Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

18

Performance Analysis on Detection of Cyberbullying in Code-Mixed Language on Social Media

Megha P. Shah

Assistant Professor

B P College of Computer Studies,

Sector-23, KSV University, Gandhinagar, Gujarat.

Dr. Bhadresh Pandya

Shree Maneklal M Patel Institute of Science And Research, Sector-23,

KSV University, Gandhinagar, Gujarat.

Abstract

Bullying that takes place online is known as cyberbullying. Social media's explosive expansion has made a lot of individuals, particularly young people, more vulnerable to cyberbullying. By using machine learning, we can identify linguistic patterns in the posts that involve cyberbullying and create a model that can automatically identify cyberbullying content. With the rapid growth of social media, cyberbullying has emerged as a major concern, affecting individuals' mental and emotional well-being. The detection and mitigation of cyberbullying are vital for creating safer digital spaces. This survey paper provides a comprehensive review of recent advancements in the detection of cyberbullying on social media platforms. It explores various approaches, including traditional machine learning methods and natural language processing (NLP) techniques. This study aims to explore ground-breaking methods for

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

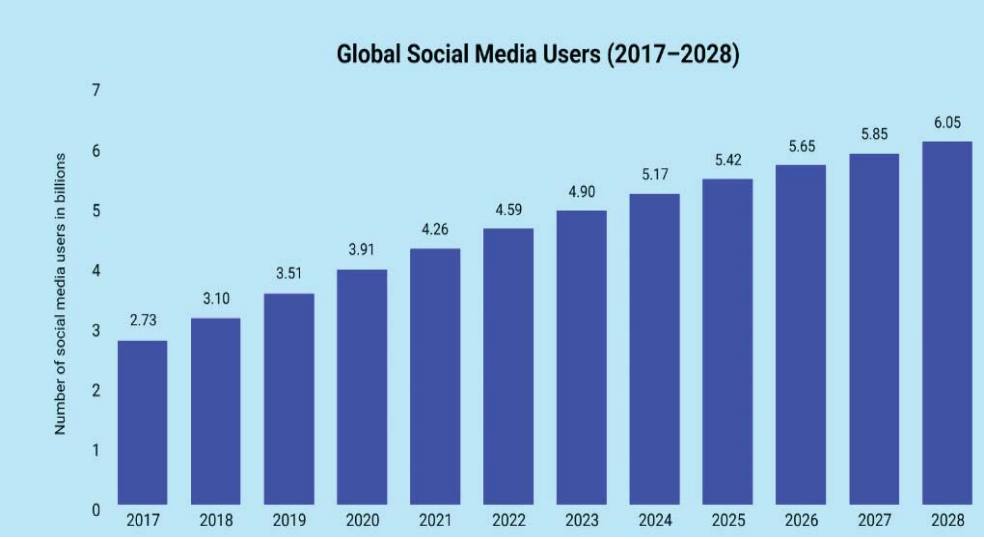
understanding and automatically detecting occurrences of cyberbullying across various social media platforms, including tweets, comments, and messages.

The survey also delves into the unique challenges associated with detecting cyberbullying, such as handling diverse languages, code-mixed text, and the evolving nature of abusive language. Special attention is given to the detection of cyberbullying in multilingual and code-mixed environments, where standard models may struggle to understand linguistic nuances.

Key Terms: lower code-mixed languages, natural language processing, social networking, machine learning, cyberbullying.

1. Introduction

The world's largest platform for connecting, exchanging, and exchanging ideas, content, pictures, videos, opinions, and daily updates is the Internet. Social media is vast and interactive because of its diversity. Some of the most popular are Facebook, WhatsApp, Instagram, Linked In, Twitter, and YouTube. Almost everyone utilizes social media these days. Data indicates that 5.17 billion people used social media networking sites in 2024; by 2028, that figure is predicted to rise to 6.05 billion.



Recent statistics indicate that cyberbullying has affected 27% of individuals at some point. Among teenagers, the incidence rate rises to 43%. A significant 88% of adolescent internet users report witnessing cyberbullying incidents. Girls are more frequently targeted by cyberbullying compared to boys. Overall, 54% of females and 44% of boys have encountered cyberbullying. Those subjected to this behaviour often experience health issues as a consequence. 64% of cyberbullying victims report increased rates of teenage melancholy, apprehension, low self-esteem, emotional distress, mental health concerns, and behavioural challenges.

