2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10306709

HSDH: Detection of Hate Speech on social media with an effective deep neural network for code-mixed Hinglish data

Rohit Kumar Kaliyar
School of Computer Science Engineering and Technology
Bennett University, India
er.rohitk06@gmail.com

Anurag Goswami
School of Computer Science Engineering and Technology
Bennett University, India
anurag.goswami@bennett.edu.in

Ujali Sharma
School of AI-DS
Indira Gandhi Delhi Technical University for Women, India
Ujali.sharma.us415@gmail.com

Kanika Kanojia
School of AI-DS
Indira Gandhi Delhi Technical University for Women, India
Kanikakj07@gmail.com

Mohit Agrawal
School of Computer Science Engineering and Technology
Bennett University, India
26mohit@gmail.com

***Abstract*—The phenomenal rise of social media platforms like Twitter, Facebook, Instagram, and Reddit has led to the blending of native languages or regional tongues with English for the purpose of improving communication in linguistically open geographic regions around the world. There are many ways in which Holocaust denial can lead to an increase in violence, from direct assault to purging out of compassion. Online, people are very hostile to one another. Distinguishing between language that incites hatred and language that is disparaging is a fundamental challenge in the categorization and tracking of extremely toxic lexical features. Our research focuses on identifying harmful tweets composed in Hinglish, a fusion of Hindi and the Roman alphabet. We propose a system in this paper for classifying tweets as either abusive, neutral, or offensive. The help of Hindi-English offensive tweet dataset is comprised of tweets written in the code-transferred language of Hindi and is further subdivided into three groups: neutral, abusive, and hateful. We studied the abusive and hate speech dataset with transfer learning and pre-trained the proposed model on Hinglish-processed English tweets. With our proposed model, we were able to improve accuracy to 98.54 percent.**

***Index Terms*—Hate Speech, Social Media, Deep Neural Network, Classification**

I. Introduction

Potentially harmful online content has become a subject of great concern due to the explosive growth in internet access by people of all ages, races, and socioeconomic backgrounds in the modern technological world, where people spend an inordinate amount of time [4] on social media. In light of their efforts to curb hate speech, social media giants like Facebook and Twitter are facing mounting pressure to address user concerns about potential infringements on their right to free expression. Many different national and international laws and governmental inquiries are being made of these online communication platforms. Therefore, the social media behemoths were charged with deciding what to air and what not to air. Baseline content was developed in response to the need for restraints; it defines hate speech and examines its differences from other forms of abusive language. Offensive content is all too common on social media, and it poses a threat to a tolerant society. With the help of Figure 1, we can see a few instances of online hate speech. Figure 2 displays all of the Scopus documents that discuss hate speech.

The proliferation of offensive material online is in large part due to the use of "par Hinglish," which consists of Hindi words written in the Roman script rather than the Devanagari script. The grammar of Hinglish is independent of the speaker's accent [13], making it a true "two-way street." As a result of its widespread use and influence in the area, Hinglish has morphosyntactical and lexical features that are distinct from standard Hindi. There are a lot of different ways that words can be spelled in Hinglish, and many different ways that they can be interpreted, making automatic classification of Hinglish extremely difficult[10].

In this paper, we opt to solve the problem of identifying offensive Hinglish tweets by developing a deep learning

model to analyze the textual data and classify it as either non-offensive, abusive, or hate-inducing. Using a dataset of manually annotated Hinglish tweets, the performance of the framework is assessed. The method has two distinct phases. This article provides a lexical translation of the Hinglish text into Roman English words, and it investigates the semantic relationship between Hindi-English code-switched language and native English. We then use a convolutional neural network (CNN) to transfer knowledge and assess the efficacy of tweets that are semantically similar but syntactically distinct after being obtained through transliteration and translation.

