

FADOHS: FRAMEWORK FOR DETECTION AND INTEGRATION OF UNSTRUCTURED DATA OF HATE SPEECH ON FACEBOOK USING SENTIMENT AND EMOTION ANALYSIS

¹ALETI PREETHI,²MEDARAMETLA NAGA SRIJA,³KAMSANI TEJASHWINI,⁴VARIKILLA SNICE,⁵MS.ASMITA PANKAJ AMBEKAR

^{1,2,3,4}Students, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous), Hyderabad Telangana, India 500100

⁵Associate Professor, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous), Hyderabad Telangana, India 500100

ABSTRACT

Hate speech refers to expressions that target individuals or communities based on their race, origin, religion, sexual orientation, or other characteristics. While hate speech can be delivered in various formats, both online and offline, the rise of social media platforms has significantly amplified its occurrence and intensity. The focus of this research is to detect and analyse unstructured data from selected social media posts that propagate hate in comment sections. To tackle this issue, we propose a novel framework, FADOHS, which integrates data analysis and natural language processing techniques to raise awareness among social media providers about the widespread nature of hate speech. Specifically, sentiment and emotion analysis algorithms are applied to examine recent posts and comments. Posts that are suspected of containing harmful language are processed before being input into a clustering algorithm for further evaluation. The experimental results show that the FADOHS framework outperforms current methods in precision, recall, and F1 scores by approximately 10%.

Keywords: hate speech, social media, sentiment analysis, emotion analysis, natural language processing, FADOHS framework, clustering algorithm, precision, recall, F1 score.

INTRODUCTION

Mark Zuckerberg, CEO of Facebook, once stated, “Hate speech and racism do not have a place on Facebook.” While Facebook has implemented various artificial intelligence (AI) techniques to combat hate speech, challenges still persist. The company reported that in Q1 2018, it removed 2.5 million pieces of hate speech, with 38% of them being flagged by its technology. However, AI alone struggles with the core question: What exactly constitutes hate speech? This question leads to various definitions, such as: “Hate speech is public expressions which spread, incite, promote or justify

hatred, discrimination, or hostility toward a specific group,” and “Hate speech is a direct attack on people based on characteristics like race, ethnicity, religion, sexual orientation, gender identity, and disability.” Facebook has acknowledged that AI is not yet advanced enough to distinguish between someone promoting hate and someone merely discussing an experience. Furthermore, hate speech can also be subtly encouraged, such as by discussing controversial topics designed to provoke hateful comments. Despite efforts to address this issue, hate comments remain prevalent, and covert discriminatory practices continue to thrive on Facebook, often going undetected by



algorithms. This ongoing issue has prompted various studies on hate speech classification methods, such as one that utilizes morpho-syntactic features, sentiment polarity, and word-embedded lexicons to classify hate speech in Italian.

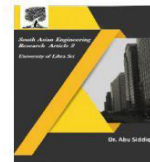
Building on this body of work, our research focuses on identifying and analyzing hate speech, particularly covert forms of hate expressed in Facebook comments on controversial topics. We begin by identifying a set of Facebook pages that discuss topics such as immigration, race, and religion. These “seed” pages are then used to construct a network through Facebook's "follow" relationship. By applying graph analysis, we identify influential pages that spread hate speech and crawl their latest posts and comments. We utilize sentiment and emotion analysis algorithms to detect posts with highly negative tones, particularly those that may incite hatred. These posts are then clustered using the K-means algorithm based on their topics. Our framework successfully identifies sensitive topics that are likely to promote hate speech.

The contributions of this study include the development of a semi-automatic method for discovering pages discussing sensitive topics, an automatic method for clustering posts based on these topics, and a new framework for detecting hate speech. Experimental results show that our framework outperforms existing methods, achieving improvements of approximately 10% in precision, recall, and F1 scores. The rest of the paper is organized as follows: Section II provides an overview of related works in hate speech detection, Section III details the system architecture, Section IV presents the experiments and

results, and Section V concludes with future work and insights.

