Multilingual Cyber-Bullying Detection System

[1]Prof. Prajakta A. Kuchewar, [2]Pranay R. Nagrale, [3]Aditya A. Meshram, [4]Sugat A. Moon, [5]Ritesh N. Nagpure

[1]Assistant Professor, [2,3,4,5]B.Tech Student
Department of Computer Science & Engineering
K.D.K. College of Engineering, Nagpur, India.

***Abstract*- Cyberbullying may be a genuine issue that influences numerous individuals online, particularly on social media stages. Cyberbullying can happen totally different dialects, and it is vital to identify and avoid it in a multilingual setting. In this paper, we propose a Multilingual Cyberbullying Detection System (MCDS) that can recognize cyberbullying in five languages: English, Hindi (Hindi and Hindi-English code-mixed), Marathi, Bengali and Tamil. We utilize a combination of Natural Language Processing and Machine Learning methods to classify content messages as cyberbullying or not. We evaluate our system on several datasets and achieve high accuracy and F1-score and Recall. Our System can be amplified to other languages and domains, and can help to protect clients from online harassment and cyberbullying or abuse. The results of our experiments have shown an accuracy up-to 95% and F1-score up-to 94%.**

***Keywords*-Cyberbullying, Machine Learning, Multilingual Cyberbullying Detection for Indian languages.**

INTRODUCTION

With the advent of net and technology, social media has emerged as a chief a part of our lifestyles. It enables us preserve in touch with each other with the use of various packages with just a few taps and/or swipes. it's far a steady source of entertainment. humans have started out feeling greater sociable regardless of their current situation, despite the fact that they may be at home or at paintings. With our smartphones and drugs the social media structures are without problems accessible, there has been a upward thrust in the range of users over the past few years. the global virtual report created by means of Dave Chaffey in 2018 shows the following records associated with internet person – there are around 4.021 billion net customers, three.196 billion social media users and 5.one hundred thirtyfive billion cellular cellphone users. But, social media has its own problems and demanding situations. as an instance, social media may additionally incorporate a variety of antisocial behaviour, inclusive of cyberbullying, cyber stalking, and cyber harassment. those behaviours have now turn out to be element our lives and are not best bounded to juveniles, however any character may be a sufferer of it.

1.1 Cyberbullying

Cyberbullying is a shape of online harassment that includes sending or posting harmful or malicious messages, photographs, or movies to or approximately any other person. Cyberbullying can occur in any language, culture, or context, and it can have terrible outcomes for the nicely-being and protection of the goal.

1.2 Effects of Cyber-bullying

Cyberbullying couldhave real poor influences at the mental and emotional properly-being of the sufferer. among the common affects of cyberbullying incorporate:

- **Low self-esteem:** Cyberbullying can make victims feel inferior, embarrassed**,** or unworthy, which can lead to low self-esteem and self-esteem.
- **Depression and Anxiety:** Cyberbullying can make victims feel sad, hopeless, and anxious, which can lead to depression and anxiety disorders.
- **Suicidal ideation:** Cyberbullying can make victims feel hopeless, helpless, and worthless, which can lead to suicidal thoughts and behaviours.
- **Poor Academic Performance:** Cyberbullying can affect the victim's concentration, motivation, and attendance, which can lead to poor academic performance and dropout rates.
- **Social Isolation:** Cyberbullying can cause victims to feel lonely, isolated, and disconnected from peers, which can lead to social withdrawal and avoidance.

Most existing strategies for identifying cyberbullying center on English content and frequently depend on manual explanation, predefined catchphrases, or opinion examination. These strategies have a few restrictions, such as moo exactness, tall taken a toll, and destitute generalizability to other dialects and spaces. Subsequently, there's a got to send

a modern approach competent of recognizing cyberbullying in different dialects, utilizing progressed strategies such as machine learning, characteristic dialect preparing, and profound learning. Such an approach would distinguish and avoid cyberbullying over distinctive stages, societies and settings, subsequently securing the online security and nobility of millions of clients.