II. Related Work

A. Hate Speech Detection

For the past 20 years, the problem of recognizing hateful speech and abusive language on the internet has always been a contentious issue in the scientific world. 'Smokey,' a decision tree-based classification with 47 semantic and syntactic textual features, was created [43]. When 'Smokey' was trained on a diverse handful of 720 dynamic web postings manually marked (as "okay" "maybe," or "flame") and then analyzed on 502 additional messages, it did a good job categorizing non-inflammatory communications but entirely failed to recognize flame texts achieving accuracy with just 88.2 %on an assignment with an 86.1 % huge percentage baseline. Gradually shifting from characteristics based purely on the language used in user-generated content [5] proposed an approach that additionally considers the individuals' publishing behaviors in addition to identifying people who are abusive.

Buckels et al. (2014) [36], on the other hand, tried to derive destructive personality characteristics in digital user activity. This is especially important for quick prudence of online communication, as Yin et al. (2009) [35] and Papegnies et al. (2017) [37] point out, with the aforementioned proposing a few descriptors (at the morphometric, idiomatic expressions, and predictive analysis stages) which can be used to distinguish when players on a French Massively Multiplayer Online gaming website switch from describing competitive match issues to making ridiculous offensive statements. Specific diagnostics samples, including the HATECHECK test scenarios, were used to evaluate models together in a variety of NLP activities, including language processing interpretation [2], computational linguistics [3], and language processing [4]. Yet, they've only been used sporadically in hate speech identification.

Neural networks have been deliberately programmed to anticipate approval ratings in certain preliminary studies (Lawrence et al., 1996; Allen and Seidenberg, 1999) [38]; Post (2011). Warstadt et al. (2018) [39] employ transfer learning whereby an unsupervised model is fine-tuned based on acceptance prognostication. The labeled datasets have not been really shared publicly, which has become a recurring issue with some of this research. Ross et al. (2016) [40], on the other hand, analyzed 541 German tweets, focusing on topics such as observer and descriptor dependability, and what data should be offered to observers. Waseem (2016) [41] addresses great similarities while presenting a sample of 6,909 English tweets labeled by CrowdFlower users and expanding on a subsequent dataset (Waseem and Hovy, 2016) [42].

Several scholars have investigated text vocabulary recognition for various linguistic pairings and accent variants over these last decades [14][15][16]. The FIRE works together to develop a series based on language recognition of code-mixed web searches for information extraction in English and Indian dialects [17][18]. The First and Second Concurrent Tasks on Text Recognition in code-Switched Data [44][45] demonstrate the need for automated code-switched encoding and give comparisons of various language recognition systems. The preferred method from the second version among these task interdependence employs a logistic regression model, with SPAENG scoring 97.30 % just on token-level F1-score.

A comparison of monolingual and multilingual BERT models was undertaken. In five distinct intermediate fine-tuning studies, Marathi monolingual models outperformed multilingual BERT versions [46]. Through freezing the BERT encoder layers, and were able to examine sentence embeddings among these models. demonstrated how monolingual MahaBERT-based models produce rich interpretations when particularly in comparison to multilingual sentence embeddings. Conversely, such embeddings are therefore not a fairly straightforward process and therefore do not perform effectively with out-of-domain social network datasets. To effectively categorize tweets [47], the authors implemented a comprehensive text mining method and the Naive Bayes machine learning classification algorithm in two sets of data (tweets Num1 and tweets Num2) acquired from Twitter. The proposed method performed well enough in terms of many metrics depending on the confusion matrix, including those of the accuracy metric, which attained 87. 23% on the first set of data and 93. 06% on the second set of data.

In this paper, we propose a new approach, grounded in advanced deep learning architecture, for assessing the quality of hate speech articles shared across social media platforms. The method will be tested in Section III, and its results will be described in Section III. The benefits of our proposed approach include:

- It automatically extracts useful features with the architecture of deep learning.
- It provides more accurate results as compared to existing benchmarks.

III. Methodology

This subsection describes the approach and structure of our proposed model.

A. Dataset

This section describes the information collected for this study.