Related Works
Few studies have fully explored the extent of hate speech on social media using automated tools. This gap in research motivates our framework, which presents a novel, in-depth approach to hate speech detection. Notable related work includes studies on overt hate and covert disrespect on social media platforms. One such study by Ben-David and Matamoros-Fernandez examines hate speech through network and multimodal analyses, retrieving data from Facebook pages associated with hate speech. Our proposed framework similarly uses Facebook graph APIs and sentiment/emotion analysis tools, building upon this approach.

Other studies have utilized the VADER tool, a rule-based sentiment analysis model, to analyze social media data. While we also use VADER for sentiment analysis, we complement it with the JAMMIN tool for emotional analysis, specifically to track posts with negative comments. Further research has developed typologies for categorizing hate speech, such as one study that created classifiers for Italian based on morpho-grammatical features and sentiment lexicons. Unlike these studies, our approach focuses on detecting hate speech specifically in Facebook comments on divisive topics. Another study examined the challenges faced by Facebook and Twitter in detecting hate speech, utilizing a crowd-sourcing tool to assess the quality of service for platform providers. While their tool aimed at identifying content violating platform policies, we found it inefficient compared to our more direct method of filtering hate-filled posts.



Research on platform racism, which explores racial prejudice emerging from social media platforms, also informs our study. A study examining the link between social media activity and offline hate crimes underscores the role of platforms like Facebook in propagating hate speech. Our work complements this by targeting and identifying negative comments posted on pages that promote hate. Moreover, several studies have explored optimization algorithms for automatic hate speech detection. However, our framework improves upon these methods by not only identifying unstructured data promoting hate speech but also categorizing them into clusters based on topic, enhancing the detection process. Additionally, we tested the quality of our framework using OpenAI's GPT-2 model, aiming to optimize our dataset and improve hate speech detection accuracy.

This literature review lays the foundation for our study and highlights the potential for further research. Our primary contribution is the development of a reliable system for identifying social media pages discussing sensitive topics, categorizing posts, and integrating unstructured data that spread hate speech.

LITERATURE SURVEY

Zuckerberg (2010) discusses the intersection of hate speech and social media, specifically on Facebook, highlighting the role of the platform in the spread of harmful rhetoric related to the refugee crisis. This work emphasizes how the platform has become a breeding ground for controversial debates, particularly in the context of migration and refugees.

Facebook (2018) reports on its efforts to remove hate speech, disclosing that it removed 2.5 million pieces of hate speech content in the first quarter of 2018. This provides insight into the scale of the issue on the platform and the company's response to curbing harmful content.

ILGA-Europe (2018) explores the broader issue of hate crime and hate speech, providing a detailed examination of how these issues manifest in society and across social media platforms. The document outlines the legal and social implications of hate speech and emphasizes the need for stronger regulation and enforcement, especially on platforms like Facebook.

Facebook (2020) outlines its community standards and the policies in place to detect and remove harmful content, including hate speech. The platform's official stance on what constitutes hate speech and its commitment to maintaining a safe environment for users are central to understanding the company's approach to this issue.

CNBC (2020) highlights the limitations of Facebook's artificial intelligence in detecting hate speech, noting that while the platform's algorithms can detect nudity, they still struggle with accurately identifying hate speech. This study is crucial in evaluating the technological challenges involved in the automatic detection of harmful content.

Chinnasamy and Manaf (2018) examine how social media has become a tool for political hatred, particularly



in the context of Malaysia's 2018 general election. This paper discusses how political actors use social media platforms to spread negative and divisive rhetoric, contributing to the broader phenomenon of political polarization and hate speech.

Matamoros-Fernández and Farkas (2021) provide a systematic review and critique of the relationship between racism, hate speech, and social media. This study critically analyzes existing literature on the topic, exploring how online platforms, particularly Facebook, have been used as spaces to propagate racist ideologies and hate speech.

Del Vigna et al. (2017) present an approach for detecting hate speech on Facebook by analyzing user posts. Their method incorporates natural language processing techniques, and they explore the potential of automated systems in identifying and categorizing hate speech across the platform. This study forms the foundation for more advanced AI-based detection systems.

