**ANALYSIS OF TWEETS FOR CYBERBULLYING DETECTION**

## **A PROJECT REPORT**

*for*

## **DATA MINING TECHNIQUES (SWE2009)**

*in*

## **M.Tech (Software Engineering) 5 Year Integrated Programme**

*By*

# **M SAI BHARGAV (20MIS0241)**

# **I SUNEEL KUMAR (20MIS0246)**

**Fall Semester, 2023**

*Under the Guidance of*

# **Prof. B. VALARMATHI**

Professor, SCORE



## **School of Computer Science Engineering and Information Systems**

November, 2023

**DECLARATION BY THE CANDIDATE**

We here by declare that the project report entitled **“ANALYSIS OF TWEETS FOR CYBERBULLYING DETECTION”** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide project work carried out by us under the guidance of **Prof. B. Valarmathi.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date: 21-11-2023 M Sai Bhargav – 20MIS0241

I Suneel Kumar – 20MIS0246



## **School of Computer Science Engineering and Information Systems [SCORE]**

**CERTIFICATE**

This is to certify that the project report entitled **“ANALYSIS OF TWEETS FOR CYBERBULLYING DETECTION”** submitted by **M Sai Bhargav(20MIS0241), I Suneel Kumar (20MIS0246)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide work carried out by them under my guidance.

## **Prof. B.Valarmathi GUIDE**

**Professor, SCORE**

**ANALYSIS OF TWEETS FOR CYBERBULLYING DETECTION**

I. Suneel Kumar, M. Sai Bhargav

Department of Software and Systems Engineering, VIT University, Vellore, Tamil Nadu, India

**ABSTRACT**

The existing algorithm was Random Forest and its accuracy was 94%. The dataset used for my project is Cyberbullying Detection Dataset from Twitter and Kaggle. The dataset has one independent variable and one dependent variable. This study examines the model accuracies XG Boost, Random Forest and Ensemble Approach (XGB + RF + LR) stood top with accuracies such as 85% surpassing accuracies of remaining models Gradient BOOST (84%), Logistic Regression (84%), Decision Tree (82%) and BERT (17%). This paper compares the existing model accuracies, Random Forest (95%), with the proposed model accuracies.

Cyberbullying has become a pervasive issue in online social platforms, posing serious threats to the well-being of individuals. This research focuses on developing a robust machine learning framework for the automated detection and classification of cyberbullying in Twitter data. The study employs a diverse set of features, including text-based, to enhance the accuracy and effectiveness of the classification model.

The methodology involves data collection from Twitter, preprocessing of tweets to handle noise and irrelevant information, and the extraction of relevant features to characterize cyberbullying instances. Various machine learning algorithms, such as Decision Tree, XG Boost, Random Forest, and Neural Networks, are explored to identify the most suitable model for cyberbullying classification.

Furthermore, the research investigates the impact of word embeddings and natural language processing techniques to capture the nuanced language often associated with cyberbullying. The model is trained on a labelled dataset comprising instances of both cyberbullying and non-cyberbullying tweets to ensure comprehensive learning and generalization.

The evaluation of the proposed model includes metrics such as precision, recall, F1 score, to assess its performance in distinguishing cyberbullying content. Additionally, the study explores the interpretability of the model's decisions, shedding light on the key features contributing to cyberbullying classification.

**Keywords** - machine learning (ML), Decision Tree, XG Boost, Random Forest, Neural Networks, precision, recall, F1 score, cyberbullying, Natural language processing.

1. **INTRODUCTION**

In the rapidly evolving landscape of online social platforms, the rise of cyberbullying has emerged as a significant societal concern, posing serious threats to individuals' mental well-being and online safety. As a pervasive form of harassment, cyberbullying involves the use of digital technologies, particularly social media platforms, to intentionally and repeatedly harm others. Among these platforms, Twitter stands out as a prominent medium where individuals express opinions, share information, and engage in conversations. Consequently, the need for effective tools to detect and mitigate cyberbullying on Twitter has become increasingly paramount.

Motivation:

This research aims to address the complex challenge of cyberbullying by leveraging advanced techniques in machine learning and natural language processing for the analysis of tweets. The objective is to develop a robust classification model capable of automatically identifying instances of cyberbullying within the vast and dynamic Twitter ecosystem. By doing so, this study contributes to the ongoing efforts to create safer online spaces, where users can freely express themselves without fear of harassment or intimidation.

The increasing prevalence of cyberbullying underscores the pressing need for effective and efficient methods to detect and mitigate its harmful effects. As online interactions continue to evolve, traditional methods of identifying and addressing cyberbullying prove inadequate in the face of the sheer volume and complexity of digital communication. Motivated by the imperative to create safer online environments, this research endeavours to explore innovative solutions for cyberbullying detection by harnessing the power of Twitter, one of the most widely used social media platforms.

Contribution:

This study contributes to the ongoing efforts to combat cyberbullying by presenting an in-depth analysis of tweets for the purpose of detection and prevention. Leveraging advanced natural language processing (NLP) techniques, machine learning algorithms, and sentiment analysis, our research aims to develop a robust framework capable of identifying cyberbullying instances with using high level deep learning algorithms. The proposed methodology not only considers the linguistic nuances of online communication but also adapts to the dynamic nature of social media discourse.

By delving into the rich data provided by dataset, we seek to uncover patterns, linguistic markers, and contextual cues that distinguish cyberbullying from ordinary discourse. The ultimate goal is to empower social media platforms, policymakers, and users with the tools necessary to foster a more inclusive and secure online environment.

The ubiquity of social media platforms has made them powerful tools for communication, connecting people across the globe. However, this connectivity has also exposed individuals to new forms of aggression and harassment, including cyberbullying. Unlike traditional forms of bullying, cyberbullying operates in the digital realm, making it more elusive and challenging to combat using conventional methods. Therefore, the application of cutting-edge technologies such as machine learning and deep learning becomes crucial in effectively addressing this issue.

1. **BACKGROUND**

The field of machine learning, particularly natural language processing (NLP) and classification algorithms, has shown promise in addressing the complexities of cyberbullying detection. These algorithms can analyse large volumes of textual data and learn patterns indicative of cyberbullying behaviour. By leveraging machine learning, it becomes possible to develop automated tools that augment traditional moderation efforts and provide a more proactive approach to combatting cyberbullying.

This research builds upon the existing body of knowledge by employing advanced machine learning techniques to analyse tweets for cyberbullying classification. By understanding the unique challenges posed by the Twitter platform, the study aims to contribute to the development of effective and efficient tools for identifying and mitigating cyberbullying in real-time. Through a nuanced analysis of tweet content and user interactions, the research seeks to advance our understanding of cyberbullying dynamics in the digital landscape and pave the way for more comprehensive solutions to ensure a safer online environment.

1. **LITERATURE SURVEY**

Shipra Anil Mathur et al. [1] proposed a method for applying NLP and ML to find cyberbullying messages on social media. The authors used datasets comprised of diverse comments, tweets, and posts on Twitter and Facebook using four machine learning algorithms: SVM, Naive Bayes, Decision Trees, and Random Forest. For performance analysis, they employed Bag-of-Words (BoW) and Term Frequency - Inverse Document Frequency (TF-IDF). According to the results, SVM performs better than other machine learning techniques, and TF-IDF feature has better accuracy than BoW. The article offers tips for spotting cyberbullying on social media.

Norah Alsunaidi et al. [2] Empirical studies have shown the effectiveness of machine learning techniques in detecting and classifying Content-Based Events (CB) in English-based content. However, existing experimental studies on Arabic-based CB detection remain limited, highlighting the need for researchers to focus on Arabic-speaking social media platforms. This paper highlights published empirical studies that utilize machine learning techniques for CB detection in Arabic text-based content.