*1) **HASOC2019 dataset**:* The HASOC2019 dataset [1]was developed to assist in the detection of derogatory and vile terms in Indo-European languages. The goal of HASOC is to

[1] https://hasocfire.github.io/hasoc/2019/

Fig. 1. Examples of hate speech on Social Media (Source: Facebook)

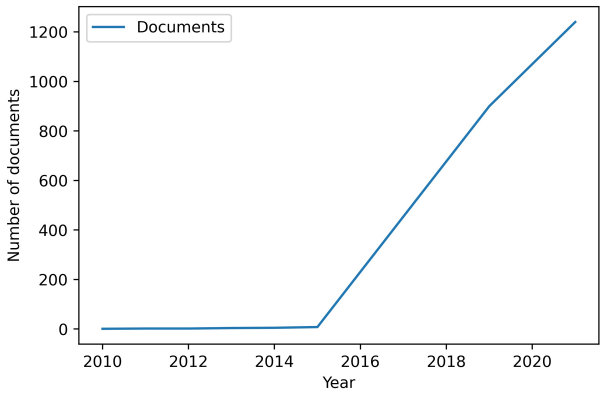

Fig. 2. The number of research documents related to hate speech news between 2010 and 2021 in the Scopus Database

advance the linguistic characterization of hate speech through study and technology. The HASOC track's first session develops tools for spotting hate speech in Hindi, German, and English. Utilizing Facebook and Twitter, three datasets were created and released to the public. On-platform objectionable text was found using the HASOC 2019 code-mixed Hindi dataset. The data enables the development and evaluation of machine learning systems under close supervision. The prevalence of harmful and undesirable content online creates a significant problem for civilizations. Objective debates are undermined by offensive language, such as insulting, harmful, disparaging, or obscene material targeted at yet another individual and viewable to others. This kind of rhetoric is becoming even more prevalent online and, therefore, can cause conversations to turn more extreme. To influence public sentiment, intelligent and critical debate is necessary. The democratic process may be threatened by objectionable content. Open societies must simultaneously come up with a suitable response to such content that avoids enforcing strict surveillance laws. Therefore, as a result, several social networking sites keep an eye on user posts. As a result, there is the urgent provision of tools that can detect dubious content automatically. In order to eradicate abusive behavior in specific media, online communities, social media businesses, and technology firms have already significantly leveraged technology and procedures to recognize objectionable language.

B. Pre-processing

Cleaning and preparing text data is referred to as "text preprocessing." In its most basic form, preprocessing refers to all the modifications applied to raw data prior to feeding it to a machine learning or deep learning algorithm. Examples of such alterations include eliminating HTML elements, extra white space, special characters, and lowercase text overall, converting number words to numeric form, and removing numerals.

TABLE I
HYPER-PARAMETERS FOR HSDH

Hyper-parameter	Description or Value
No. of Convolutional layer	6
No. of Max-pooling layer	2
Kernel-sizes	1, 3, and 5
No. of Dense layer	4
No. of filters in conv-layers	1024,512,256,128,,64,32
No. of filters in dense-layers	1024,512,128,2
Loss function	binary_crossentropy
Activation function	ReLU
Optimizer	Adam
Metrics	Accuracy
Batch-size	256, 32
Batch-Normalization	Yes
Number of Epochs	10
Dropout	0.1

C. HSDH: Architecture of our proposed deep learning model

Our proposed deep learning network has been tested in this study using datasets of actual hoaxes. Our deep neural network's layered structure is depicted in Figure 3. In this approach, the HASOC dataset was used, and we employed a very deep convolutional neural network to recognize abusive text. After preprocessing the data, we divided it into train and

