Educational Administration: Theory and Practice
2024, 30(5) 14810-14816
ISSN: 2148-2403
https://kuey.net/

Research Article

AI-Driven Foul Message Flagging

Prof. Aniket Kore[1*], Simran Bardhan[2], Rishita Merchant[3], Yogesh Jha[4]

[1*]Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra aniket.kore@djsce.ac.in
[2]Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra simranbardhan13@gmail.com
[3]Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra merchant.rishita2912@gmail.com
[4]Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra jhayogesh30@gmail.com

Citation: Prof. Aniket Kore (2024), AI-Driven Foul Message Flagging, *Educational Administration: Theory and Practice*, *30(5)* 14810-14816
Doi: 10.53555/kuey.v30i5.7446

ARTICLE INFO

ABSTRACT

The surge in online communication has posed challenges in maintaining civility, prompting the integration of AI for foul language detection. While AI can effectively identify offensive language, its limitations with sarcasm and slang necessitate a hybrid approach with human oversight. This paper advocates for configurable filters that blend AI's computational power with human expertise to adapt to diverse online communities. Through a review of existing literature and case studies, it proposes guidelines for transparent and accountable AI-powered flagging systems, emphasizing ongoing collaboration between AI and human moderators. This research contributes to fostering a cleaner and more civil online environment by acknowledging the need for both technological advancement and human intervention in mitigating offensive content.

Keywords: Natural Language Processing (NLP), Joint-Learning models, Multilingual BERT, Cross-Lingual hate speech detection, AI-Driven content moderation, Foul language flagging, Linguistic sensitivity, Digital safety, Ethical implications.

I. INTRODUCTION

The first critical aspect of this endeavour lies in hate speech detection. Hate speech, characterized by its discriminatory or derogatory nature towards individuals or groups based on attributes such as race, ethnicity, gender, religion, or sexual orientation, poses a significant threat to online discourse and community cohesion. AI-powered algorithms, trained on vast amounts of text data, can effectively identify patterns associated with hate speech, enabling automated systems to flag and remove such content promptly. However, the effectiveness of hate speech detection algorithms relies heavily on the quality and diversity of the training data, as well as the ability to adapt to emerging forms of discriminatory language.

Another challenge in foul message detection pertains to the nuanced nature of communication, particularly in discerning sarcasm. Sarcasm, often employed as a form of humour or satire, can confound AI algorithms due to its subtlety and context-dependent interpretation. Current AI models struggle to accurately detect sarcastic remarks, leading to potential misclassification of benign content as offensive. Addressing this challenge requires advancements in natural language processing (NLP) techniques and the integration of contextual clues to enhance the algorithm's understanding of sarcasm within different cultural and linguistic contexts.

Finally, the effective flagging of offensive messages necessitates a balanced approach that combines AI-driven automation with human oversight. While AI algorithms can efficiently detect explicit profanity and some forms of hate speech, they are not infallible and often overlook the nuances of language and cultural context. Human moderators play a crucial role in refining and calibrating AI-powered flagging systems, providing context-specific insights and addressing instances where automated detection falls short. By integrating human judgment with AI capabilities, online platforms can develop configurable filters that adapt to the unique needs and preferences of diverse communities, thereby fostering a more civil and inclusive online environment.

II. REVIEW OF LITERATURE

The paper labelled as [1] focuses on developing an online hate classification system for multiple social media platforms. It thoroughly explores dataset composition and labelling methods, employing various machine learning algorithms such as LR, NB, SVM, XGBoost, and FFNN. Feature representation includes Bag of Words, TF-IDF, Word Embeddings, and BERT. A central aspect is the use of a large, labelled dataset for abusive language detection, although the specificity to Twitter and subjective human annotation are acknowledged as limitations. Nevertheless, the paper provides valuable insights into the construction of robust hate classification systems suitable for diverse social media landscapes.

