1.3	<0.001

Table 10. *Logistic regression model of Direct Hate Speech Discourse*

Table 11 illustrate the outcomes of the logistic regression model examining Indirect Hate Speech discourse in relation to participation shift modes. Contrary to the previous experiment, this form of hate speech typically does not originate from a third party external to the ongoing conversation, since the present $log(OR)$ are negative.

[2]OR = Odds Ratio, CI = Confidence Interval

Furthermore, our analysis reveals a negative correlation between indirect hate speech and certain types of turn claiming (A0-X0) as well as three variants of turn usurping (AB-XA, AB-XB, and AB-XY). This observation hints at the likelihood that this form of discourse often entails a degree of shared personal knowledge of the meaning of both parts.

P-shift	**log(OR)**[3]	**95% CI**[3]	**p-value**
A0-X0	-0.69	-1.1, -0.30	<0.001
AB-XA	-0.39	-0.59, -0.18	<0.001
AB-XB	-0.42	-0.59, -0.25	<0.001
AB-XY	-0.47	-0.72, -0.23	<0.001

Table 11. *Logistic regression model of Indirect Hate Speech Discourse*

The results presented in Table 12 outline the outcomes of a logistic regression analysis investigating Offensive Speech discourse in connection with participation shift modes. Notably, the table reveals a consistent association between offensive speech and two types of turn-shifting in conversations that are less conducive to peace. Specifically, it is evident that offensive speech aligns with the claiming of a turn by an external party (A0-X0 and A0-XY), as well as with all forms of turn usurping (AB-X0, AB-XA, AB-XB, and AB-XY). Additionally, the table also recognizes two types of turn receiving (AB-BA and AB-BY), though with lower occurrence.

P-shift	**log(OR)**[3]	**95% CI**[4]	**p-value**
A0-X0	1.7	1.2, 2.2	<0.001
A0-XY	1.3	0.60, 1.9	<0.001
AB-AY	1.4	0.44, 2.2	0.002
AB-BA	1.1	0.76, 1.5	<0.001
AB-BY	0.85	0.33, 1.3	<0.001
AB-X0	1.3	0.59, 1.9	<0.001
AB-XA	1.1	0.76, 1.4	<0.001
AB-XB	1.4	1.1, 1.7	<0.001
AB-XY	1.3	0.97, 1.7	<0.001

Table 12. *Logistic regression model of Offensive Hate Speech Discourse*

Lastly, Table 13 shows the outputs on a logistic regression analysis between Counter-Speech discourse and Participation Shifts. This model summary shows that counter-speech strongly and positively correlates with the kind of discourse where the speaker sends a message to all and then addresses a particular subject (A0-AY) in a turn continuation mode. On the other hand, three forms of turn usurping (AB-XA, AB-XB, and AB-XY) and two forms of turn claiming (A0-X0 and A0XA) are negatively correlated with counter speech. This type of discourse possibly avoids conflict.

[3]OR = Odds Ratio, CI = Confidence Interval

Notably, talking directly to one person (AB-BA) is negatively linked with counter-speech, indicating that such conversations may not be private and are typically meant for a general audience.

P-shift	**log(OR)**[4]	**95% CI**[4]	**p-value**
A0-AY	1.1	0.58, 1.6	<0.001
A0-X0	-0.84	-1.2, -0.45	<0.001
A0-XA	-0.50	-0.77, -0.23	<0.001
AB-BA	-0.44	-0.66, -0.22	<0.001
AB-XA	-0.36	-0.56, -0.16	<0.001
AB-XB	-0.68	-0.85, -0.51	<0.001
AB-XY	-0.56	-0.80, -0.32	<0.001

Table 13. *Logistic regression model of Counter Discourse*

This analysis on the logistic regression revealed some nuanced insights on the relationships between the multiple P-shifts with the various types of discourse. Statistically significant associations were identified when conducting this detailed analysis, and those helped us spotlight any patterns within the conversation studied.