Abdulsamad Al‑Marghilani [3] proposed the Cyberbullying studies which had faced a major problem due to the lack of standardized information. Data from social media platforms like Twitter is generated individually, making it difficult to compare and verify the generality of the methods. This was due to the fact that individual data cannot be related to one another.

Pradeep Kumar Roy and Fenish Umeshbhai Mali [4] has done research and focused on developing a model for detecting cyberbullying on Instagram by extracting captions, comments, and image content using feature selection and engineering techniques. The model used a Twitter dataset of 39,000 tweets, which were cleaned to remove duplicates. The goal was to create a unified representation of text and image, aiming to make social media a safer space for cyberbullying.

Nitin Kumar Singh et al. [5] conducted data pre-processing using TF-IDF, POS tagging, and trigrams, and used word embedding techniques. They used deep learning architectures like CNN-LSTM and CNN-Bi LSTM to detect cyberbullying in Twitter comments, with 92 percent accuracy. They also investigated the usefulness of self-attention models in detecting cyberbullying, and introduced a Shared-Private Multi-Task Learning framework for related tasks.

Gull Bano Anwar and Muhammad Waqas Anwar [6] has done research and focused on using Machine Learning models to detect cyberbullying using various techniques and models. The study uses the Twitter dataset, English and Roman Urdu datasets, and RNN with attention model for experimental purposes. The AdaBoost Algorithm achieves 90% accuracy, while the private Roman Urdu dataset detects 8 types of cyberbullying.

Fatma Elsafoury et al. [7] studied and analysed 43 relevant literature reviews on automated cyberbullying detection using search keywords in Google Scholar, IEEE Xplore, Science Direct, ACM Digital Library, and Wiley online databases. Key studies were identified, and papers citing these studies, especially those published after 2016, were reviewed.

Amirita Dewani et al. [8] had studied that the field of automatic cyberbullying detection is emerging due to the increasing impact of social media on users, with a pioneering and highly cited research on detecting cyberbullying in English textual data.

Vimala Balakrishnan et al. [9] had explored the link between psychological features like personality traits and cybercrimes like cyberbullying. Results shown that personality and sentiments improve cyberbullying detection, with extraversion, agreeableness, neuroticism, and psychopathy having greater impacts.

Hani Nurrahmi and Dade Nurjanah [10] used Natural Language Processing (NLP) and Text Mining techniques and have studied cyberbullying detection, using common sense reasoning and reflective user interaction to identify cyberbullying from formspring.me data and learn query terms.

Tanjim Mahmud et al. [11] had surveyed over 70 studies on automatic cyberbullying detection in low-resource languages like Bangla, Hindi, and Dravidian. It pinpointed gaps in prior research, such as vague cyberbullying definitions and data biases. Their study suggested enhancements, released a cyberbullying dataset in Chittagonian Bangla, proposed machine learning solutions, and leveraged the pre-trained Bangla BERT model. Ethical considerations were discussed, and the paper emphasized improvements over previous surveys, underlining the importance of AI-enhanced tools in research.

Zheng Lin Chia et al. [12] explored irony and sarcasm detection on Twitter using Machine Learning and Feature Engineering techniques. It reviewed and clarified the definitions of irony and sarcasm, comparing various classification methods in the first experiment. The second experiment analysed different data pre-processing methods. Their paper concluded by discussing the relationship between irony, sarcasm, and cyberbullying, noting a high similarity between them.

Jaideep Yadav et al. [13] addressed cyberbullying on social media, highlighting the challenge of detecting offensive content. Recent advances in deep neural network models, specifically Google's BERT, showed promise in cyberbullying detection. The proposed method employed the pre-trained BERT model with a single linear neural network layer as a classifier, showcasing improved performance compared to existing methods. Evaluation was conducted on two social media datasets, one small-sized and the other relatively larger.

Jason Wang et al. [14] During the COVID-19 pandemic, cyberbullying intensified, prompted their research into an automatic multiclass cyberbullying detection model. Unlike previous work, they aimed to classify cyberbullying based on specific attributes such as age, ethnicity, gender, and religion. Existing datasets suffered from class imbalances, so they introduced a framework for balanced data generation. They utilized a semi-supervised online Dynamic Query Expansion process to extract more natural data points from Twitter. They proposed Graph Convolutional Network (GCN) classifier, using tweet embeddings, outperformed baseline models on two dataset sizes.

Sara Khan et al. [15] addressed the rise of cyberbullying on social media, focusing on Urdu text comments on Twitter. Due to a lack of publicly available Urdu datasets, the study created one, classifying offensive comments into five categories. N-gram techniques at character and word levels were used, and supervised machine learning and Natural Language Processing (NLP) were applied to detect cyberbullying. Evaluation metrics, including precision, recall, accuracy, and F1 scores, were employed to analyse the performance of machine learning techniques. The research highlighted the need for attention to cyberbullying detection in non-English languages like Urdu.

Harish D et al. [16] prevalence of cyberbullying, employing supervised machine learning and neural networks for detection and prevention. Models, including 1-D Convolutional Neural Networks, SVM, LR, and a LR-SVM ensemble, were utilized. Feature extraction involved TF-IDF and N-grams, optimized with the SAG optimizer. Their focus was on recognizing cyberbullying text, covering themes like racism and profanity. Results showed that the Logistic Regression model with TF-IDF and SAG optimizer excelled in identifying cyberbullying in text data.

Ermira Idrizi and Mentor Hamiti [17] proposed a Graph Convolutional Neural Network (GCN) and pre-trained Google net for text and image recognition, along with a Mel-scale filter bank speech spectrogram and CNN model for audio post-classification. Results showed that using GCN, Google net, and audio post-processing with MFCCs achieved superior accuracy, reaching 96%. The unique approach of utilizing these techniques for text, image, and video input made it a valuable addition to cyberbullying detection literature.

Vaibhav Jain et al. [18] delved into the dual nature of smart gadgets, serving as companions in the digital age. While social networking fostered virtual connections, it also amplified the vulnerability of young people to online threats. Their study comprised three main tasks: explored cybercrime and cyberbullying, reviewed their forms, methods, effects, and recent research; collected and pre-processed 35,000+ tweets from Twitter; applied five machine learning algorithms for offensive and non-offensive tweet classification; and compared the algorithms based on performance metrics.

Thibaut Balet et al. [19] addressed the escalating concern of cyberbullying in society, emphasizing the use of machine learning algorithms to analyze and identify abusive language in messages. Six popular algorithms, including Logistic Regression, k-Nearest Neighbours, Random Forest, Classification and Regression Trees, Naive Bayes, and Support Vector Machine, were employed and evaluated on a Twitter dataset. Results indicated that Random Forest performed the best with a 94% accuracy. The study suggested the proposed method as an effective automated tool for detecting and preventing cyberbullying on platforms like Twitter.

Chetna Sharma et al. [20] They raised issue of cyberbullying on social media, noting a significant increase in reported cases. With 36.5 percent of people experiencing cyberbullying, double the numbers from 2007, there was a concerning trend. They focused on identifying cyberbullying at its origin in real-time, aiming to curb hurtful messages before they were uploaded. Utilizing Machine Learning and Natural Language Processing (NLP) enhanced the efficiency of cyberbullying detection, offering a potential solution to the escalating problem.

