A Review on Hate Speech on Social Media, Impact on Society and Counter Measures

Akshay Baviskar[1] Dr. Manish Vyas[2]

Abstract: Social media has emerged as a powerful communication platform worldwide. The unprecedented growth of digital information has led to many crucial and challenging issues. Numerous technologies have provided a platform to the users to interact people from various countries, cultures and ethnicities. Although these platforms provide an opportunity to connect anywhere worldwide and the constructive growth of this inclusive communication are noticeable. Due to the nature of invisibility, mystery, and accessibility of such web applications, users can mask their identities and engage in xenophobic, racist, and sexist discourses with ease and impunity. This is known as dis-inhibition effect in social media in which social users are not capable to communicate face-to-face. This paper presents a review on machine learning and deep learning based approaches for identifying hate speech on social media.

Keywords: Hate Speech, Social Media, Countermeasures, Machine Learning, Deep Learning.

INRRODUCTION

As the social media web applications are so accessible, harassing behaviors are evolving into new patterns every day which are extremely risky [1]. 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 [2]. 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 [3].

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 [4].

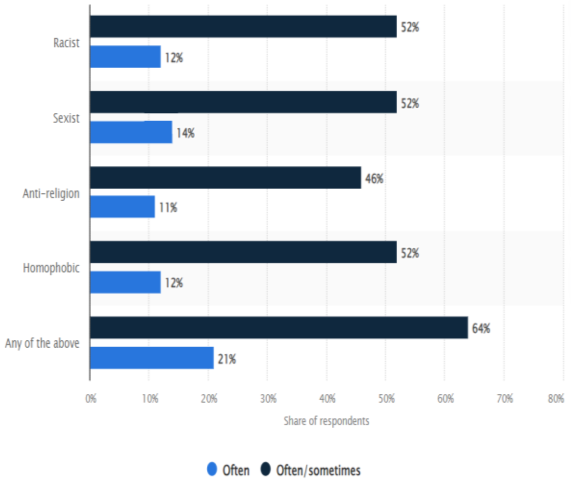

Fig.1. Rise in Hate Speech on Social Media (Source: https://www.statista.com/statistics/945392/teenagers-who-encounter-hate-speech-online-social-media-usa/}

Recently, Amnesty International1 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 [5].

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 [6].

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 [7]. 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 [8].

II. EXISTING WORK

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 [9]. 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 [10] 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 [11].

In [12] 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

III. COUNTERMEASURES AND SIGNIFICANECE OF STUDY

The presents study aims at analyzing the necessity to design automated tools for identification of hate speech online, especially pertaining to social media platforms. Automated detection of hate speech on social media platforms is crucial for several reasons [13]:

- **Scale and Speed:** Social media platforms have millions of active users generating vast amounts of content every day. Manual moderation cannot keep up with this scale and speed. Automated detection helps process and flag content in real-time, making it more efficient and effective [14].
- **User Protection:** Hate speech can be harmful, offensive, and psychologically damaging to individuals and communities. Automated detection helps identify and remove such content quickly, protecting users from potential harm.
- **Reducing Toxicity:** Hate speech creates a toxic online environment that can discourage user engagement and participation. Automated detection helps reduce the prevalence of hate speech, fostering a more positive and inclusive online community [15].
- **Legal Compliance:** Many countries have laws against hate speech, and social media platforms are responsible for ensuring compliance with these regulations. Automated detection can help identify and remove illegal content promptly.

- **Maintaining Platform Reputation:** Social media platforms aim to provide a safe and enjoyable experience for users. Failing to detect and control hate speech can damage a platform's reputation, leading to a loss of users and advertisers.
- **Resource Optimization:** Manual moderation is labor-intensive and costly. Implementing automated detection systems can optimize resource allocation and reduce the need for extensive human intervention.
- **Improving User Experience:** Hate speech can negatively impact user experience by driving away users or encouraging toxic behavior. Automated detection helps create a more welcoming and respectful online environment [16]
- **Complementing Human Moderation:** Automated detection systems can't replace human moderation entirely, but they can complement it by flagging potentially problematic content for review, making the moderation process more efficient.