PROPOSED METHODOLOGY

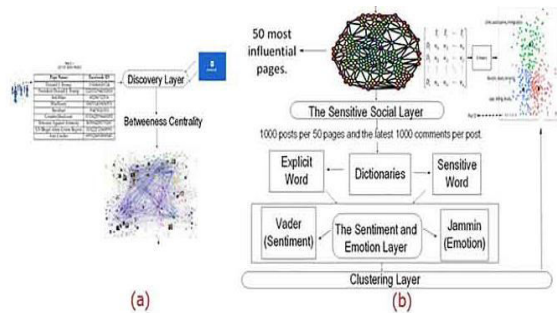
This research aims to develop a framework (FADOHS) designed to first detect hate speech on Facebook and then integrate unstructured data through clustering techniques using sentiment and emotion analysis. The framework will identify key elements from the unstructured data found in posts and comments on Facebook pages that are suspected of promoting hate speech. To evaluate the effectiveness of various data analysis and natural language processing strategies, we will implement FADOHS in four distinct stages. Ultimately, the goal is to showcase the

power of hate speech clustering by utilizing hybrid methods of data analysis and natural language processing to categorize information into different groups based on the intensity of hatred being expressed.

The architecture of the system comprises four layers. The discovery layer serves as the initial stage, followed by layers focused on sensitive social data collection, sentiment and emotion analysis, and the final clustering layer. These components work together to form a comprehensive system for detecting and analysing hate speech on social media platforms.

In the discovery layer, we address the challenges involved in detecting hate speech online and propose potential solutions based on our framework. To assist in the identification of sources that might promote hate speech, we interviewed individuals familiar with American politics and compiled a list of American celebrities or entities likely to engage in such behaviour through their posts. The Facebook Graph API was employed to extract pages that received "likes" from followers, which served as our "seeds" for further exploration. This process was repeated for each extracted page, extending to a third layer of pages, thus creating a three-level social graph. This resulted in a directed graph comprising 17,176 pages and 46,968 "likes" (or "Follows"), reflecting a one-way relationship. It is crucial to note that promoting hate speech in this context does not necessarily mean that a seed page directly bullies individuals; rather, it indicates the presence of sensitive topics. A preliminary review of the pages revealed that many were simply commercial brands with little to no association with hate

speech promotion. Therefore, it would be inaccurate to assume that all pages in the graph are inherently promoting hate speech, as some could represent the expression of free speech.



CONCLUSION

In this study, we propose FADOHS, a framework designed to identify and integrate unstructured data from Facebook pages that are suspected of promoting hate speech. The goal is to identify the typical topics discussed on these pages, which often stir negative emotions among followers. This task is particularly challenging because non-personal Facebook pages and accounts usually avoid using explicit terms to evade removal from the platform or criticism. Despite this, many pages still manage to provoke hate speech by discussing controversial topics while keeping their language relatively mild. Our proposed framework addresses this issue by clustering posts and comments, detecting frequently discussed topics that generate hate speech, and identifying hate speech within them.

FADOHS combines graph analysis, sentiment and emotion analysis, and clustering techniques to effectively analyze posts that may contain hate speech. The process begins by analyzing a small set of pages known to discuss sensitive topics that could provoke hate comments. From this, we construct a three-level social graph and use graph analysis to identify key pages. We then apply sentiment and emotion analysis using predefined

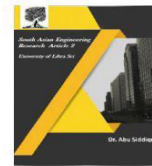
dictionaries to isolate posts with a significant level of negativity in their comments. This results in the identification and integration of unstructured data from pages that are likely to promote hate speech. The next critical step involves categorizing this data, which is done using the K-means clustering algorithm. We test different configurations to uncover meaningful groups of topics. Afterward, we manually analyze the posts within each group and assign labels to each cluster. By comparing the manual labels with the cluster centroids, we confirm that the variables align, thus validating the effectiveness of our approach.