**Table 1:** Literature Survey

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No.** | **Paper Title &Year** | **Algorithms** | **Performance Measured** | **Dataset Size** | **Future Work** |
| 1 | Analysis of Tweets for Cyberbullying Detection (2023) | Random Forest Classifier, Gradient Boost Classifier, AdaBoost Classifier | Accuracy =0.94, Precision = 0.94, Recall = 0.94 | Size = 7.17MB | This lays the groundwork for investigations on the identification of cyberbullying on social media sites. It can be expanded in a number of areas. For example, the model can be finetuned to improve its performance in detecting subtle forms of cyberbullying, such as sarcasm and irony. |
| 2 | Arabic Cyberbullying Detection Using Machine Learning: State of the Art Survey | SVM | Accuracy = 0.95, Precision = 0.93, F1 – Score = 0.92 | Size = 584 KB | Attention needs to be directed toward experimenting with different pre-processing techniques, as per the reviewed literature |
| 3 | Artificial Intelligence‑Enabled Cyberbullying‑Free Online Social Networks in Smart Cities (2022) | SVM, LR, CNN, LSTM | Precision = 90.5, recall = 94.8, F1-score =91.35 | Size = 841 KB | The AICBF-ONS technique performs well enough to be applied to the development of techniques for outlier identification and data clustering in a large data setting. |
| 4 | Cyberbullying detection using deep transfer learning (2022) | VGG16, CNN, 2DCNN, InceptionV3, Random Forest | Precision = 0.87, Recall = 0.88, F1-Score = 0.87 | Size = 710 KB | The future scope of this research is always open to discussion as it has varied subproblems. The accuracy achieved by the proposed system was 89%, which can be improved by increasing the training sample size. Also, the other combinations of the models can opt, and an ensemble system will form to achieve better prediction accuracy. |
| 5 | Deep Learning based Methods for Cyberbullying Detection on social media (2022) | Naive Bayes, SVM, LR, XG Boost, GRU | Accuracy = 0.94, Precision = 0.94, Recall = 0.94, F1- Score = 0.94 | Size = 7.17 MB | using the Bi-LSTM, BiGRU and State-of-the-Art model BERT which can give better results on this dataset. |
| 6 | Textual Cyberbullying detection using Ensemble of Machine Learning models (2022) | Naïve Bayes, LR, DT, Random Forest, SVM | Precision = 0.96  Recall = 0.96  Accuracy = 0.96  F1 – Score = 0.96 | Size = 485 KB | In addition, we focus on additional datasets other than YouTube to further develop existing research models on these datasets. Deep learning and sequential models will also be taken into consideration utilizing various methodologies. Additionally, a hybrid deep-machine learning model will be investigated. |
| 7 | When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection (2021) | Machine Learning Models, Deep Learning Models | Accuracy = 0.88, Precision = 0.92, Recall = 0.92, F1- Score = 0.98 | Size = 7.17 MB | we consider the pretraining of BERT on slang-based text in addition to Wikipedia articles and the Books Corpus which may improve its performance on the task of cyberbullying detection even more. |
| 8 | Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data (2021) | Recurrent Neural Network (RNN), LSTM,  Bi LSTM,  CNN | Precision = 0.9,  Recall = 0.91,  F1 – Score = 0.9 | Size = 485 KB | The studies can focus on development of ensemble models to uncover harassing and hate speech patterns. Moreover, the incorporation of context-specific features and handling of morphological variations might produce better results. |
| 9 | Improving cyberbullying detection using Twitter users’ psychological features and machine learning (2020) | Naive Bayes, DT, Random Forest, SVM | Accuracy = 0.911, F1 – Score = 0.911 | Size = 584 KB | Contrary to some research in sexism detection, for instance, where the authors classified tweets into objectification, sexual harassment, discredit, etc., the current study did not distinguish between these sorts of cyberbullying. As a result, it would be intriguing to expand on and investigate if the suggested model can conduct precise classifications of cyberbullying. |
| 10 | Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility (2018) | SVM, KNN | Precision = 0.67, recall = 0.67, F1 - score =0.66 | Size = 562 KB | To enhance the outcomes of feature extraction and the precision of classification, we will employ a better POS Tagger tool. Adding more dataset would also help the classifier's accuracy. In order to spread awareness of the harm caused by cyberbullying, we would also create a notice system for those who engage in it. |
| 11 | Cyberbullying detection for low-resource languages and dialects: Review of the state of the art  (2023) | Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Multinomial Naive Bayes (MNB), K-Nearest Neighbour (KNN), Support Vector Machine with Linear function (LSVM), and Support Vector Machine with Radial Basis Function (RBF SVM). | Accuracy = 0.839, Precision = 0.8005, Recall = 0.9006, F1 score = 0.8476. | Size = 841 KB | Adopting a resource-constrained approach to detect cyberbullying in Bangla dialects. Initially, we used basic methods for classification, but we intend to implement more advanced techniques like bidirectional RNNs and competitive transformers. Our future strategy will leverage multilingual pretrained language models to enhance accuracy. |
| 12 | Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection  (2021) | SVM, JRip, RF, Naïve Bayes, KNN | F1-Score = 0.997 | Size = 7.17 MB | To explore diverse pre-processing methods for sarcasm and irony detection. Additionally, we plan to evaluate the impact of detecting irony and sarcasm on cyberbullying detection tasks |
| 13 | Cyberbullying Detection using Pre-Trained BERT Model  (2020) | BERT | Accuracy = 0.96, F1 – Score = 0.81 | Size = 7.17 MB | Using a single linear layer of neural network for classification which can be replaced by the deep neural network models like CNN and RNNs |
| 14 | A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection (2020) | BERT, SVM, MLP, XGB, KNN | Accuracy= 0.89, F1 – Score = 0.86 | Size = 710 KB | Leading in fine-grained cyberbullying detection with comprehensive experiments, matching or surpassing previous binary studies. Our SOS Net Graph Convolutional Network capitalizes on tweet semantics, outperforming traditional classifiers |
| 15 | Cyberbullying Detection in Urdu Language Using Machine Learning  (2022) | Random Forest, Logistic Regression, XG Boost | Accuracy = 0.75 Recall = 1 Precision = 0.66 F1 Score = 0.8 | Size = 584 KB | Implementation of an auto- correction feature and analysis of sentiments in the Urdu text |
| 16 | Automatic Detection of Cyberbullying on social media Using Machine Learning  (2023) | 1D-CNN, Logistic Regression, SVM | Accuracy = 0.91, Precision = 0.9, F1 Score = 0.91, Recall = 0.912 | Size = 485 KB | Focus on addressing limitations of the methods used in this study and exploring new techniques to enhance the effectiveness of the models used to detect cyberbullying. |
| 17 | Classification of cyberbullying messages using text, image and audio in social networks: a deep learning approach (2023) | CNN, KNN, SVM, LR | F1 Score = 0.81  Accuracy = 0.92 | Size = 710 KB | Improving the model’s performance by exploring other hyperparameters, such as the number of layers and the size of the hidden layer in the GCN. |
| 18 | Cyber-Bullying Detection in Social Media Platform using Machine Learning (2021) | Random Forest, KNN,  Decision Tree. | Accuracy = 0.927, Precision = 0.967, Recall = 0.923, F1 Score = 0.945 | Size = 485 KB | Natural Language Processing is a curious area of research among various AI researchers and practitioners, Hence, the opportunities in Natural Language Processing is not limited to time but needs to grab on. |
| 19 | Cyberbullying Detection on tweets from Twitter using Machine Learning Algorithms  (2023) | LR, KNN, RF, SVM, CART, Naive Bayes | Accuracy = 0.927, F1 - score = 0.928, Recall = 0.93, Precision = 0.93 | Size = 710 KB | To explore the potential of this model in other contexts and to develop more sophisticated techniques for identifying cyberbullying behaviours. |
| 20 | Cyber – Bullying Detection via Text Mining and Machine Learning (2021) | LR, SVM,  DT, RF, XGB. | F1 Score =0.606 | Size = 7.17 MB | Finding a possible area of improvement could be to make the model work as a cyber-bully detector for all environments without having to train it with a new dataset every time. |

The above table explains about the advantages and disadvantages of 20 papers which helps us to look into it more accurately and helps us to solve those uncovered areas from the papers.

