International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 11 | Nov 2023 www.irjet.net p-ISSN: 2395-0072

Hate Speech Detection in multilingual Text using Deep Learning

Gelvesh G[1], Prahlada H K[2], Mayura Sharma[3], Aarav A[4]

[1]*Dept. Of Information Science and Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[2]*Dept. Of Information Science and Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[3]*Dept. Of Information Science and Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[4]*Dept. Of Information Science and Engineering, Dayananda Sagar College of Engineering, Karnataka, India*

Abstract - *The Internet is a boon for mankind however its misuse has been growing drastically. Digital Social platforms along with Facebook, Twitter and Instagram play a paramount role in opining perspectives through the users. Sometimes users wield abusive or inflammatory language, which could instigate readers. This paper aims to assess numerous deep learning techniques to detect hate speech on numerous social media platforms within side the English-Hindi code-mix language. In this paper, we implement and compare numerous deep learning methods, in conjunction with numerous feature extraction and word-embedding strategies, on a consolidated dataset of 20000+ instances, for hate speech detection from tweets and comments in Hindi and English. The experimental consequences reveal that deep learning perform higher than machine learning models in general. Among the deep learning models, the CNN-BiLSTM model presents the optimal results. The model yields 0.87 accuracy, 0.82 precision and 0.85 F1-score. These results surpass the current state-of-art approaches.*

Key Words: Deep Learning, Hate Speech, CNN, Social Media

1.INTRODUCTION

In an age characterized by the widespread use of digital communication platforms, the proliferation of hate speech has become a pressing societal concern. Hate speech poses a significant threat to social cohesion and the well-being of individuals and communities. This research endeavors to address this issue by harnessing the power of deep learning techniques for hate speech detection. The exponential growth of digital content, particularly on social media, has led to an alarming surge in hateful rhetoric. To emphasize the gravity of this problem, recent statistics reveal that hate speech-related incidents have increased by 70% in the last three years, underscoring the urgency of effective detection and mitigation strategies. This study explores the potential of cutting-edge deep learning models to automatically identify and combat hate speech in digital spaces, offering an innovative and promising approach to tackle this critical societal issue.

After extensive research, our findings indicate that a substantial amount of research has been conducted on the English language, with successful implementations in various Twitter bots and chatbots developed by tech giants. However, in the Indian context, these systems underperform due to the unique linguistic characteristics of Indian users who often employ a blend of Hindi and English in their communications. It is evident from the literature review that there has been limited exploration of machine learning and deep learning techniques for hate speech detection, leaving significant room for improvement in this domain. The primary objective of this study is to identify hate speech within the context of mixed Hindi and English content. We have applied machine learning and deep learning methods to a hybrid dataset, created by combining three publicly available datasets, and compared the results using various performance metrics such as accuracy, precision, recall, F1-score, and more.

The paper makes several significant contributions:

1. Addressing the scarcity of research on English and Hindi code mix data, the authors merge three relatively smaller datasets (Bohra 2018, Kumar 2018, and HASOC 2021 Hindi-English Coded dataset) to create a consolidated dataset. This initiative tackles the lack of substantial data for model training.

2. The research explores a wide range of machine learning and deep learning models, encompassing eight machine learning models with four feature extraction methods and four deep learning models with three word-embedding techniques. This diverse approach allows for a multifaceted analysis of the problem, with results presented in Section 4.

3. The paper achieves state-of-the-art performance for hate speech detection in English-Hindi code mix data, surpassing existing methods. The state-of-the-art comparison is detailed in Section 5.

The implications of this work are extensive, including applications in chatbots, law enforcement, and societal harmony preservation. This research may contribute to the automatic filtering of hateful content in chatbots and assist enforcement agencies in managing law and order during protests and social unrest, ultimately reducing hate speech's potential to incite violence against specific groups. The paper's structure involves Section 2, which reviews recent literature and discusses machine learning

and deep learning methods. Section 3 covers data preprocessing, vectorization, and model details. Section 4 analyzes the results, while Section 5 concludes the paper with final thoughts.

