Unveiling Patterns of Hate Speech in the Portuguese Sphere: A Social Network Analysis Approach*

Catarina Pontes[1], António Fonseca[1], Sérgio Moro[1,4], Fernando Batista[2,3], Ricardo Ribeiro[2,3], Catarina Marques[6], Paula Carvalho[3], Cláudia Silva[7], and Rita Guerra[5]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[3] INESC-ID, Lisboa, Portugal
[4] University of Jordan, Amman, Jordan
[5] ISCTE-Instituto Universitário de Lisboa and Center for Psychological Research and Social Intervention (CIS-ISCTE), Lisboa, Portugal
[6] ISCTE-Instituto Universitário de Lisboa and Business Research Unit (BRU-ISCTE), Lisboa, Portugal
[7] ITI-LARSyS and IST, Lisboa, Portugal

Abstract. This study investigates the proliferation of hate speech on Twitter (now known as X) within the Portuguese-speaking community, utilizing Social Network Analysis (SNA) to examine tweets from 2021 to 2022 flagged for hate speech content. Our analysis, grounded in a dataset annotated for direct and indirect hate speech, offensive speech, and counter-speech, aims to unravel the linguistic patterns and the specific targets of online hate. Employing network science methodologies, we dissect the structural dynamics of these interactions, focusing on the influence of central nodes and the formation of communities that facilitate the spread of hate speech. Key findings highlight a prominent volume of hate speech directed towards minority and vulnerable groups, with significant discourse around LGBT rights, racism, xenophobia, and discrimination against the Roma. The study underscores the critical role of online communities and influencers in the dissemination of hate speech and suggests a pressing need for effective monitoring and intervention strategies. Our research contributes to understanding the complexities of online hate speech and offers insights for developing tools to counteract its spread, aiming to foster a safer and more inclusive online discourse.

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

* This work was funded in part by the European Union under Grant CERV-2021-EQUAL (101049306).

1 Introduction

Twenty-First Century society has the answers to all questions in a distance of a click where 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 come some problems that only few could anticipate.

Nowadays there is a growing problem related to hate speech, 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 [7]. This study seeks to analyze a set of tweets that were selected based on certain keywords that may suggest the presence of hate speech. In the dataset, there are tweets published between 2021 and 2022 and with those, we aim to identify patterns and trends in the use of the language in those tweets. The main goal is to get important insights into how hate speech is disseminated in the Portuguese online environment so that we can apply actions in key places to mitigate it in the best way possible.

With the increasing usage of social media, Social Network Analysis 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 [6]. This technique has been applied in various fields, including marketing [17] and risk analysis [14]. 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. 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 [12].

Hate Speech is defined 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 [4]. It is often targeted at vulnerable or minority groups [8], and its prevalence has led to increased research in the field, particularly in the areas of regulation, computational linguistics, and discourse analysis [8] [18].

2 Literature review

In this literature review, we focus on the analysis of hate speech on social media using networks for this purpose. This part of our work is very important to find information about methodologies and challenges prevalent in this field. We limited our search to studies that explore the analysis of "Hate Speech" and "Social Networks", excluding topics of "Deep Learning" and "detection" as it is not the focus of our work.

Because of the growing problem which is the propagation of hate speech on the Internet in [11], it is 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 height 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 study [10]. 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 [5], it focuses 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.

[1] investigates 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 analyze 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 [2], in [20] it was analyze the information propagation path using Gephi's tools. The retweet relationship was directly related to topic diffusion behavior 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, [16] suggests an interdisciplinary study combining computation with philosophy and sociology to better understand and model their behaviors. The conduct of 5 opinion leaders was analyzed 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 analyzed 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.

Some network metrics, such as various types of centrality metrics (n-degree, eigenvector, k-shells, betweenness, and closeness), were used by [19] 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 analyzed 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.

The impact of fake news about minorities on the existence and rise of hate speech directed at those same minorities was explored by [3]. They apply a three-step routine to analyze that, which consists of one survey to analyze 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.

The dimension and the authors of online hate, harassment, and abusive speech opposing Iranian emigrants were investigated by [9]. This study is based on two pillars: qualitative interviews and a quantitative analysis of related individuals' Instagram accounts. The quantitive approach explores how violent 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 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 [13] principal aim is to create an intelligent system that can identify and monitor hate speech on Twitter. For this study, we are particularly interested in the automatic Social Network Analyzer that uses graph theories to identify social structures. This analyzer 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.