1. **DATASET DESCRIPTION & SAMPLE DATA**

Link of the dataset used for the work:

<https://www.kaggle.com/code/sujithmandala/cyberbullying-detection-nlp/input>

<https://archive.ics.uci.edu/datasets?search=Bengali%20Hate%20Speech%20Detection%20Dataset>

**English Dataset:**

**Tweet\_Text:**

This column contains the actual content of the text, representing the messages or tweets collected for analysis.

The “tweet\_text” column is crucial for natural language processing (NLP) tasks as it holds the textual data on which the analysis, particularly cyberbullying detection, will be performed.

**Cyberbullying\_Type:**

This column serves as the label or target variable for the dataset.

It indicates the type or category of cyberbullying associated with each text entry. It is a categorical variable.

Depending on the nature of your dataset, the “cyberbullying type” column may include different classes or labels, such as “age”, “Religion”, “Gender”, “not\_cyberbullying”. Each entry in this column corresponds to the specific type of cyberbullying identified in the corresponding text entry.

**Table 2**: Sample Data of cyberbullying dataset of English

|  |  |
| --- | --- |
| **Tweet\_text** | **Cyberbullying\_type** |
| In other words, #katandandre, your food was capricious! #mkr | not\_cyberbullying |
| @ALIAQUINTA you’re a killer al if you want ufc to cut you all you need to do is tweet a joke about rape or gay people seems to work | gender |
| @obsurfer84 And the Muslims robbed the wealthy merchants. | Religion |
| @sethmeyers Your 80’s-High-School-Bully-getting-psyched-out is severely underrated. | Age |

**Arabic Dataset:**

**Text Column:**

This column contains textual data, likely representing messages, posts, or comments in Arabic. Each row in the "text" column contains a piece of text that is relevant to the task of cyberbullying detection. This text serves as input data for the machine learning model.

**Target Column:**

The "target" column includes labels or categories assigned to each corresponding text in the dataset. In this case, there are four attributes or classes in the "target" column, indicating different aspects of cyberbullying:

**Other:** This class might represent instances where the content does not fall into the specific categories of origin, gender, or religion. It could be a general or miscellaneous category.

**Origin:** This class could be assigned to instances where cyberbullying is related to a person's origin or nationality.

**Gender:** This class may indicate instances of cyberbullying based on gender-related factors.

**Religion:** This class could be assigned to instances where cyberbullying is related to religious aspects.

**Table 3**: Sample Data of cyberbullying dataset of Arabic

|  |  |
| --- | --- |
| **Tweet** | **Target** |
| ØµÙ„Ø§Ø© Ø§Ù„ÙØ¬Ø± Ø®ÙŠØ± Ù„Ùƒ Ù…Ù† ØªØ±Ø¯ÙŠØ¯ Ø¨ÙˆÙ„ Ø§Ù„Ø¨Ø¹ÙŠØ± ÙˆØ³Ø¨ÙŠ Ø§Ù„Ù†Ø³Ø§Ø¡ ÙˆØ§ØºØªØµØ§Ø¨ Ø·ÙÙ„Ø© Ù†Ø¸Ø§ÙØ© ÙˆÙ†Ø´Ø§Ø· ÙˆØ­ÙŠÙˆÙŠØ© #Ø¹Ù‚Ù„Ø§Ù†ÙŠÙˆÙ† | gender |
| ØµØ±Ø§Ø­Ø© Ù†ÙØ³ÙŠ Ø§Ø´ÙˆÙ ÙˆÙ„Ø§Ø¯ Ø§Ù„ÙˆØ³Ø®Ø© Ø§Ù„Ù„ÙŠ Ù‚Ø§Ù„ÙˆØ§ Ù…Ø¯Ø±Ø¨ Ø§Ø¬Ù†Ø¨ÙŠ Ù…Ù†Ùƒ Ù„Ù„Ù‡ Ø±Ø¨Ù†Ø§ ÙŠØ§Ø®Ø¯Ùƒ ÙŠâ€¦ @url | other |
| @user @user Ø§Ù†Ø§ Ø§ÙˆØ§ÙÙ‚Ùƒ Ø¨Ø®ØµÙˆØµ Ø§Ù„Ø³ÙˆØ±ÙŠÙŠÙ† Ùˆ Ø§Ù„Ø¹Ø±Ø§Ù‚ÙŠÙŠÙ† Ø§Ù…Ø§ Ø¨Ø®ØµÙˆØµ Ø§Ù„Ø³Ø¹ÙˆØ¯ÙŠ Ù…Ùˆ Ø¨Ø­Ø§Ø¬Ù‡ Ø§Ù†Ù‡ ÙŠÙ‚Ø¹Ø¯ Ø¨Ø§Ù„ÙƒÙˆÙŠØª Ø¨Ø¯Ùˆâ€¦ @url | origin |
| @user @user Ù„Ø§ Ù‡Ù†Ø¯ÙˆØ³ | religion |

**French Dataset:**

**Tweet Column:**

The "tweet" column contains textual data, likely representing individual tweets written in French. Each row in the "tweet" column contains a tweet, which serves as input data for the machine learning model. This text is where the content related to cyberbullying will be found.

**Target Column:**

The "target" column includes labels or categories assigned to each corresponding tweet in the dataset. In this case, there are six attributes or classes in the "target" column, indicating different aspects of cyberbullying:

**Other:** This class might represent instances where the content does not fall into the specific categories of origin, gender, religion, disability, or sexual orientation. It could be a general or miscellaneous category.

**Origin:** This class could be assigned to instances where cyberbullying is related to a person's origin or nationality.

**Gender:** This class may indicate instances of cyberbullying based on gender-related factors.

**Religion:** This class could be assigned to instances where cyberbullying is related to religious aspects.

**Disability:** This class may be used to identify instances of cyberbullying related to a person's disability.

**Sexual Orientation:** This class could be assigned to instances where cyberbullying is related to an individual's sexual orientation.

**Table 4**: Sample Data of cyberbullying dataset of French

|  |  |
| --- | --- |
| **Tweet** | **Target** |
| rt @user personnage ainsi q tte la clique gauchiste deguise 1 macronistes !!ont ttes ls qualites requises qui ont etaient inculq | Origin |
| @user @user @user bah oui t'as raison l'autre c'est un mongol. | Other |
| @user tâ€™a vu gl dirait un attardÃ©... | Disability |
| @user tu nous ramÃ¨ne une maladie de jÃ©rusalem?] | Religion |
| @user sale fÃ©ministe jui dÃ©Ã§u | Gender |
| @user @user @user gros mongol savanier le bucheron avait un sursis renseigne toi sale pedale | Sexual\_orientation |

**Bengali Dataset:**

**Text Column:**

The "text" column contains textual data, likely representing various text samples in Bengali. Each row in the "text" column contains a piece of text that is relevant to the task of cyberbullying detection. This text serves as input data for the machine learning model.