This results demonstrate that certain speech types are more likely to occur in specific P-shift modes, suggesting a structured interplay between these two elements in the dialogue. Direct Hate Speech tends to involve external parties, while Indirect Hate Speech is more internal to the participants. Offensive Speech is associated with external turn-shifting and usurping, highlighting its disruptive nature. In contrast, Counter-Speech is positively linked with inclusive turn continuation and negatively with usurping and direct addressing, indicating its conflict-avoidant characteristics. These findings not only deepen our comprehension on the conversational structure but also provides a foundation for possible further exploration into the mechanics of interaction. After comparing the logistic regression outcomes for each speech type experience, we can understand how specific P-shifts align with different forms of speech, which enriches our analysis of conversational dynamics and the underlying patterns on discourse behaviour.

CHAPTER 5

Conclusions and Future Work

With this detailed analysis on Portuguese hate speech on Twitter, we can understand its dynamics and all the complexities that come with it. Throughout this analysis we highlighted how this platform can serve both beneficial and harmful purposes. On one hand, it gives a free speech place where people can be their-self's and it also acts as a way for people being informed, however, it also can act as a powerful toll for hate speech dissemination. Because of this darker side, this analysis is a great way to characterize and regulate hate speech based on user's profiles and form of interaction in order to prevent or anticipate it in the future.

With the data retrieved from Twitter API it was possible to replicate the network, which allowed us to see and comprehend how were the users connected. This process was highly important since with a structured network we can, not only create some visualization to help us evaluate their connections, but also take advantage of the network's metrics to analyse key influencers that may act as a dissemination door. Using SNA in this study was particularly significant since it provided an approach to acknowledge as the relations between the users and its part of the network, creating deeper insights about how hate speech is spreading across the network and how can we effectively monitored to control it.

The Portuguese online community was analysed and it was concluded that there are some users who are more present in conversations that may indicate hate speech. This users were part of the giant component of our network which was analysed in the most detail because it is the most connected and the most robust, and it is here that the most dynamic conversations take place.

In addition, our logistic regression analysis highlighted the relationship between P-shifts and different types of discourse, revealing patterns in how hate speech, counter-speech and offensive speech are used in the middle of the conversation. The results helped us understanding the dynamics behind the conversations which makes us better able to make informed decisions when it comes to mitigate online hate speech. With the outcome it was possible to characterize the different types of speech based on the P-shifts. In our network, indirect hate speech typically does not originate from a third party which is external to the conversation, suggesting the existence of a high level of shared personal knowledge between its participants. On the contrary, direct hate speech is predominantly used by external users, which indicates its occurrence outside the context of direct interaction between the main participants. In the case of offensive speech we found it more diverse, involving a mix of turn-shifting, turn-claiming, and turn-receiving behaviours, highlighting its complex nature within conversations. Counter-speech, on the other hand, tends to avoid conflict and is typically used to reach a general audience, as it is

negatively correlated with private, direct exchanges and positively associated with broader, and more inclusive discourse.

This study directly addressed the research questions posed. First, hate speech constitutes approximately 12% of interactions within the analyzed Twitter network, spreading predominantly through clusters of interconnected users (RQ1). Second, hate speech and counter-speech were characterized by their linguistic patterns and dynamics: indirect hate speech often arises in familiar contexts, direct hate speech is typically initiated by external users, and counter-speech tends to target broader audiences with inclusive discourse (RQ2). Finally, the dissipation of hate speech is influenced by key network nodes (opinion leaders) and user interactions; counter-speech proves effective in interrupting hate speech propagation when deployed strategically (RQ3).

Regarding the limitations of the process, we would like to point out the process of retrieving tweets through the API, as it has no way of filtering only for PT-PT, because even if we filter only for those published in Portugal with Portuguese language, we will certainly have a lot of PT-BR that could come from Brazilian immigrants in Portugal. In addition, the choice of keywords was also limiting because, especially in indirect hate speech, the tweet does not need to have a so-called offensive word for it to be so. Nevertheless, the data set proved to be quite diverse and therefore suitable for study.

For future work we believe combining social network analysis and forecasting algorithms can have a better outcome, but above all it is necessary to educate our community to be more defensive rather than offensive when faced with hate speech. In fact, if this step were carried out completely well, we would not need to analyse it because it would not exist. That is why we will continue to fight for a community free of hate speech so that it is no longer relevant in any environment but in particular, due to the study, online. It is crucial for all stakeholders involved, including users, the actual platform, policymakers, and civil society, to work together to create a safer and more inclusive online environment for everyone.