In this paper, we propose a modern approach to detect multilingual cyberbullying which we call Multilingual Cyberbullying Discovery Framework (MCDF). MCDF may be a framework competent of recognizing cyberbullying in five dialects: English, Hindi, Marathi, Bengali and Tamil. It employments two strategies, to be specific machine learning and characteristic dialect handling, to classify the input information as bullying or not. We have created a model and conducted tests with it to distinguish cyberbullying on diverse datasets from different dialects. Our exploratory comes about appear that the machine learning-based strategy works well in all dialects. Our framework is versatile, strong, and versatile to diverse dialects and spaces, and can be utilized as an apparatus to combat cyberbullying.

I.RELATED WORK

In [17] Rohit Pawar, Rajeev R. Raje described a multilingual cyberbullying detection system. In which they have created a multilingual cyberbullying detection system for English, Hindi & Marathi languages. They used machine learning technique for it. In [18] Anita Saroj, Sukomal Pal have introduced a dataset for hate speech and offensive content detection in Indian language and Indian context and tested a number of text classification techniques to recognize hate speech and offensive posts to validate the dataset.

In [19] Adaikkan Kalaivani and Durairaj Thenmozhi presents the paper for the hate speech and offensive language identification for the HASOC 2021 subtask1 shared task in the Forum for Information Retrieval Evaluation (FIRE) 2021. They have experimented with different approaches such as machine learning techniques, pre-trained BERT-based models for English, Hindi and Marathi languages. In [20] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed S. Akhtar, Manish Shrivastava have created an annotated corpus of Hindi-English code-mixed text, consisting of tweet ids and the corresponding annotations. They have also presented the supervised system used for detection of Hate Speech in the code-mixed text. The corpus consists of 4575 code-mixed tweets annotated with hate speech and normal speech and got the best accuracy of 71.7% when all the features are incorporated in the feature vector using SVM as the classification system.

In [21] Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, and Rashmi Dhumal have proposed a semi-supervised approach in detecting cyberbullying based on the five features that can be used to define a cyberbullying post or message using the BERT model. While considering just one of the features which was sentimental features the BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models. In [22], the authors investigated the automated identification of posts on social media related to cyberbullying by considering two features BoW and TF-IDF. Four machine learning algorithms were used to identify bullying text and SVM for both BoW and TF-IDF. They presented a system to cyberbullying detection in Bengali language.

In [23], the authors have proposed an automatic cyberbullying system for Arabic language on social media. In [24], Özel et al. prepared a dataset from Instagram and Twitter messages written in Turkish and then applied machine learning techniques SVM, decision tree (C4.5), Naïve Bayes Multinomial, and k-Nearest Neighbours (kNN) classifiers to detect cyberbullying. They have applied information gain and chi-square feature selection methods to improve the accuracy of classifiers. They observed that when both words and emoticons in the text messages are considered as features, cyberbully detection improves. Among the classifiers, Naïve Bayes Multinomial was the most successful one in terms both classification accuracy and running time and they achieved 84% accuracy using it.

II. MACHINE LEARNING TECHIQUES

a) An Overview

Machine Learning (ML) based classification models are used for detecting cyberbullying. ML is mainly classified into three categories: i) Supervised Learning: in this approach, the mathematical model is built based on data which contains both set of inputs and desired outputs [1]; ii) Unsupervised Learning: in this approach, the model takes set of data as input, and try to find out structure (e.g., grouping or clustering of the data) [1]; and iii) Reinforcement Learning: this approach is concerned with taking suitable actions so as to maximize the reward in particular situation [1].

b) Performance Metrics

Following are the typical performance metrics that are used to evaluate and compare performance of various classifications techniques [16]. In this work, we have used these four metrics to assess the performance of our system:

- **Accuracy:** The accuracy score is a metric used to evaluate the performance of a machine learning model. It is defined as the ratio of the number of correct predictions to the total number of predictions made by the model. It's calculated as:

Accuracy = $(TP + TN) / T$

- **Precision:** Precision is a metric that measures the proportion of true positive predictions made by a model out of all the positive predictions it made. It is calculated as the ratio of the number of true positives to the sum of true positives and false positives. It's calculated as:

Precision = $TP / (TP + FP)$

- **Recall:** In machine learning, recall is a metric that measures the proportion of true positive predictions made by a model out of all the actual positive samples. It is calculated as the ratio of the number of true positives to the sum of true positives and false negatives. It is calculated as:

Recall = TP / (TP+FN)

- **F1-Score:** The F1 score is the harmonic mean of precision and recall, which gives more weight to low values. The F1 score ranges from 0 to 1, where 0 means the worst possible result and 1 means the perfect result. F1-Score is calculated by the following formula:

F1 = 2 * (Precision * Recall) / (Precision + Recall)

where:

TP: N$umber\ of\ True\ Positive$
TN: $Number\ of\ True\ Negavtive$
FP: $Number\ of\ False\ Positive$
FN: $Number\ of\ False\ Negavtive$
T: $Total\ number\ of$ sentences

III. PROPOSED SYSTEM

The main aim of this paper is to detect cyberbullying behaviour in English, Hindi (Hindi, Hindi-English code mixed), Marathi, Bengali & Tamil texts appearing on different online forums. Our system employs principles of ML, and NLP and thus, we have followed following steps for training and testing the ML models. We created out ML model using python's ML framework i.e., scikit-learn.

A. Data Collection: We have collected labelled dataset from different resources. We collected Bengali hate speech dataset from Kaggle, Marathi, English, Hindi, Tamil dataset from HASOC, Git hub, You tube comments, Facebook comments and Twitter comments available on different sources. The Bengali dataset contains 30,000 rows, combined English dataset contains 36,811 rows, Hindi dataset is having 19,416 rows, Marathi dataset contains 18,315 rows, Hindi-English code-mixed dataset is having 18,885 rows and Tamil dataset contains 5,503 rows. The dataset is having two columns, one contains texts and other is having labels for the text i.e. Cyberbullying or non-cyberbullying.

B. Data Integration: After data collection we have merged the dataset of different languages to form a single multilingual dataset.

C. Data Preprocessing: In data preprocessing we have lower-cased the text, removed all the punctuations from text, removed the stop-words from each text (stop-words like I, is, are, you, me, for, did, do) of each language accordingly also stemming of words is also done with the help of nltk library. Stemming is a natural language processing technique that is used to reduce words to their base form, also known as the root form. The process of stemming is used to normalize text and make it easier to process. It is an important step in text pre-processing, and it is commonly used in information retrieval and text mining applications [5][6]. Stemming algorithms are used to produce morphological variants of a root/base word. A stemming algorithm reduces the words "chocolates", "chocolatey", and "choco" to the root word, "chocolate" and "retrieval", "retrieved", "retrieves" reduce to the stem "retrieve" [7].

D. Feature Extraction: ML algorithms are unable to learn from raw texts. Therefore, feature extraction is required to train the classifier models. In this step we have used various feature embedding techniques like TFIDF, Count Vectorizer, Fasttext embeddings, and Bert Tokenizer. TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic that is used to reflect how important a word is to a document, in a collection or corpus (paragraph). It is often used as a weighing factor in searches of information retrieval, text mining, and user modelling [2]. In simple terms, it is a measure of how frequently a word appears in a document, weighted by the number of documents in which it

appears. The more frequently a word appears in a document, the higher its TF-IDF score. Conversely, the less frequently a word appears in a document, the lower its TF-IDF score [3]. FastText is an open-source, free, lightweight library developed by Facebook's AI Research (FAIR) team. It is used for text classification**,** word representation, and text similarity computation**.** FastText is an extension of the word2vec model that provides embedding to the character n-grams. It is designed to efficiently handle large amounts of text data and provides tools for text classification, word representation, and text similarity computation [4]. Count Vectorizer is a tool used to vectorize text data, meaning that it will convert text into numerical data that can be used in machine learning algorithms. It is a part of the scikit-learn library in Python. BERT is a natural language processing model that stands for Bidirectional Encoder Representations from Transformers**.** It was developed by Google researchers in 2018 and has achieved state-of-the-art results on many language tasks, such as text classification, question answering, sentiment analysis, and more. BERT is based on the Transformer architecture, which uses attention mechanisms to learn the contextual relations between words in a text. Unlike previous models, BERT is bidirectional, meaning it can process both the left and right context of each word. This allows BERT to capture the full meaning of a sentence and handle complex linguistic phenomena, such as polysemy and coreference.