After the introduction section, section 2 presents literature review, followed by a review of prior research on threat detection in social networks. In section 3, we provide a Methodology that includes qualitative data, such as text analysis techniques used in previous studies. In Section 4, we present results and discussion. Next, in section 5 presents the conclusion at the end.

1.1 Background

Cyberbullying lacks a singular definition and has been explored from various perspectives in the literature, resulting in multiple proposed meanings. It is a form of harassment facilitated by Information and Communication Technologies (ICT), encompassing text-based data, messaging apps, and various social media platforms [1]. Another definition characterizes cyberbullying as “an aggressive, intentional act conducted by an individual or group, using electronic means of contact, repeatedly and over time, targeting a victim who cannot easily defend themselves.”

2. Literature Review

The literature review section will offer a comprehensive overview of the existing research on cyberbullying detection through natural language processing (NLP) techniques. It will examine various studies that have applied NLP methodologies to address the challenge of identifying hate speech. These studies have employed a range of approaches, including machine learning algorithms and keyword-based strategies. Some researchers have prioritized the development of annotated datasets specifically for cyberbullying detection, while others have investigated

the effectiveness of word embedding and language models in capturing the underlying meaning of hateful content. This review will synthesize these different methodologies, highlighting their contributions and limitations in the field.

Primary goal of the paper[13] for detection of cyberbullying in Hindi-English code-mixed language on social media using NLP and Machine learning algorithms.

Different machine learning models like Naïve Bayes, Random Forest, SVM and Logistic Regression are used to identify toxic comments in Assamese language [11].

Some authors used hybrid approach for cyberbullying comments written in Tamil-English, Bengali-English and Kannada-English and got accuracy of 88.1%.[17].

One paper examines "transliterated English and Sinhala code-mixed data" sourced from Facebook, with a strong emphasis on meticulous annotation to ensure the accuracy of the dataset. Different models ranging from traditional conventional to advanced architectural models like CNN, RNN were used and got accuracy of 82% [19].

Some papers also provide survey and discuss about how effectively neural network performed well compare to machine learning algorithms [1][3].

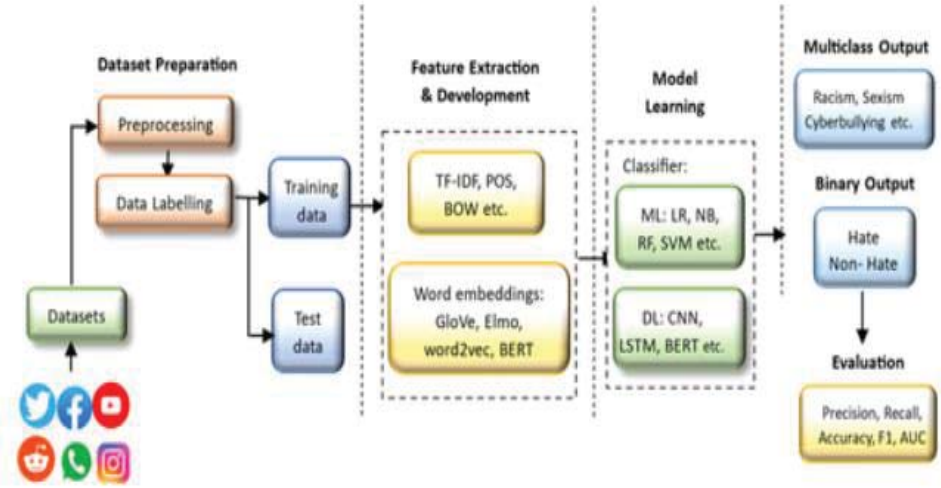
3.Research Methodology

This paper focuses on solutions for detecting cyberbullying across various social media platforms such as WhatsApp, Twitter, and YouTube. The subsequent functional block diagram outlines the two primary components of the cyberbullying detection framework:

1.Natural Language Processing

2. Machine Learning

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar



After pre-processing, the dataset is split into training and testing sets. The next step involves addressing two key aspects of text selection:

1. Count Vectorizer
2. Term frequency- Inverse frequency

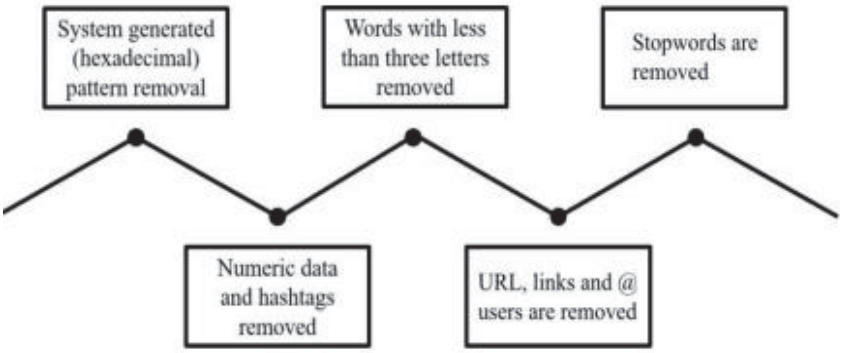
A. Data Extraction

The process of converting semi-structured or unstructured data into structured data is known as data extraction. Stated differently, this procedure makes it possible to transform unstructured or semistructured data into structured data. Meaningful insights from structured data can be used in reporting and analytics.

B. Data Cleaning

Cleaning the data is necessary before running it through several ML models. as seen in the figure below. Since these procedures don't help with the categorization phase, they must be eliminated from the data.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar



When raw data is imported from social media platforms, it includes various characters and encodings. At this stage, punctuation marks, special symbols, hashtags, retweet indicators, numbers, hex codes, and URLs are removed, as they do not impact the sentence's meaning. Words with fewer than three letters are also excluded. Additionally, sentences are converted to lowercase to avoid repetition.

C. Preprocessing Techniques

After data cleansing, we employ techniques from natural language processing (NLP) because Machine learning algorithms are unable to directly interpret or analyze raw text on their own and comprehend its meaning. Instead, we utilize pre-processing methods like tokenization, lemmatization, and vectorization to convert sentences into a structured format that is more interpretable for the algorithms.

D. Splitting data

The datasets are divided into training and testing sets . To ensure the system's applicability in real-time scenarios, The datasets designated for testing purposes are sourced from platforms through text mining. Both sets undergo various machine learning models and preprocessing techniques.

E. Feature Selection

After dividing the data, our focus shifts to extracting the key textual features. This method helps assess the accuracy of the vector representations created. It effectively captures semantic similarities between closely related words and those with varying degrees of similarity. Techniques such as Count Vectorization and TF-IDF are utilized to select features during this phase.

F. Machine Learning Algorithms

Throughout the stages of preprocessing and feature selection, several machine learning models were evaluated to compare their functionalities. The study utilized the following classifiers:

1. Support Vector Machine (SVM)

This classification algorithm works by fitting data to find the most suitable hyperplane that divides it into different classes. Once the hyperplane is established, the classifier uses features to predict the class of new data points.

The objective of this classification technique is to model the data and determine the optimal hyperplane that effectively separates the data into different classes. Once the hyperplane is established, the classifier utilizes specific properties to predict the class of new instances. The margin, representing the separation error, measures the distance between the two decision boundaries of the hyperplane. A larger margin typically results in fewer misclassifications.

2. K-nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a simple and intuitive text classification technique. It classifies new data points by comparing them to a set of known data points and selecting the class of the closest ones, using a similarity measure to determine the proximity. This method classifies a new sample based on how far away it is from its neighbour. As a result,

it looks for the training set's K-nearest neighbours and adds an item to the class whose K-nearest neighbours are the most common.

A. Logistic Regression

Logistic regression is a supervised machine learning technique used to predict categorical outcomes based on a set of independent variables. This approach uses a statistical method to model the relationship between the inputs and the categorical dependent variable, allowing us to make predictions about new data based on patterns identified in the training data set. Logistic regression is used to categorize data into two possible classes, such as 0 or 1, Yes or No, or True or False. Rather than assigning a definite class, it calculates the probability that a given data point falls into class '1', indicating the likelihood of belonging to that category.

B. Random Forest Classifier

The Random Forest classifier is composed of multiple decision trees, each providing a class prediction. The overall prediction is based on the majority vote from all the trees. While some trees may make errors, the majority are likely to provide correct predictions, which helps improve the model's accuracy.

5. Classifier

Among the ensemble boosting algorithms is Adaboost. Boosting algorithms use the errors of previous, weaker models to try and create a stronger learner or model. It makes an effort to lower errors that result from its inability to recognize patterns in the data. The AdaBoost algorithm builds a group by sequentially adding one-level decision trees, called weak learners, to the model.