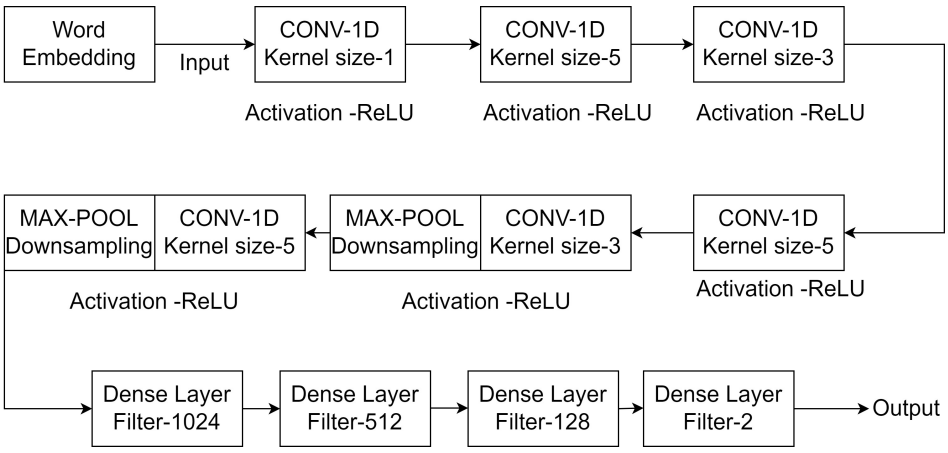

Fig. 3. Proposed Model

TABLE II
COMPARISON WITH EXISTING BENCHMARKS USING THE DATASET-PHEME

Author	Model	Accuracy (%)
Spertus et al. [43]	Machine Learning [Features-BOW]	87.20%
Alaoui et al. [47]	Machine Learning [Naive Bayes]	88.20%
Velankar et al.[46]	BERT-based approach	93.06%
Solorio et al. [44]	Featured-based Approach [ML]	96.20%
Molina et al. [45]	Logistic Regression [ML]	97.30%
Our proposed model	HSDH	**98.54%**

test sets in an 80:20 ratio, and then we divided the test set into a test and validation set in a 50:50 ratio. Our sequential model receives an embedding layer as input that has a dimension of 100 and an input length of 2000. First, two one-dimensional convolutional layers are generated, each with 1024 filters, a kernel size of 1, and a Relu activation function applied. These layers are then standardized using BatchNormalization after activation function application.

Then two one-dimensional convolutional layers with 1024 filters, a kernel size of 1, and the application of the Relu activation function are standardized using batch normalization and then added. In the model with 32 filters and a kernel size of 5, five further 1-D convolutional layers are added after that. These layers are standardized using BatchNormalization, and Relu Activation is applied once more after normalization. The model now has five more 1-D convolutional layers with 64 filters, three kernel sizes, and the Relu Activation function. These layers are standardized using BatchNormalization, and Relu Activation is applied once more after normalization. The outputs from this layer are then fed into the subsequent 1-D convolutional layer, which uses the Relu activation function, a kernel size of 5, and 64 filters.

To standardize this layer BatchNormalization technique is used and then again Relu activation is applied, then outputs from this layer are fed as input to the next 1-D convolutional Layer with 128 filters, kernel size as 3, and relu activation following BatchNormalization and applying MaxPooling1D with pool size as 2 to downsize the layer, here strides will be default taken as pool size. The MaxPooling Relu activation function is further applied. This type of layer is again made two more times, then seven more 1-D convolutional layers are added to the model with 128 filters, kernel size of 5, and relu activation following BatchNormalization to standardize the layer and applying relu activation after standardization. Five more 1-D convolutional layers are added to the model with filters set to 256, and kernel size set to 3 following the Relu activation function, further standardizing the layer using BatchNormalisation and again applying Relu activation. Then the output from the latest layer will be fed to the next 1-D convolutional layer having 256 filters, kernel size 5, and Relu activation function, further following BatchNormalisation. This type of layer is added two more times in the model.