Bibliography

[1] S. M. C. Moro. *Feature selection strategies for improving data-driven decision support in bank telemarketing*. Doctoral thesis, Iscte - University Institute of Lisbon, 2015. URL http://hdl.handle.net/10071/9688.

[2] D. R. Gibson. Participation Shifts: Order and Differentiation in Group Conversation. *Social Forces*, 81(4):1335–1380, June 2003. ISSN 0037-7732, 1534-7605. doi: 10.1353/sof.2003.0055. URL https://academic.oup.com/sf/article-lookup/doi/10.1353/sof.2003.0055.

[3] . Hidayati, . Aflina, and . Arifuddin. Hate speech on social media: A pragmatic approach. *KnE Social Sciences*, page 308–317, March 2021. ISSN 2518-668X. doi: 10.18502/kss.v5i4.8690. URL http://dx.doi.org/10.18502/kss.v5i4.8690.

[4] Dr Ramakrishna Hegde. Review paper on hate speech detection. *Engineering and Technology Journal*, 06(12), December 2021. ISSN 2456-3358. doi: 10.47191/etj/v6i12.05. URL http://dx.doi.org/10.47191/etj/v6i12.05.

[5] Tim O'Reilly and Sarah Milstein. *The Twitter Book*. O'Reilly Media, Inc., 2009. ISBN 0596802811.

[6] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533144. URL http://dx.doi.org/10.1145/3531146.3533144.

[7] European Commission against Racism and Intolerance. Ecri report on portugal, 2018. URL https://rm.coe.int/fifth-report-on-portugal/16808de7da.

[8] Comissão para a Igualdade e Contra a Discriminação Racial (Commission for Equality and Against Racial Discrimination). Relatório anual 2022 [annual report 2022], 2023. URL https://www.cicdr.pt/documents/57891/0/Relat%C3%B3rio+Anual+CICDR+2022.pdf/589b161f-05d8-471f-bf12-f3857bfad171.

[9] ILGA Portugal – Intervenção Lésbica, Gay, Bissexual, Trans e Intersexo. Relatório anual 2020-2022 [annual report 2020-2022], 2023. URL https://ilga-portugal.pt/files/uploads/2023/10/F_Relatorio-Observatorio-Discriminacao-Contra-Pessoas-LGBTI-2020-2022.pdf.

[10] Bhikhu Parekh. Is there a case for banning hate speech? In Michael Herz and Peter Molnar, editors, *The Content and Context of Hate Speech: Rethinking Regulation and Response*, pages 37–56. Cambridge University Press, Cambridge, New York, 2012.

[11] United Nations. Understanding hate speech. URL https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. Accessed: November 25, 2024.

[12] Assembleia da República. Código penal português, 2024. URL https://diariodarepublica.pt/dr/lexionario/termo/crime-incitamento-ao-odio-a-violencia. Artigo 240.º, n.º 2.

[13] Bill Kuechler and Vijay Vaishnavi. On theory development in design science research: anatomy of a research project. *European Journal of Information Systems*, 17(5):489–504, October 2008. ISSN 1476-9344. doi: 10.1057/ejis.2008.40. URL http://dx.doi.org/10.1057/ejis.2008.40.

[14] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. 2000. URL https://api.semanticscholar.org/CorpusID:1211505.

[15] António Fonseca, Catarina Pontes, Sérgio Moro, Fernando Batista, Ricardo Ribeiro, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. Analyzing hate speech dynamics on twitter/x: Insights from conversational data and the impact of user interaction patterns. *Heliyon*, 10(11):e32246, June 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e32246. URL http://dx.doi.org/10.1016/j.heliyon.2024.e32246.

[16] Information processing and management of uncertainty in knowledge-based systems — link.springer.com. https://link.springer.com/book/9783031740022. [Accessed 09-09-2024].

[17] ChenLiming. page 328–345. Cambridge University Press, March 2019. doi: 10.1017/9781108349383.025. URL http://dx.doi.org/10.1017/9781108349383.025.

