A DATA-DRIVEN STRATEGY FOR ONLINE HATE SPEECH SPREADER IDENTIFICATION USING MODIFIED PAGERANK

Smita Ghosh and Shiv Jhalani
Department of Mathematics and Computer Science, Santa Clara University,
Santa Clara, California, USA

ABSTRACT

Social media platforms have become breeding grounds for the dissemination of misinformation and harmful content, including hate speech. This research paper aims to tackle the urgent problem of hate speech circulation on Online Social Networks by primarily focusing on the early identification of users who are prone to spreading it. To achieve this objective, a novel data-driven metric is introduced, referred to as 'Hate Speech Potential'. Additionally, an innovative approach is proposed that leverages a modified version of the PageRank algorithm, termed the 'Hate Speech Potential Rank' algorithm, to effectively detect and identify malicious users within a network. In a vast network with billions of nodes, the rapid spread of content makes timely detection and mitigation crucial. By assessing a user's past behaviour of sharing or publishing hate speech, their 'Hate Speech Potential' can be determined, enabling the identification of sources and spreaders of such content. The modified PageRank algorithm considers both the user's individual characteristics and the influence of their neighbourhood, thereby capturing a more comprehensive picture of their sharing patterns. A pre-trained machine learning model was employed to accurately classify hate speech posts. By combining the predicted labels and user characteristics and implementing the modified PageRank algorithm, this paper aims to gain deeper insights into the dynamics of information dissemination within a social network, thereby contributing to a better understanding of user sharing behaviour and facilitating the development of effective strategies for addressing hate speech. K-Means clustering was used in experimental evaluations, demonstrating the effectiveness of the proposed approach.

KEYWORDS

Hate Speech Spreader Detection, PageRank, Social Networks, Machine Learning

1. INTRODUCTION

In the past decade, the popularity of Online Social Networks (OSNs) has significantly increased due to the widespread availability of the internet and the affordability of electronic devices. This surge in users has led to a substantial growth in content consumption and the amount of time individuals spend on these platforms. In recent years, hate speech has emerged as a growing criminal issue [12], not only in face-to-face interactions but also in online communication. According to the Encyclopedia of the American Constitution [17], hate speech refers to speech that specifically targets and attacks an individual or a group based on various attributes, including but not limited to race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. This definition highlights the discriminatory nature of hate speech, which seeks to demean, marginalize, or incite hatred towards individuals or communities based on their inherent characteristics. On an OSN, individuals create disturbances in the online experience by generating an attraction towards low-quality content and actively participating in its

David C. Wyld et al. (Eds): SIPR, NCWC, BIBC, SOFEA, DSML, ARIA, NLP, CSEN -2023
pp. 199-211, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131916

dissemination throughout the network. There are several contributing factors to this concerning trend. Firstly, the internet, particularly social networks, provides an environment where individuals are more prone to adopting aggressive behaviour due to the shield of anonymity it offers [8]. Secondly, people are increasingly inclined to express their opinions openly on online platforms, inadvertently fuelling the propagation of hate speech. Given the immense harm that such prejudiced communication can inflict on society, both governments and social network platforms stand to benefit from the development and implementation of detection and prevention tools.

In a large network of billion nodes, the speed of content propagation is very high [22], and as harmful impact of hate speech poses a significant challenge on social media platforms it becomes necessary to implement effective measures for damage control and recovery. Simply identifying and blocking hate speech content provides only a temporary solution. Instead, a more effective approach is to develop systems capable of detecting and profiling the individuals who disseminate hate speech, thereby addressing the root of the problem. By focusing on identifying and monitoring the content polluters who share hate speech, more comprehensive and sustainable measures can be implemented [7]. In this paper, a data driven metric called 'Hate Speech Potential (HSP)' is proposed and by using a modified PageRank algorithm, it helps in identifying spreaders of hate speech on an OSN. Fig. 1 shows a social network of 15 nodes and their HSPR labels. By setting a threshold value of 0.8, 2 nodes are identified as nodes of interest. This identification aids OSN providers in effectively monitoring a vast network consisting of billions of nodes.