C. Multinomial NB Classifier

Natural language processing (NLP) is the key area for the probabilistic machine learning method known as the Multinomial NB classifier. The system classifies text, like newspaper articles or pieces of mail, using Bayes' theorem. It calculates the probability of each possible

tag for a given sample and assigns the tag with the highest probability. The system is based on the assumption that each feature being classified is independent of the others.

G. Evaluation Phase

Machine learning packages written in Python are used to implement the bullying detection techniques. To assess the performance of a classifier, we calculate True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP). These four values are used to create a confusion matrix. The classifier's effectiveness is then evaluated using various performance metrics. In text classification, several common performance measures are considered, including the following metrics:

Precision:

Precision, also known as the positive predictive value, represents the percentage of correctly identified positive instances among all the predicted positives.

Precision=TP/TP+FP

Recall:

Recall, also referred to as sensitivity, is the percentage of actual positive instances that are correctly identified as positive by the classifier.

Recall= TP/TP+FN

F-Measure:

The F-Measure, also known as the F1 Score, is the harmonic mean of precision and recall. It balances precision and recall by giving them equal weight in its calculation.

Measure=2*precision*recall/precision + recall

Accuracy:

Accuracy is the proportion of correctly identified instances, including both true positives and true negatives, among the total number of instances.

Accuracy= correct Predictions / Total Predictions

4. Result and Discussion

The following Table depicts languages, techniques, level of LID and performance is shown as below.

Sr	Year	Reference	Languages Used	Techniques	Performance
1	2016	[13]	Hindi-English-Gujarati, Hindi-English	NB	Accuracy-97%
2	2016	[15]	Spanish-English	BLSTM-CNN	F1 score-95%
3	2017	[25]	Kannada-English	MNB, SVM, RF, LR	Accuracy-95%, Precision-96%, Recall-95%
4	2017	[22]	Konkani-English	SVM, RF	Accuracy-94%
5	2017	[26]	Tamil-English, Malayalam-English	SVM	Accuracy-95.47%, F1-94.78%
6	2018	[5]	Hindi-English	RF, SVM, KNN, LR	Accuracy-67%
7	2018	[8]	Spanish-English, Turkish-German, Dutch-English	Dictionary-based	F1-96%

8	2018	[12]	Telugu-English	NV, RF	Accuracy-91%
9	2019	[10]	Bengali-English	LSTM, SVM	Accuracy-92%
10	2019	[18]	Spanish-French-English	CRF	Accuracy-98%, recall-96%, precision-95%, F1-95%
11	2020	[16]	Gujarati-Hindi-English	NB-KNN-LR-RF	Accuracy-92%
12	2020	[24]	Hindi-English	BLSTM-	F1-94%
13	2021	[20]	Hindi-English	Rule-based	F1-88%
14	2022	[23]	Hindi-English	Decision Tree	F1-97%
15	2017	[14]	Arabic-English	SVM	95%
22	2024	[11]	Assamese-English	SVM	94%
23	2022	[17]	Bengali-English, Tamil-English, Kannada-English	Hybrid	88.1%
24	2024	[19]	Sinhala-English	BERT	82%
25	2024	[1]	Arabic, Hindi-English	BERT, RF, LR	95%
26	2024	[3]	English	RF	94%

SVM was the most commonly used method by researchers for language identification tasks. Its ability to build effective classifier models and deliver strong performance makes it a popular choice [26]. According to the selected research, SVM has shown significant effectiveness. Veena et al. [26] achieved 93% accuracy for word-level Malayalam-English and 95% accuracy for Tamil-English code-mixed language identification using a linear kernel SVM classifier. Several Machine learning techniques were combined with Word2Vec embedding for Hindi-English by Chaitanya et al. [5]. Their investigations showed that the SVM with the Skip-gram had the greatest accuracy, at 67.24%. One of the SVM variations, the Support Vector Classifier (SVC), was implemented by Kazi et al. [16]. The outcome demonstrated that the SVM using N-gram features and an RBF kernel had the highest accuracy, at 92%.