With a small number of dropouts, the accuracy will progressively increase and the loss will gradually decrease. Dropout was chosen as a parameter because it makes the classification model simpler and prevents over-fitting. The activation function has been determined to be ReLU (Rectified Linear Unit).Decision-making functions with enhanced non-linear properties can zero out negative activation map values locally without affecting other convolution layer fields, removing unwanted information. The ReLU equation is given by:

$$\sigma = max(0, z) \tag{1}$$

In order to evaluate the effectiveness of a classification model, we have employed binary cross-entropy as a loss function. Additionally, it rises as the projected probability moves away from the label. In binary classification (number of classes M equals 2), cross-entropy can be calculated as:

$$L = -\left(ylog(p) + (1-y)log(1-p)\right) \tag{2}$$

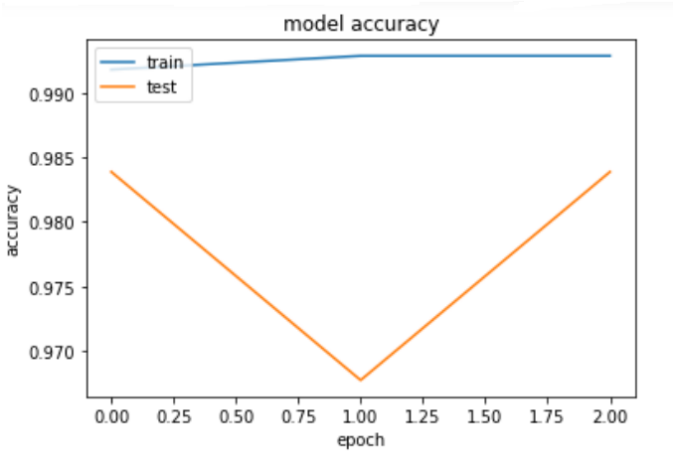

Fig. 4. Training and Validation Accuracy with HSDH

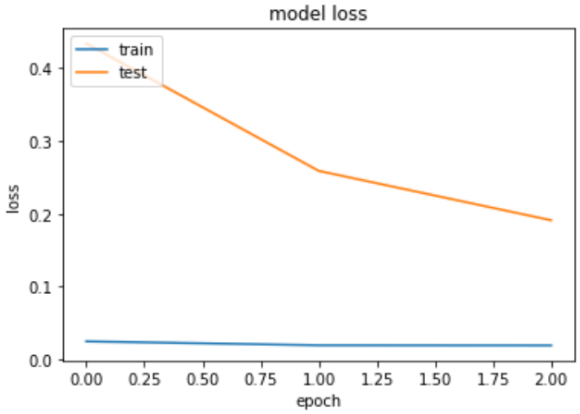

Fig. 5. Cross Entropy Loss with HSDH

We have chosen Adam to be an optimizer in our network. For our suggested model, we have thought about the ideal hyperparameters (see Table 1 for additional information). A model that minimizes a predetermined loss function is produced through hyperparameter optimization, which locates a tuple of hyperparameters.

NVIDIA DGX-1 V100 hardware was used to conduct the research in this study. The computer has 128 GB of RAM, 1000 TFLOPS speed, 5120 tensor cores, and 40600 CUDA cores.

IV. Results and Discussion

The real-world fake news dataset HASOC2019 was used to tabulate the experimental and evaluation results. Experimental findings show that when compared to other available detection models for detecting hate speech, our proposed model provides state-of-the-art results.

The accuracy and cross-entropy loss using the HASOC dataset are displayed in Figures 4 and 5. We can see that our suggested model performs quite admirably over a period of 20 epochs.

In the case of HASOC, cross-entropy loss is hardly noticeable. Using HASOC, our suggested model performed well, with an accuracy of 98.54%. Our suggested model has significantly increased the accuracy of hate speech detection using social media data, according to classification results. With the help of additional real-world fake news datasets, we have verified our proposed model. With other datasets as well, we have produced results that are compelling.

V. Conclusion and Future Work

With the help of our suggested deep learning model, we were able to get excellent results because it was able to capture both phrase-level representations and temporal semantics while maintaining optimum accuracy. The experimental results demonstrated empirically how well the suggested approach addressed the challenge of detecting hate speech.

