A Review on Machine Learning Techniques for Identification of Abusive and Hateful Speech

Priyanka Goswami[1], Prof. Preetish Kshirsagar[2]

Abstract: Natural language processing has a large number of applications especially for semantic web. This study provides a comprehensive assessment of contemporary machine learning and deep learning algorithms for hate speech identification. Examples of such occurrences include hate speech, abusive language, threats, and derogatory remarks. Hate speech constitutes abuse that is not limited to a single gender; it affects all individuals. In the present context, comprehending the dynamic patterns (incidents, geographical predominance, demography, etc.) is essential for formulating ways to analyse hate speech activities. Social media platforms function as an information system that aggregates and categorises hate speech data from diverse sources, mostly users. This aggregated information is analysed to discern insightful patterns from vast quantities of social media data, which cannot be monitored continuously. Contextual interdependence across diverse lexicons in data will be essential for identifying hate speech. Current research reveals a scarcity of studies focused on hate speech detection concerning user behaviour. This study addresses hate speech as an online exponential issue aimed at damaging targeted individuals. Such incidents exacerbate social disparities and asymmetries by rendering online spaces unfriendly and inaccessible.

Keywords: Natural Language Processing, Radical and Hateful Content, Social Media, Contextual Dependency, Classification Accuracy.

1. Introduction

In an increasingly digital world, online platforms have become powerful tools for communication and expression. However, they have also provided a fertile ground for the spread of radical and hateful content. The scale, speed, and anonymity afforded by the internet make manual monitoring insufficient. This has necessitated the adoption of advanced computational techniques like Natural Language Processing (NLP) to automatically identify, filter, and mitigate such harmful content. NLP offers a scalable, data-driven solution to detect nuanced and evolving language patterns used in hate speech and radical ideologies. Social media and its number of users is increasing each year. So is the case of hate speech on social media. Online hate speech is complex, multifaceted, and often context-dependent. It spans racial slurs, xenophobia, misogyny, religious intolerance, and political extremism. Radical content may be disguised through euphemisms, coded language, or sarcasm, making it challenging to detect using traditional keyword-based methods. NLP, with its ability to understand context, sentiment, syntax, and semantics, provides a sophisticated framework to decode such messages. This makes it essential for building systems that can distinguish between free speech and content that incites violence or discrimination.

Hate speech on social media needs to be filtered as hate speech may lead to:

- Serious psychological problems in users.
- Depression and suicide in users.
- Violence

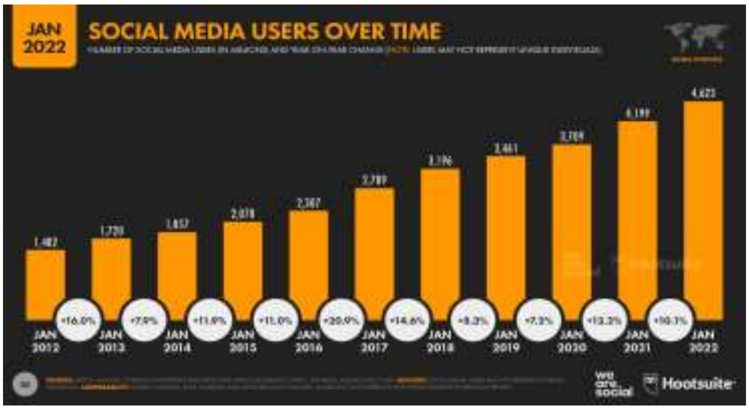

Fig.1 Increase in number of social media users

As the social media web applications are so accessible, harassing behaviors are evolving into new patterns every day which are extremely risky. Therefore, it is necessity of the current era to study and analyze such antisocial patterns in social media. In social media or social network, any user can use offensive language to express hatred towards an individual or a group of people. The motive of such users' is to insult, humiliate, harass, derogation or giving threat over social media or network. Facebook, Twitter, Instagram are continuously improving their policies and providing a new ways to users to eliminate hateful content from the website [2].