E. Training & Testing: We have trained and tested various models using the cyberbullying dataset. For each embedding or feature extraction technique we have tested different models. We have used various ML models like Logistic regression, Random Forest Classifier, Decision Tree Classifier, Extra Trees Classifier, AdaBoost Classifier, XG Boost Classifier, Naive Bayes Classifier. Also Used Bert Transformer Model. Did the hyperparameter tunning to get more accurate results.

IV. RESULTS

Following are the results of the tested ML models and their performance.

The first table shows the overall performance of different models using different feature extraction technique. The Count Vectorizer technique used below is having max features of 3,000, TFIDF technique also implemented with max features 3,000. The FastText technique is implemented where dimension parameter is 1,500 and the results are obtained. BERT model is having loss of 44.28% due to limited computation resources it was not possible for us to lower the loss. Out of all following tested models the XG Boost, Ada Boost and Extra Trees Classifier models have given the best results among all models.

The second table shows the performance of each of three models for each language. The second table shows the performance of models for completely unknown dataset i.e. test dataset of each language individually. The test dataset of English language consists of 3,683 rows, Hindi test dataset contain 1,943 rows, Hindi-English code-mixed test dataset contain 1890 rows. Marathi, Bengali and Tamil language test dataset contain 1,833, 3,001 and 552 rows respectively.

Table 1: Overall performance of Machine learning algorithms using different feature extraction techniques

Embedding	Classifier	Accuracy	Precision	F1 score	Recall
Count Vectorizer	Gaussian NB	71.86	70.08	72.20	74.45
	Multinomial NB	72.56	74.10	70.80	67.78
	Logistic Regression	77.86	84.04	75.03	67.76
	Decision Tree	64.17	99.67	42.59	27.08
	Random Forest	77.65	80.20	76.05	72.31
	AdaBoost	71.84	89.71	62.67	48.16
	Extra Trees	77.72	80.65	76.00	71.85
	XG Boost	76.30	89.65	70.78	58.47
TFIDF	Gaussian NB	73.54	77.42	72.27	70.24
	Multinomial NB	74.20	77.84	71.61	66.31
	Logistic Regression	78.50	84.01	76.01	69.40
	Decision Tree	64.16	99.67	42.58	27.07
	Random Forest	78.09	81.56	76.22	71.54
	AdaBoost	72.09	90.09	63.03	48.47
	Extra Trees	78.26	81.26	76.45	71.88

	XG Boost	77.40	88.51	72.93	62.01
FastText	Gaussian NB	92.39	96.06	91.95	88.18
	Logistic Regression	94.56	95.01	94.45	93.90
	Decision Tree	91.79	91.66	91.67	91.68
	Random Forest	94.30	94.42	94.20	93.99
	AdaBoost	94.64	94.67	94.55	94.44
	Extra Trees	94.31	94.22	94.22	94.23
	XG Boost	95.01	94.90	94.95	94.99
Bert Tokenizer	BERT Sequence Classifier	78.99	79.17	78.30	77.46