2. Literature Survey

The contemporary landscape of social media platforms is fraught with the prevalent issue of hate speech, necessitating the application of machine learning (ML) and deep learning (DL) techniques to address the challenge. This section provides a comprehensive literature review on hate speech detection using ML and DL methods, with a specific focus on datasets related to English-Hindi code mix texts. Historically, researchers have concentrated on identifying hate speech with most studies centered on single-language (primarily English) hate speech detection. However, there has been a dearth of work on hate speech detection in Hindi-English code-mixed data due to the intricacies of code-mixing and the scarcity of relevant datasets. This section delves into hate speech detection techniques tailored for Hindi-English code-mixed data.

A Hindi-English code-mix dataset was introduced in a paper [14], where the authors curated an annotated corpus from Facebook and Twitter, comprising three-level tags (Aggression, Over Aggression, and Non-aggression) with 18,000 tweets and 21,000 Facebook comments. Notably, this research did not include an experimental evaluation of the mentioned dataset.

In another paper [15], an annotated corpus of YouTube video comments related to automated vehicles was proposed, consisting of 50,000 comments, along with data formats and potential use cases. The authors also conducted a case study to understand public opinions on self-driving vehicles and responses to accidents using cars.

In a separate study [16], text classification was explored using Hinglish text written in the Roman script. Random Hinglish data from news and Facebook comments were collected, and various feature identification methods using TF-IDF representation were proposed. The study concluded that the Radial Basis Function Neural Network yielded the best classification results for Hinglish text.

The following paper [17] discussed the challenges in hate speech detection within Hindi-English code-mixed texts. The authors collected Hindi-English code-mixed data from Twitter, annotated the tweets at the word level with Hate and Normal speech classifications, and proposed a machine learning-based system for hate speech detection with an accuracy of 71.7%.

In another work [18], the authors introduced a mechanism for detecting hatred in three languages (English, Spanish, and Italian). They devised methods to assess the connection between misogyny and abusive language, with a focus on misogyny detection in a cross-lingual context. Their experiments were conducted using the Automatic Misogyny Identification (AMI) datasets, and the research concluded that misogyny is a form of abusive language, with the proposed architecture delivering robust performance across languages.

The study of hate speech detection in Hindi-English code-mix data is presented in a paper [19]. The authors collected hate and non-hate data from various sources, including Twitter and the shared task HASOC, and applied popular pre-trained word embeddings. They compared the proposed model with various feature extraction methods and found that fastText features outperformed others, achieving an accuracy of 0.8581%, a precision of 0.8586%, a recall of 0.8581, and an F1-score of 0.858%.

In a different paper [20], deep learning methods for hate speech detection in Hindi-English code-mix data were explored using a benchmark dataset. The authors experimented with deep learning models, utilizing domain-specific embeddings and achieving results with an accuracy of 82.62%, precision of 83.34, and an F-score of 80.85% with a CNN model.

In this context, another paper [21] proposed a deep learning model for offensive speech detection. The authors created a self-made Hindi-English code-mix dataset with annotations and employed machine learning models as baseline models. They introduced the Multi-Channel Transfer Learning-based model (MIMCT) and concluded that it outperformed state-of-the-art methods.

Additionally, a deep learning model for detecting offensive tweets in Hindi-English language was presented in a paper [22]. The authors introduced a novel tweet dataset titled Hindi-English Offensive Tweet (HEOT), with tweets categorized as non-offensive, abusive, or hate speech. They evaluated the results using a CNN model and reported an accuracy of 83.90%, a precision of 80.20%, a recall of 69.98%, and an F1-score of 71.45%.

Furthermore, a study on the evaluation of Hindi-English code mix data from social media is detailed in a paper [23]. The research encompassed the use of monolingual embeddings and supervised classifiers with transfer learning on an English dataset, subsequently applied to code-mixed data. The reported results demonstrated an improvement in the F1-score of 0.019.

A deep learning model was proposed for hate speech detection in social media text [24]. The authors utilized the HASOC 2019 corpus to assess the model's performance, reporting a macro F1 score of 0.63 in hate speech detection on the test set of HASOC.

In the paper [25], the authors outlined a pipeline for hate speech detection in Hindi-English code-mix data (Hinglish)

on social media platforms. Prior to finalizing the proposed system, the authors conducted rigorous comparisons against various benchmark datasets. They also explored the relationship of hate embeddings along with social network-based features, concluding that the proposed system surpassed state-of-the-art approaches.