  Despite these benefits, it's important to acknowledge that automated hate speech detection systems may have limitations and biases. Striking a balance between freedom of speech and combating hate speech is an ongoing challenge, and continuous research and development are required to improve the effectiveness and fairness of these systems. Additionally, transparent guidelines and user feedback mechanisms should be established to address potential errors and appeals.

  As there is a very thin fuzzy boundary between hate speech/non-hate speech with opinionated speech being touted as freedom of speech, it is necessary to have minimal false positive and false negative.

Hence, it is necessary to machine learning/deep learning aiming to mitigate the challenges identified in existing standard literature. This work aims at designing a probabilistic approach for identification of hate/non-hate speech which can identify texts with a probability based point of view to emulate human thinking process. The approach would aim at achieving the objectives cited subsequently [17].

IV. COMMONLY EMPLOYED MACHINE LEARNING AND DEEP LEARNING MODELS

The commonly used state of the art models are presented in this section [18].

Logistic Regression and Support Vector Machines (SVM): Logistic Regression and SVM are classic linear models that can be effective for binary classification tasks like hate speech detection. They work well when the decision boundary between hateful and non-hateful content is relatively simple.

Naive Bayes: Naive Bayes is a probabilistic model that works well for text classification tasks, including hate speech detection.
It assumes independence between features, making it computationally efficient and often suitable for large datasets.

Decision Trees: Decision Trees are tree-like structures that recursively split the dataset based on feature values.
They are interpretable and can capture complex relationships in the data, making them useful for hate speech detection.

Random Forests: Random Forests are ensembles of decision trees, combining the strength of multiple models to improve overall accuracy and robustness.
They are less prone to overfitting and can handle a larger feature space.
Gradient Boosting Models: Gradient Boosting models, such as XGBoost or LightGBM, sequentially build a series of weak learners to create a strong predictive model.
They are powerful and often perform well in practice, but may require careful tuning of hyperparameters.

Recurrent Neural Networks (RNNs): RNNs are neural network architectures designed for sequential data, making them suitable for analyzing text.

They can capture dependencies in language and context over time, making them effective for hate speech detection tasks.
Long Short-Term Memory Networks (LSTMs) and Gated Recurrent Units (GRUs):

LSTMs and GRUs are specialized RNN architectures capable of learning long-term dependencies in sequential data.
They are particularly useful when hate speech detection requires understanding the context and relationships between words in a sentence.

Convolutional Neural Networks (CNNs): CNNs, commonly used for image processing, can also be adapted for text classification tasks.
They excel at capturing local patterns and can be effective in identifying key features in hate speech detection.

Transformer Models (e.g., BERT, GPT): Transformer-based models, like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have achieved state-of-the-art results in various natural language processing tasks, including hate speech detection.
They leverage attention mechanisms and pre-training on large datasets to understand contextual nuances.

Ensemble Models: Combining multiple models into an ensemble, such as a voting classifier or stacking, can often improve overall performance by leveraging the strengths of different algorithms.
Ensemble models can enhance robustness and generalization.

In practice, the choice of a machine learning model for hate speech detection depends on factors like the complexity of the task, the nature of the data, and the desired trade-off between interpretability and predictive power. Researchers and practitioners often experiment with different models to find the most suitable one for a specific context [18].

CONCLUSION

It can be concluded that successfully implementing machine learning systems for hate speech detection on social media requires a comprehensive approach that involves data preparation, model selection, training, evaluation, and ongoing refinement to stay effective in dynamic online environments. Additionally, ethical considerations are crucial to ensure responsible and unbiased use of such models. This paper presents a comprehensive review on the existing approaches to identify hate speech on social media platforms.

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] A. Arango, J. Pérez, and B. Poblete, ‘‘Hate speech detection is not as easy as you may think: A closer look at model validation (extended version),’’ Inf. Syst., vol. 105, Mar. 2022, Art. no. 101584.

[17] A. Schmidt and M. Wiegand, ‘‘A survey on hate speech detection using natural language processing’’ in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10

[18] 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.

[19] 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.