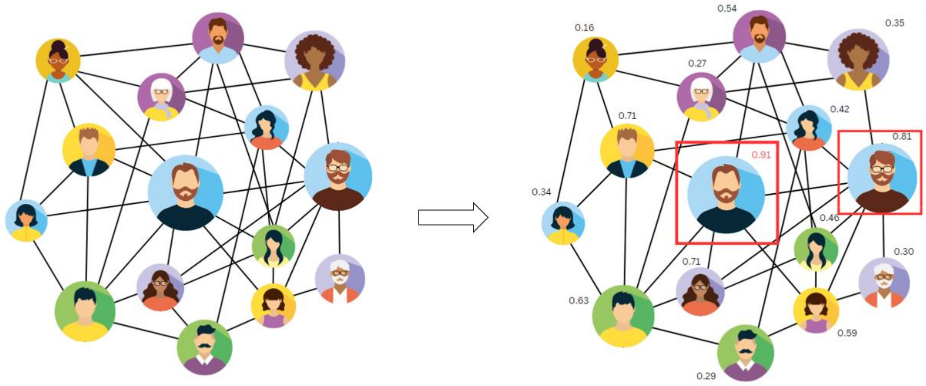

Figure 1. HSPR generated Social Network

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results [23]. PageRank operates by assessing both the quantity and quality of links pointing to a webpage, providing a rough estimation of its significance. The PageRank algorithm captures the essence of how users disseminate information within an OSN. Studies show that people tend to adopt information from people they know and trust [11]. The PageRank algorithm, originally developed to rank web pages based on their importance and relevance, can be adapted and utilized in the context of identifying hate speech spreaders online. By drawing parallels to the way PageRank assigns significance to websites with more inbound links, a similar concept can be applied to pinpoint individuals who actively propagate hate speech through their network connections. In this paper, a modified PR algorithm called Hate Speech Potential Rank (HSPR) is proposed which uses the HSP to generate ranking of users on an OSN. Section 4.3 and Section 4.4 describe HSP and HSPR in detail. Users who have a HSPR, exceeding a certain threshold value are identified as nodes of interest. By identifying the pool of potential sources and spreaders, which initially consists of billions of users, as a small collection, the task of monitoring and controlling the dissemination of hate speech becomes more manageable for the

OSN provider. This paper is further divided into the following sections: Section 2 provides a comprehensive overview of existing research and approaches in the field. Section 3 describes why Twitter was chosen as the social media platform for this paper. Section 4 presents an innovative approach for detecting hate speech spreaders online, enabling targeted interventions and mitigation strategies. Section 5 demonstrate the effectiveness of the proposed approach in identifying hate speech spreaders. Section 6 concludes the findings of the paper and Section 7 describes the future steps for the proposed idea.

2. Related Work

2.1 Hate Speech Detection

In the context of online platforms, including popular forums like Facebook, YouTube, and Twitter, hate speech is widely recognized as a detrimental form of communication. These platforms acknowledge the harmful nature of hate speech and have implemented policies aimed at its removal. Such policies reflect a commitment to fostering safer and more inclusive online environments by mitigating the presence and impact of hate speech content. By actively addressing and removing hate speech, these platforms aim to protect their users from the negative consequences associated with such content, promoting a healthier online discourse [3, 4, 6]. The growing prevalence of hate speech on the Internet [16] has raised significant societal concerns. As a result, there is a strong motivation to delve into the research and development of automatic hate speech detection systems. The urgency to address this issue stems from the need to effectively identify and combat the harmful effects of hate speech in online platforms. By studying and developing automated techniques for hate speech detection, the objective is to enhance online safety, foster inclusive environments, and mitigate the adverse effects of hate speech on individuals and society at large.

Recent approaches in the field have shown promising results in the detection of hate speech within textual content. These approaches employ various techniques and methodologies to analyse and identify patterns, linguistic cues, and contextual information indicative of hate speech. By leveraging machine learning algorithms, natural language processing techniques, and domain-specific knowledge, these approaches aim to effectively classify and flag instances of hate speech, enabling better content moderation and fostering a safer online environment. The continuous development and refinement of such approaches hold great potential for addressing the challenges posed by hate speech and promoting responsible and respectful communication online [13]. In [10] the authors train multi-class classifiers such as Linear SVM, Logistic Regression to distinguish between offensive and hate speech. In the realm of text classification problems, neural network approaches have emerged as state-of-the-art techniques. To enhance the classification of hate speech, [24] explores the adaptation of an ensemble method specifically designed for neural networks. By leveraging the strengths of multiple neural network models, the proposed ensemble method aims to improve the accuracy and effectiveness of hate speech classification. This approach acknowledges the advancements in neural networks and seeks to harness their potential for more robust and reliable hate speech detection