In our upcoming study, we'll try to use different metadata for more precise classification and graph-based analysis to determine the precise path that hate speech news pieces take as they spread.

References

[1] Mathur, Puneet, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. "Did you offend me? classification of offensive tweets in hinglish language." In Proceedings of the 2nd workshop on abusive language online (ALW2), pp. 138-148. 2018.

[2] Bohra, Aditya, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. "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, pp. 36-41. 2018.

[3] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate- speech." In Proceedings of the first workshop on abusive language online, pp. 85-90. 2017.

[4] Warner, William, and Julia Hirschberg. "Detecting hate speech on the world wide web." In Proceedings of the second workshop on language in social media, pp. 19-26. 2012.

[5] Yuvaraj, Natarajan, Victor Chang, Balasubramanian Gobinathan, Arulprakash Pinagapani, Srihari Kannan, Gaurav Dhiman, and Arsath Raja Rajan. "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification." Computers & Electrical Engineering 92 (2021): 107186.

[6] Sowmya, V. B., Monojit Choudhury, Kalika Bali, Tirthankar Dasgupta, and Anupam Basu. "Resource creation for training and testing of transliteration systems for indian languages." In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC & 39;10). 2010.

[7] Sreelakshmi, K., B. Premjith, and K. P. Soman. "Detection of hate speech text in Hindi-English code-mixed data." Procedia Computer Science 171 (2020): 737-744.

[8] Kumari, Kirti, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, and Nripendra Pratap Rana. "Bilingual Cyber-aggression detection on social media using LSTM autoencoder." Soft Computing 25, no. 14 (2021): 8999-9012.

[9] Kumar, Ritesh, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. "Benchmarking aggression identification in social media." In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp. 1-11. 2018.

[10] Veerasamy, Sevagen, Yash Khare, Abhijit Ramesh, S. Adarsh, Pranjal Singh, and T. Anjali. "Hate Speech Detection using mono BERT model in custom Content-Management-System" In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp.1681-1686. IEEE, 2022.

[11] Mandl, Thomas, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. "Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages." In Proceedings of the 11th forum for information retrieval evaluation, pp. 14-17. 2019.

[12] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." In Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1, pp. 512-515. 2017.

[13] ElSherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. "Peer to peer hate: Hate speech instigators and their targets." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1. 2018.

[14] Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. "Multilingual and multi-aspect hate speech analysis." arXiv preprint arXiv:1908.11049 (2019).

[15] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P. and Sanguinetti, M., 2019, June. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation (pp. 54-63).

[16] Mathur, Puneet, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. "Did you offend me? classification of offensive tweets in hinglish language." In Proceedings of the 2nd workshop on abusive language online (ALW2), pp. 138-148. 2018.

[17] Vashistha, Neeraj, and Arkaitz Zubiaga. "Online multilingual hate speech detection: experimenting with Hindi and English social media" Information 12, no. 1 (2020): 5.

[18] Islam, Tanvirul, Nadim Ahmed, and Subhenur Latif. "An evolutionary approach to comparative analysis of detecting Bangla abusive text." Bulletin of Electrical Engineering and Informatics 10, no. 4 (2021): 2163-2169.

[19] Park, Hyunju, and Hong Kook Kim. "Verbal Abuse Classification Using Multiple Deep Neural Networks." In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 316-319. IEEE, 2021.

[20] Ahuja, Ravinder, Alisha Banga, and S. C. Sharma. "Detecting abusive comments using ensemble deep learning algorithms." In Malware Analysis Using Artificial Intelligence and Deep Learning, pp. 515-534. Springer, Cham, 2021.

[21] Haoxiang, Wang. "Emotional Analysis of Bogus Statistics in Social Media." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 178-186.

[22] Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. "All you need is" love" evading hate speech detection." In Proceedings of the 11th ACM workshop on artificial intelligence and security, pp. 2-12. 2018.

