International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-10, Issue-1, February 2023
https://doi.org/10.55524/ijirem.2023.10.1.22
Article ID IJIRD-1236, Pages 123-125
www.ijirem.org

Cyberbullying: Hate Comment Classifier

Anvita Karne[1], Vaishnavi Thakare[2], and Sonal Fatangare[3]

[1,2] Student, Department of Computer Science & Engineering, Rasiklal M. Dhariwal School of Engineering, Pune, India
[3]Assistant Professor, Department Computer Science & Engineering, Rasiklal M. Dhariwal School of Engineering, Pune, India

ABSTRACT- "Hate speech" refers to objectionable statements that may endanger societal harmony and targets a group or a person based on inborn qualities (such as race, religion, or gender). The issue of hate speech has been continuously growing on social media platforms lately. Our research focuses on creating a robust comment classifier that will categorize comments according to their toxicity. A series of activities or issues with getting a software to automatically categorize input comments into categories depending on the toxicity of the comment are referred to as comment classification. Our model combines LSTM and BERT with additional language processing methods. The application determines the category of toxic comments and shows the proportion of toxicity using an algorithmic technique.

KEYWORDS- Artificial Intelligence, LSTM, BERT, Cyber Security, Machine Learning.

I. INTRODUCTION

Rarely would someone walk up to a crowd and declare their hatred for everyone who belongs to a particular race, ethnicity, or religion. This does happen frequently online, where the anonymity of a profile gives individuals the courage to post provocative views on social media.

Social media is an invaluable tool for bringing people together and enabling networking, news sharing, and the expression of individual viewpoints. Unfortunately, it is frequently accompanied by negativity, including hate speech. Hate speech is defined as inciting violence or expressing hatred in public toward an individual or group on the basis of race, religion, sex, or sexual orientation. It is typically believed to involve expressing hostility toward a particular person or group based on their race, color, national origin, sex, handicap, religion, or sexual orientation. Other types of hate speech include uploading violent or otherwise offensive photographs on social media.

No specific set of guidelines exists that can reliably distinguish hate speech for what it is. Fundamentally, hate speech involves intimidating, harassing, or encouraging violence against others because of their identity, which might include race, skin color, ethnicity, religion, gender, or sexual orientation. I despise all people of a given race, for instance, is an example of a generic hate speech remark. Others may be directed at a specific individual or generalize about all people in the group to which the individual belongs. From stating one's views about a particular group or person to threatening them with violence, including words about wishing to kill them, hate speech can take many different forms.

Any comment that spreads, encourages, or incites hatred, violence, or prejudice against a person or community is referred to as hate speech. Additionally, hate speech and violence driven by hatred continue to be serious global concerns. Because they are afraid of reprisals, believe they won't be treated seriously, or lack faith in the legal system, victims seldom disclose crimes to the authorities. As a result, there is no legal organization available, which prevents the debaucher from being dealt with. According to our research, hate speech is frequently seen in news stories, discussion forums, blogs, and online comments.

The primary function of our model is to identify all hateful content and categorize it according to how toxic it is.

II. LITERATURE SURVEY

In the literature, several concepts are typically related with the definition of online hatred. Online hatred is a multidisciplinary phenomenon that has been researched through a variety of theoretical lenses and conceptual frameworks, including social psychology, Human-Computer Interface, politics, and legislation/regulatory elements. For example, in "CNN based Hate-o-Meter: A Hate Speech Detecting Tool" (Chaudhari, Parseja and Patyal 2020) they have used Neural networks, Convolutional Neural Networks, Flask API. The dataset used are Tweeter Sentiment Analysis Dataset. It had an accuracy of 80%. Furthermore in [2] "(F. M. Plaza-Del-Arco 2021) A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis" which was published in IEEE 10.1109/ACCESS.2021.3103697, they have used Transformer based approach. A Spanish tweets dataset is used. It has an accuracy of 75%.

[3] In "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding with Deep Learning and BERT" which was published in IEEE 10.1109/ACCESS.2021.3100435, Weak Supervised two - path bootstrapping approach is used. 10k tweets before and after election day (USA) were used as a testing dataset.

[4] "Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning" published in IEEE Transactions on Affective Computing, they have discussed about emotion detection with the help of correlation learning. The tech used is Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE).

[5] "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model" published in IEEE10.1109/ACCESS.2022.3144266 discusses in depth how sentiment analysis is carried out using GCR-NN model.
[6] In "Text Analysis for hate speech detection using backpropagation neural networks" published in 2018 International Conference on Control Electronics, renewable energy and communications (ICCEREC) discusses about how neural networks and back propagation is used for hate speech detection.
[7] "Detecting Online Hate Speech Using Context Aware Models" published in International Conference on Computer Science, Engineering and Applications (ICCSEA) uses Context based Aware models for NLP.
In "Early Detection of Online Hate Speech Spreaders with Learned representations" which was presented in 24th International Conference on World Wide Web discusses about how Lexical representations and Random Forests are used in online hate speech detection.