Due to large number of web and social application users, abundant amount of data is generated which is noisy and challenging to find hidden patterns. On social media many users are openly posting abusive words for women, and promoting hatred through social posts [2].

Recently, Amnesty International[1] published a report *"Stop online abuse on ToxicTwitter"*. Previously they have published the report *"Toxic Twitter – A Toxic Place for Women"* which clearly indicates that people can be threatening directly based on religion, caste, color, gender etc. Report also suggests that Twitter has no check to protect users against harassment. Therefore, it is very important to fight against online abuse and hate speech. In oxford, hate speech is defined as *prejudice, threat, derogation, animosity,* typically against a person, women or group of people[2].

In general, degrading the image of a person and online threatening is increasing and being replicating online. It is very complex to understand the definition of sexism, but it may sound "social", "negative", "humour", "insulting", "offensive", "derogative" etc. In other words, hate speech can be as malicious and violating which can affect and harm people in numerous way including professional life, carrier opportunities, household-parenting character, sexual image, life growth and expectation are few of them.

In current era, hate speech in online social media is widest spectrum of diverse behaviors and attitudes which is having dangerous results for the society. Thus, the main motive of the research work is to detect hate speech in a broad form. Through this study, our motive will be to study explicit misogyny to other understanding form that involve implicit hate speech behaviors.

In the previous study and to best of my knowledge, no previous work has addressed the analysis and detection of this implicit behavior in social network and applications conversations. Thus, through this research work, my aim is to understand the people attitude expressed in social media conversations. From the conversation and posts over social media users' beliefs and behavior can be predicted. In this research work, my main motive is to extract data written in English language and the proposed method and conclusion extracted can be directly applied to other languages also.

Literature is the field of hate speech analysis in terms of threat, derogation animosity is growing day by day. Multiple evidence is available based on hate speech which use classification based on NLP and machine learning approaches. Researchers used Twitter for extracting data for analysis purpose [5]. Analysis of various data have dependency among various lexicons and for this purpose contextual polarity needs to be addressed. Lack of contextual information needs to be carried further for better understanding. In [6] authors design a contextual information-based methods for analyzing the impact on performance. Authors analyzed the contextual impact and analyzed automatically for Twitter dataset. Numerous experiments have conducted which are based on transformer for contextual information analysis. In [7] authors described the various classes of hate speech using advanced layer of DNNs. Authors used the bidirectional capsule networks, which also analyze the impact of contextual information with forward and backward directions of the input data.

2. Related Work

This section highlights the existing literature which is focusing on the importance of the contributions of this work. The noteworthy contribution in the domain has been highlighted with the salient features of the work done.

Numerous authors have defined various methods of hate speech and some reviews and surveys of hate speech detection issues are also discussed which are available in [10], [11], [12], [13], [14], and [15]. In the research conducted by [11], authors have given the methods of hate speech detection. They applied the approach for an informatics perspective which helps the users to analyze hate speech in social domains. It is considered the second survey on this topic after that of [7], which provided a short overview of hate speech detection within NLP. According to [16], various feature extraction methods are explored by authors. The survey by [11] explained the comparative analysis of various existing hate speech approaches with each other on the basis of common features. Authors also provide a summarized version of statistics on detection methods. A case study discussion on the hate speech terminologies needed to explore is also given including the features involved in hate speech domain. Research on bullying is also conducted in later stage in which they explained different datasets of English for hate speech detection. In another study by [10], a more reliable, accurate, and comprehensive classification of anger-linked social media messages for detecting hate speech was established. This approach helps to identify the anger discourse on social media platforms. With the help of the proposed methodology users' can ensure the various classes of anger which eventually leads to extensive participation in hate crimes.

Similarly, in [14] authors have explained the various classes of hate speech. They explained and classify six hate speech classification and detection models used on a variety of social media sites. The proposed model is based on NLP, data analytics and machine learning domains. Comparative analysis and various between various methods are also discussed in this study.

In a further work by [13], a study using NLP technique is conducted. Various approach like dictionaries, bag-of-words, and n-gram are explained and discussed for hate speech detection. A comparative study to analyze hate speech automatically on social media authors explained the various methods which can be used to detect hate speech on online social media sites.