In Phadte and Wagh's study [22], CRF outperformed SVM and Random Forest approaches by 94% in accuracy. Gundapu and Mamidi [12] evaluated four machine learning techniques—Naïve Bayes, Random Forest, Hidden Markov Model (HMM), and CRF—and found that CRF achieved the highest accuracy at 91.28%. Additionally, CRF demonstrated its effectiveness in detecting multilingual code-mixed data with 98% accuracy and a 95% F1 score in Mishra and Sharma's study [18].

Gupta et al. [13] used a supervised learning approach combined with the edit distance method. Their findings revealed that integrating the N-gram Markov model with Naïve Bayes yielded impressive results, particularly in identifying languages based on misspelled words.

Muthuthanthri & Smith [19] employed BERT algorithm and achieved 82% accuracy. Amiruzzaman et al.[1] achieved strong accuracy with 94% using Random Forest algorithm in Arabic dataset while got 60% accuracy in Hinglish Dataset using Logistic Regression.

5. Conclusion

This thorough literature review presents the current state of research on code-mixed languages and offers a framework for future investigation. It encompasses forty primary studies published from 2016 to 2023.

The results indicated that the multichannel CNN combined with BLSTM and CRF achieved remarkable performance in addressing code-mixed language issues in several neural network-based studies. For non-neural network methods, using SVM and CRF is recommended. Additionally, the transformer-based approach has proven to be one of the most dependable techniques for handling code-mixed languages due to its exceptional performance.

Acknowledgement

"I would like to extend my heartfelt thanks to my advisor, Dr. Bhadresh Pandya, for his invaluable guidance and unwavering support throughout this research. I also wish to express my gratitude to my colleague for their assistance in data collection."

Conflict of Interest:

The authors would like to confirm that there is no conflict of interests associated with this publication and there is no financial fund for this work that can affect the research outcomes.

References

[1] Amiruzzaman, A., Rahman, A., Farjana, A., & Chowdhury, H. R. (2024). Multilingual cyberbullying classification for social platforms. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)* (pp. 904–909). Dhaka, Bangladesh. https://doi.org/10.1109/ICEEICT62016.2024.-10534579

[2] Azeez, N. A., & Misra, S. (2023). Identification and detection of cyberbullying on Facebook using machine learning algorithms. *Journal of Cases on Information Technology, 23*(4). https://doi.org/10.4018/JCIT.296254

[3] Balaji, P. G., Katariya, P. P., Sruthi, S., & Venugopalan, M. (2024). Cyberbullying detection on multiclass data using machine learning and a hybrid CNN-BiLSTM architecture. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–6). Chikkaballapur, India. https://doi.org/10.1109/ICKECS61492.2024.10616957

[4] Balakrisnan, V., & Kaity, M. (2023). Cyberbullying detection and machine learning: A systematic literature review. *Artificial Intelligence Review, 56*(Suppl 1), 1375–1416. https://doi.org/10.1007/s10462-023-10553-w

[5] Chaitanya, I., Madapakula, I., Gupta, S. K., & Thara, S. (2018). Word-level language identification in code-mixed data using word embedding methods for Indian languages. In *Proceedings of the International Conference on Advances in Computing, Communications, and Informatics (ICACCI)* (pp. 1137–1141). Bangalore, India, September. https://doi.org/10.1109/ICACCI.2018.8554501

[6] Chakraborty, P., & Seddiqui, Md. H. (2019). Threat and abusive language detection on social media in Bengali language. In *Proceedings of the 1st International Conference on Advances in Science, Engineering, and Robotics Technology. *

[7] Chu, C. C.-F., So, R., Li, S. S.-W., Kwong, E. K.-L., & Chiu, C.-H. (2023). A framework for early detection of cyberbullying in Chinese-English code-mixed social media text using natural language processing and machine learning. In *2023 5th International Conference on Natural Language Processing (ICNLP)* (pp. 298–302). Guangzhou, China. https://doi.org/10.1109/ICNLP58431.2023.00061

[8] Claeser, D., Felske, D., & Kent, S. (2018). Token-level code-switching detection using Wikipedia as a lexical resource. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology* (pp. 192–198). Cham, Switzerland: Springer.