3 Data Collection

Twitter data potentially related to four target groups (Roma, Racism, Xenophobia, LGBTQI+) was accessed using the Twitter API. We compiled a list of

259 keywords associated with these target groups to retrieve tweets containing these keywords. To select potential targets, we consider first only the unambiguous words corresponding to 174 entries. The ambiguous words were not selected in this first retrieval since they can have different meanings depending on the context. Ambiguous words were associated with insults from a predefined list with approximately 800 entries. Data collection was limited to a two-year span from January 1, 2021, to December 31, 2022. The language was filtered to Portuguese, predominantly resulting in Portuguese from Brazil (pt-br) instead of Portuguese from Portugal (pt-pt). To ensure geographical relevance, the collection was narrowed to tweets posted in Portugal. Additionally, conversations to which the tweets belong were retrieved, focusing only on those with a parent tweet published in Portugal. The conversation tweets were also collected using the API with its specific filter. A third extraction involved creating a list of unique conversation IDs and retrieving tweets with those IDs to link them to the parent tweet that initiated the conversation. The final dataset consists of conversations only, with a parent tweet published in Portugal, resulting in 29531 tweets.

4 Data Annotation

The set of 29531 tweets was 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 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 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 have a small representation. However, the values rise when looking into Indirect Hate Speech and Counter Speech.

5 Social Network Analysis

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 from the user who is replying 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.

Table 1. Whole Graph Metrics.

Number of Nodes	9 952
Number of Edges	24 532
Average Degree	2.465
Network Diameter	7
Average Path Length	5.53
Average clustering coefficient	0.063
Modularity	0.883

Our network has 9952 nodes and 24532 edges and most of the graph is connected, but around it, we can see some isolated groups. Figure 1 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, this graph is analysed in Table 1. 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.

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 [15], meaning that they cover a large area but they are not well connected.

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.

Degree: 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's 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 behavior of the complex system, offering insights into its structure and dynamics.

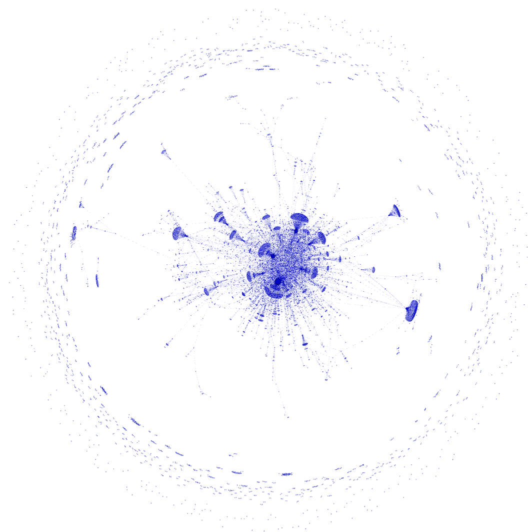

Fig. 1. Graph Representation of all the Conversations Data.

5.1 Giant Component Analysis

The giant component, as represented in Figure 2, contains approximately 69% of all the nodes in the original graph and 76% of the edges, which translates into 6852 nodes and 18676 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.

Betweeness 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 opinion leaders. In our network, we can highlight a few users that fit in that description, as shown in Figure 3. Looking at the relative magnitude of the Degree Centrality (normalized in the dataset), shown in Figure 4, we can see as well that these users are well involved in the discussions,

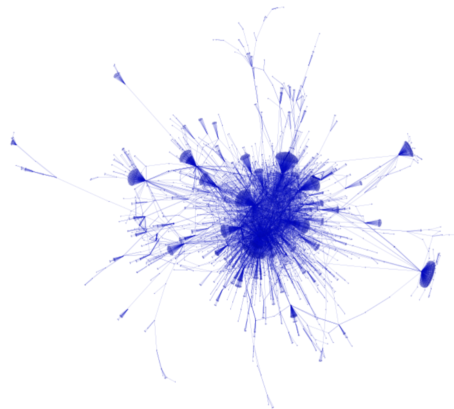

Fig. 2. Graph Representation of the Central Core.

however their daily Twitter usage is not relatively very large, neither their popularity measured by the relative number of followers, in comparison with other users in our dataset. To support this analysis we performed betweenness centrality correlations with centrality degree, as well as with the number of followers and the number of tweets posted, which were 0.89, 0.02 and 0.04, respectively.

