OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 06 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

DETECTING ONLINE HARASSMENT BASED ON SOCIAL MEDIA TEXT BY USING ENSEMBLE LEARNING

Hamna Iqbal
Department of Computer Science, University of Southern Punjab, Multan.

Muhammad Sabir*
Department of Computer Science, University of Southern Punjab, Multan.

Areeba Razzaq
Department of Computer Science, University of Southern Punjab, Multan.

Jahanzeb Munir
Department of Information Technology, The Islamia University of Bahawalpur.

****Corresponding author: Muhammad Sabir (**muhammadsabir@usp.edu.pk**)***

Article Info

Abstract

Social media is now a tool for passing information, as well as a place to communicate but at the same time, a platform for cyberbullying, hate speech, and Offensive Language. In order to struggle this increasing problem, methods associated with machine learning, for example ensemble learning, are being employed to identify offensive material on such sites. A process with the name of Ensemble learning uses the predictions of many models of classifiers for instance Logistic Regression, SVC and Random Forest in order to classify and reduce the errors as much as possible. When researchers preprocess text from the social media, they were able to preprocess out such features as “Hate Speech,” “Cyberbullying,” and “Offensive Language.” With this approach, a study was accomplished up to 93% accuracy, meaning that ensemble learning is efficient at distinguishing online harassment and can be optimized more by other language model progressions plus sentiment analysis.

Keywords:

Online Harassment, Ensemble Learning, Social Media, Machine Learning, Cyberbullying.

Introduction

In addition, the effect of other developing communication platform, such as social media which has also changed the nature of communication and sharing information, has given rise to a significant issue: cyberbullying. (Azeez & Fadhal, 2023) It is also defined as online harassment On social media, using Use of foul language and conducting on self in a manner that may cause harm to another person has become dangerous behaviors. (Sultan et al., 2023) Cyberbullying means the use of Information technology devices to pass on threats of harm or threats of confined space has appropriate means to stalking, threatening, and harassing persons. The effects of cyber bullying and stalking are dangerous and disturbing to society with an urgent need to take appropriate actions positive social media use. Since, the online behaviors are not fixed, conventional methods of detection often prove inadequate to overcome these difficulties, this work puts forward an ensemble learning based method to enhance the reliability and accuracy of detecting online harassment.

The increasing cases of cyber harassment can be supported by a Pew Research Centre study, which noted that cases of cyberbullying, trolling, and sexual harassment have escalated with over half of the US adulthood population being victims. This dread can be backed by a study done by (Alam et al., 2021), where there was a percentage increase of 2% of cyber harassment cases from the year 2007 to 2019 because of the mobile devices and social networks usage among the users of the digital platform. Cyberbullying, as one of the most common subtypes of harassment over the internet, has wide-spread effects on the victims that are often negative and long-term: fear, anxiety and depression, up to self-harm and suicides (Semangern et al., 2019). This is not good for individuals, but hate speech on the internet affects entire communities as this has singled out as the reason for the increased violence against minorities all over the world (Khairy et al., 2023).

This research focuses on one form of computer mediated abuse; cyber bullying by proposing an ensemble learning model to identify posts on social media that are likely to pose harm to users. The study is based on a dataset obtained from Kaggle used to categorize text into, “Hate Speech,” “Cyberbullying,” and “Offensive Language” after applying basic Natural Language Processing methodologies such as tokenization, stop word removal, and stemming. (Sreevidya et al., 2024) The multiple classifier system, the ensemble model using logistic regression, support vector machine, decision tree and random forest classifiers in the soft voting method accuracy in the classification of abusive content.(Alqahtani & Ilyas, 2024) The issue is linked to the fact that attention to such forms of cyber abuse is shifting to social networks and other online platforms where conventional filtering approaches are no longer capable of drawing line between flippant and aggressive language as most discourse on social media is informal (Khairy et al., 2023) This research work presents an enhanced technique for automated identification of cyber harassment and its impact, which can be highly parapsychological and sociological for women and youths.

ii. Related Work:

(Hegde et al., 2023) This paper discusses the application of the ML models that are used for the detection of cyberbullying on SNS with the help of NLP, feature extraction, SA, and CV. The Support Vector Machine (SVM) model is found to have the highest increase in the accuracy of identifying cases of cyberbullying by the analysis done on different algorithms in the current study. In sum, the research using positive approaches in formulating strategies that could help in the prevention or reduction of cyberbullying aims at enhancing the safety of social media users’ online environment.

It is therefore important to have efficient detection mechanisms in order to eradicate the vice of cyber bullying. For this purpose, ensemble learning is a machine learning technology that applies two or more

models together in order to improve the total precision and resilience. These can be recognized using techniques which employ natural language processing which include name entity recognition, the sentiment analysis algorithm, and part-of-speech tagging. Many domains such as spam filtering and text classification are indicative of the potential of ensemble learning.(Azeez & Fadhal, 2023) .

(Bai & Malempati, 2023) Although online insulting is still a growing problem, Bai and the Mallampati attempt to address it in their paper titled "Ensemble Deep Learning (EDL) for Cyber- bullying on Social Media." In this practice, two approaches of the Ensemble Deep Learning model are used for classification and analysis of the word, image, and video data. BERT is applied to the textual data, while CNNs, RNNs and DBNs are explored for the other. To improve the classification of the bullying communications they introduce sentiment analysis in general, and Aspect based Sentiment Analysis (ABSA) specifically.