The literature review, referencing paper [2], provides a comprehensive analysis of the challenges in distinguishing offensive language from hate speech online. While it recognizes the paper's valuable contribution in addressing this issue, particularly through the introduction of a novel dataset from Twitter, it also notes potential limitations such as platform specificity and biases introduced by human annotation methods. Despite these concerns, the review acknowledges the significance of the dataset's diverse content and the insights gained from employing machine learning algorithms alongside human annotation. Additionally, the review highlights specific obstacles in differentiating hate speech from profanity, emphasizing the complexities involved in automated identification and the nuanced nature of offensive language. Overall, it underscores the importance of ongoing interdisciplinary research efforts to effectively combat online hate speech and promote a safer digital environment.

The literature review, attributed to paper [3], presents a comprehensive survey on troll detection methods, offering insights into existing approaches and the prevailing challenges in the field. Despite the absence of a specific dataset mentioned, the review provides valuable background information on the complexities associated with troll detection, outlining the various obstacles researchers encounter. By summarising existing methods, the paper offers a consolidated view of the current landscape of troll detection techniques, facilitating a better understanding of the methodologies employed in this domain. Moreover, the review highlights the ongoing challenges faced by researchers, shedding light on areas that require further investigation and innovation. Overall, the paper serves as a valuable resource for researchers and practitioners alike, offering a holistic overview of troll detection methodologies and paving the way for future advancements in the field.

The literature review, citing paper [4], examines hate speech detection models, emphasising their multilingual capabilities and employment of zero-shot learning with LASER embeddings. Despite recognizing challenges like limited data for low-resource languages, paper [4]'s contribution in assembling a multilingual Twitter dataset and its innovative approach of joint learning with knowledge injection are praised. Overall, the review underscores the importance of developing reliable hate speech detection models for diverse languages and highlights the potential of collaborative learning strategies to address this task effectively.

The literature review, attributed to paper [5], offers a comprehensive evaluation of resources and benchmark corpora available for hate speech detection on social media platforms. By examining various annotated corpora, benchmarks, and lexica, the review aids researchers in selecting suitable datasets for their studies. Following Kitchenham's guidelines, the review rigorously searched Google Scholar and Books for relevant English and multilingual content across two distinct periods, ensuring a thorough analysis of available resources in the field.

Paper [6] conducts a targeted analysis of tweets surrounding George Floyd's death in May 2020, with a specific focus on French tweets. Employing annotation on Yandex Toloka, the study develops French hate speech detection models leveraging multilingual BERT, CamemBERT, and transfer learning with HateXplain. By categorizing tweets into hate, offensive, or normal categories, the research underscores the importance of multilingual racial hate speech detection using transfer learning. Demonstrating the efficacy of this approach across languages, the study contributes valuable insights into the detection and mitigation of hate speech in diverse linguistic contexts.

The paper cited as [7] explores predictive features for detecting hate speech on Twitter, with a specific emphasis on distinguishing between hateful symbols and individual expressions. Drawing from a dataset sourced from Twitter, the study investigates characteristics that are indicative of hate speech, offering valuable insights into the complex dynamics of online discourse. By focusing on this differentiation, the research contributes significantly to the understanding of hate speech detection, providing nuanced perspectives essential for researchers and practitioners engaged in computational social science.

In paper [8], the focus lies on exploring user network features for hate speech detection on Twitter and assessing different model architectures for combining embedding types. With a dataset comprising 16,000 tweets labelled as racist, sexist, or neutral, the study compares deep learning architectures (including FastText, CNNs, LSTMs) with baselines such as character n-grams, TF-IDF, and Bag-of-Words vectors. Evaluation metrics such as precision, recall, and F1 score are employed to gauge model performance. The research aims to advance hate speech detection on Twitter by leveraging deep learning techniques to achieve superior accuracy compared to existing methods.