5.2 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 represent small conversations between users and constitute around 30% of the whole graph.

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.

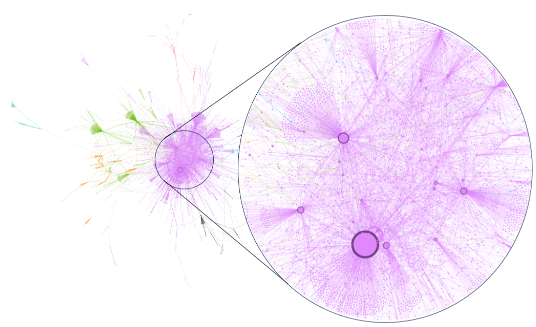

Fig. 3. Graph Representation of the Betweenness Centrality on Central Core.

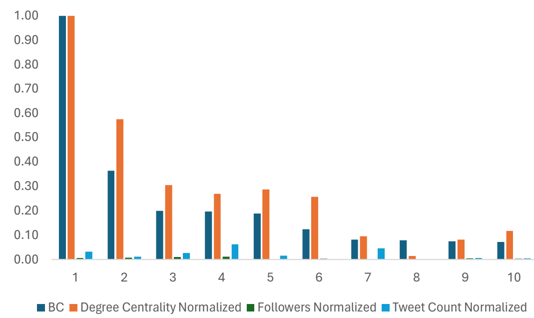

Fig. 4. Relative Magnitude of Betweenness Centrality, Degree Centrality, Number of Followers and Number of Tweets Posted of the top 10 nodes

5.3 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 analyzed users with the highest betweenness centrality are present in this graph, represented in Figure 5, which may indicate a huge problem for the dissipation of hate in social media.

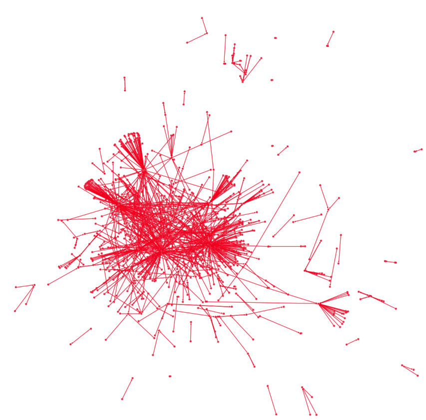

Fig. 5. Graph Representation of Hate Speech in Central Core.

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.

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.

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

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. 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.

6 Conclusions

Our study of hate speech on Twitter, using network science, shows how complicated and connected online interactions can be. We looked at how users on Twitter interact with each other and found patterns in harmful speech. Our results revealed a complex structure of social networks on Twitter, highlighting the existence of three distinct groups that have different behaviors: the central core, the peripheries and the isolated users that are not connected with central core. This characterization reveals the complexity of interaction dynamics and the possibility of opinion bubbles forming or hate speech spreading within these groups.

We also found that users who connect different groups, known as having high betweenness centrality, can help spread hate speech. To fight against hate speech effectively, we need to use multiple strategies. These should focus on both stopping users who spread hate speech and cutting off the paths that help spread it. This approach is key to reducing hate speech on social media and making online spaces better for everyone.

Lastly, our research contributes to the broader academic and practical domains by elucidating on the dynamics of hate speech propagation in portuguese online social networks. The insights obtained from this study offer valuable knowledge for researchers, policymakers, and practitioners committed to make efforts to address online toxicity and promote healthier and peaceful digital communities. By understanding the mechanisms supporting hate speech dissemination, we can develop more targeted interventions to prevent its proliferation, ultimately promoting a safer and more inclusive online discourse.

