Computational Methods for Detecting Transformer-Based Models Algorithm Fairness in Societal Bias Over Training Epochs

Samar Shabbir[*], Ahmad Salman Khan[*], Ahtisham Ahmad[*], Muhammad Waqas[**], Mansoor Qadir[***], Mubina Zaka[****], Afshan Ishaq[*****]

[*] Department of Software Engineering, University of Lahore, Pakistan
[**] Department of Electrical Engineering, Iqra National University, Pakistan
[***] Department of Computer Science, CECOS University of IT & Emerging Sciences, Pakistan
[****] Department of Computer Science & Information Technologies Hazara University Mansehra.
[*****] Department of Electronics, University of Engineering & Technology Abbottabad Campus, Pakistan

Abstract- Social media sites play a significant role in shaping public opinion in today's dynamic world. One of the challenges to these beneficial features of internet conversations is hateful sentiment. This research addresses the critical need to detect and mitigate hate speech, emphasizing the importance of fostering an inclusive and respectful online environment leveraging a Twitter dataset to explore hate-speech detection, classifying text into various categories of hate, violence, dehumanization, and demonization. Natural Language Processing (NLP) methods detecting hateful sentiment are tailored with unique challenges faced by people. The evaluation process involves rigorous testing of various models to assess their effectiveness in identifying hateful sentiment within social media sites such as X (Twitter) dataset. The proposed approach aims to contribute to ongoing efforts in combating hate speech within the online community. Analyzing the intricacies of hateful sentiment is imperative for safeguarding people freedom and establishing a digital environment that ensures safety and inclusivity in public discussions. The testing accuracies for LSTM achieved at 87.27%, BERT demonstrated a precision of 91.38%, DistilBERT exhibited a testing accuracy of 91.47%, RoBERTa performed at 91.05%, Hybrid RNN showcased an accuracy of 90.71%, and XLNet emerged with the highest testing accuracy at 91.68%.

Index Terms- NLP, Hateful Sentiment, Detection, BERT, XLNet

I. INTRODUCTION

In today's highly connected world, social media websites like Facebook, Instagram and Twitter play a very crucial role in spreading news and information. Twitter has become an exceptional place for public figures, news outlets, and the general people to get their voices heard and debate important issues. Individuals are vulnerable to hate speech, which not only endangers their safety but also impedes the free flow of information. To create societies that value human rights, equality and peaceful coexistence, the elimination of hate speech is crucial. Domestic violence is one of the reasons that is related to health, welfare and human rights. Identification of urgent situations amidst the abundance of online content imposes difficulty. Several key objectives, including creation of a unique dataset from social media, is considered a benchmark for accuracy, annotating it with multiple classes, and conducting a series of comprehensive experiments using different deep learning architectures [1-4]. Intimate partner violence is another significant global public health concern, impacting millions of individuals. Research suggests that approximately one in four women experiences severe violence in their lifetime. This research is motivated by the need to address the absence of artificial intelligence systems capable of automatically detecting experiences shared by victims on social media. Although victims of intimate partner violence often turn to platforms such as Twitter for support, there has been a lack of initiative in utilizing social media to address this important public health issue. The NLP pipeline that has been developed shows a level of performance that is on par with human capabilities. It exhibits minimal bias, especially when it comes to words related to gender and race [5].

Exposure to sexist speech has far-reaching consequences that extend beyond online platforms, significantly impacting the lives of women and infringing upon their freedom of speech [6]. This study presents a novel objective, aiming to understand and analyze the various manifestations of sexism in online discussions, spanning from overt hate speech to subtle forms of expression. Drug addiction has become another bigger problem in the US, thus it's crucial to find reliable methods to identify drug-abuse risk behavior in a broad population of Twitter users. The system would utilize a significant amount of unlabeled data to automatically improve annotated data and effectively monitor the changing patterns of drug abuse on Twitter. The model is assessed using a dataset of three million tweets related to drug abuse, which includes geo-location data. The evaluation showcases its efficacy in identifying risk behaviors associated with drug abuse [7].

This research is a performance evaluation of hate speech detection on Twitter utilizing Advanced NLP techniques and assessing various models. The findings offer valuable insights for researchers, practitioners, and platform developers who aim to address hateful sentiment. This study aids in the creation of an online forum for public conversation that is more inclusive. A significant advancement in hate speech detection on Twitter is presented. It offers a comprehensive methodology, evaluating various NLP models, aiming to foster a comprehensive understanding that extends beyond technical aspects to consider

the broader societal implications of online conversations. The study's findings have the potential to be valuable for researchers, practitioners, and platform developers who are focused on improving the safety and inclusivity of online spaces. This investigation aims to define the activities in the life cycle of detecting hate speech using NLP. The main contributions w.r.t. detection are:

- Analyze the literature that is currently available on Natural Language Processing based on hate speech identification.
- Examine how NLP approaches may be used to train machine learning algorithms and models to detect hateful sentiment in text.
- Propose which machine learning model or algorithm best uses natural language processing (NLP) approaches to identify hate speech in text.

II. Related work

One of the main concerns about using social media today is the impact communication has, whether positive or bad, on individuals or society at large. The number of articles published in the social media field demonstrates the significance of this emerging research area, such as sentiment and social network analysis drawing attention from corporations, governments, to even the academic sector. Analysis and comprehension of social networks represent a significant research issue. Despite the growing number of victims who disclose their experiences on social media, there is a lack of research on extracting actionable insights in the domain of domestic violence [1]. The proposed method entails the identification of multiple classes from DV social media posts using advanced deep-learning models. The effectiveness of cyberbullying detection is utilized with the initial dataset consisting of more than 30,000 tweets obtained from the University of Maryland. The study presented in [2] investigates word embedding-based machine learning methods, including Distributed Bag of Words (DBOW) and Distributed Memory Mean (DMM). It evaluates the effectiveness of Word2Vec Convolutional Neural Networks (CNNs) in classifying online hate. The two datasets were used for a number of algorithms to determine the optimal classification strategy for the data given. These algorithms included Logistic Regression, Linear SVC, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Linear SVC showed the highest efficacy across both datasets, while Bernoulli Naive Bayes had the lowest performance. The study delved into different Doc2Vec models, such as DBOW, DMM, and a hybrid of DBOW and DMM. The hybrid DBOW + DMM model produced the most favorable outcomes for both datasets, surpassing the accuracy of the individual models. Trigram (DBOW) and bigram (DMM) vectors were used to train a neural network using the results of the Doc2Vec studies.