III. LIVE SURVEY

In this regard, research has emphasized the significance of differentiating between simply abusive or offensive speech and hateful speech in order to provide both a differentiated orientation for manual tagging and a more nuanced basis for automated classification (Davidson et al., 2017; Fortuna et al., 2019; Founta et al., 2018). Despite the fact that a number of non-deep learning algorithms have been used in the past (see, for example, Nobata et al., 2016; Ibrohim and Budi, 2019), neural networks have recently emerged as the state-of-the-art for text classification and hate speech detection. This is especially true when used in conjunction with different word-embedding techniques to represent the text data in a vector space (Kim, 2014; Kshirsagar et al.,). Most contemporary approaches use user data and other metadata in the analysis to increase model accuracy (Mathew et al., 2019; Ribeiro et al., 2018; Waseem and Hovy, 2016; Miro-Llinares et al., 2018; Stoop et al., 2019). However, the focus of these models is on recognising hostile user accounts and situations, rather than hateful content per se. Furthermore, a model's reliance on metadata limits its applicability to other datasets or online platforms (Meyer and Gamback, 2019).

IV. FUTURE SCOPE

In the future, the authors hope to expand the project by adding multiple languages support to assist persons who do not understand English. The method used to create the tool can be tested and optimized for improved outcomes. Crowdsourcing functionality can be added to our application, allowing users to report hate speech directly through the tool, and the model will optimize itself. There is opportunity for improvement in terms of tool input. We may instruct the tool to read text from.txt,.docx, and.pdf files. As the tool can also be used for legal data reporting, we can include a section in the future that displays all the sources that have been reported for having high hate speech content. This material will be useful to legal groups who conduct hate crime investigations.

V. CONCLUSION

In conclusion, the hate-based classifier based on LSTM and BERT models has shown promising results in detecting hate speech and offensive language. The LSTM model was able to capture the temporal nature of the text and the BERT model was able to capture the contextual meaning of the words, which will result in high accuracy and F1 scores.
However, there is still room for improvement in the hate-based classification models. One limitation of these models is that they heavily rely on training data and may not generalize well to different datasets or languages. Additionally, it is important to consider the ethical implications of using these models, such as potential bias or censorship.
Overall, hate speech detection is an important task for maintaining a safe and inclusive online community, and the development of advanced models like LSTM and BERT provides a useful tool for this purpose. With further research and refinement, these models can contribute to the fight against hate speech and promoting a more respectful online environment.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] A. Chaudhari, A. Parseja and A. Patyal, "CNN based Hate-o-Meter: A Hate Speech Detecting Tool," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 940-944, doi: 10.1109/ICSSIT48917.2020.9214247.
[2] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, doi: 10.1109/ACCESS.2021.3103697.
[3] H. S. Alatawi, A. M. Alhothali and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT," in IEEE Access, vol. 9, pp. 106363-106374, 2021, doi: 10.1109/ACCESS.2021.3100435.
[4] J. Deng and F. Ren, "Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.3034215.
[5] E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," in IEEE Access, vol. 10, pp. 9717-9728, 2022, doi: 10.1109/ACCESS.2022.3144266
[6] N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis For Hate Speech Detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Bandung, Indonesia, 2018, pp. 159-165, doi: 10.1109/ICCEREC.2018.8712109.
[7] Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

[8] Jai Rakshita & Goel, Devanshi & Sahu, Prashant & Kumar, Abhinav & Singh, Jyoti. (2021). Profiling Hate Speech Spreaders on Twitter.
[9] Aouchiche, Imane & Boumahdi, Fatima & Madani, Amina & Remmide, Mohamed Abdelkarim. (2023). Hate Speech Prediction on Social Media. SN Computer Science. 4. 10.1007/s42979-023-01668-6.
[10] Moin Ahmed, Mohit Goel, Raju Kumar, Aruna Bhat, "Sentiment Analysis on Twitter using Ordinal Regression", 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp.1-4, 2021.
[11] Mitushi Raj, Samridhi Singh, Kanishka Solanki, Ramani Selvanambi, "An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques", SN Computer Science, vol.3, no.5, 2022.
[12] Sepideh Saeedi Majd, Habib Izadkhah, Shahriar Lotfi, "Detection of Multiple Emotions in Texts Using Long Short-Term Memory Recurrent Neural Networks", 2022 8th International Conference on Web Research (ICWR), pp.29-33, 2022
[13] Zhuqing Yang, Liya Zhou, Zhengjun Jing, "A Novel Affective Analysis System Modeling Method Integrating Affective Cognitive Model and Bi-LSTM Neural Network", Computational Intelligence and Neuroscience, vol.2022, pp.1, 2022.
[14] Kapil, Prashant & Ekbal, Asif. (2020). A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems. 106458. 10.1016/j.knosys.2020.106458.
[15] R. Jayakrishnan, G.N. Gopal, and M.S. Santhikrishna, “Multiclass emotion detection and annotation in malayalam novels,” 2018 Int. Conf. on Comput. Commun. and Inform. (ICCCI), Jan. 2018.
[16] M. Ahlgren. 40C Twitter Statistics & Facts. Accessed: Sep. 1, 2021.[Online].Available: https://www.websitehostingrating.com/twitterstatistics/
[17] C. Chen, R. Zhuo, and J. Ren, “Gated recurrent neural network with sentimental relations for sentiment classification,” Inf. Sciences, vol. 502, pp. 268–278, Oct. 2019.
[18] R. Alshalan and H. Al-Khalifa, ‘‘A deep learning approach for automatic hate speech detection in the Saudi Twittersphere,’’ Appl. Sci., vol. 10, no. 23, p. 8614, Dec. 2020.