References

1. Arce-García, S., Menéndez-Menéndez, M.I.: 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 p. e320106 (Dec 2022). https://doi.org/10.3145/epi.2023.ene.06, https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86993
2. Bastian, M., Heymann, S., Jacomy, M.: 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 (Mar 2009). https://doi.org/10.1609/icwsm.v3i1.13937, http://dx.doi.org/10.1609/icwsm.v3i1.13937
3. Blanco-Herrero, D., 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. In: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality. pp. 949–955. ACM, León Spain (Oct 2019). https://doi.org/10.1145/3362789.3362842, https://dl.acm.org/doi/10.1145/3362789.3362842
4. De Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate Speech Dataset from a White Supremacy Forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 11–20. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/W18-5102, http://aclweb.org/anthology/W18-5102
5. DiCicco, K., Noor, N., Yousefi, N., Maleki, M., Spann, B., Agarwal, N.: Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Social Media Platforms. vol. 3406, pp. 30–42 (2023), https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161629095&partnerID=40&md5=96ede7d24b3c4ef96fb208d46743d6e9
6. Edlund, J.E., Nichols, A.L. (eds.): p. 328–345. Cambridge University Press (Mar 2019). https://doi.org/10.1017/9781108349383.025, http://dx.doi.org/10.1017/9781108349383.025
7. Hegde, D.R.: Review paper on hate speech detection. Engineering and Technology Journal **06**(12) (Dec 2021). https://doi.org/10.47191/etj/v6i12.05, http://dx.doi.org/10.47191/etj/v6i12.05
8. Izquierdo Montero, A., Laforgue-Bullido, N., Abril-Hervás, D.: Hate speech: a systematic review of scientific production and educational considerations. Revista Fuentes **2**(24), 222–233 (2022). https://doi.org/10.12795/revistafuentes.2022.20240, https://institucional.us.es/revistas/fuente/24_2/Art_9.pdf
9. Kargar, S., Rauchfleisch, A.: State-aligned trolling in Iran and the double-edged affordances of Instagram. New Media & Society **21**(7), 1506–1527 (Jul 2019). https://doi.org/10.1177/1461444818825133, http://journals.sagepub.com/doi/10.1177/1461444818825133

10. Kim, J.Y., Kesari, A.: Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. Journal of Online Trust and Safety **1**(1) (Oct 2021). https://doi.org/10.54501/jots.v1i1.13, http://dx.doi.org/10.54501/jots.v1i1.13
11. Nguyen, L., Rastogi, N.: Graph-based Approach for Studying Spread of Radical Online Sentiment. In: Companion Proceedings of the ACM Web Conference 2023. pp. 1373–1380. ACM, Austin TX USA (Apr 2023). https://doi.org/10.1145/3543873.3587634, https://dl.acm.org/doi/10.1145/3543873.3587634
12. Nie, Z., Waheed, M., Kasimon, D., Wan Abas, W.A.B.: The Role of Social Network Analysis in Social Media Research. Applied Sciences **13**(17), 9486 (Aug 2023). https://doi.org/10.3390/app13179486, https://www.mdpi.com/2076-3417/13/17/9486
13. Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and Monitoring Hate Speech in Twitter. Sensors **19**(21), 4654 (Oct 2019). https://doi.org/10.3390/s19214654, https://www.mdpi.com/1424-8220/19/21/4654
14. Rahim: p. 2665–2665. Springer New York (2018). https://doi.org/10.1007/978-1-4939-7131-2_101152, http://dx.doi.org/10.1007/978-1-4939-7131-2_101152
15. Ravazzi, C., Tempo, R., Dabbene, F.: Learning influence structure in sparse social networks. IEEE Transactions on Control of Network Systems **5**(4), 1976–1986 (Dec 2018). https://doi.org/10.1109/tcns.2017.2781367, http://dx.doi.org/10.1109/TCNS.2017.2781367
16. Riquelme, F., Rivera, D., Serrano, B.: Analyzing the far-right political action on Twitter: the Chilean constituent process. Social Network Analysis and Mining **12**(1), 161 (Dec 2022). https://doi.org/10.1007/s13278-022-00990-w, https://link.springer.com/10.1007/s13278-022-00990-w
17. Sweet, T.: Social Network Analysis, p. 434–444. Routledge (Nov 2018). https://doi.org/10.4324/9781315755649-32, http://dx.doi.org/10.4324/9781315755649-32
18. Tontodimamma, A., Nissi, E., Sarra, A., Fontanella, L.: Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics **126**(1), 157–179 (Jan 2021). https://doi.org/10.1007/s11192-020-03737-6, http://link.springer.com/10.1007/s11192-020-03737-6
19. Torregrosa, J., Panizo-Lledot, A., Bello-Orgaz, G., Camacho, D.: Analyzing the relationship between relevance and extremist discourse in an alt-right network on Twitter. Social Network Analysis and Mining **10**(1), 68 (Dec 2020). https://doi.org/10.1007/s13278-020-00676-1, https://link.springer.com/10.1007/s13278-020-00676-1
20. Weng, Z., Lin, A.: 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 (Dec 2022). https://doi.org/10.3390/ijerph192416376, https://www.mdpi.com/1660-4601/19/24/16376