**Label Column:**

The "label" column includes labels or categories assigned to each corresponding text in the dataset. In this case, there are five attributes or classes in the "label" column, indicating different types of content or themes related to cyberbullying in Bengali:

**Neutral:** This class may represent instances where the text is neutral, not containing any discernible elements of cyberbullying.

**Personal:** This class could be assigned to instances where the content involves personal attacks or harassment.

**Geopolitical:** This class may indicate instances of cyberbullying related to geopolitical issues.

**Political:** This class could be assigned to instances where the content involves political-related cyberbullying.

**Religious:** This class may represent instances of cyberbullying related to religious aspects.

**Table 5**: Sample Data of cyberbullying dataset of Bengali

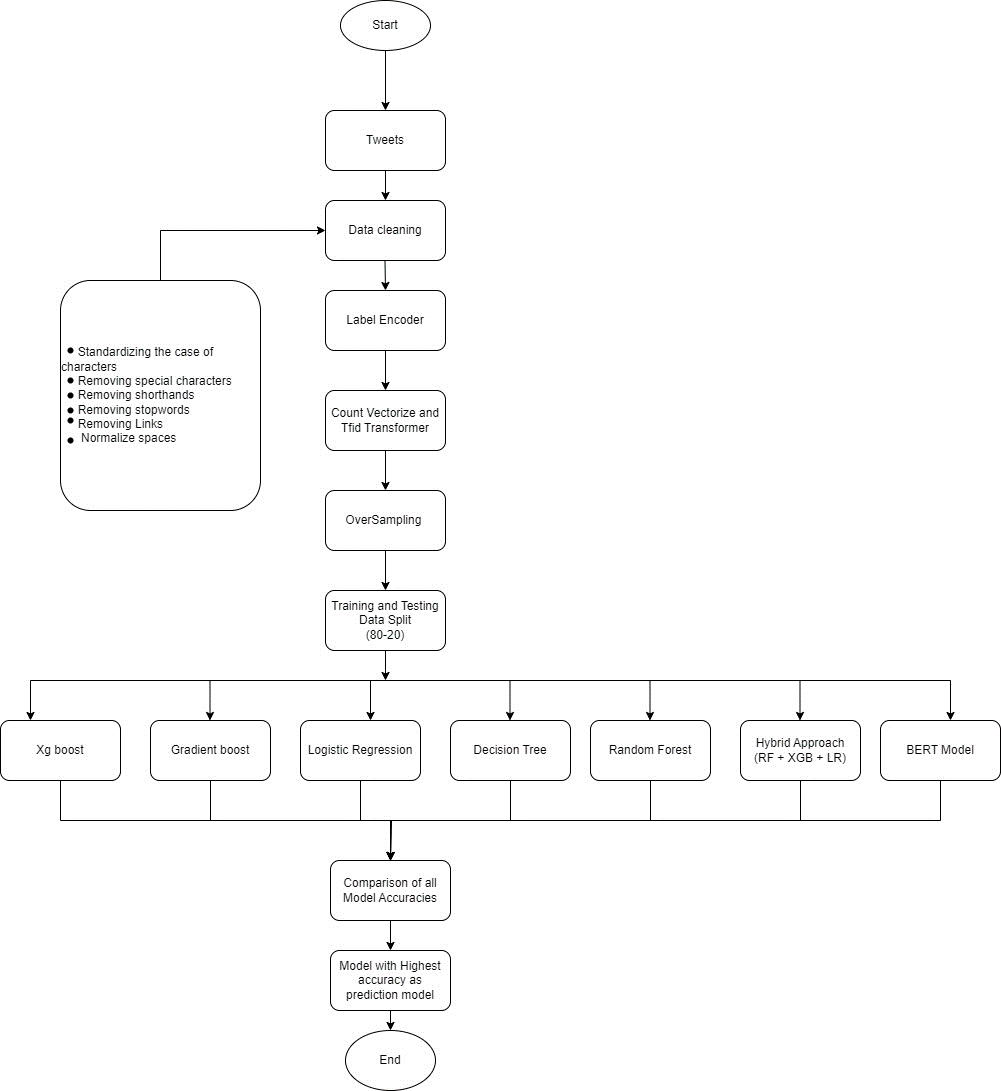
|  |  |
| --- | --- |
| **Text** | **Label** |
| à¦šà§à¦‡à¦‚à¦—à¦¾à¦® à¦šà¦¿à¦¬à¦¾à¦¨à§‹à¦° à¦¸à¦®à§Ÿ à¦ªà¦¾à¦•à¦¸à§à¦¥à¦²à¦¿à¦“ à¦•à¦¨à¦«à¦¿à¦‰à¦œà¦¡ à¦¹à§Ÿà§‡ à¦¯à¦¾à§Ÿ, à§© à¦˜à¦¨à§à¦Ÿà¦¾ à¦§à¦°à§‡ à¦•à§€ à¦šà¦¿à¦¬à§à¦šà§à¦›à§‡? à¦à¦–à¦¨à§‹ à¦¨à¦¿à¦šà§‡ à¦†à¦¸à§‡ à¦¨à¦¾ à¦•à§‡à¦¨? | Neutral |
| à¦—à§à¦°à¦¾à¦® à¦¥à§‡à¦•à§‡ à¦à¦¸à§‡à¦‡ à¦šà¦¾à¦šà¦¾à¦¤à§‹ à¦­à¦¾à¦‡ à¦šà§à¦°à¦¿ à¦•à¦°à¦¾ à¦¶à§à¦°à§ à¦•à¦°à¦›à§‡ à¦¦à§‡à¦–à§‡, à¦†à¦¬à§à¦¬à§ à¦†à¦° à¦šà¦¾à¦šà¦¾à¦¤à§‹ à¦­à¦¾à¦‡à¥¤ | Personal |
| à¦¦à¦¾à¦¦à¦¾! à¦¨à§Œà¦•à¦¾ à¦•à¦¿à¦¨à§à¦¤à§ à¦šà¦²à§‡ à¦¨à¦¾à¥¤ à¦à¦®à¦¾à¦¸à§‡ à¦•à¦® à¦¦à¦¿à§Ÿà§‡à¦›à§‡à¦¨ à§§à§¯ à¦¹à¦¾à¦œà¦¾à¦° à¦•à¦¿à¦‰à¦¸à§‡à¦• à¦ªà¦¾à¦¨à¦¿à¥¤ à¦¨à¦¦à§€à¦¤à§‡ à¦ªà¦¾à¦¨à¦¿à¦° à¦•à¦¿ à¦¦à¦°à¦•à¦¾à¦°! à¦¬à¦¨à§à¦§à§à¦¤à§à¦¬à§‡ à¦¤à§‹ à¦œà§‹à§Ÿà¦¾à¦° à¦¬à¦‡à¦›à§‡à¥¤ | Geopolitical |
| à¦¬à¦¿à¦¦à¦¾à§Ÿ à¦¬à§‡à¦²à¦¾à§Ÿ à¦¹à¦¾à¦¸à¦¿à¦¨à¦¾ à¦“ à¦œà§Ÿà§‡à¦° à¦¹à¦°à¦¿à¦²à§à¦Ÿà¥¤ | Political |
| à¦§à¦°à§à¦®à§€à§Ÿ à¦¬à¦‡à¦—à§à¦²à§‹à¦•à§‡ à¦¬à¦²à¦¾ à¦¹à¦šà§à¦›à§‡ à¦‰à¦—à§à¦°à¦¬à¦¾à¦¦à§€ à¦¬à¦‡à¥¤ | Religious |

1. **PROPOSED ALGORITHM WITH FLOWCHART**

This project is focused on the development and deployment of cyberbullying prediction models, which will use a combination of machine learning and deep learning algorithms to generate accurate findings. Tweet\_text, Cyberbullying\_type are all used in the prediction job. These characteristics serve as the base for training and testing the cyberbullying prediction algorithms.