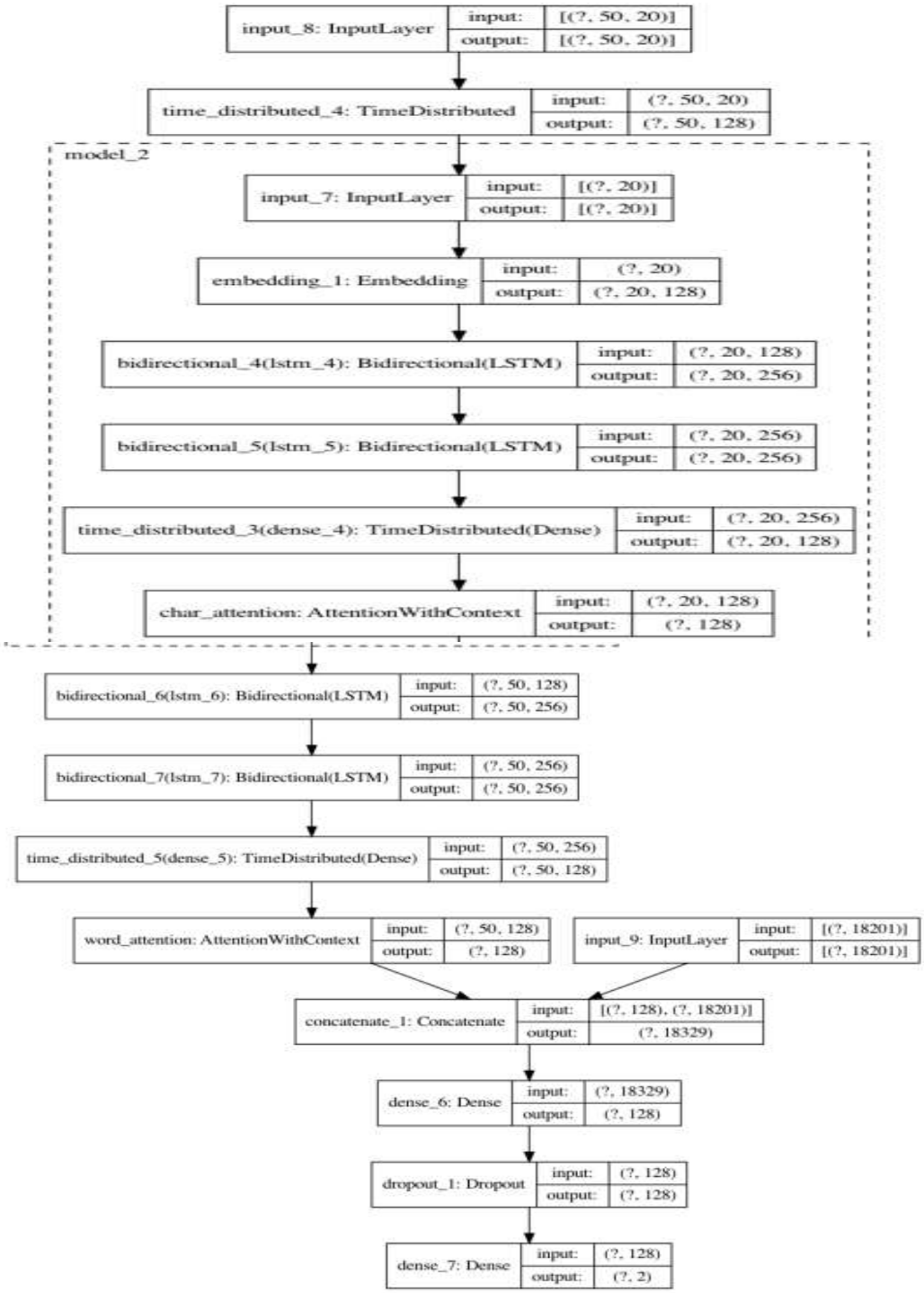
Paper [9] acknowledges the challenges in automatic aggression detection and highlights the significance of further investigation in this area. With a dataset comprising 18,000 tweets and 21,000 Facebook comments, it serves as a valuable resource for researchers. The study collected and labelled Hindi-English social media