The study in [3] is based on a combination of Support Vector Machines and Recurrent Neural Network models. These models analyze various features such as word embeddings, sentiment, and irony. The findings highlight the complex nature of the task, especially in detecting hidden forms of aggression, which exposes the shortcomings of commonly employed methods. This research in [4] suggests a method for categorizing hate material into six different groups using a customized LSTM-GRU model. specialists have extracted and annotated tweets to create a highly controlled dataset. A proactive social media-based intervention and support framework must include the model because of its efficacy, which also makes it a vital part of large-scale cohort studies and population-level surveillance [5]. The numerous ways that sexist views and actions are visible in social network talks, especially on Twitter, are thoroughly examined in this research [6]. A system utilizing machine learning is introduced, enabling a comparison between conventional approaches and neural network-based methods. It achieves an accuracy rate of 74% in effectively detecting sexist expressions. The MeTwo corpus is a significant resource for Spanish texts, consisting of 3600 tweets that have been labeled as either sexist or not. These labels were determined through majority votes from three annotators. The authors aim to classify sexist expressions into different facets, addressing the various aspects of women that are targeted.

A method to collect drug-abuse risk behavior tweets on a broad scale is presented in this research [7]. The method blends data crowdsourcing methods with supervised machine learning. The outcomes show how well the suggested models work, with an accuracy of 86.53%, recall of 88.6%, and an F1-value of 86.63%. These results significantly outperform traditional models and signify a noteworthy breakthrough in cutting-edge results. The authors of the work in [8] used an MT-DNN multi-task learning network to participate in the DA-VINCIS competition. The authors tackled the challenge by implementing several preprocessing techniques. Four experiments were carried out with different setups, resulting in significant findings. In Subtask 1, the authors achieved the greatest F1 score (74.80%), recall (74.09%), and precision (75.52%). Subtask 2's outcomes are as follows: 39.20% for F1, 37.79% for precision, and 43.88% for recall.

The problem of hate speech by white supremacists on social media is examined in [9], which also looks into the application of natural language processing and deep learning to automatically identify such content on Twitter. Bidirectional Encoder Representations from Transformers (BERT) and a bidirectional Long Short-Term Memory (BiLSTM) model with domain-specific word embeddings from a white supremacist corpus are the two models that are analyzed. BERT obtained an F1-score of 0.80, whilst the BiLSTM model obtained an F1-score of 0.75. Furthermore, BERT establishes itself as state-of-the-art by surpassing the domain-specific approach by 4 points, demonstrating its superiority.

Cyberbullying in contemporary societies, with a specific emphasis on Twitter is essential for developing detection methods that do not depend on victims' interactions [10]. Employing seven machine learning classifiers—Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM)—the researchers assembled a worldwide dataset of 37,373 tweets. Performance criteria, such as accuracy, precision, recall, and F1 score, were used to assess the classifiers on the global dataset. With a median accuracy of almost 90.57%, the experimental findings show how successful logistic regression (LR) is. LR outperformed other classifiers, achieving the highest F1 score of 0.928. With a precision of 0.968, the SGD classifier was the most accurate, while the SVM classifier had the best recall of 1.00. The research offers a cyberbully detection

model that combines Word2Vec and TF-IDF feature extraction methods with a variety of classifiers. With an F1 score of 0.928 and a classification accuracy of 90.57%, the LR model showed remarkable performance. While LR regularly outperformed other classifiers, especially with bigger data sets and optimum prediction times, SGD and LGBM showed comparable performance.

The issue of identifying offensive content on social media and suggesting using a combination of Recurrent Neural Network (RNN) models as an ensemble classifier is inevitable. The approach integrates user-related information, including users' attitudes towards racism or sexism, using vectors of word frequencies extracted from text [11]. By developing a deep learning architecture that takes advantage of word frequency vectorization and is compatible with any language, the work significantly advances the subject. To guarantee the reliability of the results, the trials were repeated and the averaged outcomes were used. While the rising problem of toxic online content, specifically hate speech, within the context of the expanding internet user population is examined in [12]. The system utilizes word embeddings and explores both LSTM and Bidirectional LSTM (Bi-LSTM) neural networks. By applying an early stopping criterion based on the loss function during training, an LSTM network was able to attain an 86% accuracy rate. Several LSTM and Bi-LSTM models, including basic deep neural networks and stacked networks, are tested. An LSTM model achieves an accuracy of 86%, showcasing improvements in overall accuracy particularly for the "hate speech" category when compared to baseline models.

The issue of hate speech and machine learning techniques for detecting it is provided with effective solutions for addressing this problem [13]. This method demonstrates superior performance compared to ten commonly used machine learning methods, resulting in an enhanced average F1-score of 94.2% across two well-established tweet datasets. The goal of this study is to build and validate a vote ensemble learning strategy that will address the difficulties posed by the vague and context-dependent character of Twitter slang. This approach showcases exceptional performance when compared to alternative learning methods. Nevertheless, the study recognizes a drawback in the bag of words representation. Dealing with the ongoing issue of spam, current machine learning and black-listing techniques have managed to achieve an accuracy rate of around 80%. However, they face difficulties when it comes to spam drift and the creation of false information in real-world situations. To overcome these obstacles, a novel deep learning-based strategy is presented in this study [14]. The method proposed utilizes WordVector Training Mode to acquire the syntax of each tweet, creating a binary classifier using the learned representation dataset. The experimental evaluation involved analyzing a 10-day dataset of real tweets.

The detection of situational tweets during disasters where a combination of situational and non-situational information may be found is another concern. The findings suggest that deep learning models, particularly BLSTM with attention mechanisms using crisis word embeddings, are more effective than traditional methods in identifying situational tweets with diverse content in times of disaster. For the objective of recognizing situational tweets in Hindi during catastrophes, this work is the first attempt to apply deep learning algorithms. The available evidence indicates that CNN exhibits strong performance when applied to Hindi tweets, highlighting the significance of deep learning models in different language domains and disaster scenarios [15]. Mohapatra utilized techniques that encompassed the extraction of distinctive attributes, including term frequency-inverse document frequency (TF-IDF), word embeddings, and n-grams [18]. Some of the researchers have used supervised machine learning (ML) based text categorization techniques in past years to categorize hate speech material [19 – 31], [32 – 34]. Bhavesh uses a bag of words to streamline the classification of hate speech in data extracted from Twitter, with the aim of simplifying the process [32]. Analyses method for identifying hate speech on the internet and separating it from other forms of offensive language is presented in [35]. The feature extraction process involves the conversion of all tweets to lowercase and their subsequent stemming utilizing the Porter stemmer. After calculating their respective TF-IDF values assign weights to the unigram, bigram, and tri-gram features during their creation. Several characteristics, such as the quantity of characters, words, and syllables contained in each tweet, are extracted. Mentions, hashtags, retweets, and URLs are commonly denoted through the utilization of binary as well as count indicators [38].

[36] used a classification method "fine-grained" that divides the 'Offense' class into three more subclasses: 'Profanity' which means using absurd language without offending anyone, 'Insult' which means profanity directed at an individual and the harshest form of hate speech 'Abuse' in which negative characteristics are attributed to a group of individuals. [37] claims that some elements—such as racism, violence, gender inequality, and so forth—have a definite connection to hate speech. Using the benefits of supervised classification methods, it hopes to establish lexical baselines for this purpose. They've considered about three labeled categories: "HATE" has 2399 cases of hate speech, "OFFENSIVE" has 4836 instances of offensive (not hate speech) and "OK" has 7274 instances with no offensive material at all. [38] labelled the data into groups i-e HS, Offensive, Neither. [39] used binary classification method to differentiate being a hate and non-hate speech. One approach to detecting hate speech is classification, which may be done using the following three categories: Abusive, (HS) and Non-HS [40]. The machine learning algorithms are incapable of deciphering classification rules from unprocessed text. Algorithms like these require numerical characteristics in order to comprehend classification rules [41 – 47]. Therefore, among the most important stages in text classification is feature engineering. This phase is used to extract essential features from unprocessed text and to display the extracted features numerically. Different studies have used different ways to describe features, based on dictionary, Bag-of-words and TFIDF.