[18] Tracy Sweet. *Social Network Analysis*, page 434–444. Routledge, November 2018. ISBN 9781315755649. doi: 10.4324/9781315755649-32. URL http://dx.doi.org/10.4324/9781315755649-32.

[19] Rahim. page 2665–2665. Springer New York, 2018. doi: 10.1007/978-1-4939-7131-2_101152. URL http://dx.doi.org/10.1007/978-1-4939-7131-2_101152.

[20] Stefano Faralli and Paola Velardi. Special issue on social network analysis. *Applied Sciences*, 12(18):8993, September 2022. ISSN 2076-3417. doi: 10.3390/app12188993. URL http://dx.doi.org/10.3390/app12188993.

[21] Zhou Nie, Moniza Waheed, Diyana Kasimon, and Wan Anita Binti Wan Abas. The Role of Social Network Analysis in Social Media Research. *Applied Sciences*, 13(17):9486, August 2023. ISSN 2076-3417. doi: 10.3390/app13179486. URL https://www.mdpi.com/2076-3417/13/17/9486.

[22] Jacques Angot and Emmanuel Josserand. *Analyzing Social Networks*, page 312–331. SAGE Publications Ltd, 2001. doi: 10.4135/9781849208970.n14. URL http://dx.doi.org/10.4135/9781849208970.N14.

[23] Alex Primo. O aspecto relacional das interações na web 2.0. *E-Compós*, 9, jun. 2007. doi: 10.30962/ec.153. URL https://e-compos.org.br/e-compos/article/view/153.

[24] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL http://aclweb.org/anthology/W18-5102.

[25] Alberto Izquierdo Montero, Noemi Laforgue-Bullido, and David Abril-Hervás. Hate speech: a systematic review of scientific production and educational considerations. *Revista Fuentes*, 2(24):222–233, 2022. ISSN 15757072, 21727775. doi: 10.12795/revistafuentes.2022.20240. URL https://institucional.us.es/revistas/fuente/24_2/Art_9.pdf.

[26] Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1): 157–179, January 2021. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-020-03737-6. URL http://link.springer.com/10.1007/s11192-020-03737-6.

[27] Hossam Faris, Ibrahim Aljarah, Maria Habib, and Pedro Castillo. Hate speech detection using word embedding and deep learning in the arabic language context. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications, 2020. doi: 10.5220/0008954004530460. URL http://dx.doi.org/10.5220/0008954004530460.

[28] Kristiawan Nugroho, Edy Noersasongko, Purwanto, Muljono, Ahmad Zainul Fanani, Affandy, and Ruri Suko Basuki. Improving random forest method to detect hatespeech and offensive word. In *2019 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, July 2019. doi: 10.1109/icoiact46704.2019.8938451. URL http://dx.doi.org/10.1109/ICOIACT46704.2019.8938451.

[29] Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology amp; Politics*, 17(1):66–78, December 2019. ISSN 1933-169X. doi: 10.1080/19331681.2019.1702607. URL http://dx.doi.org/10.1080/19331681.2019.1702607.

[30] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n71. URL http://dx.doi.org/10.1136/bmj.n71.

[31] Nidhi Rastogi Le Nguyen. Graph-based Approach for Studying Spread of Radical Online Sentiment. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1373–1380, Austin TX USA, April 2023. ACM. ISBN 978-1-4503-9419-2. doi: 10.1145/3543873.3587634. URL https://dl.acm.org/doi/10.1145/3543873.3587634.

[32] Jae Yeon Kim and Aniket Kesari. Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. *Journal of Online Trust and Safety*, 1 (1), October 2021. ISSN 2770-3142. doi: 10.54501/jots.v1i1.13. URL http://dx.doi.org/10.54501/jots.v1i1.13.

[33] K. DiCicco, N.B. Noor, N. Yousefi, M. Maleki, B. Spann, and N. Agarwal. Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Social Media Platforms. volume 3406, pages 30–42, 2023. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161629095&partnerID=40&md5=96ede7d24b3c4ef96fb208d46743d6e9.

[34] Sergio Arce-García and María-Isabel Menéndez-Menéndez. Inflaming public debate: a methodology to determine origin and characteristics of hate speech about sexual and gender diversity on Twitter. *El Profesional de la información*, page e320106, December 2022. ISSN 16992407, 13866710. doi: 10.3145/epi.2023.ene.06. URL https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86993.