[9] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology, 20*(2), Article 10, March. 22 pages. https://doi.org/10.1145/3377323

[10] Das, S. D., Mandal, S., & Das, D. (2019). Language identification of Bengali-English code-mixed data using character and phonetic-based LSTM models. In *Proceedings of the 11th Forum on Information Retrieval Evaluation (FIRE)* (pp. 60–64). Kolkata, India, December. https://doi.org/10.1145/3368567.3368578

[11] Dutta, S., Neog, M., & Baruah, N. (2024). Assamese toxic comment detection on social media using machine learning methods. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)* (pp. 1–8). Vellore, India. https://doi.org/10.1109/ic-ETITE58242.2024.10493331

[12] Gundapu, S., & Mamidi, R. (2018). Word-level language identification in English-Telugu code-mixed data. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, Hong Kong.

[13] Gupta, B., Bhatt, G., & Mittal, A. (2016). Language identification and disambiguation in Indian mixed-script. In *Proceedings of the International Conference on Distributed Computing and Internet Technology (ICDCIT)* (Vol. 9581, pp. 113–121). Cham: Springer. https://doi.org/10.1007/978-3-319-28034-9_14

[14] Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Saint Joseph University*, Lebanon; *Telecom ParisTech*, France. November.

[15] Jaech, A., Mulcaire, G., Ostendorf, M., & Smith, N. A. (2016). A neural model for language identification in code-switched tweets. In *Proceedings of the 2nd Workshop on Computational Approaches to Code Switching* (pp. 60–64). Austin, TX, USA.

[16] Kazi, M., Mehta, H., & Bharti, S. (2020). Sentence-level language identification in Gujarati-Hindi code-mixed scripts. In *Proceedings of the IEEE International Symposium on Sustainable Energy, Signal Processing, and Cyber Security (iSSSC)* (pp. 1–6). Gunupur, Odisha, India, December. https://doi.org/10.1109/iSSSC50941.-2020.9358837

[17] Mathur, K., Mehta, K. N., Shivakumar, K., & D, U. (2022). Detection of cyberbullying on social media code mixed data. In *2022 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (pp. 16–23). Zurich, Switzerland. https://doi.org/10.1109/CCEM57073.2022.00011

[18] Mishra, A., & Sharma, Y. (2019). Language identification and context-based analysis of code-switching behaviors in social media discussions. In *Proceedings of the IEEE International Conference on Big Data (BigData)* (pp. 5951–5956). Los Angeles, CA, USA, December. https://doi.org/10.1109/BigData47090.2019.9006032

[19] Muthuthanthri, M., & Smith, R. I. (2024). Hate speech detection for transliterated English and Sinhala code-mixed data. In *2024 4th International Conference on Advanced Research in Computing (ICARC)* (pp. 155–160). Belihuloya, Sri Lanka. https://doi.org/10.1109/ICARC61713.2024.10499768

[20] Nguyen, L., Bryant, C., Kidwai, S., & Biberauer, T. (2021). Automatic language identification in code-switched Hindi-English social media text. *Journal of Open Humanities Data, 7*, 1–13. June.

[21] Nafis, N., Kanojia, D., Saini, N., & Murthy, R. (2023). Towards safer communities: Detecting aggression and offensive language in code-mixed tweets to combat cyberbullying. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)* (pp. 29–41). Association for Computational Linguistics. July 13, 2023.

[22] Phadte, A., & Wagh, R. (2017). Word-level language identification system for Konkani-English code-mixed social media text (CMST). In *Proceedings of the 10th Annual ACM India Compute Conference (ZZZ-Compute)* (pp. 103–107). Bhopal, India. https://doi.org/10.1145/3140107.3140132

[23] Phadtare, C., Rajpara, K., & Shah, K. (2022). Cyber-bullying detection in Hinglish languages using machine learning. *International Journal of Engineering Research & Technology, 11*(5), May.

[24] Shekhar, S., Sharma, D. K., & Sufyan Beg, M. M. (2020). An effective bi-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and Roman Hindi. *Computación y Sistemas, 24*(4), December. https://doi.org/10.13053/cys-24-4-3151

[25] Sowmya Lakshmi, B. S., & Shambhavi, B. R. (2017). An automatic language identification system for code-mixed English-Kannada social media text. In *Proceedings of the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)* (pp. 1–5). Bengaluru, India. https://doi.org/10.1109/CSITSS.2017.8447784

[26] Veena, P. V., Kumar, M. A., & Soman, K. P. (2017). An effective way of word-level language identification for code-mixed Facebook comments using word embedding via character embedding. In *Proceedings of the International Conference on Advances in Computing, Communications, and Informatics (ICACCI)* (pp. 1552–1556). Udupi, India, September.