In Machine Learning field, Classification is a significant component. Several classification techniques are available, including Decision Tree, Logistic Regression, Support Vector Machines and Artificial Neural Network. A group of trees is commonly referred to as a forest can be used for classification as well as regression tasks. [36] recommended Logistic Regression as a method of categorization. Cross-validated F1 seems to improve with the addition of emoji features, but performance drops when evaluated on development data. [43] states that in

order to better analyze the data, SVM can map it from one dimension to another, either non-linearly or linearly. It searches for the linear optimum division hyper-plane inside this additional dimension to differentiate tuples from sets. [39] employs an ensemble model by incorporating the estimates of Random Forest, Bidirectional Long Short-Term Memory (BiLSTM) and Support Vector Machine (SVM). The final prediction was determined by taking the majority vote. An additional ensemble model was employed that leverages the confidence scores of Random Forest (RF) as well of the Support Vector Machine (SVM) and algorithms to determine the probability of a given instance, belonging to either class 0 or class 1. The mean of the confidence values and the binary value of the BiLSTM were used to get the final forecast in [32] employed machine learning techniques such as Logistic Regression, Random Forest and Support Vector Machines to categorize instances of hate speech within tweets. Based on the findings, it can be inferred that utilizing unprocessed Data and models of machine learning with default parameters, the Random Forest model with the bag of words approach yielded the most efficient results, with an F1 Score:0.6580 and an Accuracy Score:0.9629. [18] employed Random Forest (RF) and Support Vector Machine (SVM) as models for classification and assessed their performance by utilizing metrics such as precision, F1-score, accuracy, and recall. [16] employed Deep learning techniques CNN+GRU, LSTM and LSTM + Attention (aLSTM) and evaluated data through Precision, Recall and F1 metrics.

III. MATERIALS AND METHODS

This study delves into comprehending and tackling hate speech on Twitter, taking into account the ever-changing nature of this social media site. The study utilizes a dataset obtained from Twitter and applies sophisticated Natural Language Processing (NLP) methods to categories material into three distinct groups: hate speech, inflammatory language, and neutral content:

A. Research Design

The main objective is to comprehensively evaluate and minimize occurrences of hate speech, acknowledging the significant impact that platforms such as Twitter have on molding public discourse. The study rigorously evaluates multiple NLP models, including LSTM, BERT, Distil BERT, Roberta, Hybrid RNN, and XLNet, using both training and testing datasets to assess their efficacy in detecting subtle hate speech. The study actively contributes to the continuing efforts to combat hate speech by emphasizing the importance of creating an inclusive online environment. The statement emphasizes the effectiveness of advanced NLP techniques in tackling the specific difficulties presented by hate speech on social media.

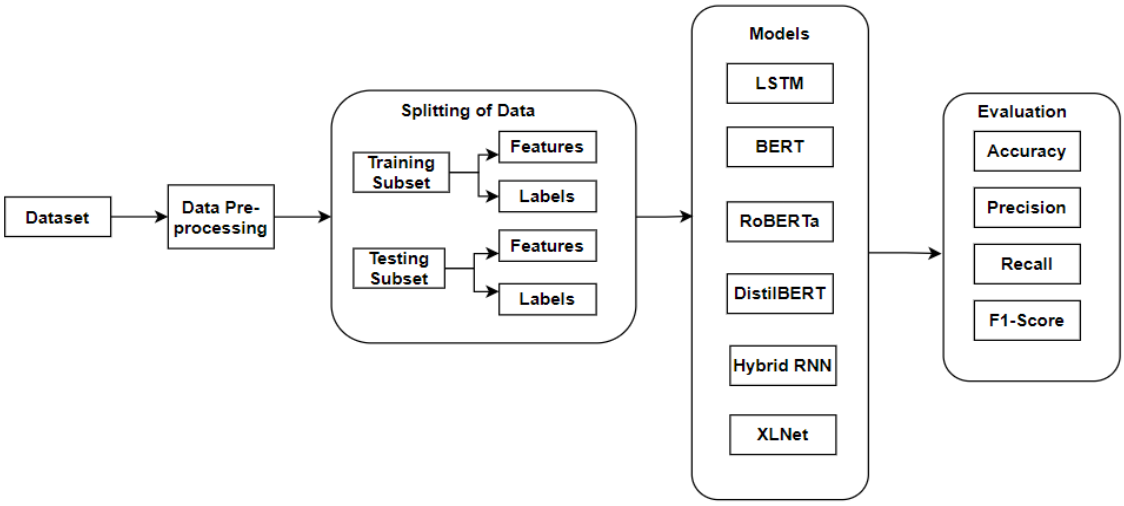

Fig. 1 Proposed Architectural Design

B. Data Analysis

LSTM, BERT, Distil BERT, Roberta, Hybrid RNN, and XLNet are crucial components in the data analysis for this thesis that examines hate speech detection on Twitter. These NLP algorithms will be utilized to analyze the Twitter dataset and make predictions about the classification of text into hate speech, offensive language, or neutral content.

1. Long Short-Term Memory (LSTM)

LSTM is a specific kind of recurrent neural network (RNN) that is specifically developed to tackle the issue of the vanishing gradient problem commonly encountered in standard RNNs. It is proficient in capturing distant relationships in sequential data, making it well-suited for analyzing word sequences in sentences or tweets. LSTM can be employed to represent the context and connections among words in a tweet, facilitating the comprehension of the intricate language employed in hate speech.

2. BERT

BERT is a transformer-based model that comprehensively analyses the complete context of a word by examining the surrounding context on both the left and right sides across all layers of the model. It demonstrates exceptional proficiency in capturing the semantic significance of words and their interconnections within a phrase. To enable hate speech

identification, BERT can undergo fine-tuning by training it on a task-specific la-belled dataset.

3. DistilBERT
DistilBERT is a condensed iteration of BERT, created to enhance computing efficiency while preserving a comparable level of effectiveness. It is appropriate for situations where there is a shortage of computational resources, yet it nevertheless offers strong and reliable methods for text categorization jobs such as identifying hate speech.

4. RoBERTa
RoBERTa represents an advancement over BERT by eliminating the Next Sentence Prediction objective and employing larger mini-batches and learning rates during training. This model demonstrates excellent performance in multiple NLP benchmarks and can be effectively utilized for hate speech detection.