2.2. Hate Speech Spreader Detection

Hate speech can have severe negative consequences, leading to harassment, discrimination, and even incitement to violence. Identifying and addressing hate speech spreaders helps mitigate the potential harm inflicted on targeted individuals or groups. Also, hate speech can quickly spread and gain momentum on social media platforms, leading to its amplification and wider dissemination. By identifying hate speech spreaders early on, platforms can prevent the rapid propagation of harmful content and limit its reach. In [7] the authors used n-grams and Voting Classifiers (VC) to detect hate speech spreaders. The proposed models utilize a combination of traditional char and word n-grams with syntactic n-grams as features extracted from the training set. These features are fed to a VC that employs three Machine Learning (ML) classifiers namely, Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) with hard and soft voting. The authors in [20], describe a deep learning model based on a Convolutional Neural Network (CNN). The model was developed for the Profiling Hate Speech Spreaders. In another paper [9], the authors used lexical and psycho-emotional information to detect hate speech spreaders on Twitter. The detection of hate speech spreaders plays a crucial role in addressing the harmful impact of hate speech in online platforms. Driven by the pressing need to address the issue of hate speech spreader detection, this paper focuses on the identification of individuals who actively spread hate speech. By pinpointing these individuals, targeted interventions can be implemented, leading to the adoption of appropriate measures to effectively mitigate the negative consequences of hate speech. Recognizing the importance of this task, the paper aims to contribute to the advancement of hate speech spreader detection and prevention strategies, ultimately fostering a safer and more inclusive online environment. Hate speech spreader detection is still in its early stages of research and development. The complex nature of hate speech, varying contexts, evolving online platforms, and the need for accurate and efficient detection methods pose significant challenges. Therefore, ongoing research efforts are essential to advance the field and develop effective techniques that can robustly and reliably identify hate speech spreaders.

2.3. Pagerank Algorithm

PageRank (PR) is an algorithm employed by Google Search to determine the ranking of web pages within their search engine results. It operates by assigning a numerical value to each web page, known as its PageRank score, which is influenced by the quantity and quality of links pointing to that page. The underlying principle of PageRank is that pages with a higher number of reputable and relevant inbound links are deemed more important and are thus positioned higher in search results. By analysing the link structure of the web, PageRank provides a valuable measure of a page’s authority and significance, helping users discover the most relevant and trustworthy content based on their search queries [23].

The PageRank algorithm, originally devised for ranking web pages based on their importance and relevance, can be adapted and applied to the identification of hate speech spreaders online. Just as PageRank assigns higher importance to websites with more inbound links, a similar approach can be employed to identify users who disseminate hate speech through connections in the network. This paper modifies the PageRank algorithm into a Hate Speech Potential Rank algorithm, where the nodes represent individuals in an online social network, and the edges denote connections between users, such as followers or friends. By assigning weights to these connections based on the influence and impact of the users, the algorithm can calculate a Hate Speech Potential Rank (HSPR) for each individual. Users with higher HSPR values are more likely to be hate speech spreaders, enabling the identification of key actors in the propagation of harmful content. Section 4.4 describes the HSPR algorithm in detail.

2.4. K-Means Clustering

In the realm of hate speech detection and spreader identification, prior research has predominantly relied on supervised machine learning techniques, where labelled datasets are utilized to train models for classification tasks. While these approaches have demonstrated effectiveness, they often encounter challenges in coping with the dynamic and rapidly evolving nature of online content. To address this challenge, the research under consideration adopts an innovative unsupervised methodology. To evaluate the effectiveness of this unsupervised approach, K-Means clustering, a widely recognized technique for grouping data points based on similarity, is applied to the results generated by the 'Hate Speech Potential' metric and the 'Hate Speech Potential Rank' algorithm. This unsupervised evaluation approach yields valuable insights into the grouping and distribution of users displaying potential proclivities for hate speech propagation. This method, distinct from traditional supervised techniques, offers a complementary perspective within the broader field of hate speech research, revealing latent patterns and contributing to a more comprehensive understanding of hate speech dynamics within online social networks.

3. CHOOSING TWITTER

Twitter can be a valuable social media platform for identifying hate speech spreaders due to several reasons. Firstly, new data suggests that hate speech is on the rise on Twitter [2, 5]. Secondly, Twitter is a widely used platform with a large user base, providing a diverse pool of individuals to analyse for hate speech dissemination. Thirdly, Twitter’s public nature allows for easier access to user-generated content, making it relatively straightforward to observe and monitor conversations and interactions. Additionally, Twitter’s network structure, with its follower-followee relationships, offers insights into the connections and influence among users, aiding in the identification of potential spreaders. Lastly, Twitter’s real-time nature facilitates the timely detection and response to hate speech incidents, enabling prompt intervention and mitigation strategies. Overall, Twitter’s popularity, public accessibility, network structure, and real-time nature make it a suitable platform for identifying hate speech spreaders and implementing effective countermeasures.