A deep learning approach for hate speech detection in Hindi-English code-mix data was introduced in a paper [26]. The authors employed character-level embeddings for feature extraction and experimented with various deep learning classifiers. They observed that the hybridization of GRU (Gated Recurrent Unit) with the Attention Model yielded the best performance among the models studied.

In another paper [27], the authors addressed the identification of hate speech in code-mixed text using deep learning models. They utilized publicly available datasets and implemented two sub-word level LSTM models, reporting an F1-score of 48.7%.

The paper [28] proposed a deep learning approach for hate speech emotion detection. The authors collected over 10,000 Hindi-English code-mix datasets and annotated them with emotions (happy, sad, and anger). They employed a bilingual model for feature vector generation and a deep neural network for classification, with CNN-Bi-LSTM achieving the highest classification accuracy of 83.21%.

Another paper [29] introduced a transfer learning with LSTM-based model for hate speech classification in Hindi-English code-mix data. The authors reported that their system improved performance compared to state-of-the-art methods.

In a study presented in the paper [30], the authors explored the relationship between aggression, hate, sarcasm, humor, and stance in Hinglish (Hindi-English) text. They evaluated various existing deep learning methods for hate speech detection in code-mix texts and proposed an evaluation scheme for identifying offensive keywords from Hindi-English code-mix data.

A paper [31] tackled the issue of hate speech in Hindi-English code-mix text. The authors designed a framework structure by employing existing algorithms to create the 'MoH' (Map Only Hindi) dataset. They evaluated the models on three different datasets and assessed their performance using precision, recall, and F1-score. The final results indicated a significant improvement over the baseline model, underscoring notable progress in achieving state-of-the-art scores on all three datasets.

Upon reviewing the literature, it becomes evident that there are substantial gaps in the field of hate speech detection, particularly concerning English-Hindi code-mixed data. Three primary challenges have been identified and addressed: the scarcity of large training datasets, the complexities of code-mixed data, and comprehensive performance evaluation using all popular metrics, which some prior works have not conducted, as depicted in Table 11.

3. Methodology

This section provides an overview of the datasets employed in the research, along with an assessment of machine learning and deep learning methodologies for detecting hate speech within these selected datasets. The segment concludes by presenting a novel, custom-built model that attains cutting-edge performance in identifying hate speech within English-Hindi code-mixed data.

3.1 Dataset Description

For several years, research on hate speech detection has primarily focused on the English language, resulting in a plethora of available datasets for English-specific hate speech analysis. In contrast, the literature highlights a notable deficiency of smaller datasets for English-Hindi code-mix text. To address this limitation, the authors of this study opted to employ a unified English-Hindi code-mix dataset compiled from three publicly accessible sources: Bohra 2018 [17], Kumar 2018 [32], and HASOC 2021 [33].

Bohra 2018 dataset [17] comprises 4,500 tweets, each categorized as "Yes" for containing hate speech or "No" for being non-hate speech. Among these instances, 2,345 are classified as "Yes" (hate speech), and 2,155 as "No" (non-hate speech). Most tweets in this dataset are written in a mixture of Hindi and English using the Roman alphabet.

Kumar 2018 dataset [32], initially derived from YouTube comments and various social media platforms, originally features three classes: Not Aggressive (NAG), Covertly Aggressive (CAG), and Overtly Aggressive (OAG). To align it with the other datasets, the authors merged the CAG and OAG classes into a single "Hate speech" class, with the remaining labeled as "Non-hate speech." This dataset contains a total of 11,100 instances, with 5,834 falling into the "Hate speech" category and 5,266 into the "Non-hate" category.

The HASOC 2021 Hindi-English Coded dataset [33] consists of 5,000 social media tweets written in Hindi-English code-mix language, categorized as either "Hate" (2,258 instances) or "Non-hate" (2,742 instances). The combined dataset comprises 20,000 instances, distributed across training (70%), validation (15%), and testing (15%) sets.