The paper suggests an ensemble stacking machine learning model for the detection of cyber bullying on the Twitter platform. Four feature extraction methods are used in the model: Bag of Words, TF-IDF, Word2Vec and GLOVE along with five Machine Learning algorithms, including Decision Tree, Random Forest, Linear Support Vector Classification, Logistic Regression and K- Nearest Neighbors. (D. S. Aabdalla & Vasumathi, 2024)

This paper analyses the domain of cybersecurity and artificial intelligence while utilizing Naive Bayes together with Bi-LSTM to identify cyberbullying linked to religion, age, ethnicity or gender on the platform of Twitter. (Orelaja et al., 2024) The work draws Sentiment140 dataset, which is originally constructed for the specific purpose of selective sentiment classification and then modified for the purpose of identifying cyberbullying. The results thus show how both models work and also how the Bi-LSTM model is able to identify more complex cyberbullying incidences.

(Hoque & Seddiqui, 2024) In this paper, we have concentrated on detecting of cyberbullies in Bengali language which is low resource language in terms of number of available tools and research in natural language processing field. The authors test a number of pre-processing options, feature selection algorithms, and machine learning algorithms for classification of texts containing cyberbullying. Some of the approaches used by the participants included; the classical machine learning models, including the SVM, MNB, RF, and LR, deep learning models such as LSTM, BILSTM, CNN-BILSTM, and a transformer-based pre-trained model (BERT).

(Tolba et al., 2021) Regarding issues rising from substantially imbalanced data on SM locations, Tolba et al., working on the detection of online harassment, proposed the use of hybrid ensemble methods. Three models of word-embedding (word2vec, Glove, SSWE) and nine methods for handling the class imbalance problem are described and analyzed by the authors who focus on feature representation, unbalanced data handling, and supervised learning strategies.

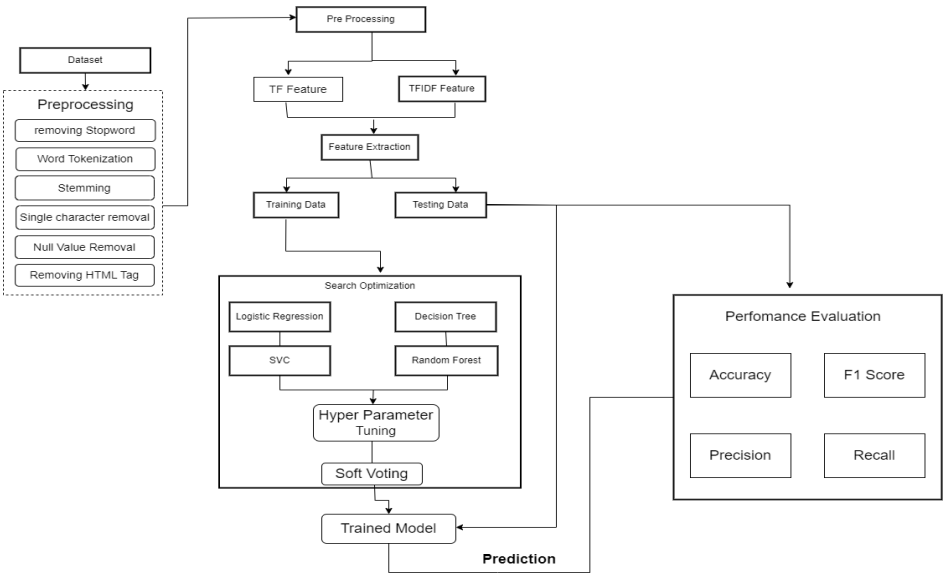
(Davidson et al., 2019) The study offers an ensemble learning base in formulating general rules that may be used in determining online abuse in text from any social media. The model is trained with a set of machine learning models to classify post as harassing or non-harassing based on the text features derived and extracted from the social media postings and cleaned using NLP tools.

The paper also assesses the identification of toxic messages that is taken with reference to social media discourse and approval of ensemble learning. (Hartmann et al., 2019) The proposed model will be trained in logistic regression followed by a random forest classifier since it is an ensemble of methods. This paper aims at enhancing both the classifiers' performance through ensemble learning and consequently develop a more efficient model for screening out unfriendly linguistic expressions in engagements with social media.

(Alqahtani & Ilyas, 2024) The objective of the paper is to analyze six various types of cyberbullying in the tweets. The research also points out the challenges that are associated with definition of cyberbullying on the social media sites including the complexities and the variety of the hurtful acts. The research aims at improving the level of accurate identification of cyberbullying through the strength of ensemble classifiers and thus the efficiency of the investigation and response procedures of social media.

iii. The Proposed Model:

In this study, the approach used in constructing the ensemble learning model for identifying cyber harassment from texts posted on the social media. The section presents how the dataset was obtained from Kaggle website, the cleaning process of data, feature engineering techniques, the algorithms employed and the assessment metrics utilized.