4. PROPOSED MODEL

A three-phase model is presented in this paper. In Phase 1, the emphasis is on the identification of a reliable and robust classifier that effectively and accurately categorizes a post as ’hate speech’. In Phase 2, the predicted labels are utilized to calculate the initial rank HSP for each user within the social network. Section 4.3 provides a detailed description of the HSP metric. During Phase 3, the HSP calculated in Phase 2 serves as the initial ranks for the modified HSPR algorithm, which is then executed to assign final HSPR values to every user in the network. This process ensures a comprehensive evaluation of each user’s hate speech potential based on their prior HSP values and the algorithm’s ranking mechanism.

4.1. Hate Speech Classifier

The paper’s primary objective is to identify hate speech spreaders. Given the existence of current state-of-the-art classifiers, a pre-trained model was employed as a classifier to generate the predicted labels. This approach leverages the capabilities of an already established and high-performing classifier to accurately detect hate speech, enabling the subsequent identification of individuals responsible for its dissemination. In this paper, the HateSonar Model used in [15] was employed to detect hate speech in this paper. Fig 2 shows the steps of this phase.

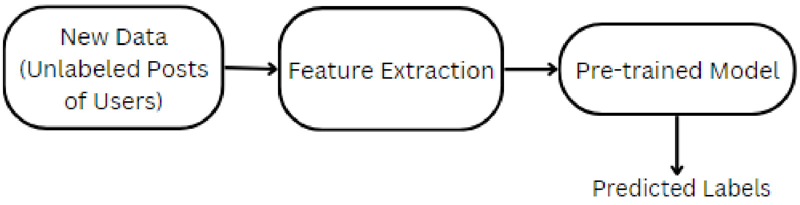

Figure. 2. Generating Predicted Labels using a Pre-trained Classifier

4.2 Tweet List Generation

The choice of Twitter as the social media platform for this study was motivated by the reasons outlined in Section 3. However, due to Twitter's privacy policy and data collection limitations, obtaining a dataset with a complete history of tweets for a specific user posed a challenge. The Tweepy[19] Python library provides convenient access to the Twitter API to collect tweets and relevant data. However, Tweepy has a limitation of retrieving data only for the past 7 days. Therefore, an alternative approach proposed in [14] was used in this paper to generate the tweet lists.

An existing dataset of labelled tweets [21], was used to generate a random list of tweets for each user on the network. This dataset contained tweets that were labelled as 'hate-speech', 'offensive' or 'neither'. A random number generator was used to determine the length of each list for each user. Based on this random number, tweets were selected randomly from the existing dataset and assigned to users. Fig 3 illustrates an example of tweet list generation using this approach. To transform the dataset into a binary classification problem, all tweets labelled as 'hate-speech' or 'offensive' were grouped together to form the 'hate-speech' class, while tweets labelled as 'neither' constituted instances of the other class. This conversion allowed for the classification task to distinguish between the 'hate-speech' class and the 'non-hate-speech' class. Table 1 describes this dataset.

Table 1. Heading and text fonts.

#Hate_Speech Posts	**#Not_Hate_Speech Posts**
20620	4163

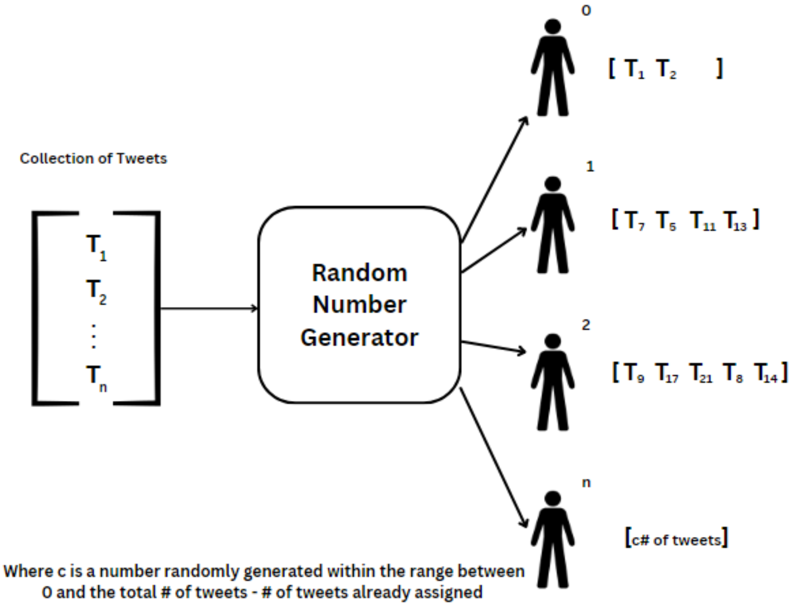
Collection of Tweets
T_1
T_2
⋮
T_n
Random Number Generator
0
[T_1 T_2]
1
[T_7 T_5 T_{11} T_{13}]
2
[T_9 T_{17} T_{21} T_8 T_{14}]
n
[c# of tweets]
Where c is a number randomly generated within the range between 0 and the total # of tweets - # of tweets already assigned