Misogyny on social media is also spreading and to analyze their impact on social media user is also very important. In [17] authors studied about various methods for analyzing and classifying the nature of misogyny in social media. For this purpose, they consider Twitter as a platform. Approaches like deep learning and machine learning are used to analyze the misogyny behavior in social media Twitter.

Table.1 Comparative Analysis of Baseline Approaches

S.No.	Dataset	Approach	Performance
1	Kaggle Hate Speech and Offensive Language dataset	TDF-ID with Naïve Bayes	Ac=72.27%
2	Kaggle Hate Speech and Offensive Language dataset	TDF-ID with KNN	Ac=85.76%
3	Kaggle Hate Speech and Offensive Language dataset	TDF-ID with Logistic Regression	Ac=90.46%
4	Kaggle Hate Speech and Offensive Language dataset	TDF-ID with Decision Trees	Ac=82.43%
5	Open Super-large Crawled Aggregated corpus (OSCAT)	SVM	F-1 Score :(mean for both datasets) 73%
6	Open Super-large Crawled Aggregated corpus (OSCAT)	Logistic Regression (LR)	74%
7	Open Super-large Crawled Aggregated corpus (OSCAT)	Random Forest	65%
8	Open Super-large Crawled Aggregated corpus (OSCAT)	Bagging (Ensemble Approach)	70%
9	Open Super-large Crawled Aggregated corpus (OSCAT)	RNN with BOW	66%
10	Open Super-large Crawled Aggregated corpus (OSCAT)	LSTM with BOW	67%

11	Open Super-large Crawled Aggregated corpus (OSCAT)	BERT	77%
12	DE-TRAIN	CNN	Ac=78.11%
13	DE-TRAIN	Bi-LSTM	Ac=71.04%
14	DE-TRAIN	mBERT	Ac=66.31%

Despite its potential, NLP faces significant challenges in identifying hateful and radical content. Language is inherently ambiguous and constantly evolving, especially in online communities [18]. Cultural and regional differences further complicate interpretation. Moreover, users often evade detection by using slang, acronyms, or multilingual texts. Training models that are fair, unbiased, and privacy-compliant is also a key concern [19]. Therefore, continuous model updating, diverse dataset curation, and ethical oversight are critical to improving NLP-based detection systems [20]. Integrating NLP into social media platforms, comment sections, and forums can significantly reduce the proliferation of harmful content. It enables proactive content moderation, protects vulnerable communities, and promotes healthier digital discourse [21]

Conclusion:

It can be concluded that integrating NLP into social media platforms, comment sections, and forums can significantly reduce the proliferation of harmful content. It enables proactive content moderation, protects vulnerable communities, and promotes healthier digital discourse. Looking forward, incorporating explainable AI, multimodal analysis (text, speech, and images), and community-specific language models will enhance the robustness and transparency of detection systems. NLP thus stands as a crucial pillar in the global effort to combat online hate and radicalization. from previous discussions that the study entails identifying benchmark datasets and models for detecting hate speech and coming up with a machine learning/deep learning model which would be accurate, as well as have low time complexity. The analysis presented in this paper can be used to develop further algorithms with the aim of achieving higher accuracy. While challenges remain, continued research and ethical deployment of NLP models are essential in creating safer online environments that respect both freedom of expression and community well-being.

References

[1] M. F. Wright, B. D. Harper, and S. Wachs, ``The associations between cyberbullying and callous-unemotional traits among adolescents:The moderating effect of online disinhibition," *J. Personality Individual Differences*, vol. 140, pp. 41_45, Apr. 2019.

[2] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz and L. Plaza, "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data," in *IEEE Access*, vol. 8, pp. 219563-219576, 2020, doi: 10.1109/ACCESS.2020.3042604.

[3] S. Khan *et al.*, "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network," in *IEEE Access*, vol. 10, pp. 7881-7894, 2022, doi: 10.1109/ACCESS.2022.3143799.