comments, focusing on Indian topics, annotating them for aggression type and effect. An improved annotation scheme was implemented to address initial discrepancies, resulting in higher agreement levels and enhancing the quality of the aggression-annotated corpus of Hindi-English code-mixed data.
In the study cited as [10], novel deep learning techniques are explored to effectively capture textual cues related to hate speech, alongside methods for identifying sarcasm and misinformation within hate speech. Leveraging a public dataset of Hindi-English code-mixed tweets, the research also employs a larger dataset to train domain-specific word embeddings. Three deep learning models—CNN-1D, LSTM, and BiLSTM—are implemented for hate speech detection, with their performance compared to previous statistical approaches like SVM and Random Forest. Through rigorous evaluation using 10-fold cross-validation on a benchmark dataset, the study contributes valuable insights into hate speech detection from code-mixed Hindi-English tweets, offering promising avenues for future research in this domain.

III. METHODOLOGY

A. Architectural Diagram

Fig. 1. Architectural Diagram



B. Proposed Work

AI-driven foul message flagging aspires to cultivate a more civil and inclusive online environment. By shifting through mountains of text, AI can identify a spectrum of offensive language, from hate speech spewing venom to everyday swears that grate on some users. This not only improves the overall user experience by fostering a more respectful online space but also frees up human moderators to tackle complex situations where AI might stumble. Additionally, AI can categorize content based on language, ensuring a safer experience for younger users or those sensitive to offensive language. This system also offers a crucial advantage in the fight against online negativity – the ability to continuously learn and adapt to new forms of offensive language, including slang and evolving hate speech tactics. By combining AI's detection prowess with human oversight, online platforms can strive toward a cleaner, more positive digital space. Our system boasts a robust, multilingual dataset (social media, forums) enriched with data augmentation for cultural understanding. Transfer learning and cross-lingual NLP techniques ensure seamless handling of diverse languages. Advanced NLP and machine learning minimize false positives/negatives, effectively addressing

sarcasm and ambiguity in text. Designed for scalability and real-time monitoring, the system utilizes precision-driven NLP for accurate detection.

C. Implementation

A) Datasets

The dataset comprises Hinglish tweets and comments meticulously annotated for hate speech, derived from a plethora of social media platforms and online communities. Through detailed categorization of hate speech instances, the dataset provides a comprehensive foundation for hate speech detection. To bolster sarcasm detection capabilities, data augmentation techniques were strategically employed, resulting in the generation of additional instances of sarcastic tweets and comments within the Hinglish hate speech dataset. These augmentation methods, including paraphrasing and context modification, were intricately designed to diversify the training data, thus facilitating exposure to a broader spectrum of sarcastic expressions. The augmented dataset functions as a pivotal component of the training pipeline, with the overarching goal of refining the model's sarcasm detection performance. By incorporating this enriched dataset, the model is poised to accurately identify sarcastic content embedded within Hinglish text across an array of online platforms and forums, thereby significantly enhancing its utility and effectiveness in real-world applications.