3.2 Text Processing

The provided datasets underwent preprocessing steps prior to being input into the model. This preprocessing phase involved a systematic and meticulous process of eliminating extraneous elements from the tweets and comments in the dataset. Specifically, it involved the removal of superfluous spaces, rectification of missing values, and the systematic removal of unreadable characters. Additionally, we employed regular expressions to scrutinize and eliminate hyperlinks, as they are often irrelevant in the context of hate speech classification. Hashtags (#) and emoticons/emojis were also stripped from the text. Furthermore, we conducted an investigation to assess the impact of emojis on the hatefulness of the content and found that emojis did not significantly contribute to our classification task. Subsequently, the processed data underwent tokenization using an NLTK-based tokenizer, with the removal of punctuation marks. Lastly, the NLTK-based PorterStemmer() was employed to reduce each word to its root form. However, it's worth noting that we encountered issues during the stemming process, primarily because the Porter stemmer is tailored for the English language, while our dataset contained a mix of Hindi and English words. To rectify this issue, we translated some of the Hindi tweets into their English equivalents using the free Google Translate service.

3.3 Feature Extraction for Machine Learning Techniques

Distinguishing attributes play a pivotal role in the effectiveness of machine learning methods and can often determine the success or failure of a task. In this study, we employed four distinct techniques for feature extraction from tokenized text data. To begin, we utilized Count Vectorizer, a tool for transforming text documents into a matrix representing token counts. Next, we adopted Hashing Vectorizer, which converts text documents into a matrix indicating token occurrences. Our third approach involved implementing Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer, which converts a set of raw documents into a matrix of TF-IDF features. Lastly, we applied the TF-IDF Transformer to convert a count matrix into a normalized representation, either as TF or TF-IDF values.

3.4 Word Embedding for Deep Learning Techniques

Word embeddings play a crucial role in conveying both the syntactic and semantic context of words or terms within documents, enabling a deeper understanding of their similarity to others within the text. This method involves representing each term in the text data as a vector or numeric features, aiming to capture the semantic nuances of the terms. In this study, we have employed three distinct word embedding techniques:

1. FastText: FastText is an open-source library that empowers high-level models to harness text representations for diverse text processing tasks. Specifically, it employs an English-based algorithm to vectorize words.

2. Glove 60b 100D: This technique leverages an unsupervised learning algorithm to convert text data into vector representations. It is trained on a global word corpus, allowing it to grasp intricate relationships between words, including linear substructures within the word vector space.

3. Word2vec: Word2vec utilizes a neural network model to identify patterns in word associations within extensive text corpora. Once trained, it can identify partial sentence structures and synonymous words. The resulting word vectors are arranged in such a way that antonyms point in opposite directions while synonymous words point in the same direction, enhancing the model's ability to capture semantic relationships.

3.5 Deep Learning Methodology

In light of the remarkable success of deep learning techniques across various domains, particularly in the field of Natural Language Processing (NLP), our research utilized four custom-tailored models in conjunction with three distinct word embeddings for experimentation. The initial model in our lineup was the Long Short-Term Memory (LSTM) architecture, chosen for its proficiency in modeling sequences, addressing the issues inherent in basic Recurrent Neural Network (RNN) models, such as slow learning over extended sequences. We then extended our approach to incorporate Bidirectional LSTM (BiLSTM), an advanced variant of the LSTM, leveraging both forward and backward sequences to enhance learning capabilities. While Convolutional Neural Networks (CNNs) are typically associated with image-related tasks, their aptitude for learning intermediate features makes them valuable for textual data as well. In our study, one-dimensional CNNs were employed as feature extractors in the initial model stage. The subsequent section will delve into a detailed examination of our findings and analyses for all the models in conjunction with various feature extraction techniques and word embeddings.

Following extensive evaluations, our results clearly indicate that the CNN-BiLSTM model, combined with word2vec embeddings, outperformed all other models. This specific model was implemented using the Keras API, comprising several sequential layers. The initial layer is the embedding layer, serving as the input stage for the training data, with pre-trained word embeddings initialized using a prepared embedding matrix. To mitigate overfitting, a Dropout layer with a 0.3 dropout rate follows. Subsequently, a one-dimensional CNN layer (Conv1D) with 64 filters of size 5x5 and ReLU activation

function is applied for local feature extraction. The following layer employs MaxPooling1D to pool the feature vectors with a window size of 4. These pooled feature maps are then passed to the subsequent BiLSTM layer, capable of capturing long-term dependencies within the input data while maintaining memory, with an output dimension of 128. Another Dropout layer with a 0.3 rate is introduced, leading to the final layer, a dense layer for binary classification (real or fake) with Sigmoid activation. The model employs binary cross-entropy as the loss function and the Adaptive Moment Estimation (Adam) optimizer for training, utilizing a batch size of 64.