A. Data Pre-Processing

First one is preprocessing in which the raw data is cleaned and prepared for the model. The following steps were applied to the dataset:

- **Removing Stop Words:** These are words such as ‘and,’ ‘the,’ and ‘is’ which the use of a stop word list filter eliminates as they contribute the sentence meaning minimally. These words are discarded for the purpose of noise that is required to be trimmed down in order to enhance the performance of the text processing model. Excluding these things, the model targets more important words that would help in identifying such patterns as harassment or abusive language.
- **Word Tokenization:** Tokenization is the process of segmenting the sentence into discrete words or “tokens.” Tokenization is critical during the feature extraction since it presents the model with an opportunity to examine each of the words in detail. Tokenization assists the system in being able to identify important terms that are most frequently used in a text to pinpoint patterns that the system needs to use in identifying online harassment.
- **Stemming:** Stemming reduces words to their base form for example “running” will be “run” This helps to simplify the text by cheapening related words as one word. Such normalization is useful to avoid repetition and enhances the model’s performance in terms of the generalization because by reducing the Word Vectors it enhances the computation at the same time.
- **Single Character Removal:** Single graphic characters like single letters, or simple symbols are usually without much interpretive significance in textual studies. These are regarded as noise which if included in the data set, will distort or affect the results significantly. While removing them filters

some noise that may exist in a text string and thus make it easier for the model to learn correlations that are more significant.

- **Null Value Removal:** Lack of values or/and null values present in the entries do not add useful or worthwhile information to the existing datasets. If left in, they may introduce errors, or may reduce the accuracy of the model, or unsatisfactory solutions may be generated. These are eliminated to allow only accurate data to be fed to the model and thus improving its reliability in the outputs generated.
- **Removing HTML Tags:** Some text data from social media or by web scraping is attached to the HTML tags which are not part of the text's content, for example, , <a>. It removes all these tags which may interfere with the pure textual content that is useful in the analysis hence making the model to have minimal or no interference of noise when it is identifying the important linguistic features.

B. TF-IDF:

Next to data pre-processing, is a part of the text transformation procedure that should be performed before towing the TF-IDF algorithm. It includes procedures like elimination of Common English words, word breaking down into words, and performs actions on stemming which reduces differences of a word.

Term Frequency (TF): The name given to the act of counting the occurrences of a given word in a document. The idea is based on the hypothesis that the content of the text may be identified from the most frequently used words.

Term Frequency-Inverse Document Frequency (TF-IDF): This method assigns relative frequencies of terms to multiply for a given word by a factor that will depends on the size of the corpus, a factor that will enhance the weights assigned to such terms especially in the case where some terms in the document are more important due to their rarity in the entire work. Applying the value of percentage to a Textual classification problem is often more efficient as compared to applying the TF-IDF.

C. Ensemble Classification

Ensemble learning is a technique of the machine learning in which more than one model is used to generate the final result on the assumption that the error rates are reduced by combining the outputs of different models. With reference to the given diagram, Ensemble Learning is probably achieved under the "Soft Voting" technique. This approach makes use of a large set of classifiers trained on the same set of training data; the voting process is conducted by weighted voting.

Ensemble learning is a group of methods that use a set of models which when combined yield a higher accuracy and better generality than when each mode is used individually.

Accuracy: Measures the proportion of the total number of instances where the target variable was correctly identified – both the positive and negative ones – out the total number of instances predicted.

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+TN+FP+FN}} \qquad \text{(Eq.10)}$$

Where:

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Precision: Calculates the number of true positive which represents the instances that are classified as harassment and indeed harassment.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \qquad \text{(Eq11)}$$

Recall: Tells about model's ability to get best fit in actual positive instances that is harassment.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \qquad \text{(Eq12)}$$

F1 Score: The formula derived mean between both precision and recall which is referred to as the harmonic mean.

$$\boldsymbol{F1} = \mathbf{2} \times \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision + Recall}} \qquad \text{(Eq13)}$$

Logistic Regression: This method of analysis Logistics Regression was observed to be offering high levels of accuracy more especially where we had observed high levels of precision since this is an algorithm that uses linear decision boundaries. However, it seems to slightly lower the recall especially when it comes to the more complex forms of harassment patterns in the text.

SVC: In fact, the feature of Support Vector Classifier using RBF kernel shown ability to model more complex decision region which is high recall. However, this led to also sometimes a negative outcome whereby the level of accuracy was lowered due to concurrency of profit margin and misidentification.

Decision Tree: It was observed that the Decision Tree model was very frail and vulnerable to the noise and over fitting more so when trained on small features. It performed well in detecting harassment cases that would have been overlooked by conventional approaches; the true recall was high albeit at the cost of low precision and many false positives.

Random Forest: As mentioned in the Random Forest section, Random Forests were able to concern precision and recall because constructing number of decision trees and averaging results of all of them minimize overfitting and gave acceptable performance whenever a number of features interacted.

Ensemble Model Performance

When the hyper parameters were optimized for every classifier separately, all these models were ensemble with Soft Voting. Whereas in soft voting, the final decision is obtained by a weighted average of the prediction probability of each model rather than a simple voting system which takes into account the confidence level of each model in the final decision.

Soft Voting:

Soft voting is a voting system which is employed in ensemble learning where multiple models or classifiers are integrated to designate the final decision chances of probabilities assigned for it. Soft voting, in unison, instead of making a single value prediction of a class, each single classifier model like the logistic regression, decision trees or support vector machine gives out a probe for each class that is likely. The final prediction is done by combining these probabilities from all the models and choosing the class which has on average the highest probability.

iv. Results and Discussion

4.1 Dataset

The given dataset contains count values for hate speech, offensive language, and neither classes with the final class of labels (for example, 1 or 2). The number of instances which fall under each category is represented in each row. The class column specifies the primary tag of the content that belongs to the given line above the column.

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	\
0	0	3	0	0	3	2	
1	1	3	0	3	0	1	
2	2	3	0	3	0	1	
3	3	3	0	2	1	1	
4	4	6	0	6	0	1	

	tweet
0	!!! RT @mayasolovely: As a woman you shouldn't...
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Figure 1 Dataset

This balanced dataset was then divided into the training dataset, which consisted of 70% and the testing dataset which consisted of 30%.