a) Aggressive Content Dataset

Fig. 2(a). Trained Processed Aggression Dataset

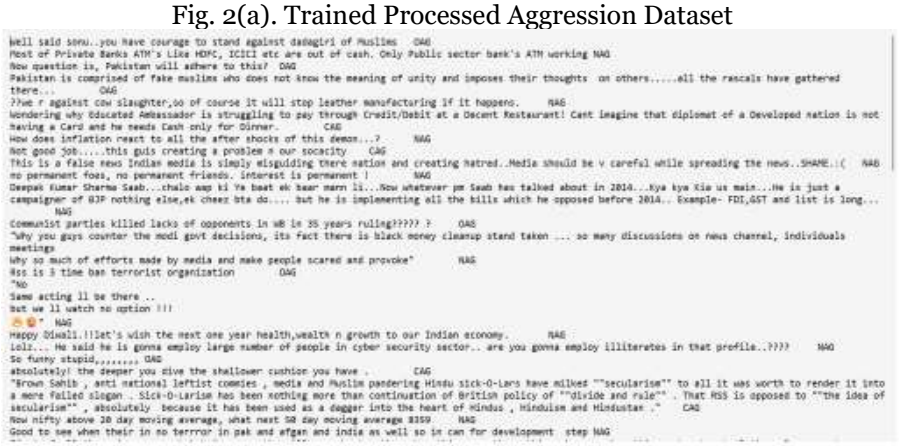


Fig. 2(b). Validation Processed Aggression Dataset

b) Hate Speech Dataset for Hinglish

Fig. 3. Hate Speech Dataset for Hinglish

c) English and Spanish Hate Speech Dataset

Fig. 4(a). English Hate Speech Dataset

Fig. 4(b). Spanish Hate Speech Dataset

d) Sarcasm Content Dataset

Fig. 5. Processed Sarcasm Hinglish Dataset

Text	Category
Triple Talaq par Burbak Kuchh nahi bolega	NO
Batao ye uss site pr se akki sir ke verdict nikaal laaye jaha he ajay ki ek bi movie hit nai	YES
Hindu baheno par julam bardas nahi hoga @TripleTalaq Hindu daram par lago hoga hamari Hindu baheno ki soraksa ke liye	NO
Naa bhai.. aisa nhi hai.. mere handle kame se bhi kuchh hona nhi hai.. politics se mera door door tak ka naata nhi hai	NO
#RememberingRajiv aaj agar musalman auraten triple talaq ki wajah se itni jaleel ho rahi hai to iss ka sabse bada reason Rajib gandhi hai.	NO
are cricket se sanyas le liya kya viru aur social service suru kardiya.khel hi bhul gaye.2 innings 0 n 0	NO
Sohail bahi hame bhi treen kardo muje bhi shock he cricket	NO
bhai triple talaq se aap kya samjhte hai samjhaye aap zara ..agar triple talaq pta hota apko toh aisa nhi kehte..	NO
Bachcho ki death par politics ke bajay unke liye prayer aur unki family ko console kiya jana jyada zaruri he.. thodi humanity bhi zaruri he	NO
#@spn_cricket lijiye S Thakur ne phir pahni no.10 jarsi, inka kharab din chalu ho gaya..@BCCI#Slsvind	NO

Show 10 per page 1 2 10 100 500 520 525

B) Models

1. Data Collection and Augmentation: The foundation lies in data. We collect multilingual data from various online communities (social media, forums) encompassing both positive, negative, and sarcastic text. To account for cultural nuances and expand dataset size, we employ data augmentation techniques. This can involve techniques like synonym replacement, back-translation, and random shuffling to create variations of existing data points, fostering a more robust and generalizable model.

2. Text Classification with Ensemble Learning: The core of the system lies in utilizing multiple models for text classification.

(i) Aggressive Content Dataset

- Categorical Classification: The problem deals with classifying text into three categories: aggressive, non-aggressive, and neutral. This suggests a multi-class classification problem.
- Softmax Activation: The final layer of the neural network uses a "softmax" activation function. This function is commonly used in multi-class classification problems to convert the network's output into probabilities for each class.
- Categorical Cross-Entropy Loss: The code uses the "categorical_crossentropy" loss function. This loss function is specifically designed for multi-class classification problems.

(ii) Hate Speech Detection

- Hierarchical Attention Network (HAN): HAN is a neural network architecture designed to handle hierarchical structures in data, such as text. In hate speech detection, HAN can effectively capture the hierarchical relationships between words, sentences, and documents, enabling more accurate classification.
- Transformer (e.g., BERT): Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), are state-of-the-art architectures for natural language processing tasks. In hate speech detection, BERT can leverage its bidirectional context understanding to capture subtle linguistic cues indicative of hate speech.
- Word-Level Long Short-Term Memory (WLSTM): A regularization term is part of the Lasso Regression linear regression model. It does both feature selection and regularization by adding the absolute values of the coefficients to the loss function. The less significant features are shrunk towards zero, which promotes sparse solutions. Lasso Regression can be helpful for handling multicollinearity and feature selection. When working with high-dimensional datasets, it is efficient.