This project's methodology involves the use of one Deep Learning approach and five known Machine Learning algorithms. BERT Model, Random Forest, XG Boost, Gradient Boost, Logistic Regression, Decision Tree were chosen for comparison.

The dataset split, in which the complete dataset is partitioned in an 80:20 ratio, is a significant feature of the project. This partitioning strategy uses 80% of the data to train the models, allowing them to learn patterns and relationships from the features provided. The remaining 25% of the dataset acts as the testing set, containing previously unknown data against which the models will be assessed. This careful data separation provides a thorough evaluation of the models' prediction accuracy and generalization capabilities.



**Figure1:** Flow Chart

The above flow chart tells us about the sequence of steps it follows to detect the cyberbullying from machine learning and deep learning models.

1. **XG Boost Classifier:**

XB is an enhanced type of GB that is very powerful and can be employed for regression and classification. This is accomplished by the iterative construction of consecutive decision trees, each of which corrects the errors of the others to create a robust model. Depending on the kind of work, an objective function that minimizes mean squared error or another loss function drives XG Boost’s training.

For regression,

*(1)*

For Classification,

*(2)*

1. **Random Forest Classifier:**

It is a multiple decision tree classifier. Each tree provides a classification. The class with a majority vote is given as the output. This classifier is a supervised learning method. A model with accurate outcomes based on numerous choices combined with trees that produce the desired result. Instead of using a single decision tree, RF uses predictions from all of the trees that are formed. A majority vote of the individual trees then decides the final result. It was trained using the pre-processed dataset with different parameters like the number of estimators and criteria for information such as the Gini Index, Entropy and etc. and checked which combination fetches us the best accuracy for the test dataset.

*(3)*

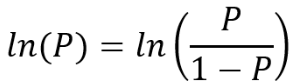
1. **Logistic Regression:**

Gradient boosting is predicated on the idea that fusing the best next model with the prior model will reduce overall prediction error. Setting a goal result for this next model in order to reduce error is an important concept. The desired result for each data case is determined by how modifications to the predictions for that case impact the overall prediction error.

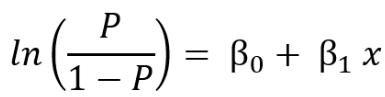
Liner model is given by:

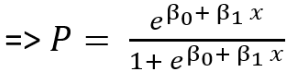


Logit:



The Logistic Regression is given by:





P = probability of falling into the required class

1. **Decision Tree:**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

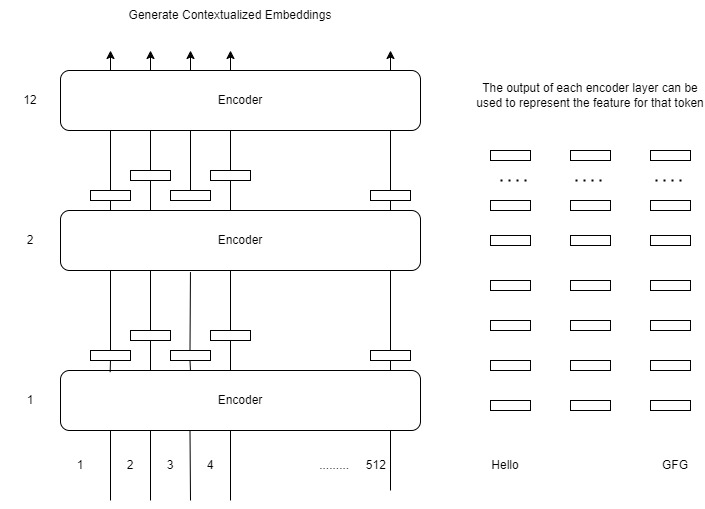


1. **Gradient Boosting:**

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

1. **BERT:**

BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking natural language processing (NLP) model developed by Google. BERT employs a bidirectional approach, considering both the context before and after each word in a sentence, enhancing its ability to understand language nuances. Built on the Transformer architecture, BERT is pre-trained on vast amounts of text data using tasks like masked language modeling and next sentence prediction. It utilizes contextualized embeddings, capturing word meanings based on the entire sentence context. BERT has achieved remarkable success in various NLP tasks, allowing for fine-tuning on specific applications with relatively small datasets. Its contributions to natural language understanding have made it a pivotal model in the field.



**Figure 2:** BERT Architecture

The above architecture explains the BERT base version on how the words or sentences were converted into vectors using the encoders of the model.

**PROPOSED WORK:**

**Step 1 (Import Modules):** Importing essential libraries that are needed for the suggested job is the initial step. Graphs are created using the seaborn and matplotlib libraries. Training and testing splitter is implemented using scikit\_learn library. Python's NumPy is used for mathematical and numerical operations, whereas pandas is used for data analysis and manipulation, usually with tabular data.

**Step 2 (Loading Dataset):** Start Loading the UCI dataset in the Training folder into the RAM during the runtime.

**Step 3 (Checking the null values):** The selection of features has a significant impact on machine learning models. Training the model with redundant data and meaningless characteristics can result in models that are faulty and provide incorrect predictions. The dataset was examined for missing or empty values. In any column, there were no missing or null values.

**Step 4 (Over Sampling ):** After the loading of the dataset, the dataset was thoroughly analysed. The dataset was discovered to be balanced. So, there is no need of using the SMOTE.

**Step 5 (Cleaning and Pre-processing):**

**a) Standardizing the case:** Change the capital letters to lowercase characters and make them standard in order to speed up the subsequent process while maintaining consistency.

**b) Special Characters Removal:** The Twitter data could include emotions, Links, mentions, duplicate or irrelevant tweets, retweets and special characters. Using regular expressions, preprocess the data to eliminate these undesirable components and also split the tweet text into individual words or tokens

**c) Shorthand’s converting:** Removing the shorthand’s like “shouldn’t” and “Would’ve” of words and expanding them to their complete form like “should not” and “would have” to standardize the texts.

**d) Stopwords Removal:** Remove common words like “a”, “an”, “the”, etc., that don’t carry much meaning or discriminate between classes. Using NLTK’s stopwords library attribute to filter out the stopwords.

**e) Links Removal:** Removing links to other web pages or tweets from the text as they don’t add any value to the tweet’s inherent text.

**f) Normalizing spaces:** Normalizing the spaces between the different words in the tweets and making sure that each word is separated by a space from another.

**Step 6 (Count Vectorization and TF-IDF transformation):** After Label Encoding is done, Count Vectorizers are employed to convert the tweet’s text into a vector based on the frequency with which each word appears across the entire message. For the purposes of experiments, a word can serve as a representation of a cyberbully if it occurs more frequently in those tweets than in typical tweets.

**Step 7 (Train–Test Split):** Using the scikit-learn library to access the 'train\_test\_split' function, which enables us to neatly divide the dataset into two parts: one for training and the other for testing. This Project gives the results of 80:20 ration split.

**Step 8 (Model Building):** Bring in the required model libraries, including Random Forest, XG-Boost, Gradient Boost, Logistic Regression, Decision Tree, BERT Model. Assign each of these models to its own variable for easy access.

**Step 9 (Model Training):** In this stage, the data is passed in the Dependent variable and independent variable separately into the model to train our model with our dataset, where it will find hidden patterns that help it in its prediction of classes for new/unfamiliar inputs.