5. Hybrid RNN
Hybrid RNN entails the fusion of diverse recurrent layers, such as LSTM, with other varieties of neural network layers. Hybrid models are capable of capturing both immediate and prolonged connections in the data, rendering them appropriate for analyzing the sequential characteristics of text in tweets.

6. XLNet
XLNet is a transformer-based model that builds upon BERT by taking into account all potential word permutations within a sentence. It captures both forward and backward context, similar to BERT, but with improved computational efficiency. XLNet is a valuable tool for identifying hate speech due to its capacity to analyze intricate word relationships.

C. Evaluation

Thoroughly examining the findings of our study is the last phase in our research process. Thorough evaluation is crucial when working with machine learning algorithms, especially those developed for heart disease prediction. The evaluation process is crucial in determining the model's performance and identifying its strengths and weaknesses. When evaluating binary classification problems, like predicting the presence or absence of heart disease, we depend on various metrics. The metrics offer valuable insights into the effectiveness of the model and allow us to identify its strengths and weaknesses. Commonly used evaluation metrics in this context include accuracy, precision, recall, and F1-score. These metrics provide a thorough understanding of the model's performance and help in making informed decisions about its reliability and applicability in real-world situations.
Accuracy is a metric that indicates the percentage of correctly classified samples in a dataset. The calculation is determined by the following equation:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

The precision metric measures the proportion of correctly detected positive samples to all samples classified as positive. The value given above is calculated by using (2):

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

The recall formula to represent the proportion of identified high priority to the actual is given by (3):

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

The F1-score is a statistical metric that offers a balanced assessment of a model's effectiveness. It achieves this by calculating the harmonic mean of precision and recall. The formula for computing the F1-score is given in (4):

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

D. Processed Diagram

Preprocessing approaches are strongly advised for the task of detecting hate speech on Twitter. In order to improve the effectiveness of the NLP algorithms, the dataset will go through essential preprocessing procedures like noise elimination, tokenization, and stemming. Subsequently, the data will be accurately partitioned into training and testing sets to ensure a thorough evaluation of the model's performance. Throughout the training and evaluation phase, a variety of techniques such as LSTM, BERT, Distil BERT, Roberta, Hybrid RNN, and XLNet will be utilized to evaluate the performance of the hate speech detection models. The algorithms will undergo refinement and assessment utilizing measures such as precision, recall, and F1 score to obtain a thorough grasp of their efficiency. Fig.2 shows the processed diagram for detection.

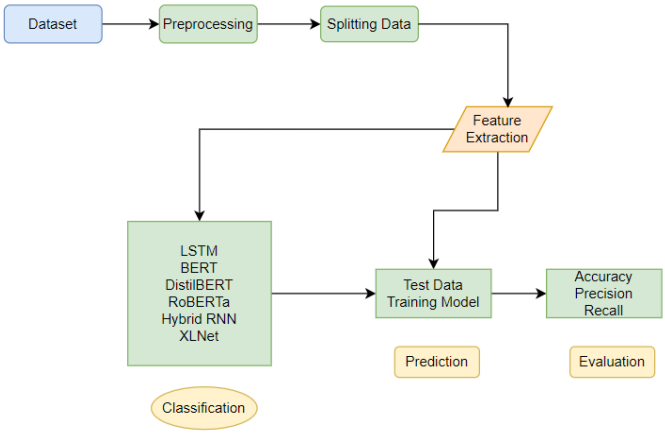

Fig. 2 Processed diagram for detection

IV. IMPLEMENTATION AND RESULTS

This section conducts a classification task using natural language processing (NLP) techniques to predict the presence of hate speech on X. The code employs Pytorch and Hugging Face's transformers library to handle pre-trained transformer-based models such as BERT, XLNet, Distil BERT, and Roberta. In addition, the scikit-learn library is utilized for model evaluation, while the Keras API of TensorFlow is employed for constructing a neural network model. Seaborn and matplotlib.pyplot are utilized for data visualization.
Here is a description of every library:

- Torch: PyTorch serves as the predominant framework for deep learning.
- BertTokenizer and BertForSequenceClassification: These components are utilised for working with BERT models, which are based on Bidirectional Encoder Representa-tions from Transformers. The tokenizer converts text into tokens, while BertForSe-quenceClassification is a pre-trained BERT model that has been fine-tuned specifically for sequence classification tasks.

- Matplotlib: This Python library serves the purpose of creating visual representations of data in a graphical format. It offers a diverse set of functions to produce different types of graphs, both static and interactive.
- Seaborn: Seaborn is a collection of tools for creating visually appealing and technically advanced visualizations of statistical data.
- Collections: You can find other alternatives to lists, tuples, and dictionaries in the Python package known as Collections.
- Scikit-learn: Machine learning may be done with the Python package. For preparing data, choosing a model, and evaluating it, it provides several functions and methods.
- XLNetTokenizer and XLNetForSequenceClassification: Similar to the BERT compo-nents, these are used for working with XLNet models, a generalized autoregressive pretraining model.
- DistilBertTokenizer, RobertaTokenizer and DistilBertForSequenceClassification: These components are utilized for interacting with DistilBERT and RoBERTa models. The tokenizers transform text into individual tokens, whereas DistilBertForSe-quenceClassification is a pre-trained Distil BERT model that has been specifically adjusted for sequence classification tasks.
- tensorflow.keras.models and tensorflow.keras.layers: These components are integral to TensorFlow's Keras API and are utilized for constructing a neural network model. Sequential models consist of a linear stack of layers that are used to construct neural networks. The architecture of the neural network is defined using layers such as Embedding, LSTM, Dense, and Dropout. These layers indicate the potential for employing a straightforward LSTM-based model for the NLP task.
- Train_test_split: To verify a model, a Scikit-learn method separates data into training and testing sets.
- Confusion_matrix: A Scikit-learn function that compares the predicted and actual class labels in order to assess how accurate a classification model is.4.1. Loading the Dataset:
- The read_csv() method in Pandas will then be used to import the csv dataset file into a DataFrame.

A. Data Preprocessing

The text data is divided into smaller units known as tokens. Tokenizers such as BertTokenizer, XLNetTokenizer, DistilBertTokenizer, and RobertaTokenizer are utilized for this purpose. Every token in the text represents a significant element.

B. Training and Testing

The preprocessed and arranged data are divided into three groups by the code using the train test split() function from the Scikit-Learn library: training, testing, and validation. The revised version presents the data distribution as follows: 80% for training, 10% for testing, and 10% for validation. By allocating a larger portion of the data for training the model, we can effectively evaluate its performance and validate its generalization on new data.

C. Feature Extraction

The process of feature extraction plays a crucial role in natural language processing. It involves converting raw text data into a numerical representation that is well-suited for machine learning models. Within the given code, feature extraction is mainly achieved by employing transformer-based models such as BERT, XLNet, DistilBERT, and RoBERTa, which involve tokenization. Tokenization is the process of breaking down text into smaller units called tokens. These models are designed to capture contextual information by assigning each token a numerical vector, referred to as an embedding. The embeddings function as comprehensive and insightful characteristics, capturing semantic connections and contextual subtleties within the text. Feature extraction plays a crucial role in allowing machine learning models to comprehend and make predictions using the intricate patterns and complexities found in the data. For instance, it is necessary for hate speech detection on Twitter.