(iii) Sarcasm Detection

- ZSC (Zero Shot Classification): Zero-shot classification is a machine learning technique where a model is trained to recognize classes it has never seen during training. In sarcasm detection, zero-shot classification can be employed to categorize text instances as sarcastic or non-sarcastic without the need for labelled sarcastic data. By leveraging pre-trained language models, such as GPT (Generative Pre-trained Transformer), zero-shot classification can effectively identify sarcastic expressions based on their contextual understanding.

IV. CONCLUSION

The novel ensemble method proposed for hate speech and sarcasm detection in textual data represents a sophisticated approach that amalgamates several established models, including BERT, HAN, and WLSTM, with a zero-shot model to enhance efficiency. By leveraging pre-trained models, the method ensures heightened accuracy in detecting both hate speech and sarcasm, thereby contributing to a more nuanced understanding of text data. Additionally, the incorporation of data augmentation techniques serves to account for diverse cultural nuances, enriching the model's adaptability across different linguistic contexts.
Furthermore, AI-powered flagging mechanisms, such as rule-based filtering and sentiment analysis, are employed to categorize identified text types, enabling the system to take appropriate actions. For instance, hateful content identified by the ensemble method is promptly flagged for removal, while instances of flagged sarcasm are contextualized to prevent misinterpretations and foster accurate comprehension.

V. FUTURE SCOPE

In terms of future scope, expanding the project to detect languages in regional languages presents a significant opportunity to enhance its applicability and inclusivity. By incorporating language detection capabilities for regional languages, the ensemble method can effectively cater to a broader user base, facilitating a more comprehensive analysis of text data across diverse linguistic landscapes.
Additionally, developing a user interface (UI) that supports multilingual languages would further augment the accessibility and usability of the system. A well-designed UI capable of seamlessly accommodating multiple languages would enhance user experience and ensure that individuals from different linguistic backgrounds can effectively interact with the system.
Furthermore, integrating additional features such as aggregation datasets and the analysis of various emotions would enrich the functionality of the ensemble method. Aggregation datasets could provide valuable insights by consolidating data from multiple sources, thereby improving the robustness of the model. Similarly, analysing a wide range of emotions beyond hate speech and sarcasm could offer a more nuanced understanding of text data, enabling the system to detect and respond to diverse emotional expressions effectively.
Overall, these future scopes hold immense potential for advancing the capabilities of the project, ultimately contributing to its effectiveness in promoting a positive and respectful online environment. By embracing language diversity, refining the user interface, and incorporating additional features, the ensemble method can evolve into a comprehensive solution tailored to meet the evolving needs of online content moderation and sentiment analysis.

REFERENCES

[1] Salminen, J., Hopf, M., Chowdhury, S.A. et al. Developing an online hate classifier for multiple social media platforms. Hum. Cent. Comput. Inf. Sci. 10, 1 (2020).https://doi.org/10.1186/s13673-019-0205-6
[2] Shervin Malmasi & Marcos Zampieri (2018) Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence, 30:2, 187-202, DOI: 10.1080/0952813X.2017.1409284
[3] Tomaiuolo, Michele & Lombardo, Gianfranco & Mordonini, Monica & Cagnoni, Stefano & Poggi,. (2020). A Survey on Troll Detection. Future Internet. 12. 31. 10.3390/fi12020031.
[4] Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, Information Processing & Management, Volume 58, Issue 4, 2021, 102544, ISSN 0306-4573,
[5] Poletto, F., Basile, V., Sanguinetti, M. et al. Resources and benchmark corpora for hate speech detection: a systematic review. Lang Resources & Evaluation 55, 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8
[6] Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023. Multilingual Racial Hate Speech Detection Using Transfer Learning. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 41–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
[7] Talat, Zeerak and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” North American Chapter of the Association for Computational Linguistics (2016).
[8] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Investigating deep learning architectures for hate speech detection on Twitter.
[9] Kumar, R., Reganti, A. N., Bhatia, A., & Maheshwari, T. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data.
[10] Kamble, S., & Joshi, A. (2018). Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models.