Figure. 3. Tweet List Generation for Each User

4.3. HSP Formulation

In this paper, a new data-driven metric called ’Hate Speech Potential’ (HSP) is proposed to assess the level of hate speech involvement of users. This metric is similar to the metric proposed in [1]. It is represented on a scale of 0 to 1, where 0 indicates no involvement in hate speech and 1 indicates the highest level of hate speech participation. The formulation of the HSP metric involves the following steps:
(1) Using the classifier in Section 4.1 to label the posts of each user.
(2) Determining the ratio of the number of hate speech posts to the total number of posts made by a user within a specific timeframe

$$\mathbf{HSP(U) = |Hate\ Speech\ Posts|/|List\ of\ Posts|}$$

where HSP(U) denotes the Hate Speech Potential of user U, |Hate Speech Posts| denotes the total number of tweets that were classified as hate-speech by the classifier and |List of Posts| refers to the total number posts that the user makes in a given time period. The tweet list generated in Section 4.2 is used to calculate the HSP of every user.

4.4. Hate Speech Potential Rank (Hspr) Algorithm

In this paper, a modified PR algorithm called ’Hate Speech Potential Rank’ (HSPR) algorithm was proposed. The PageRank algorithm, originally developed for ranking web pages according to their significance and relevance, can be adapted and applied to the detection of hate speech spreaders in online platforms. Similar to how PageRank assigns greater importance to websites with more incoming links, a similar approach can be employed to identify users who propagate hate speech through their connections in the online social network. In this modified algorithm, individuals are represented as nodes in the network, while connections between users, such as followers or friends, are represented as edges. This approach leverages the network structure to identify users who have a higher likelihood of spreading hate speech based on their connections to influential users. The initial ranks of the nodes are no longer the traditional $1/(number\ \_o\ f\ \_nodes)$ set for a PR algorithm but instead is now the HSP value calculated in Section 4.3. The motivation behind the modified initial ranks stems from the recognition that each user possesses their own potential for generating hate speech posts. The HSP is considered a more representative measure for determining the initial ranks, as opposed to the traditional approach of assigning $1/(number\ \_o\ f\ \_nodes)$.

In the PR algorithm, the rank of a webpage increases when it has connections to popular pages. Similarly, for the HSPR algorithm, it is anticipated that the final ranks (HSPR values) will be higher for individuals who are connected to others with a high initial HSP value. This is because these individuals are exposed to more hate speech content posted by the users they follow. The influence of connections with high HSP users is expected to contribute to the higher ranks of individuals in the network, indicating their potential involvement in the spread of hate speech. In this context, a popular node refers to an individual who frequently shares a significant amount of hate speech content (a high HSP value). By mapping the principles of PageRank to hate speech detection, this algorithm provides a valuable tool for understanding and addressing the dissemination of hate speech online, ultimately contributing to the creation of safer and more inclusive digital spaces.

4.5. Input Graph For The HSPR Algorithm

A network can be visualized as a directed graph denoted as G = (V, E), where V represents the set of vertices representing users, and E represents the set of edges representing the relationships shared among them. On Twitter, these relationships are represented as directed edges, indicating that if user X follows user Y, there exists a directed edge (X,Y) in the graph. To represent a snapshot of the relationships within the Twitter network, a graph dataset [18] was utilized due to limited access to the original Twitter network. The dataset contained a two-columned edge set where a tuple of (v,u) indicated that user ’v’ follows user ’u’. A subset of this dataset was used to accommodate 400 users. Table 2 describes this dataset. Each edge represents a relationship where one user follows another user in the network, while each node represents a Twitter user

Table 2. Tweets Dataset

Number of Nodes	Number of Edges
400	36800

Algorithm 1 describes the pseudo code for the HSPR algorithm.