D. ML Algorithms: Training & Evaluation

The code uses the reduced feature set to train and evaluate many machine learning methods.

E. LSTM

LSTM is a specific kind of recurrent neural network (RNN) that is specifically developed to tackle the issue of the vanishing gradient problem commonly encountered in standard RNNs. LSTM can be employed to represent the context and connections among words in a tweet, facilitating the comprehension of the intricate language employed in hate speech. The LSTM model exhibits impressive performance during training. However, there is a discernible decline in accuracy and other metrics on the validation set, indicating the possibility of overfitting. Additional refinement or investigation of the model's structure could prove advantageous. The accuracy of the test set, which is 87.27%, suggests that the model has a decent ability to generalize to new and unseen data. Figure 3. below shows the confusion matrix for LSTM.

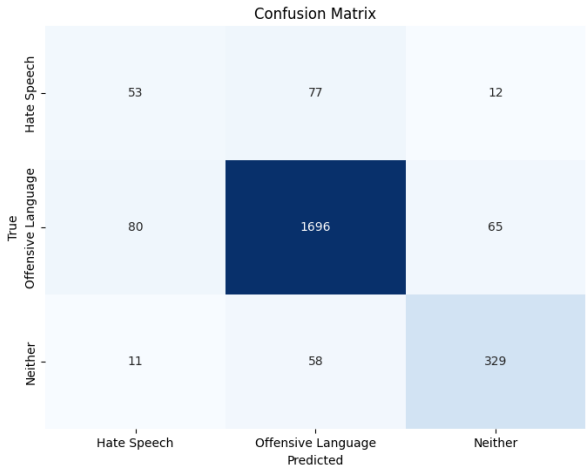

Fig. 3: Confusion Matrix of LSTM

F. BERT

BERT is a transformer-based model that comprehensively analyses the complete context of a word by examining the surrounding context on both the left and right sides across all layers of the model. It demonstrates exceptional proficiency in capturing the semantic significance of words and their interconnections within a phrase. To enable hate speech identification, BERT can undergo fine-tuning by training it on a

task-specific labeled dataset. The evaluation results demonstrate the performance of a classification model, possibly leveraging BERT, on a dataset containing three classes. The model's overall accuracy is around 91.39%, indicating its proficiency in making accurate predictions across various classes. Upon closer examination, a more in-depth analysis uncovers different metrics for precision, recall, and F1-score for each class. The model demonstrates a strong capability to accurately identify instances within the non-hate speech class (class 1), as evidenced by its high precision and recall. On the other hand, the detection of instances of hate speech (class 0) poses a difficulty, resulting in lower precision and recall values. Class 2 demonstrates strong precision and recall. The overall F1-score is 72%, highlighting the importance of carefully considering both precision and recall for all classes. The model's performance is generally strong, but there is room for improvement, especially in increasing its sensitivity to instances of hate speech.

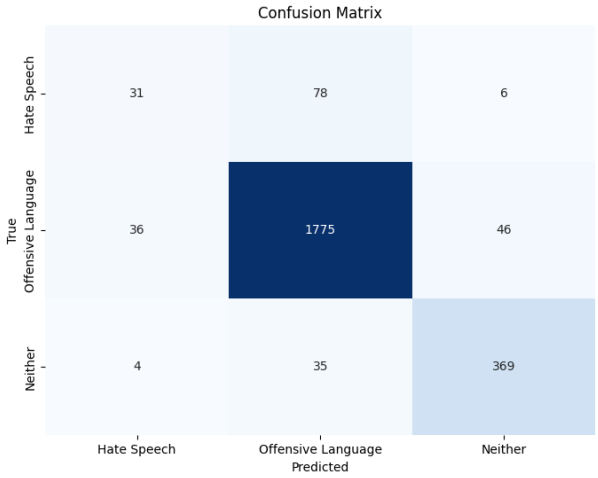

Fig. 4 Confusion Matrix of BERT

G. DistilBERT

DistilBERT is a condensed iteration of BERT, created to enhance computing efficiency while preserving a comparable level of effectiveness. It is appropriate for situations where there is a shortage of computational resources, yet it nevertheless offers strong and reliable methods for text categorization jobs such as identifying hate speech. The results presented here offer an assessment of a classification model, possibly implemented using DistilBERT, for a task that encompasses three distinct classes. The model's overall accuracy is around 91.47%, demonstrating its ability to accurately predict across hate speech, offensive language, and neutral categories. The precision of positive predictions can vary across different classes, indicating differences in accuracy. The identification of non-hate speech (class 1) demonstrates a high level of precision, reaching 95%. This indicates a strong capability to accurately detect instances falling under this category. Precision for the neutral category (class 2) is also commendable, standing at 87%. Nevertheless, when it comes to instances of hate speech (class 0), the precision is relatively lower at 44%. This suggests that there is room for improvement in accurately identifying instances within this particular category. The model demonstrates a well-balanced distribution of recall values among different classes. It achieves a macro average F1-score of 76%, indicating its overall effectiveness. Although the performance is commendable, there are potential areas for improvement, particularly in increasing the precision for identifying instances of hate speech.

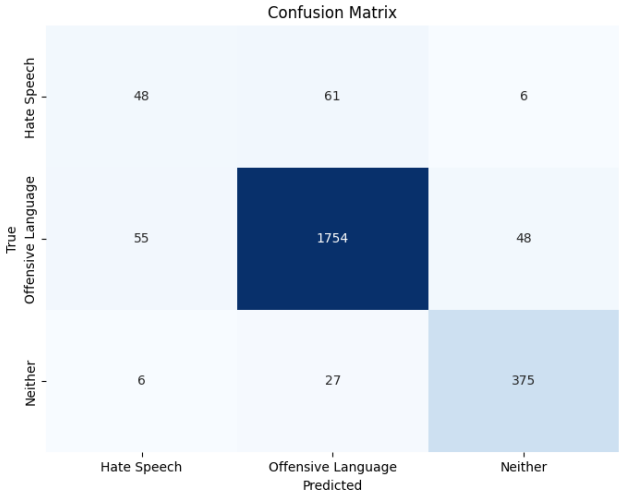

Fig. 5. Confusion Matrix of DistilBERT

H. RoBERTa

The results provided offer an evaluation of a classification model, likely utilizing RoBERTa, for a task that involves three distinct classes. The model demonstrates a high level of accuracy, around 91.05%, in accurately predicting hate speech, offensive language, and neutral categories. The measurement of accuracy in positive predictions shows variation among different classes. The precision for non-hate speech (class 1) is quite high at 96%, which demonstrates a strong capability to accurately detect instances falling under this category. The precision for the neutral category (class 2) is impressive, standing at 87%. Nevertheless, in cases of hate speech (class 0), the precision is relatively lower at 41%, indicating the possibility of enhancing the accuracy in identifying instances within this particular category. The model demonstrates a well-balanced distribution of recall values among different classes, resulting in a macro average F1-score of 76%. In general, although the accuracy is commendable, there is room for improvement, especially in increasing the precision for instances of hate speech in order to further enhance the performance of the model.