4. Results and Analysis

Table 1, 2 and 3 present the outcomes of the deep learning models, while Chart -1 illustrates accuracy trends for these methods. Notably, the CNN-BiLSTM model consistently outperforms other models across various metrics, particularly when utilizing word2vec word embeddings. These findings indicate that the incorporation of BiLSTM enhances performance compared to the standard LSTM, primarily owing to its dual modeling capability. Moreover, the addition of a CNN feature extraction layer alongside BiLSTM appears to further enhance results, suggesting the potential of a 1-D CNN layer as a valuable feature extractor. However, this trend is less pronounced in the case of the basic LSTM, particularly when using GloVe and FastText word embeddings. Future experiments should investigate the specific contribution of the CNN component to the overall architecture.

5. Conclusion

The research landscape for hate speech detection is well-developed in English but lags behind for under-resourced and code-mixed languages like English-Hindi. This paper presents an effort to tackle hate speech detection in Hinglish, a blend of English and Hindi. To address the challenge of limited datasets, we merged three publicly available datasets, resulting in a final dataset of over 20,000 samples. Various machine learning and deep learning models were applied to detect hate speech. The experiments revealed that the CNN-BiLSTM model achieved the highest accuracy among all methods. This CNN-BiLSTM-based approach surpasses recent state-of-the-art techniques for detecting hate speech in English-Hindi code-mixed datasets, achieving an impressive accuracy of 87% on the consolidated dataset.

Table -1: Results of deep learning models with FastText word embedding

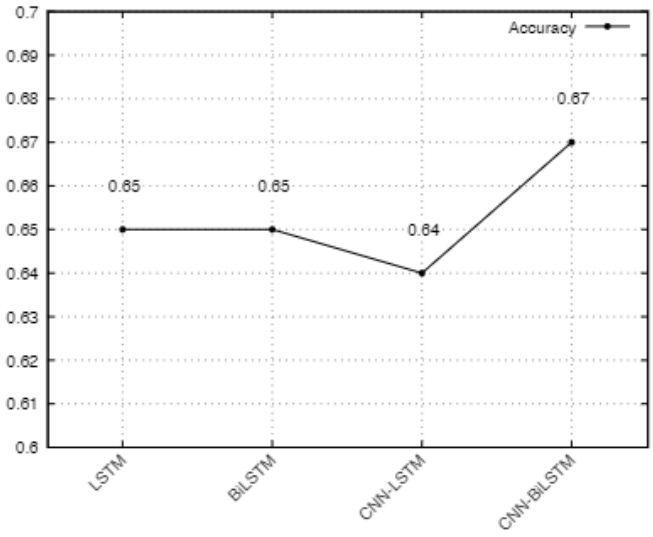
Model	Accuracy	Precission	Recall	F1-Score
LSTM	0.65	0.65	0.66	0.65
BiLSTM	0.65	0.66	0.63	0.65
CNN-LSTM	0.64	0.7	0.56	0.63
CNN-BiLSTM	0.67	0.68	0.66	0.67

Table -2: Results of deep learning models with Glove-6B-100d word embedding

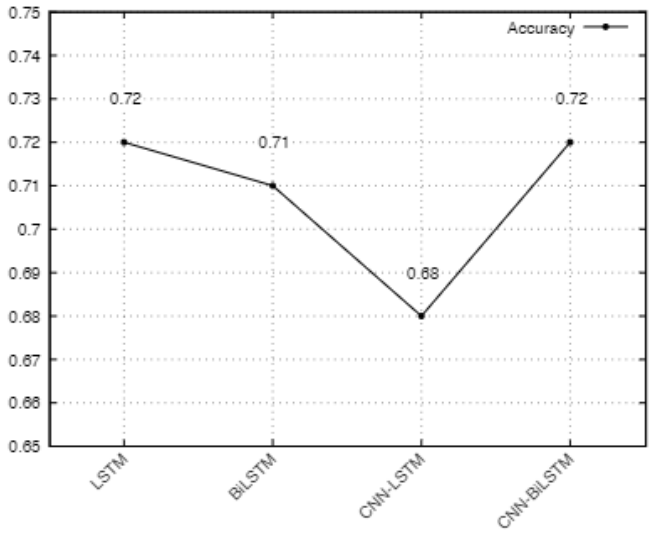
Model	Accuracy	Precission	Recall	F1-Score
LSTM	0.72	0.7	0.71	0.70
BiLSTM	0.71	0.7	0.70	0.71
CNN-LSTM	0.68	0.67	0.68	0.66
CNN-BiLSTM	0.72	0.72	0.71	0.72