Algorithm 1 Hate Speech Potential Rank (HSPR)

```
Require: Graph $G$, damping factor $d$, maximum iterations $maxIterations$
Ensure: Hate Speech Potential Rank values $HSPR$
$N \leftarrow$ number of nodes in $G$
Initialize $HSPR$ array with $N$ users, each set to HSP
for iteration $\leftarrow$ 1 to $maxIterations$ do
    Create $newHSPR$ array with $N$ users, each set to 0
    for node in $G$ do
        for neighbor in node.neighbors do
            $newHSPR[\text{neighbor}] \leftarrow newHSPR[\text{neighbor}] + \frac{HSPR[\text{node}]}{\text{neighbor.outDegree}}$
    for node in $G$ do
        $HSPR[\text{node}] \leftarrow (1 - d)\ /N + d \times newHSPR[\text{node}]$
return $HSPR$
```

5. Results and Observations

For all the experimentation, the damping factor was set to 0.85 and maxIterations was set to 100. Initial experiments were done on small graphs to assess the preliminary results. Table 3 shows the graphs that were experimented with. Graph 1 and Graph 2 were small graphs that were randomly generated. Graph 3 was the Twitter graph built in Section 4.4.1 Before feeding the initial ranks into the HSPR algorithm, the values were normalized to achieve a sum of 1. Fig. 4 depicts the small graph of 4 nodes (Graph 1) with the initial HSP values.

5.1. Increase in Rank

The experimental findings revealed a notable rise in the ranks of several users. Specifically, in Graph 1 (Fig. 4), a substantial increase in the ranks of nodes B, C, and D was observed. One possible explanation for this occurrence is that these nodes were connected to node A, who initially possessed a remarkably high Hate Speech Potential (HSP) score, indicating frequent dissemination of hate speech content by node A. Consequently, the followers of node A were exposed to harmful content, contributing to the elevation of their ranks in the network. This observation underscores the influence of influential nodes on the ranking of their followers within the context of hate speech propagation. The observed results align with the anticipated theory proposed in Section 4.4. The empirical evidence supports the hypothesis that individuals who are connected to users with high HSP values tend to have higher HSPR values themselves, indicating a greater likelihood of engaging in the spread of hate speech. These results validate the underlying principles of the proposed algorithm and reinforce its utility in identifying influential hate speech spreaders within the network.

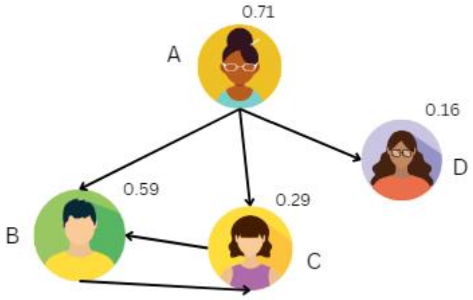

Figure. 4. Tweet List Generation for Each User

5.2. Decrease in Rank

Some of the findings also revealed a decrease in the ranks of certain nodes. For example, in Graph 1, node A exhibited a decrease in rank. This decline can be attributed to a decrease in the number or quality of incoming links to the node. In the algorithm, nodes with higher weights on their incoming links contribute more to the rank of a node. Therefore, a reduction in incoming links can lead to a lower rank. In the case of Graph 1, node A had no incoming edges, indicating that it did not follow any other nodes in the network. Consequently, the rank of node A decreased. One potential solution to address this issue is to consider the maximum value between the original HSP and the final HSPR as the final ranking for a node ($max\ (HSP, HSPR)$).

5.3. Potential Spreaders using Threshold

All the ranks obtained from the algorithm were values between 0 and 1. In choosing the threshold, the desired sensitivity of the analysis was considered. Firstly, the "fixed threshold" method was employed which involved setting a constant value between 0 and 1 as the threshold. This approach provided a straightforward and adjustable threshold that can be customized based on specific requirements or domain expertise. For a fixed value, a lower threshold would include more values above the threshold, while a higher threshold would be a more selective approach. In order to identify potential spreaders more effectively, initial experiments were conducted with a threshold value of 0.8. The HSP values vector can be seen as a probability distribution over all the users. Since the sum of all of the values will be 1, getting users with HSPR value greater than a high value threshold would be unlikely. As determining an appropriate fixed threshold value proved challenging, alternate methods were proposed. The second method proposed was the "avg_sum" method that calculated the average of the final Hate Speech Potential Rank (HSPR) values for all nodes. This approach aimed to determine a threshold based on the overall average HSPR values, considering the collective behavior of the network. Alternatively, the "avg_max_min" method was proposed which calculated the average of the maximum and minimum HSPR values among the list of ranks. By considering both the highest and lowest HSPR values, this approach aimed to find a threshold that balances extreme values within the network. The list below summarizes the three methods.
(1) fixed threshold: Set a constant value between 0 and 1 as a threshold
(2) avg_sum: Average of the final HSPR values of the nodes.
(3) avg_max_min: Average of the max and min HSPR values among the list of ranks.