[35] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):361–362, March 2009. ISSN 2162-3449. doi: 10.1609/icwsm.v3i1.13937. URL http://dx.doi.org/10.1609/icwsm.v3i1.13937.

[36] Zixuan Weng and Aijun Lin. Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, 19(24):16376, December 2022. ISSN 1660-4601. doi: 10.3390/ijerph192416376. URL https://www.mdpi.com/1660-4601/19/24/16376.

[37] Fabián Riquelme, Diego Rivera, and Benjamín Serrano. Analyzing the far-right political action on Twitter: the Chilean constituent process. *Social Network Analysis and Mining*, 12(1):161, December 2022. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-022-00990-w. URL https://link.springer.com/10.1007/s13278-022-00990-w.

[38] Daniil Volkovskii and Svetlana Bodrunova. *Incivility Balanced? Civil vs. Uncivil Speech in Online Political Discussions as Dependent on Political Parallelism*, page 65–80. Springer Nature Switzerland, 2023. ISBN 9783031314698. doi: 10.1007/978-3-031-31469-8_5. URL http://dx.doi.org/10.1007/978-3-031-31469-8_5.

[39] Sejung Park and Jiwon Kim. Tweeting about abusive comments and misogyny in South Korea following the suicide of Sulli, a female K-pop star: Social and semantic network analyses. *El Profesional de la información*, page e300405, September 2021. ISSN 16992407, 13866710. doi: 10.3145/epi.2021.sep.05. URL https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86365.

[40] Javier Torregrosa, Ángel Panizo-Lledot, Gema Bello-Orgaz, and David Camacho. Analyzing the relationship between relevance and extremist discourse in an alt-right network

on Twitter. *Social Network Analysis and Mining*, 10(1):68, December 2020. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-020-00676-1. URL https://link.springer.com/10.1007/s13278-020-00676-1.

[41] David Blanco-Herrero and Carlos Arcila Calderón. Spread and reception of fake news promoting hate speech against migrants and refugees in social media: Research Plan for the Doctoral Programme Education in the Knowledge Society. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 949–955, León Spain, October 2019. ACM. ISBN 978-1-4503-7191-9. doi: 10.1145/3362789.3362842. URL https://dl.acm.org/doi/10.1145/3362789.3362842.

[42] Simin Kargar and Adrian Rauchfleisch. State-aligned trolling in Iran and the double-edged affordances of Instagram. *New Media & Society*, 21(7):1506–1527, July 2019. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444818825133. URL http://journals.sagepub.com/doi/10.1177/1461444818825133.

[43] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19 (21):4654, October 2019. ISSN 1424-8220. doi: 10.3390/s19214654. URL https://www.mdpi.com/1424-8220/19/21/4654.

[44] Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. Hate speech dynamics against african descent, roma and lgbtqi communities in portugal. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2362–2370, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.253.

[45] Público. Professor da universidade de aveiro suspenso por comentários homofóbicos. julho 2022. URL https://www.publico.pt/2022/07/04/p3/noticia/suspenso-professor-universidade-aveiro-acusado-comentarios-homofobicos-2012418.

[46] Diário de Notícias. Racismo: "apenas por sermos brancos tivemos tamanha comoção". august 2022. URL https://www.dn.pt/sociedade/racismo-apenas-por-sermos-brancos-tivemos-tamanha-comocao-15064707.html.

[47] Chiara Ravazzi, Roberto Tempo, and Fabrizio Dabbene. Learning influence structure in sparse social networks. *IEEE Transactions on Control of Network Systems*, 5(4): 1976–1986, December 2018. ISSN 2372-2533. doi: 10.1109/tcns.2017.2781367. URL http://dx.doi.org/10.1109/TCNS.2017.2781367.

[48] Narges Motalebi, Nathaniel T. Stevens, and Stefan H. Steiner. Hurdle blockmodels for sparse network modeling. *The American Statistician*, 75(4):383–393, February 2021. ISSN 1537-2731. doi: 10.1080/00031305.2020.1865199. URL http://dx.doi.org/10.1080/00031305.2020.1865199.