4.2 Preprocessing Impact:

This balanced dataset was then divided into the training dataset, which consisted of 70% and the testing dataset which consisted of 30%.

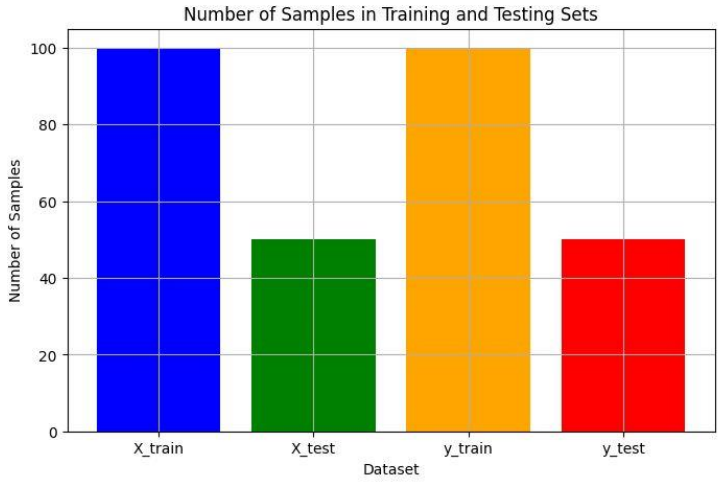


Figure 4.0 1 Preprocessing

This bar chart compares the number of samples in the training and testing datasets for both features (X_train, X_test) and labels (y_train, y_test). The training set has approximately 100 samples, while the testing set has about 50 samples, reflecting a common 2:1 train-test split. The alignment between X (features) and y (labels) ensures consistent data preparation. It highlights the dataset balance for model evaluation.

4.3 Model Training and Tuning

This section is devoted to the description of the training and tuning process of each classifier constituting the ensemble, namely, Logistic Regression, Support Vector Classifier (SVC), Decision Tree, and Random Forest, as well as the method of their integration into the stipulated ensemble model. The idea was to tune each of these models for better performance individually and also in the proposed ensemble.

4.3.1 Logistic Regression

Logistic Regression which is a linear model is also used frequently in binary classification methods was relatively productive in identifying online harassment.

Class	Precision	Recall	F1-Score	Support
Hate speech	49%	20%	28%	293
Offensive language	94%	98%	96%	3831
Accuracy	-	-	91%	4124
Macro avg	72%	59%	62%	4124
Weighted avg	91%	93%	91%	4124

Table 4.1 Logistic Regression

The classification report shows the model performs well for offensive language with a 94% precision and 96% F1-score, but poorly for hate speech with only 20% recall and 28% F1-score, indicating it struggles to detect hate speech. The overall accuracy is 91%, heavily influenced by the larger support (samples) for offensive language. The macro average highlights imbalanced performance across classes, while the weighted average reflects better overall performance due to the dominant class.

The given below bar chart shows that the Logistic Regression model achieved an accuracy of 0.91

Figure 4.2

The confusion matrix gives the performance of a classification model, most of the instances are classified correctly in the class Abusive, but there are some instances where, one class Abusive class was classified as Non-Abusive class.

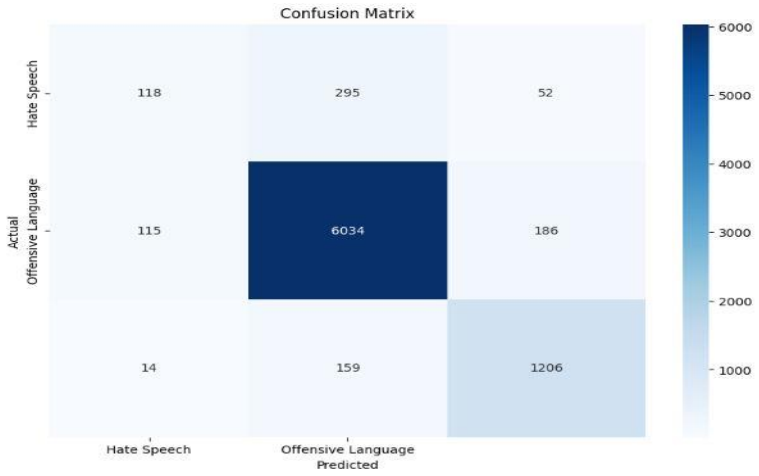


Figure 4.3 Confusion Matrix

This confusion matrix visualizes the classification results for hate speech and offensive language. The model correctly predicted 118 instances of hate speech but misclassified 295 as offensive language. For offensive language, it performed well, correctly identifying 6034 instances, with few misclassifications. The imbalance in correct predictions highlights the challenge of detecting hate speech compared to offensive language.

4.3.2 Support Vector Classifier (SVC):

SVC is more applicable on high dimensionality and high dimensionality data was used for this task; The model was scalable in distinguishing between harassment and non-harassment using features derived from the text.

Class	Precision	Recall	F1-Score	Support
Hate speech	71%	2%	3%	293
Offensive language	93%	100%	96%	3831
Accuracy	-	-	92%	4124
Macro avg	82%	51%	50%	4124
Weighted avg	91%	93%	90%	4124

Table 4.2 SVC

This classification report shows excellent performance for offensive language with 93% precision and 96% F1-score, but extremely poor performance for hate speech, with only 2% recall and a 3% F1-score,

indicating almost no hate speech detection. The accuracy is 92%, driven by the overwhelming presence of offensive language in the dataset. The macro average reflects the disparity in performance, while the weighted average is skewed by the dominant class.