5.4. Evaluation using K-Means Clustering

To assess the effectiveness of our approach in identifying #hate speech spreaders, we employ K-Means clustering as an evaluation measure. After employing our novel data-driven metric, 'Hate Speech Potential,' and the 'Hate Speech Potential Rank' algorithm to detect and pinpoint potential spreaders of hate speech, we utilize K-Means clustering to categorize these users into distinct clusters. This clustering technique allows us to group users based on their sharing patterns and 'Hate Speech Potential,' thereby enabling a comprehensive evaluation of our methodology. By inspecting the resulting clusters, we can gain insights into the distribution and concentration of users who exhibit tendencies toward hate speech propagation. This evaluation not only validates the effectiveness of our approach but also provides a nuanced understanding of the network dynamics related to hate speech dissemination, offering valuable insights for future interventions and strategies to combat online hate speech.

The clustering results closely align with the outcomes of the proposed approach for identifying #hate speech spreaders. Notably, the number of data points within each cluster obtained through

K-Means clustering closely matches the results generated by the data-driven 'Hate Speech Potential' metric and the 'Hate Speech Potential Rank' algorithm. This congruence highlights the effectiveness of the unsupervised methodology in early identification of users prone to hate speech dissemination. Furthermore, it's noteworthy that users labelled as spreaders using the proposed approach were found to be part of the same cluster. This underscores that the clustering process successfully groups users with similar characteristics and tendencies, offering empirical validation of the proposed approach's ability to discern and classify potential hate speech spreaders within the online social network.

Table 3 presents the average number of potential spreaders obtained using the methods in Section 5.3 and compares it to the result obtained by the K-Means Clustering algorithm.

Table 3. Graph Properties and Potential Spreaders

Graph Name	**Number of Nodes**	**Number of Edges**	**#Spreaders (avg_sum)**	**#Spreaders (avg_max_min)**	**#Spreaders (K-Means)**
Graph 1	4	5	2	3	2
Graph 2	20	187	8	6	9
Graph 3	400	36800	189	134	183

6. CONCLUSIONS

This paper has presented a modified PageRank algorithm tailored specifically for the identification of hate speech spreaders. By leveraging the principles of PageRank and adapting them to the context of online social networks, the algorithm demonstrates the capability to identify a pool of individuals actively involved in the dissemination of hate speech. The algorithm assigns Hate Speech Potential Rank (HSPR) values to users based on the influence and impact of their connections, enabling the identification of key actors in the propagation of harmful content. Through the application of this algorithm, the potential to contribute to the understanding and addressing of the dissemination of hate speech online was showcased. The use of the HSPR algorithm provides a valuable tool for detecting and profiling hate speech spreaders, leading to targeted interventions and the implementation of appropriate measures to mitigate the negative consequences of hate speech. It is important to consider that the ethical and legal implications of hate speech spreader detection and intervention is crucial in order to strike a balance between freedom of expression and the prevention of harm. In conclusion, the modified HSPR algorithm for hate speech spreader detection holds great promise as a valuable tool in the ongoing fight against hate speech online. By combining technical advancements with a comprehensive understanding of the social dynamics and challenges associated with hate speech, one can work towards fostering a more inclusive and respectful online environment for all users

7. FUTURE WORK

Further research and efforts are needed to continuously improve the proposed algorithm’s accuracy, scalability, and applicability across different social media platforms. In future work, several avenues can be explored to further enhance the detection and mitigation of hate speech spreaders. Firstly, incorporating user profile information such as demographics, past behavior, and social network features can provide valuable insights into the motivations and patterns of hate speech spreaders. Additionally, investigating the use of advanced machine learning techniques, such as deep learning and natural language processing, can improve the accuracy and efficiency of hate speech detection algorithms. Furthermore, exploring the integration of context-aware analysis, including temporal dynamics and regional variations, can provide a more nuanced understanding of hate speech propagation across different platforms and communities. Finally, considering the ethical and legal implications of hate speech detection and mitigation is crucial, and future research should focus on developing responsible and unbiased approaches to address these challenges. Overall, these directions for future work can contribute to the development of more effective and comprehensive strategies to combat hate speech and promote a safer online environment

8. ACKNOWLEDGEMENTS

The authors would like to thank CJ Oshiro, a junior in the department of Mathematics and Computer Science at Santa Clara University for his contribution to some of the experiments and implementation of the proposed approach. The authors would also like to thank the department of Mathematics and Computer Science at Santa Clara University for their constant support.