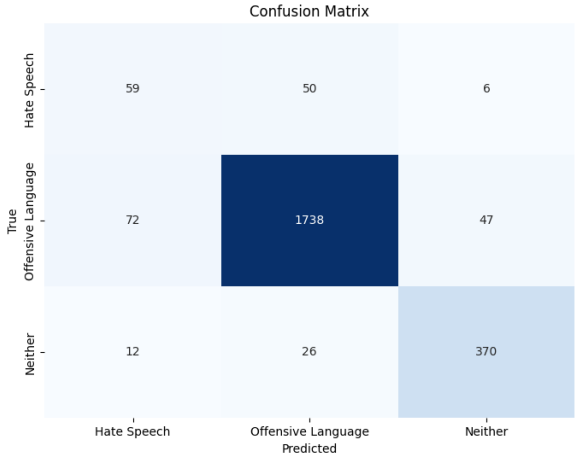

Fig. 6. Confusion Matrix of RoBERTa

I. Hybrid RNN

Hybrid RNN entails the fusion of diverse recurrent layers, such as LSTM, with other varieties of neural network layers. Hybrid models are capable of capturing both immediate and prolonged connections in the data, rendering them appropriate for analyzing the sequential characteristics of text in tweets. The evaluation results presented here provide a comprehensive overview of the classification model's performance. It is highly probable that the model employed a Hybrid Recurrent Neural Network (RNN) to accurately categories content from CF users into three distinct classes. The model exhibits an overall accuracy of around 90.72%, demonstrating its competence in making precise predictions across hate speech, offensive language, and neutral categories. It is worth mentioning that the precision rates for non-hate speech (class 1) and the neutral category (class 2) are both quite high, with 93% and 92% respectively. Nevertheless, the precision for instances of hate speech (class 0) is relatively lower, standing at 45%. This suggests that there is room for improvement in accurately identifying instances within this category. It is worth noting that the model's performance is strong for class 1 and class 2 instances, but it falls behind when it comes to hate speech instances (class 0). The F1-score of 90% demonstrates the model's high level of effectiveness. Although the performance is commendable, there is room for improvement, particularly in increasing sensitivity to instances of hate speech.

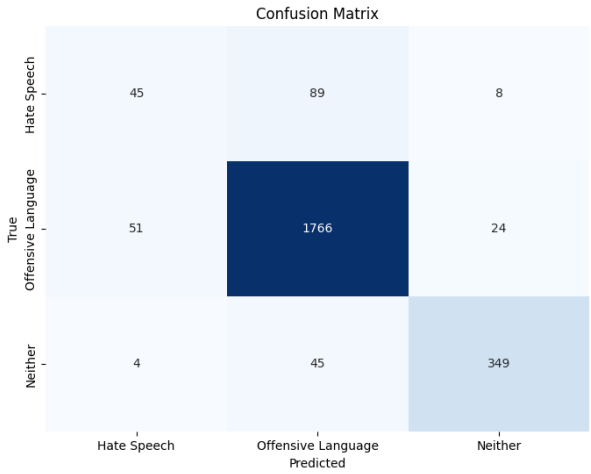

Figure 7. Confusion Matrix of Hybrid RNN

J. XLNet

XLNet is a transformer-based model that builds upon BERT by taking into account all potential word permutations within a sentence. It captures both forward and backward context, similar to BERT, but with improved computational efficiency. XLNet is a valuable tool for identifying hate speech due to its capacity to analyze intricate word relationships. The evaluation results demonstrate the efficacy of a classification model, possibly implemented using XLNet, in effectively addressing a task with three distinct classes. The model's overall accuracy is around 91.68%, highlighting its proficiency in making accurate predictions across various classes. The model demonstrates exceptional precision and recall for the non-hate speech class (class 1), highlighting its strong ability to accurately detect instances in this category. Nevertheless, there are difficulties in accurately identifying instances of hate speech (class 0), as evidenced by lower precision and recall values. Class 2 exhibits high precision and recall values. The F1-score for the macro average is 74%, highlighting the significance of maintaining a balanced approach to precision and recall for all classes. The F1-score is 91%, indicating the model's strong performance in multi-class classification tasks.

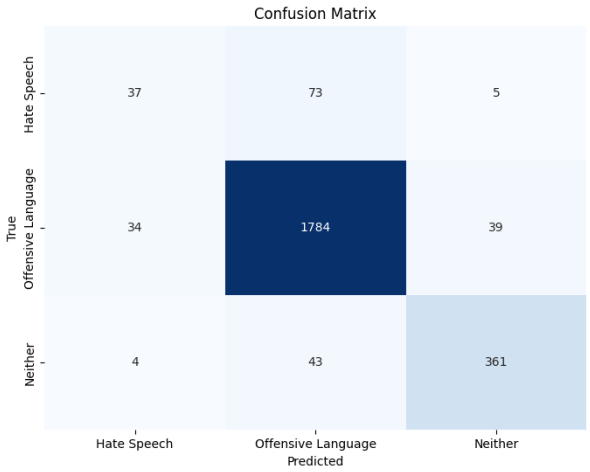

Fig. 8. Confusion Matrix of XLNet

V. DISCUSSION

Various classification models, such as LSTM, BERT, RoBERTa, DistilBERT, Hybrid RNN, and XLNet, were evaluated to categories tweets into hate speech, offensive language, and neutral content. This evaluation yielded valuable insights. The LSTM model exhibits impressive performance during training, but there are concerns regarding potential overfitting. This is evident from a decrease in accuracy on the validation set. In Figure 9, the accuracy of BERT is 91.39%, demonstrating strong precision and recall for non-hate speech instances. However, it faces difficulties in accurately identifying hate speech instances. RoBERTa achieves an accuracy of 91.05%, demonstrating strong precision in classifying non-hate speech and the neutral category. However, there is still potential for improvement in accurately identifying instances of hate speech. DistilBERT demonstrates a high accuracy rate of 91.47%, indicating its effectiveness. However, it could benefit from improvements in precision when it comes to identifying instances of hate speech. The Hybrid RNN model achieves an accuracy of 90.72%, demonstrating high precision in classifying non-hate speech and the neutral category. However, it exhibits lower precision when identifying instances of hate speech. XLNet achieves an accuracy of 91.68%, showcasing its impressive precision and recall in identifying non-hate speech.