From the below bar chart, the SVC model got a 0. 92

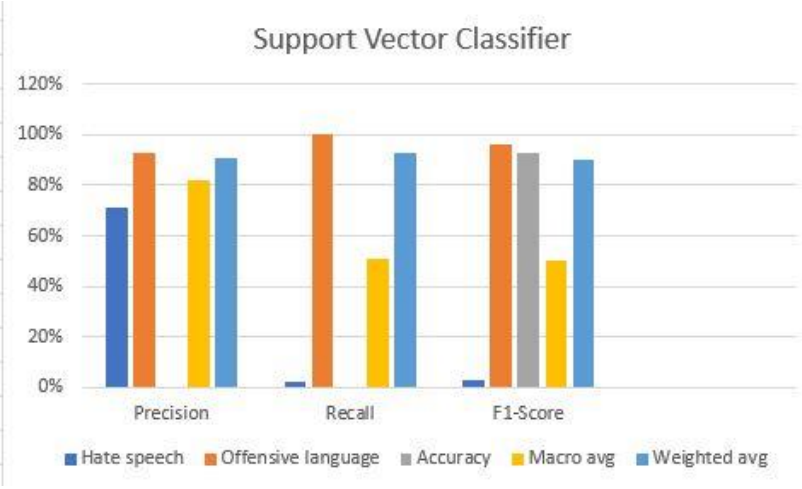


Figure 4.4

The confusion matrix gives the performance of a classification model, most of the instances are classified correctly in the class Abusive, but there are some instances where, one class Hate Speech class was classified as Offensive Language class.

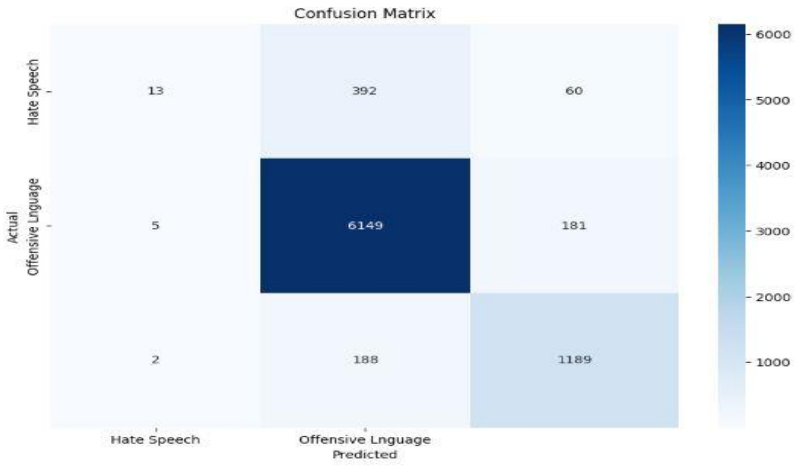


Figure 4.5 Confusion Matrix

This confusion matrix evaluates a model's classification performance for two categories: "Hate Speech" and "Offensive Language." The diagonal cells (13 and 6149) represent correct predictions, while the off-diagonal cells (e.g., 392 and 60) indicate misclassifications. The heatmap color intensity reflects the number of samples, with darker shades representing higher values.

4.3.3 Decision Tree:

Decision Tree is easy to understand and interpretable model in which, the data is split based on the values of the feature.

Class	Precision	Recall	F1-Score	Support
Hate speech	35%	30%	32%	293
Offensive language	95%	96%	95%	3831
Accuracy	-	-	91%	4124
Macro avg	65%	63%	64%	4124
Weighted avg	90%	91%	91%	4124

Table 4.2 Decision Tree

This table reviews the classification metrics for the model. It shows that the model performs well for "Offensive Language" with a high F1-score of 95%, but struggles with "Hate Speech," achieving a lower F1-score of 32%. The overall accuracy is 91%, with macro averages highlighting imbalanced performance across the two classes.

This bar chart revealed that the Decision Tree Classifier model had the accuracy of 91

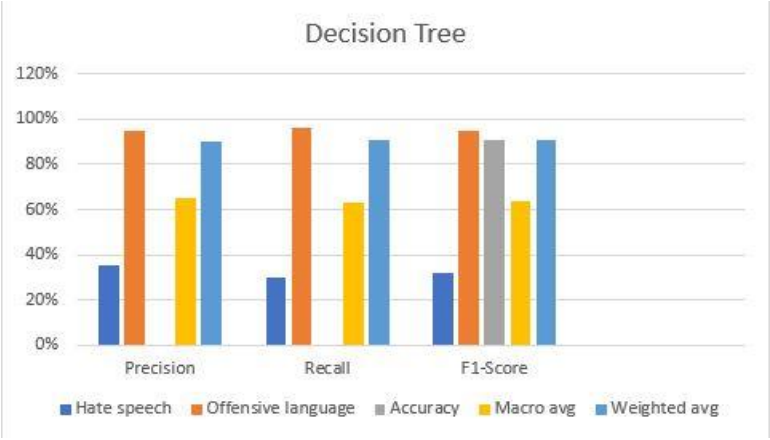


Figure 4.6

The confusion matrix gives the performance of a classification model, most of the instances are classified correctly in the class Abusive, but there are some instances where, one class Abusive class was classified as Non-Abusive class.