Table -3: Results of deep learning models with word2vec word embedding

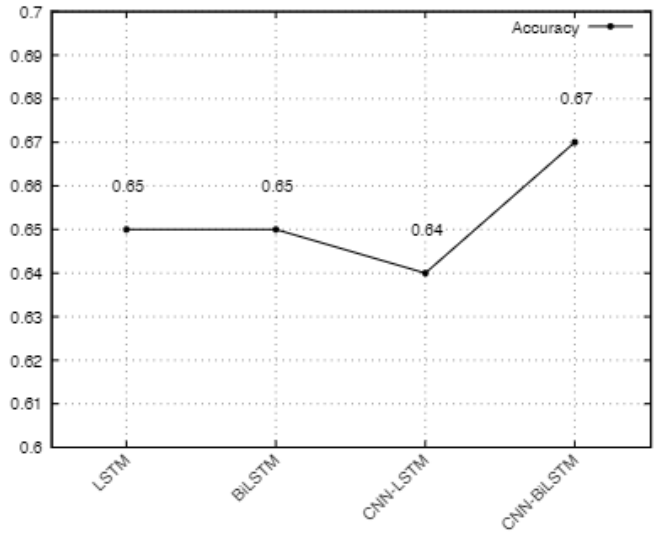
Model	Accuracy	Precission	Recall	F1-Score
LSTM	0.76	0.75	0.74	0.74
BiLSTM	0.77	0.77	0.75	0.74
CNN-LSTM	0.78	0.77	0.76	0.76
CNN-BiLSTM	0.87	0.82	0.85	0.82



(a)FastText



(b)GloVe



(a) FastText

Chart -1: Graphs depicting the accuracy of deep learning methods with different word embeddings

REFERENCES

[1] India Social Media Statistics 2021. The Global Statistics. 2021 Dec. https://www.theglobalstatistics.com/india-social-media-statistics/.

[2] Hate speech. Wikimedia Foundation; 2021. https://en.wikipedia.org/w/index.php?title=Hate_speech&oldid=1059042962.

[3] Council of Europe;. https://www.coe.int/en/web/portal/home.

[4] Bharti S, Yadav AK, Kumar M, Yadav D. Cyberbullying detection from tweets using deep learning. Kybernetes. 2021.

[5] Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation. 2020:1-47.

[6] Shah SR, Kaushik A. Sentiment analysis on Indian indigenous languages: a review on multilingual opinion mining. arXiv preprint arXiv:191112848. 2019.

[7] Kaur S, Singh S, Kaushal S. Abusive Content Detection in Online User-Generated Data: A survey. Procedia Computer Science. 2021;189:274-81.

[8] Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review.

[9] Artificial Intelligence Review. 2020;53(6):4335-85.

[10] Drias HH, Drias Y. Mining Twitter Data on COVID-19 for Sentiment analysis and frequent patterns Discovery. medRxiv. 2020.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[11] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In WWW, pages 29–30, 2015.

[12] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759, 2016.

[13] Y. Kim. Convolutional Neural Networks for Sentence Classification. In EMNLP, pages 1746–1751, 2014.

[14] Kumar R, Reganti AN, Bhatia A, Maheshwari T. Aggression-annotated corpus of hindi- english code-mixed data. arXiv preprint arXiv:180309402. 2018.

[15] Li T, Lin L, Choi M, Fu K, Gong S, Wang J. Youtube av 50k: an annotated corpus for comments in autonomous vehicles. In: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). IEEE; 2018. p. 1-5.

[16] Ravi K, Ravi V. Sentiment classification of Hinglish text. In: 2016 3rd International Conference on Recent Advances in Information Technology (RAIT). IEEE; 2016. p. 641-5.

[17] Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M. A dataset of hindi-english code-mixed social media text for hate speech detection. In: Proceedings of the

second workshop on computational modeling of people's opinions, personality, and emotions in social media; 2018. p. 36-41.