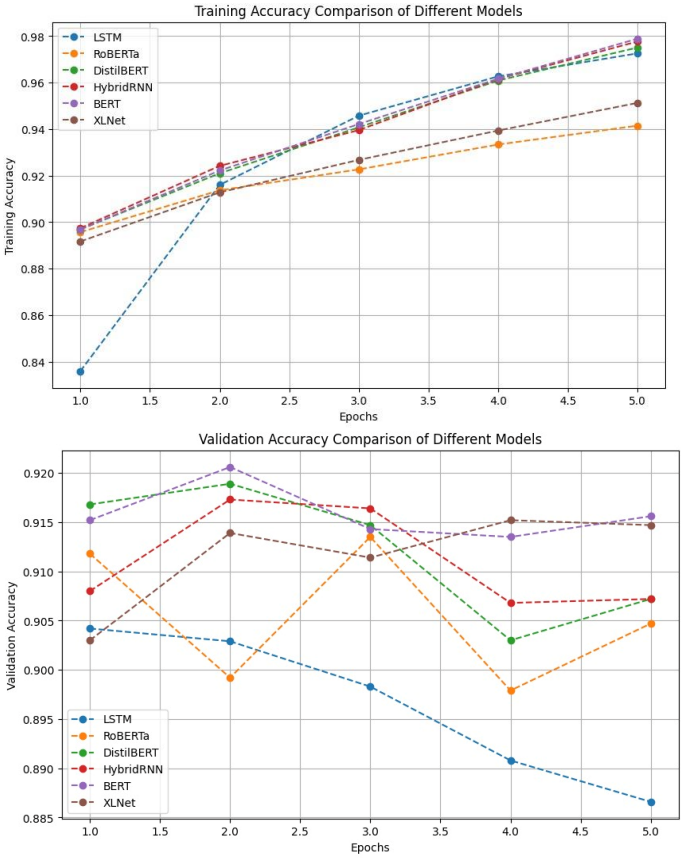

Fig. 9. Training and Validation Accuracy of different models

The comparison accuracy of various models is shown in Table 1. In general, although these models demonstrate impressive levels of accuracy, the detailed performance metrics highlight the necessity for further improvement, specifically in increasing the ability to identify instances of hate speech in all cases.

This study presents a noteworthy endeavor in tackling the pressing problem of hate speech on Twitter. The research methodology used is strong, involving a comprehensive approach to understand and address hate speech. Upon closer examination, it becomes evident that the study possesses both commendable qualities and opportunities for enhancement across different facets.

A dataset from Twitter is used to ensure relevance to the platform being investigated. The utilization of advanced Natural Language Processing (NLP) techniques enhances the depth of the research design. The study's comprehensiveness is enhanced by the inclusion of various NLP models, such as LSTM, BERT, DistilBERT, RoBERTa, Hybrid RNN, and XLNet, which offer a range of different approaches.

Table 1 Accuracy of Models.

Number	Model	Accuracy
0	LSTM	87.27
1	BERT	91.39
2	RoBERTa	91.05
3	DistilBERT	91.47
4	HyBrid RNN	90.72
5	XLNet	91.68

Taking into account the ever-changing landscape of social media rigorous research methodology is examine for hate speech on Twitter. The study assesses the effectiveness of various NLP models, such as LSTM, BERT, RoBERTa, DistilBERT, Hybrid RNN, and XLNet, in identifying hate speech. The study offers a comprehensive analysis of the various models, highlighting their respective strengths and weaknesses. It underscores the importance of further improvement and heightened awareness towards instances of hate speech. The study would benefit from providing a more explicit justification for the selection of diverse models. The thorough assessment and uniformity in visual aids enhance transparency. The study acknowledges the impressive performance and appropriately identifies areas for improvement. It also emphasizes the need for ongoing refinement. However, to enhance the critical evaluation, the study could provide specific recommendations or suggest future research directions. The research provides a systematic analysis of hate speech, offering valuable insights into the field of NLP and social media discourse.

VI. CONCLUSION

Ultimately, this study examines the complex issue of hate speech on Twitter, acknowledging the ever-changing nature of social media. The study utilizes a strong research design and a dataset from Twitter. It employs advanced Natural Language Processing (NLP) techniques to categories content into hate speech, inflammatory language, and neutral expressions. The main goal is to thoroughly assess and reduce occurrences of hate speech, recognizing the significant influence of platforms such as Twitter on shaping public discussions.

A wide range of NLP models, such as LSTM, BERT, RoBERTa, DistilBERT, Hybrid RNN, and XLNet, are thoroughly evaluated using training and testing datasets. The models demonstrate impressive accuracy, with LSTM achieving 87.27% and XLNet achieving 91.68%. A thorough examination uncovers distinct advantages and obstacles for each model, highlighting the importance of continuous improvement, especially in enhancing the ability to detect instances of hate speech. The potential bias or lack of diversity in the dataset could have an impact on the model's ability to generalize, highlighting the importance of using more representative datasets. It is crucial to thoroughly analyze algorithmic fairness due to the ethical concerns arising from the biases present in Natural Language Processing models, which often mirror societal biases. To overcome these limitations and improve the efficacy of hate speech detection, it is imperative for future research to focus on a number of important areas. By incorporating ethical AI frameworks and user feedback loops, we can improve transparency, accountability, and consider the human context that algorithms often neglect.

REFERENCES

[1] Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. "Deep Learning for Multi-Class Identification from Domestic Violence Online Posts." IEEE Access 7 2019: 46210-46224. doi: 10.1109/ACCESS.2019.2908827

[2] Ketsbaia, L., Issac, B., & Chen, X. "Detection of Hate Tweets using Machine Learning and Deep Learning." IEEE 19th International Conference

on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 2020, pp. 751-758. doi: 10.1109/TrustCom50675.2020.00103.

[3] Tommasel, Antonela, Juan Manuel Rodriguez, and Daniela Godoy. "Textual Aggression Detection through Deep Learning." In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), edited by Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, 177-187. Santa Fe, New Mexico, USA: Association for Computational Linguistics, August 2018. https://aclanthology.org/W18-4421.

[4] Ali, Mohsan, Mehdi Hassan, Kashif Kifayat, Jin Young Kim, Saqib Hakak, and Muhammad Khurram Khan. "Social Media Content Classification and Community Detection Using Deep Learning and Graph Analytics." Technological Forecasting and Social Change 188 (2023): 122252. https://doi.org/10.1016/j.techfore.2022.122252.

[5] Al-Garadi, Mohammed Ali, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. "Natural Language Model for Automatic Identification of Intimate Partner Violence Reports from Twitter." Array 15 (2022): 100217. ISSN 2590-0056. https://doi.org/10.1016/j.array.2022.100217.

[6] Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., & Plaza, L. "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data." IEEE Access 8 (2020): 219563-219576. doi: 10.1109/ACCESS.2020.3042604.

[7] Hu, H., Phan, N., Chun, S.A., et al. "An Insight Analysis and Detection of Drug-Abuse Risk Behavior on Twitter with Self-Taught Deep Learning." Computational Social Networks 6, 10 (2019). https://doi.org/10.1186/s40649-019-0071-4.

[8] Ta, Hoang Thang, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. "Multi-Task Learning for Detection of Aggressive and Violent Incidents from Social Media." Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico; Dalat University, Lam Dong, Vietnam; College of Information Science and Technology, University of Nebraska Omaha, Omaha, Nebraska, USA, 2022.