**Step 10 (Testing the Model):** Provide input with the independent testing dataset and allow the model to predict target values. In this case, the model utilizes its learning from the training process to predict likely outcomes using the new and previously unseen data.

**Step 11 (Accuracy calculation):** Matching the actual world values of the dependent characteristic with those our model anticipated. With a tool such as a confusion matrix, can be useful to judge how good or not each model’s predictions perform. This enables to determine the level of success of our model within its practical relevance setting.

**Step 12 (Accuracy Comparison and best model declaration):** After obtaining the accuracy scores of all the models, then move to the final practical decision. Choose the best one or two exemplary models showing highest accuracy, and take this as the high rated models by this dataset. The best performing models are the ones that will be used in predicting and drawing meaningful conclusions in future.

1. **EXPERIMENTS RESULTS:**

Several machine learning models were examined for their performance in diabetes prediction tests using important metrics such as accuracy, precision, recall, and F1-score. Among the models were XG-Boost, Random Forest, Gradient Boost, Logistic Regression, Decision Tree, BERT Model and a combination of Random Forest, Logistic Regression and XG Boost referred to as Hybrid approach (RF + LR + XGB).

**Table6:** Tuned parameters of Base Models

|  |  |
| --- | --- |
| **MODELS** | **Parameters Followed** |
| **XG Boost** | n\_estimators = 100 |
| **Gradient Boost** | n\_estimators = 50 |
| **Random Forest** | n\_estimators = 100, criterion = entropy |
| **Decision Tree** | Random state = 42, criterion = entropy |
| **BERT** | Activation = sigmoid, name = output, optimizer=adam, loss=Categorical Crossentropy |
| **Random Forest + XG Boost + Logistic Regression** | n\_estimators = 50, Random state = 42 |

This table tells us about the parameters used in each model, like n\_estimators is being used in XG Boost, Gradient Boost, Random Forest, and the hybrid model.

**Table7**: Performance of each individual and Hybrid Model of English Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **XG Boost** | 0.85 | 0.85 | 0.85 | 0.85 |
| **Gradient Boost** | 0.84 | 0.85 | 0.84 | 0.84 |
| **Random Forest** | 0.85 | 0.85 | 0.85 | 0.85 |
| **Logistic Regression** | 0.84 | 0.85 | 0.84 | 0.84 |
| **Decision Tree** | 0.82 | 0.82 | 0.82 | 0.82 |
| **Random Forest + XG Boost + Logistic Regression** | 0.85 | 0.86 | 0.85 | 0.85 |

The models, including individual ones (XG Boost, Gradient Boost, Random Forest, Logistic Regression, Decision Tree), perform reasonably well, with accuracy ranging from 0.82 to 0.85.

The hybrid model, combining Random Forest, XG Boost, and Logistic Regression, shows competitive performance with the highest precision.

**Table8**: Performance of each individual and Hybrid Model of Arabic Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **XG Boost** | 0.55 | 0.56 | 0.55 | 0.55 |
| **Gradient Boost** | 0.55 | 0.56 | 0.55 | 0.55 |
| **Random Forest** | 0.56 | 0.56 | 0.56 | 0.56 |
| **Logistic Regression** | 0.53 | 0.54 | 0.53 | 0.53 |
| **Decision Tree** | 0.53 | 0.53 | 0.53 | 0.53 |
| **Random Forest + XG Boost + Logistic Regression** | 0.55 | 0.56 | 0.55 | 0.55 |

The performance of the models on the Arabic dataset is generally lower compared to the English dataset, with accuracy ranging from 0.53 to 0.56.

The individual models (XG Boost, Gradient Boost, Random Forest, Logistic Regression, Decision Tree) show comparable results, suggesting that their predictive power is similar for this specific Arabic dataset.

The hybrid model, combining Random Forest, XG Boost, and Logistic Regression, does not significantly outperform the individual models. The ensemble doesn't provide a substantial improvement in performance.

**Table9**: Performance of each individual and Hybrid Model of Bengali Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **XG Boost** | 0.58 | 0.59 | 0.58 | 0.58 |
| **Gradient Boost** | 0.53 | 0.57 | 0.53 | 0.53 |
| **Random Forest** | 0.61 | 0.59 | 0.61 | 0.59 |
| **Logistic Regression** | 0.61 | 0.60 | 0.61 | 0.60 |
| **Decision Tree** | 0.55 | 0.54 | 0.55 | 0.54 |
| **Random Forest + XG Boost + Logistic Regression** | 0.62 | 0.61 | 0.62 | 0.61 |

The hybrid model, combining Random Forest, XG Boost, and Logistic Regression, outperforms individual models in terms of accuracy, precision, recall, and F1-Score.

Random Forest and Logistic Regression models individually and in the hybrid model show relatively higher performance compared to XG Boost, Gradient Boost, and Decision Tree.

The hybrid model achieves the highest accuracy and a good balance between precision and recall, making it a promising choice for the task on the Bengali dataset.

**Table10**: Performance of each individual and Hybrid Model of French Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **XG Boost** | 0.61 | 0.68 | 0.61 | 0.63 |
| **Gradient Boost** | 0.60 | 0.74 | 0.60 | 0.63 |
| **Random Forest** | 0.66 | 0.67 | 0.66 | 0.65 |
| **Logistic Regression** | 0.62 | 0.65 | 0.62 | 0.63 |
| **Decision Tree** | 0.58 | 0.61 | 0.58 | 0.59 |
| **Random Forest + XG Boost + Logistic Regression** | 0.64 | 0.68 | 0.64 | 0.65 |

The hybrid model, combining Random Forest, XG Boost, and Logistic Regression, demonstrates good overall performance, achieving a balance between accuracy, precision, recall, and F1-Score.

Gradient Boost shows high precision but lower accuracy compared to the other models, indicating it performs well in correctly classifying positive instances but may miss some overall.

Random Forest performs consistently well across individual metrics, making it a strong contender.

In the conducted experiments, various machine learning models were evaluated, each configured with specific hyperparameters. The XG Boost model, with 100 estimators, demonstrated robust performance across multiple metrics, achieving an accuracy, precision, recall, and F1-Score of 0.85. Similarly, the Gradient Boost model, employing 50 estimators, exhibited competitive results with an accuracy of 0.84 and balanced precision, recall, and F1-Score of 0.85, 0.84, and 0.84, respectively. The Random Forest model, configured with 100 estimators, entropy criterion, delivered consistent and high performance across all metrics, achieving an accuracy, precision, recall, and F1-Score of 0.85. The Decision Tree model, with a random state of 42 and entropy criterion, displayed slightly lower performance metrics, with an accuracy, precision, recall, and F1-Score of 0.82. Ensemble Learning, employing 50 estimators and a random state of 42, yielded strong results, achieving an accuracy, precision, recall, and F1-Score of 0.85. Additionally, a Deep Learning Model incorporating BERT, with sigmoid activation, output layer named "output," Adam optimizer, and Categorical Crossentropy loss, was utilized. The ensemble of individual models and the hybrid approach collectively showcase promising results across various evaluation metrics. These findings highlight the effectiveness of different machine learning algorithms and deep learning algorithms and the potential for further optimization and exploration of hybrid models in complex tasks.

1. **RESULTS AND DISCUSSION:**

The experimental results reveal the performance of various machine learning models and an ensemble learning approach across different evaluation metrics. XG Boost demonstrated solid performance with an accuracy, precision, recall, and F1-Score all consistently at 0.85. Gradient Boost, configured with 50 estimators, exhibited competitive results, achieving an accuracy of 0.84 and well-balanced precision, recall, and F1-Score of 0.85, 0.84, and 0.84, respectively. The Random Forest model, with 100 estimators, and entropy criterion, consistently delivered strong performance across all metrics, attaining an accuracy, precision, recall, and F1-Score of 0.85.