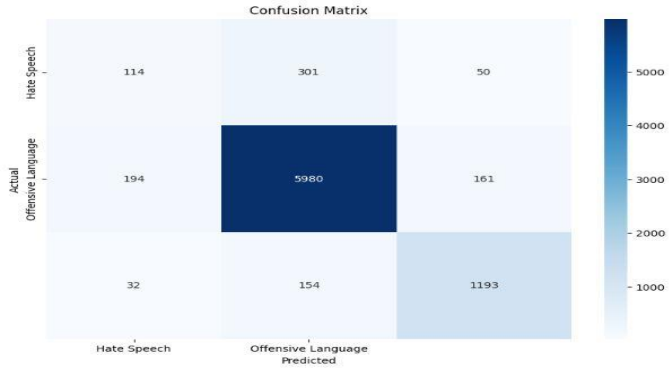


Figure 4.7 Confusion Matrix

This confusion matrix shows improved performance in predicting "Hate Speech" compared to the previous one, with 114 correct predictions and fewer misclassifications (e.g., 50 classified as "Offensive Language"). The majority of predictions for "Offensive Language" remain highly accurate (5980 correct). The overall classification appears balanced, with significant improvements in "Hate Speech" detection.

4.3.4 Random Forest:

A combination of classifiers in which the outputs for the same object from different trees are combined to form a strong classifier.

Class	Precision	Recall	F1-Score	Support
Hate speech	52%	11%	19%	293
Offensive language	94%	99%	96%	3831
Accuracy	-	-	92%	4124
Macro avg	73%	55%	57%	4124
Weighted avg	91%	93%	91%	4124

Table 4.3 Random Forest

The model achieves high performance in identifying offensive language (F1-score: 96%) but struggles with hate speech (F1-score: 19%) due to low recall (11%). Overall accuracy is 92%, heavily influenced by the larger support of offensive language examples. Macro averages highlight the disparity in performance across classes, while weighted averages reflect the dominance of the majority class.

As the bar chart on figure 3 shows, Random Forest model had an accuracy of 92.

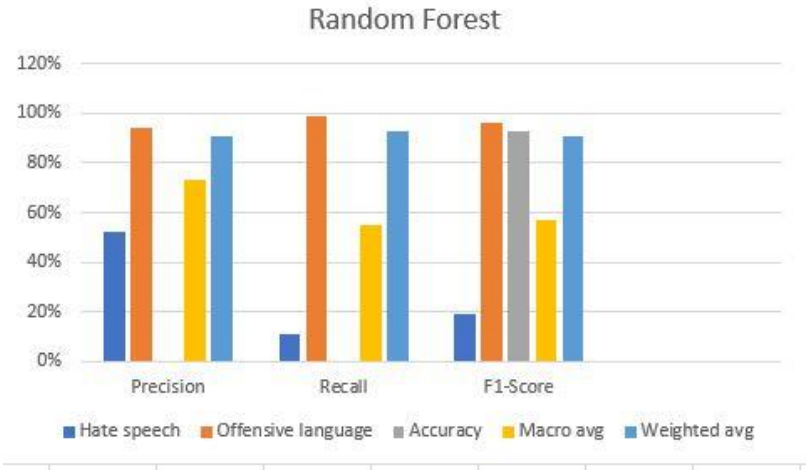


Figure 4.8

The confusion matrix gives the performance of a classification model, most of the instances are classified correctly in the class Abusive, but there are some instances where, one class Abusive class was classified as Non-Abusive class.

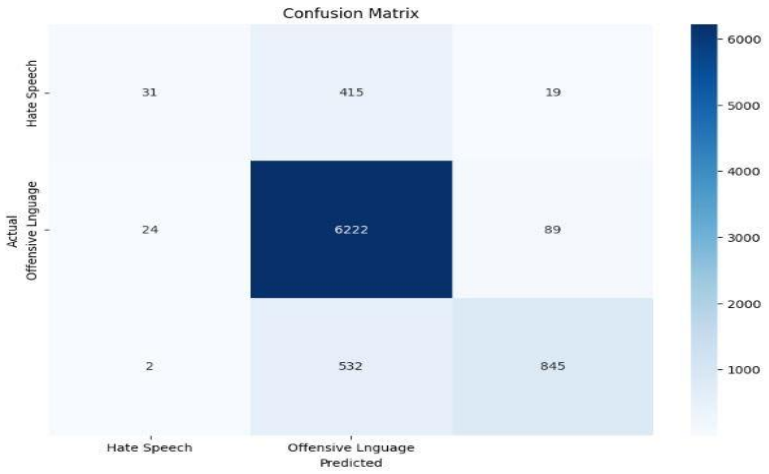


Figure 4.9 Confusion Matrix

The confusion matrix shows that the model predicts offensive language with high accuracy (6,222 correct predictions) but often misclassifies hate speech as offensive language (415 cases). It identifies only 31 instances of hate speech correctly, indicating poor performance on this class. Misclassification between the two classes highlights a challenge in distinguishing subtle differences between hate speech and offensive language.

4.4 Ensemble Model Performance

When the hyperparameters were optimized for every classifier separately, all these models were ensemble with Soft Voting on the same results, we can see that the ensemble model(SLR) displayed higher performance on all the changes indicators. Below are the results:

Model	Accuracy	Precision	Recall	F1-Score
LR	90%	94%	98%	96%
SVC	92%	93%	100%	96%
RF	92%	94%	99%	96%
SLR Model	93%	94%	99%	96%

Table 4.4 Comparison Model

The table presents the performance metrics of four different machine learning models: LR, SVC, RF, and SLR Model. All models exhibit high accuracy, but SLR model achieve highest accuracy 93% and precision, recall, and F1-score, indicating strong overall performance. However, SVC stands out with perfect recall, suggesting it's particularly adept at identifying positive cases.