Our experiments show that FADOHS can identify several pages that allegedly promote hate speech and related topics from a small set of seed pages. This work clearly demonstrates how unstructured data, such as Facebook posts, can be analyzed meaningfully through the application of this framework. The experimental results indicate that the FADOHS framework outperforms the state-of-the-art approaches, achieving approximately a 10% improvement in precision, recall, and F1 scores.

In future research, we plan to extend the application of our framework to include comments and their replies, aiming to accurately identify individuals suspected of promoting hate speech. The long-term benefits of this research could be substantial, as it may help detect cyberbullies and cyberterrorists. Additionally, we aim to conduct a more thorough examination of the emotion filtering and clustering results to determine the most reliable setup for optimizing outcomes.

REFERENCES

- [1] Zuckerberg, "Refugee Crisis: Hate Speech Has Place Facebook," Honolulu,



- HI, USA, 2010.
- [2] "Facebook Removed 2.5 Million Pieces Hate Speech 1st Quarter," Jul. 2018. [online].
- [3] "Hate Crime & Hate Speech," May 2018. [online] Available: <https://www.ilga-europe.org/what-we-do/ouradvocacy-work/hate-crime-hate-speech>.
- [4] "Community Standards Home," May 2020. [online] Available: <https://www.facebook.com/communitystandards/>.
- [5] "Facebook's Artificial Intelligence Still Has Trouble Finding Hate Speech—But it Finds a Lot of Nudity," May 2020. [online] Available: <https://www.cnbc.com/2018/05/15/facebook-k-artificial-intelligence-still-finds-it-hard-to-identify-hate-speech.html>.
- [6] S. Chinnasamy and N. A. Manaf, "Social media as political hatred mode in Ts 2018 general election," SHS Web Conf., vol. 53, pp. 2005, 2018.
- [7] A. Matamoros-Fernández and J. Farkas, "Racism hate speech and social media: A systematic review and critique," *Telev. New Media*, vol. 22, no. 2, pp. 205-224, Feb. 2021.
- [8] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me hate me not: Hate speech detection on Facebook," *Proc. 1st Italian Conf. Cybersecur. (ITASEC)*, pp. 86-95, 2017.
- [9] M. Ahmed, R. Seraj, and S. M. S. Islam, "The K-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, pp. 1295, Aug. 2020.
- [10] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-Learning environment: Clustering using K-means," *Amer. J. Distance Educ.*, vol. 34, no. 2, pp. 137-156, Apr. 2020.
- [11] Z. Lv, T. Liu, J. A. Benediktsson, and H. Du, "Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images," *IEEE Access*, vol. 7, pp. 34425-34437, 2019.
- [12] D. Kucukusta, M. Perelygina, and W. S. Lam, "CSR communication strategies and stakeholder engagement of upscale hotels in social media," *Int. J. Contemp. Hospitality Manage.*, vol. 31, no. 5, pp. 2129-2148, May 2019.
- [13] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on Facebook using sentiment and emotion analysis," *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIIC)*, pp. 169-174, Feb. 2019.
- [14] G. C. Santia and J. R. Williams, "BuzzFace: A news veracity dataset with Facebook user commentary and egos," *Proc. 12th Int. AAAI Conf. Web Social Media*, pp. 531-540, 2018.
- [15] A. Chopra, A. Dimri, and S. Rawat, "Comparative analysis of statistical classifiers for predicting news popularity on social web," *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, pp. 1-8, Jan. 2019.
- [16] B. Lin, F. Zampetti, G. Bavota, M. D. Penta, M. Lanza, and R. Oliveto, "Sentiment analysis for software engineering: How far can we go?" *Proc. IEEE/ACM 40th Int. Conf. Softw. Eng. (ICSE)*, pp. 94-104, 2018.
- [17] V. Franzoni, Y. Li, and P. Mengoni, "A path-based model for emotion abstraction on Facebook using sentiment analysis and taxonomy knowledge," *Proc. Int. Conf. Web Intell.*, pp. 947-952, Aug. 2017.