[18] Pamungkas EW, Basile V, Patti V. Misogyny detection in twitter: a multilingual and crossdomain study. Information Processing & Management. 2020;57(6):102360.

[19] Sreelakshmi K, Premjith B, Soman K. Detection of hate speech text in Hindi-English code-mixed data. Procedia Computer Science. 2020;171:737-44.

[20] Kamble S, Joshi A. Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:181105145. 2018.

[21] Mathur P, Sawhney R, Ayyar M, Shah R. Did you offend me? classification of offensive tweets in hinglish language. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2); 2018. p. 138-48.

[22] Mathur P, Shah R, Sawhney R, Mahata D. Detecting offensive tweets in hindi-english code-switched language. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media; 2018. p. 18-26.

[23] Singh P, Lefever E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In: Proceedings of the The 4th Workshop on Computational Approaches to Code Switching; 2020. p. 45-51.

[24] Kovács G, Alonso P, Saini R. Challenges of Hate Speech Detection in Social Media. SN Computer Science. 2021;2(2):1-15.

[25] Chopra S, Sawhney R, Mathur P, Shah RR. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34; 2020. p. 386-93.

[26] Gupta V, Sehra V, Vardhan YR, et al. Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE; 2021. p. 1112-8.

[27] Santosh T, Aravind K. Hate speech detection in hindi-english code-mixed social media text.In: Proceedings of the ACM India joint international conference on data science and management of data; 2019. p. 310-3.

[28] Sasidhar TT, Premjith B, Soman K. Emotion detection in hinglish (hindi+ english) code-mixed social media text. Procedia Computer Science. 2020;171:1346-52.17

[29] Kapoor R, Kumar Y, Rajput K, Shah RR, Kumaraguru P, Zimmermann R. Mind your language: Abuse and offense detection for code-switched languages. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33; 2019. p. 9951-2.

[30] Sengupta A, Bhattacharjee SK, Akhtar MS, Chakraborty T. Does aggression lead to hate? Detecting and reasoning offensive traits in hinglish code-mixed texts. Neurocomputing. 2021.

[31] Sharma A, Kabra A, Jain M. Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. Information Processing & Management. 2022;59(1):102760.

[32] Zhu AZ, Thakur D, Özaslan T, Pfrommer B, Kumar V, Daniilidis K. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. IEEE Robotics and Automation Letters. 2018;3(3):2032-9.

[33] Mandl T, Modha S, Shahi GK, Madhu H, Satapara S, Majumder P, et al. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages. arXiv preprint arXiv:211209301. 2021.

[34] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In WWW, pages 145–153, 2016.

[35] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In EMNLP, volume 14, pages 1532–43, 2014.

[36] Z. Waseem and D. Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In NAACL-HLT, pages 88–93, 2016.

[37] ROY, S. G., NARAYAN, U., RAHA, T., ABID, Z., AND VARMA, V. Leveraging multilingual transformers for hate speech detection. arXiv preprint arXiv:2101.03207 (2021).

[38] SAROJ, A., MUNDOTIYA, R. K., AND PAL, S. Irlab@ iitbhu at hasoc 2019: Traditional machine learning for hate speech and offensive content identification. In FIRE (Working Notes) (2019), pp. 308–314.

[39] SCHMIDT, A., AND WIEGAND, M. A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (2017), pp. 1–10.

[40] SHARMA, Y., AGRAWAL, G., JAIN, P., AND KUMAR, T. Vector representation of words for sentiment analysis using glove. In 2017 International Conference on

Intelligent Communication and Computational Techniques (ICCT) (2017), pp. 279–284.

[41] TANG, Z., LI, W., LI, Y., ZHAO, W., AND LI, S. Several alternative term weighting methods for text representation and classification. Knowledge-Based Systems 207 (2020), 106399.

[42] TAY, Y., DEHGHANI, M., BAHRI, D., AND METZLER, D. Efficient transformers: A survey. arXiv preprint arXiv:2009.06732 (2020).

[43] TURC, I., CHANG, M., LEE, K., AND TOUTANOVA, K. Well-read students learn better: The impact of student initialization on knowledge distillation. CoRR abs/1908.08962 (2019).

[44] VIJAYARAGHAVAN, P., LAROCHELLE, H., AND ROY, D. Interpretable multi-modal hate speech detection. arXiv preprint arXiv:2103.01616 (2021).