REFERENCES

[1] 2022. Depression Detection using Machine and Deep Learning Models to Assess Mental Health of Social Media Users., 01 pages. https: //doi.org/10.5121/csit.2022.122101 [Online; accessed 29. May 2023]

[2] 2023. Analysis finds hate speech has significantly increased on Twitter. https://phys.org/news/2023-04-analysis-speech-significantly-twitter.html.

[3] 2023. Hate Speech Policy—YouTube Help. https://support.google.com/youtube/answer/2801939. Accessed: 2022.

[4] 2023. Hateful Conduct Policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Accessed: 2022.

[5] 2023. New Data Suggests that Hate Speech is on the Rise on Twitter 2.0. https://www.socialmediatoday.com/news/New-Report-Suggests-Hate-Speech-Rising-on-Twitter/645482/.

[6] Community Standards. https://www.facebook.com/communitystandards/objectionable_content. Accessed: 2022.

[7] Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021. HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier.. In CLEF (Working Notes). 1829–1836.

[8] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet 7, 2 (2015), 223–242.

[9] Riccardo Cervero. 2021. Use of lexical and psycho-emotional information to detect Hate Speech Spreaders on Twitter. In CEUR Workshop Proc, Vol. 2936. 1883–1891.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media, Vol. 11. 512–515.

[11] Andrew Duffy, Edson Tandoc, and Rich Ling. 2020. Too good to be true, too good not to share: the social utility of fake news. Information, Communication & Society 23, 13 (2020), 1965–1979.

[12] FBI. 2015. 2015 Hate Crime Statistics. Retrieved from https://ucr.fbi.gov/hate-crime/.

[13] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) 51, 4 (2018), 1–30.

[14] Smita Ghosh, Pramita Das, Sneha Ghosh, and Diptaraj Sen. 2022. Detection of Clickbait Content Spreaders on Online Social Networks. In 2022 5th International Conference on Information and Computer Technologies (ICICT). IEEE, 23–28.

[15] Pulkit Gigoo, Rishabh Tiwari, Saikrishna Ramesh, Sanket Bendrey, and Rupali Tornekar. 2022. Sentiment Analysis of Religious Tweets. (2022).

[16] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In Proceedings of the 28th ACM conference on hypertext and social media. 85–94.

[17] John T. Nockleby. 2000. Hate Speech. Encyclopedia of the American Constitution 3 (2000), 1277–1279.

[18] Tore Opsahl. 2013. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social networks 35, 2 (2013), 159–167.

[19] Joshua Roesslein. 2018. tweepy documentation. Online. https://tweepy.readthedocs.io/en/latest/

[20] Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, Marco La Cascia, et al. 2021. Detection of hate speech spreaders using convolutional neural networks.In CLEF (Working Notes). 2126–2136.

[21] t davidson. 2023. hate-speech-and-offensive-language. https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data [Online; accessed 30. May 2023].

[22] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. science 359, 6380 (2018), 1146–1151.

[23] Wikipedia contributors. 2023. PageRank — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=PageRank&oldid=1154246079 [Online; accessed 26-May-2023].

[24] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

AUTHORS

Dr. Smita Ghosh is currently an Assistant Professor in the department of Math and Cs at Santa Clara University. She graduated with her PhD in Computer Science from The University of Texas at Dallas in May 2020. Prior to that she also has her Master’s in Computer Science from the same university. Her research interests are Social Network Analysis (composed influence maximisation, information diffusion, hypergraphs, data analysis, machine learning in social networks analysis), Data Science (deep learning, data mining, trustworthy artificial intelligence, sentiment analysis, ethics in artificial intelligence)

Shiv Jhalani is pursuing his bachelor's degree in Computer Science in the department of Mathematics and Computer Science at Santa Clara University. His primary areas of interest which also constitutes his area of research are machine learning, natural language processing and social network analysis.