[4] R. Singh *et al.*, "Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter," in *IEEE Access*, vol. 8, pp. 194027-194044, 2020, doi: 10.1109/ACCESS.2020.3030621.

[5] Singh, T., Kumari, M. Burst: real-time events burst detection in social text stream. *J Supercomput* **77**, 11228–11256 (2021). https://doi.org/10.1007/s11227-021-03717-4

[6] Singh, T., Kumari, M. & Gupta, D.S. Real-time event detection and classification in social text steam using embedding. *Cluster Comput* **25**, 3799–3817 (2022).

[7] D. K. Jain, R. Jain, Y. Upadhyay, A. Kathuria, and X. Lan, ``Deep re_nement: Capsule network with attention mechanism-based system for text classification," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 1839_1856,Apr. 2020.

[8] P. K. Jain, R. Pamula, and S. Ansari, ``A supervised machine learning approach for the credibility assessment of user-generated content," *Wireless Pers. Commun.*, vol. 118, no. 4, pp. 2469_2485, Jun. 2021.

[7] Z. Zhang, D. Robinson, and J. Tepper, ``Detecting hate speech on Twitter using a convolution-GRU based deep

neural network," in *Proc. Eur. Semantic Web Conf.* Heraklion, Greece. Cham, Switzerland: Springer, 2018, pp. 745_760.

[8] A. R. Gover, S. B. Harper, and L. Langton, ``Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality," *Amer. J. Criminal Justice*, vol. 45, no. 7, pp. 647_667, 2020.

[9] **https://www.ohchr.org/en/statements/2023/01/freedom-speech-not-freedom-spread-racial-hatred-social-media-un-experts**.

[10] J. Langham and K. Gosha, ''The classification of aggressive dialogue in social media platforms,'' in Proc. ACM SIGMIS Conf. Comput. People Res., Jun. 2018, pp. 60–63.

[11] P. Fortuna and S. Nunes, ''A survey on automatic detection of hate speech in text,'' ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.

[12] W. Dorris, R. Hu, N. Vishwamitra, F. Luo, and M. Costello, ''Towards automatic detection and explanation of hate speech and offensive language,'' in Proc. 6th Int. Workshop Secur. Privacy Anal., Mar. 2020, pp. 23–29.

[13] A. Alrehili, ''Automatic hate speech detection on social media: A brief survey'' in Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA), Nov. 2019, pp. 1–6.

[14] S. Modi, ''AHTDT—Automatic hate text detection techniques in social media'' in Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol. (ICCSDET), Dec. 2018, pp. 1–3.

[15] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, ''Machine learning techniques for hate speech classification of Twitter data: State of the-art, future challenges and research directions'' Comput. Sci. Rev., vol. 38, Nov. 2020, Art. no. 100311.

[16] F Husain, O Uzuner, "Investigating the effect of preprocessing arabic text on offensive language and hate speech detection", ACM Transactions on Asian and Low Resource Language Information Processing vol.21, no.4, pp.1-20.

[17] I Bigoulaeva, V Hangya, I Gurevych, A Fraser, "Label modification and bootstrapping for zero-shot cross-lingual hate speech detection", Language Resources and Evaluation, Springer 2023, Art.no.1198.

[18] M. Yang and H. Chen, "Partially supervised learning for radical opinion identification in hate group web forums," 2012 IEEE International Conference on Intelligence and Security Informatics, Washington, DC, USA, 2012, pp. 96-101.

[19] K. T. Mursi, M. D. Alahmadi, F. S. Alsubaei and A. S. Alghamdi, "Detecting Islamic Radicalism Arabic Tweets Using Natural Language Processing," in IEEE Access, vol. 10, pp. 72526-72534, 2022.

[20] I Ajala, S Feroze, M El Barachi, F Oroumchian, "Combining artificial intelligence and expert content analysis to explore radical views on twitter: Case study on far-right discourse", Journal of Cleaner Production, Elsevier 2022, vol.362, 132263.

[21] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," in IEEE Access, vol. 9, pp. 88364-88376, 2021.