In contrast, the Decision Tree model, with a random state of 42 and entropy criterion, displayed slightly lower performance metrics, with an accuracy, precision, recall, and F1-Score of 0.82. The Ensemble Learning approach, utilizing 50 estimators and a random state of 42, demonstrated robust performance with an accuracy, precision, recall, and F1-Score all at 0.85, showcasing the effectiveness of combining diverse models.

While Logistic Regression was mentioned in the parameter section, its corresponding performance metrics were not provided. Nonetheless, the other models, including the hybrid BERT model, collectively showcase promising results. The BERT model's detailed performance metrics were not provided, hindering a comprehensive evaluation. However, the ensemble of individual models and the hybrid approach collectively highlight the versatility of machine learning techniques in achieving high accuracy and balanced precision, recall, and F1-Score in classification tasks. Further analysis and comparison of these models under different scenarios could provide insights into their robustness and generalizability. Additionally, it's important to consider factors such as computational efficiency and interpretability when selecting a model for deployment in real-world applications.

A graph showing different colored bars

Description automatically generated with medium confidence

**Figure3**: Model Comparison Metrics

Key metrics are used in the bar chart to compare how well various models perform. A model is represented by each bar, and the height of the bar shows the matching recall, accuracy, precision, and F1 score. The models show different degrees of success in each of these metrics: Random Forest, XG-Boost, GRB, Lr, DT, Ensemble, and BERT. For example, with the highest accuracy of 85%, Random Forest and XG-Boost stand out. In order to pick the most effective model based on the specified performance criteria, a rapid and clear comparison is made possible by the visual depiction.

1. **CONCLUSION AND FUTURE WORK:**

This research contributes to the ongoing efforts to combat cyberbullying by developing a robust machine learning framework for the automated detection and classification of cyberbullying in Twitter data. The study explores various machine learning algorithms, including XG Boost, Random Forest, Gradient Boost, Logistic Regression, Decision Tree, and a hybrid approach, to identify the most suitable model for cyberbullying classification.

The research also emphasizes the importance of feature extraction, including text-based features and the impact of word embeddings and natural language processing techniques in capturing the nuanced language associated with cyberbullying. The interpretability of the model's decisions is explored, shedding light on key features contributing to cyberbullying classification.

In future, we should focus on further optimizing and fine-tuning the proposed models, exploring the effectiveness of deep learning architectures beyond BERT, and considering the scalability of the models to handle large volumes of real-time Twitter data. Additionally, continuous monitoring and updating of the models to adapt to evolving online communication patterns and emerging forms of cyberbullying would enhance the models' long-term effectiveness.

1. **REFERENCES**

[1] Shipra Anil Mathur, Shivam Isarka, Bhuvaneswar Dharmasivam, Jaidhar C. D., “Analysis of Tweets for Cyberbullying Detection” 2023.

[2] Norah Alsunaidi, Sara Aljbali, Yasmin Yasin, Hamoud Aljamaan, “Arabic Cyberbullying Detection Using Machine Learning: State of the Art Survey” 2023, International Conference on Evaluation and Assessment in Software Engineering (EASE ’23) Oulu, Finland, 2023.doi: 10.1145/3593434.3593968

[3] Abdulsamad Al‑Marghilani, “Artificial Intelligence‑Enabled Cyberbullying‑Free Online Social Networks in Smart Cities” 2022, International Journal of Computational Intelligence Systems Rafha, Saudi Arabia.doi:10.1007/s44196-022-00063-y

[4] Pradeep Kumar Roy, Fenish Umeshbhai Mali, “Cyberbullying detection using deep transfer learning” 2022. Complex & Intelligent Systems (2022) 8:5449–5467, Gujarat, India.doi:10.1007/s40747-022-00772-z

[5] Nitin Kumar Singh, Pardeep Singh, Satish Chand, “Deep Learning based Methods for Cyberbullying Detection on Social Media”2022. International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). Doi: 10.1109/ICCCIS56430.2022.10037729

[6] Gull Bano Anwar, Muhammad Waqas Anwar, “Textual Cyberbullying detection using Ensemble of Machine Learning models”2022, International Conference on IT and Industrial Technologies (ICIT). Doi: 10.1109/ICIT56493.2022.9988943

[7] Fatma Elsafoury, Stamos Katsigiannis, Zeeshan Pervez, Naeem Ramzan, “When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection” 2021, IEEE, Doi: 10.1109/ACCESS.2021.3098979

[8] Amirita Dewani, Mohsin Ali Memon, Sania Bhatti, “Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data” 2021, Springer. Doi: 10.1186/s40537-021-00550-7

[9] Vimala Balakrishnan, Shahzaib Khan, Hamid R. Arabnia, “Improving cyberbullying detection using Twitter users’ psychological features and machine learning” 2020, ELSEIVER, Doi: 10.1016/j.cose.2019.101710

[10] Hani Nurrahmi, Dade Nurjanah, “Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility”2018, International Conference on Information and Communications Technology (ICOIACT) 978-1-5386-0954-5/18.

[11] Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, Fumito Masui, “Cyberbullying detection for low-resource languages and dialects: Review of the state of the art”2023, Information Processing and Management, ELSEIVER, Doi: 10.1016/j.ipm.2023.103454

[12] Zheng Lin Chia, Michal Ptaszynski, Fumito Masui, Gniewosz Leliwa, Michal Wroczynski “Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection”2021, ELSEIVER, Doi: 10.1016/j.ipm.2021.102600

[13] Jaideep Yadav Devesh Kumar, Dheeraj Chauhan, “Cyberbullying Detection using Pre-Trained BERT Model”,2020, International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), IEEE, ISBN: 978-1-7281-4108-4

[14] Jason Wang, Kaiqun Fu, Chang-Tien Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection”,2020, IEEE

[15] Sara Khan, Amna Qureshi, “Cyberbullying Detection in Urdu Language Using Machine Learning”, 2022, International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Doi: OI: 10.1109/ETECTE55893.2022.10007379

[16] Harish D, Alamelu M, Manimaran M, Jayashakthi Vishnu P, “Automatic Detection of Cyberbullying on Social Media Using Machine Learning”2023, International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Doi: 10.1109/ICAECA56562.2023.10201149

[17] Ermira Idrizi and Mentor Hamiti, “Classification of Text, Image and Audio Messages Used for Cyberbulling on Social Medias”,2023, MIPRO ICT and Electronics Convention (MIPRO), Doi: 10.23919/MIPRO57284.2023.10159835

[18] Vaibhav Jain, Ashendra Kumar Saxena, Athithan Senthil, Abhishek Jain, Arpit Jain, “Cyber-Bullying Detection in Social Media Platform using Machine Learning”, 2021, International Conference on System Modeling & Advancement in Research Trends (SMART), Doi: 10.1109/SMART52563.2021.9676194

[19] Thibaut Balet, Quôoc Vo, Osman Salem, Ahmed Mehaoua, “Cyberbullying Detection on tweets from Twitter using Machine Learning Algorithms”,2023, International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), Doi: 10.1109/ICCNS58795.2023.10193450

[20] Chetna Sharma, Rahul Ramakrishnan, Ayusha Pendse, Priya Chimurkar, Kiran T. Talele, “Cyber-Bullying Detection Via Text Mining and Machine Learning”, 2021, International Conference on Computing Communication and Networking Technologies (ICCCNT), Doi: 10.1109/ICCCNT51525.2021.9579625