[9] Alatawi, H. S., Alhothali, A. M., & Moria, K. M. "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT." IEEE Access 9 (2021): 106363-106374. doi: 10.1109/ACCESS.2021.3100435.

[10] Muneer, Amgad, and Suliman Mohamed Fati. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter." Future Internet 12, no. 11 (2020): 187. https://doi.org/10.3390/fi12110187.

[11] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. "Effective Hate-Speech Detection in Twitter Data Using Recurrent Neural Networks." Applied Intelligence 48 (2018): 4730–4742. https://doi.org/10.1007/s10489-018-1242-y.

[12] , A., Singh, A., Bhadauria, H.S., Virmani, J., Kriti (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In: Jain, S., Paul, S. (eds) Recent Trends in Image and Signal Processing in Computer Vision. Advances in Intelligent Systems and Computing, vol 1124. Springer, Singapore. https://doi.org/10.1007/978-981-15-2740-1_17

[13] Mutanga, Raymond T., Nalindren Naicker, and Oludayo O. Olugbara. "Detecting Hate Speech on Twitter Network Using Ensemble Machine Learning." International Journal of Advanced Computer Science and Applications 13, no. 3 (2022). DOI:10.14569/IJACSA.2022.0130341.

[14] Wu, T., Liu, S., Zhang, J., & Xiang, Y. (2017). "Twitter Spam Detection Based on Deep Learning." Proceedings of the Australasian Computer Science Week Multiconference on - ACSW '17. doi:10.1145/3014812.3014815.

[15] Madichetty, S., & Muthukumarasamy, S. "Detection of Situational Information from Twitter During Disaster Using Deep Learning Models." Sādhanā 45 (2020): 270. https://doi.org/10.1007/s12046-020-01504-0.

[16] Polychronis CharitidisDoropoulos, Stavros Vologiannidis, Ioannis Papastergiou, Sophia Karakeva,Stavros. (2020). Towards countering hate speech on social media. Online Social Networks and Media.

[17] Fortuna Paula, S. J. (2020). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 6786--6794). European Language Resources Association.

[18] Mohapatra, S. a.-C. (2021). Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques. Applied Sciences, 8575.

[19] Vidgen, B. &. (2020). Detecting weak and strong Islamophobic hate speech on social media. Journal of Information Technology & Politics, 66-78. .

[20] Qureshi, K. &. (2021). Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text. IEEE Access, 109465-109477.

[21] Wiedemann Gregor, R. E. (2019). UHH-LT at SemEval-2019 Task 6 Supervised vs. Unsupervised Transfer Learning for Offensive Language Detection",. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 782--787).

[22] Modha Sandip, M. ,. (2019). DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation. In "Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 577--581). Association for Computational Linguistics.

[23] Muhammad Usman Shahid Khan, A. A. (2021). HateClassify: A Service Framework for Hate Speech Identification on Social Media. IEEE Internet Computing, 40-49.

[24] Agarwal S, C. C. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. Expert Syst Appl.

[25] Marco Siino, E. D. (2021). Detection of Hate Speech Spreaders using convolutional neural networks. Conference and Labs of the Evaluation Forum.

[26] Vashistha, N. A. (2020). Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media. information.

[27] Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In Proceedings of the 2020 IEEE Winter Conference on Applications on Computer Vision (WACV), 1459-1467.

[28] N. D. Srivastava, S. Y. (2020). Combating Online Hate: A Comparative Study on Identification of Hate Speech and Offensive Content in Social Media Text. IEEE Recent Advances in Intelligent Computational Systems (RAICS), 47-52.

[29] Frenda Simonaa, G. B.-y.-G. (2019). Online Hate Speech Against Women: Automatic Identification of Misogyny and Sexism on Twitter. Journal of Intelligent & Fuzzy Systems, .

[30] Moon Jihyung, C. W. (2020). BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media (pp. 25--31). Association for Computational Linguistics.

[31] Mathew, B. S. (2021). A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence.

[32] Bhavesh Pariyani, K. S. (2021). Hate Speech Detection in Twitter using Natural Language Processing. Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks.

[33] Sylvia Jaki, T. D. (2019). Right-wing german hate speech on twitter: Analysis and automatic detection. arXiv.

[34] Köffer Sebastian, R. D. (2018). Discussing the Value of Automatic Hate Speech Detection in Online Debates. Multikonferenz Wirtschaftsinformatik (MKWI 2018.

[35] Shervin Malmasi, M. Z. (2017). Detecting Hate Speech in Social Media. (pp. 467-472.). Proceedings of Recent Advances in Natural Language Processing (RANLP).

[36] Kent, S. (2018). German Hate Speech Detection on Twitter. 14th Conference on Natural Language Processing .

[37] Jahan, M. S. (2021). A systematic review of Hate Speech automatic detection using Natural Language Processing.

[38] Thomas Davidson, D. W. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. To appear in the Proceedings of ICWSM.

[39] Aria Nourbakhsh, F. V. (2019). An Ensemble Approach to Hate Speech Detection. Proceedings of the 13th International Workshop on Semantic Evaluation, 484-488.

[40] Sindhu Abro, S. S. (2020). Automatic Hate Speech Detection using Machine Learning: A Comparative Study. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8.

[41] Magu Rijul, J. K. (2017). Detecting the Hate Code on Social Media. Proceedings of the International AAAI Conference on Web and Social Media.

[42] Ari Muzakir, K. A. (2022). Classification of Hate Speech Language Detection on Social Media: Preliminary Study for Improvement. International Conference on Networking, Intelligent Systems and Security.

[43] Shimaa M Abd El-Salam, M. M. (2019). Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. Informatics in Medicine .

[44] UNESCO. (2018). World Trends in Freedom of Expression and Media Development: Global Report 2017/2018. United Nations Educational, Scientific and Cultural Organization.

[45] Fortuna, P. &. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys.

[46] Md Saroar Jahan, M. O. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing.

[47] Badjatiya, P. a. (2017). Deep Learning for Hate Speech Detection in Tweets. International WorldWide Web Conferences Steering Committee, (pp. 759-760).

COMPETING INTERESTS

The authors have declared that no competing interests exist

AUTHORS

First Author – Samar Shabbir, Department of Software Engineering, University of Lahore, Pakistan.

Second Author – Ahmad Salman Khan, Ph.D. (SE), Department of Software Engineering, University of Lahore, Pakistan.

Third Author – Ahtisham Ahmad, Department of Software Engineering, University of Lahore, Pakistan.

Fourth Author – Muhammad Waqas, Ph.D. (EE), Iqra National University, Peshawar, Pakistan.

Fifth Author – Mansoor Qadir, Ph.D. (CS), CECOS University of IT & Emerging Sciences, Peshawar, Pakistan.

Sixth Author – Mubina Zaka, Department of Computer Science & Information Technologies Hazara University Mansehra.

Seventh Author – Afshana Ishaq, Department of Electronics, University of Engineering & Technology Abbottabad Campus, Pakistan.

Correspondence Author – Mansoor Qadir, Ph.D. (CS), CECOS University of IT & Emerging Sciences, Peshawar, Pakistan.