[23] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.

[24] Ayo, Femi Emmanuel, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions." Computer Science Review 38 (2020): 100311.

[25] O'Keeffe, Gwenn Schurgin, and Kathleen Clarke-Pearson. "The impact of social media on children, adolescents, and families." Pediatrics 127, no. 4 (2011): 800-804.

[26] Ravi, Kumar, and Vadlamani Ravi. "Sentiment classification of Hinglish text." In 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), pp. 641-645. IEEE, 2016.

[27] Vidgen, Bertie, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. "Challenges and frontiers in abusive content detection." Association for Computational Linguistics, 2019.

[28] Kuss, Daria J., and Mark D. Griffiths. "Online social networking and addiction—a review of the psychological literature." International journal of environmental research and public health 8, no. 9 (2011): 3528-3552.

[29] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 512-515. 2017.

[30] Srivastava, Ananya, Mohammed Hasan, Bhargav Yagnik, Rahee Walambe, and Ketan Kotecha. "Role of artificial intelligence in detection of hateful speech for Hinglish data on social media." In Applications of Artificial Intelligence and Machine Learning, pp. 83-95. Springer, Singapore, 2021.

[31] Sinha, R. Mahesh K., and Anil Thakur. "Machine translation of bilingual hindi-english (hinglish) text." In Proceedings of Machine Translation Summit X: Papers, pp. 149-156. 2005.

[32] Bassignana, Elisa, Valerio Basile, and Viviana Patti. "Hurtlex: A multilingual lexicon of words to hurt." In 5th Italian Conference on Computational Linguistics, CLiC-it 2018, vol. 2253, pp. 1-6. CEUR-WS, 2018.

[33] Mathur, Puneet, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. "Detecting offensive tweets in hindi-english code-switched language." In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, pp. 18-26. 2018.

[34] Thakur, Varsha, Roshani Sahu, and Somya Omer. "Current State of Hinglish Text Sentiment Analysis." In Proceedings of the International Conference on Innovative Computing & Communications (ICICC). 2020.

[35] Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2 (2009): 1-7.

[36] Buckels, Erin E., Paul D. Trapnell, and Delroy L. Paulhus. "Trolls just want to have fun." Personality and individual Differences 67 (2014): 97-102.

[37] Papegnies, Etienne, Vincent Labatut, Richard Dufour, and Georges Linares. "Impact of content features for automatic online abuse detection." In International Conference on Computational Linguistics and Intelligent Text Processing, pp. 404-419. Springer, Cham, 2017.

[38] Allen, Joseph, and Mark S. Seidenberg. "The emergence of grammaticality in connectionist networks." The emergence of language (1999): 115-151.

[39] Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. "Neural network acceptability judgments." Transactions of the Association for Computational Linguistics 7 (2019): 625-641.

[40] Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. "Measuring the reliability of hate speech annotations: The case of the european refugee crisis." arXiv preprint arXiv:1701.08118 (2017).

[41] Waseem, Zeerak. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter." In Proceedings of the first workshop on NLP and computational social science, pp. 138-142. 2016.

[42] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.

[43] Spertus, Ellen. "Smokey: Automatic recognition of hostile messages." In Aaai/iaai, pp. 1058-1065. 1997.

[44] Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari et al. "Overview for the first shared task on language identification in code-switched data." In Proceedings of the First Workshop on Computational Approaches to Code Switching, pp. 62-72. 2014.

[45] Molina, Giovanni, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. "Overview for the second shared task on language identification in code-switched data." arXiv preprint arXiv:1909.13016 (2019).

[46] Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi. "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi." arXiv preprint arXiv:2204.08669 (2022).

[47] Alaoui, Safae Sossi, Yousef Farhaoui, and Brahim Aksasse. "Hate Speech Detection Using Text Mining and Machine Learning." International Journal of Decision Support System Technology (IJDSST) 14, no. 1 (2022): 1-20.