4.5.1 Classifier Accuracy Comparison:

The bar chart above compares SVC, Logistic Regression, Random Forest and Voting Classifier and all models prove to have high accuracy with a slight edge with the Voting Classifier.

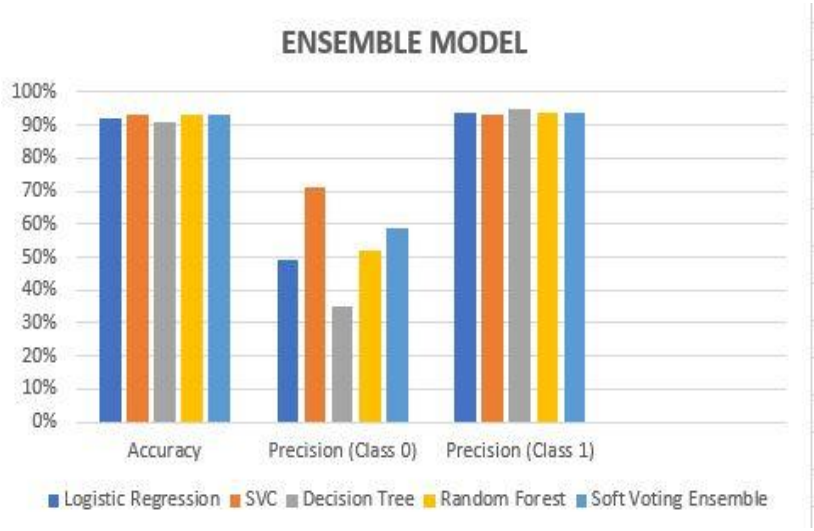


Figure 4.20 Classifier Accuracy Comparison

The bar chart shows the accuracy comparison of four different classifiers: SVC, Logistic Regression, Random Forest, and Voting Classifier. The Voting Classifier has the highest accuracy, followed by Random Forest. Logistic Regression has the lowest accuracy.

4.5.2 Voting Classifier Confusion Matrix:

As shown in the confusion matrix (figure 4.9) the classifier with the best accuracy for the Voting Classifier is the Random Forest Classifier whereby there is a high level of accuracy in predicting the majority class (label 1) by the classifier although there is an indication of misclassifications in all the classes.

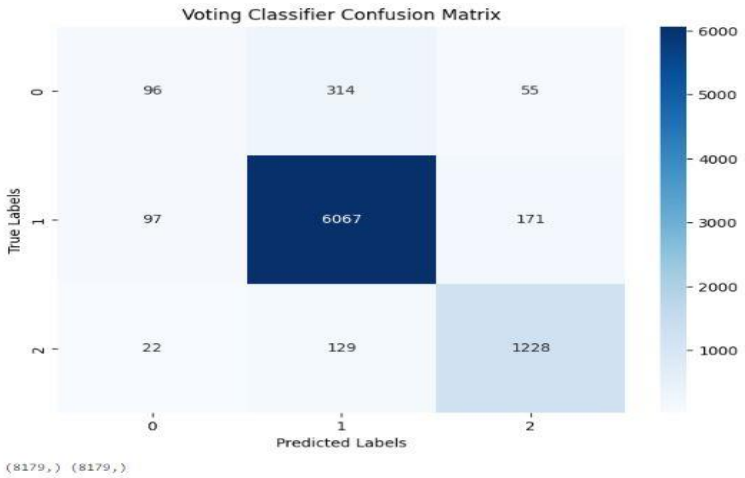


Figure 4.11 Voting Classifier Confusion Matrix

In Figure 4.11, the confusion matrix provides an overview of the classification performance for the Voting Classifier, particularly in distinguishing between Hate Speech and Offensive Language content. The confusion matrix displays counts for each prediction category, which is essential for interpreting the classifier's effectiveness. Each cell was labeled clearly, with the x-axis indicating predicted values and the y-axis showing actual values. The ensemble's success due to its "soft voting" method, which combines the strengths of individual classifiers like SVC and Random Forest to improve overall accuracy.

Conclusion:

As a result, it has been proven that an ensemble learning method can be efficiently applied for the classification of online harassment from social media text. It was possible to train a model of an increased efficiency when using a number of classifiers at once, thus it provides the identification of the various forms of harassment including hate speech, cyber bulling, and violence. Even though, the model shows high performance in a great number of aspects, there's a great potential to increase its effectiveness in identification, notably the recognition of more subtle or context-sensitive types of harassment. The major improvement in this research is the proposed Ensemble system which combines four ML classifiers including Logistic Regression, SVC, Decision Tree, and Random Forest for better recognition of the harassment. Thus, the proposed ensemble method to combine the classifiers that have the unique advantages demonstrated better performance in comparison with the individual classifiers in the tests with the accuracy of 93%. The voted model incorporated a soft voting technique meaning that the model minimized false positives and false negatives in detecting diverse forms of harassment while excluding unnecessary alarms.

Reference:

Abarna, S., Sheeba, J. I., Jayasrilakshmi, S., & Devaneyan, S. P. (2022). Identification of cyber harassment and intention of target users on social media platforms. Engineering Applications of Artificial Intelligence, 115(May), 105283. https://doi.org/10.1016/j.engappai.2022.105283

Abdrakhmanov, R., Kenesbayev, S. M., & Berkimbayev, K. (2024). Offensive Language Detection on Social Media using Machine Learning. 15(5), 575–582.

Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021). Cyberbullying detection: An ensemble based machine learning approach. Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021, Icicv, 710–715. https://doi.org/10.1109/ICICV50876.2021.9388499

Alqahtani, A. F., & Ilyas, M. (2024). An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying. Machine Learning and Knowledge Extraction, 6(1), 156–170. https://doi.org/10.3390/make6010009

Amer, A., & Yahya, A. (2022). Detecting Cybercrime: An Evaluation of Machine Learning and Deep Learning Using Natural Language Processing Techniques on the Social Network Detecting Cybercrime: An Evaluation of Machine Learning and Deep Learning Using Natural Language Processing Techni. In Research Square. https://doi.org/10.21203/rs.3.rs-2184218/v1

Azeez, N. A., & Fadhal, E. (2023). Classification of Virtual Harassment on Social Networks Using Ensemble Learning Techniques. Applied Sciences (Switzerland), 13(7). https://doi.org/10.3390/app13074570

Bai, Z. S., & Malempati, S. (2023). Ensemble Deep Learning (EDL) for Cyber-bullying on Social Media. International Journal of Advanced Computer Science and Applications, 14(7), 551–560. https://doi.org/10.14569/IJACSA.2023.0140761

Bölücü, N., & Canbay, P. (2024). Syntax-aware Offensive Content Detection in Low-resourced Code-mixed Languages with Continual Pre-training. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3653450

D. S. Aabdalla, I., & Vasumathi, D. (2024). Wavelet Scattering Transform for ECG Cardiovascular Disease Classification. International Journal of Artificial Intelligence & Applications, 15(1), 101–113. https://doi.org/10.5121/ijaia.2024.15107

Das, M., Bengal, W., Mukherjee, A., & Bengal, W. (2022). BanglaAbuseMeme: A Dataset for Bengali Abusive Meme Classification.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. 25–35. https://doi.org/10.18653/v1/w19-3504

El Koshiry, A. M., Eliwa, E. H. I., El-Hafeez, T. A., & Khairy, M. (2024). Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique. PeerJ Computer Science, 10, 1–33. https://doi.org/10.7717/peerj-cs.1961

Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. Mathematics, 11(16). https://doi.org/10.3390/math11163567

O. (2024). Comparing the effectiveness of behavioral activation in group vs. self-help format for reducing depression, repetitive thoughts, and enhancing performance of patients with major depressive disorder: a randomized clinical trial. BMC Psychiatry, 24(1), 1–12. https://doi.org/10.1186/s12888-024-05973-z

Sekwatlakwatla, S. P., & Malele, V. (2024). Model for Enhancing Cloud Computing Resource Allocation Management Using Data Analytics. 6(1), 514–526. https://doi.org/10.51519/journalisi.v6i1.679

Semangern, T., Chaisitsak, W., & Senivongse, T. (2019). Identification of risk of cyberbullying from social network messages. Lecture Notes in Engineering and Computer Science, 2019-Octob, 276–282.

Sharma, N., Chouhan, K., & Verma, V. (2022). Cyberbullying Detection On Instagram. 6(2), 18–21.

Sreevidya, N., Hamsini, A., & Vainateya, R. (2024). Ensemble Learning Based Prediction for Cyber Harassment Observations on Tweets. 17(6), 102–113. https://doi.org/10.9734/AJRCOS/2024/v17i6460

Sultan, T., Jahan, N., Basak, R., Jony, M. S. A., & Nabil, R. H. (2023). Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition. International Journal of Intelligent Systems and Applications, 15(2), 1–13. https://doi.org/10.5815/ijisa.2023.02.01

Szcześniak, M., Falewicz, A., Strochalska, K., & Rybarski, R. (2022). Anxiety and Depression in a Non-Clinical Sample of Young Polish Adults: Presence of Meaning in Life as a Mediator. International Journal of Environmental Research and Public Health, 19(10). https://doi.org/10.3390/ijerph19106065

Teng, T. H., & Varathan, K. D. (2023). Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches. IEEE Access, 11(April), 55533–55560. https://doi.org/10.1109/ACCESS.2023.3275130

Thilagavathy, A., Deepa, R., Lalitha, S. D., Raju, D. N., Ramya, R., Ramya, M., Rithika, L., & Sundari, K. K. (2023). Semantic-Based Classification of Toxic Comments Using Ensemble Learning. E3S Web of Conferences, 399, 1–8. https://doi.org/10.1051/e3sconf/202339904017

Tolba, M., Ouadfel, S., & Meshoul, S. (2021). Hybrid ensemble approaches to online harassment detection in highly imbalanced data. Expert Systems with Applications, 175(February), 114751. https://doi.org/10.1016/j.eswa.2021.114751

Tuli, S., Gill, S. S., Xu, M., Garraghan, P., Bahsoon, R., Dustdar, S., & Sakellariou, R. (2021). HUNTER: AI based Holistic Resource Management for Sustainable Cloud Computing.

Uddin, N., Uddin Ahamed, M. K., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering, 4(February), 327–339. https://doi.org/10.1016/j.ijcce.2023.09.001

Xingyi, G., & Adnan, H. M. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. International Journal of Data and Network Science, 8(1), 25–34. https://doi.org/10.5267/j.ijdns.2023.10.021