Table 2: Performance of top 3 best model for each individual language

Model	**Language**	**Accuracy**	**Precision**	**F1 score**	**Recall**
XG Boost Classifier	English	85.87	90.47	89.31	88.19
	Hindi	86.30	86.34	85.69	85.05
	Hindi-English code mixed	80.46	83.78	80.67	77.78
	Marathi	91.75	91.21	89.65	88.14
	Bengali	84.86	75.73	77.81	80.00
	Tamil	73.13	73.19	48.96	36.78
Ada Boost Classifier	English	85.82	90.36	89.28	88.23
	Hindi	85.78	86.35	85.04	83.77
	Hindi-English code mixed	80.51	83.65	80.77	78.08
	Marathi	92.03	91.50	90.00	88.54
	Bengali	84.63	74.22	78.01	82.21
	Tamil	71.68	69.89	45.45	33.67
Extra Trees Classifier	English	84.84	89.38	88.57	87.78
	Hindi	86.40	86.77	85.74	84.73
	Hindi-English code mixed	80.46	84.30	80.52	77.07
	Marathi	92.41	92.16	90.45	88.81
	Bengali	85.36	76.22	78.63	81.20
	Tamil	72.77	72.16	48.27	36.26

V. CONCLUSION AND FUTURE WORK

This paper has provided a multilingual cyberbullying detection approach for detecting cyberbullying in messages, tweets, emails and social media comments for five languages. Results of our experiments shows that XG Boost, AdaBoost and Extra Trees model outperforms all other algorithms on these datasets. Results of our study show that our systems perform well across five languages and different domains and hence, it can be used to detect cyberbullying. Many future extensions of our works are possible. These are as follows:

- We would like to validate this approach on very large datasets.
- We would like to add more dataset of Tamil language to improve the performance of this model for Tamil language.
- We would like to add more languages other than above five mentioned languages.
- We would like to provide language inputs and detect sentiment, and sarcasm associated with it.
- Explore other approaches such as Natural Language Process (NLP) and, using translator and compare performance of different approaches.
- We would like to try it with Deep Learning approach. We want to use LSTM and other technique in future.
- The BERT model can achieve more accurate results if provided with a large dataset. We can try to achieve even better results. We would like to improve the performance and lower the loss of Bert model using high computation power as well try to explore other combinations of feature extraction techniques.

REFERENCES:

[1] Machine Learning Basics. Retrieved from https://www.edureka.co/blog/what-is-machine-learning/, 2/16/19.
[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[3] https://medium.com/analytics-vidhya/understanding-tf-idf-in-nlp-4a28eebdee6a
[4] https://fasttext.cc/
[5] https://www.tutorialspoint.com/natural_language_toolkit/natural_language_toolkit_stemming_lemmatization.html
[6] https://www.geeksforgeeks.org/introduction-to-stemming/
[7] https://www.geeksforgeeks.org/python-stemming-words-with-nltk/
[8]https://hammer.purdue.edu/articles/thesis/MULTILINGUAL_CYBERBULLYING_DETECTION_SYSTEM/8035463
[9] https://scholarworks.iupui.edu/bitstream/handle/1805/24303/Pawar2019Multilingual.pdf?sequence=1
[10] https://link.springer.com/article/10.1007/s42979-022-01308-5
[11] https://ieeexplore.ieee.org/abstract/document/8833846/
[12] https://www.comparitech.com/internet-providers/cyberbullying-statistics/
[13] https://www.verywellhealth.com/cyberbullying-effects-and-what-to-do-5220584
[14] https://www.apa.org/topics/bullying/cyberbullying-online-social-media
[15] https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying
[16] https://www.security.org/resources/cyberbullying-facts-statistics/
[17] Multilingual Cyberbullying Detection System by Rohit Pawar, Rajeev R. Raje content (iupui.edu)
[18] An Indian Language Social Media Collection for Hate and Offensive Speech - ACL Anthology
[19] Multilingual Hate speech and Offensive language detection in English, Hindi, and Marathi languages (ceur-ws.org)
[20] A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection - ACL Anthology
[21] Cyber Bullying Detection on Social Media using Machine Learning | ITM Web of Conferences (itm-conferences.org)
[22] Cyberbullying Detection on Social Networks Using Machine Learning Approaches (researchgate.net)
[23] Using Machine Learning Algorithms for Automatic cyber BullyingDetection in Arabic Social Media
[24] S. Zel, E. Sara, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in turkish," in International Conference on Computer Science and Engineering, pp. 366–370, Oct 2017.