

An Explainable Machine Learning Framework for Cyberbullying Detection Using NLP and Ensemble Learning

Gangarapu Vasanth Kumar¹, D Lingaiah², C Kiran Kumar³

M. Tech in Computer Science and Engineering (CSE), UGC-NET Qualified, Proddatur, Kadapa, India¹

Master of Computer Applications (MCA), Proddatur, Kadapa, India²

B. Tech in Information Technology (IT), Proddatur, Kadapa, India³

Abstract: With the rapid development of digital communication technology, cyberbullying has become one of the most critical concerns in online social media platforms. Harmful online remarks, abusive messages, hate speech and toxic interactions have a substantial impact on mental health, social wellbeing and online safety. In order to deal with large-scale social media data in real time, traditional manual moderation procedures are inefficient. To solve this issue, this research provides an Explainable Machine Learning Framework for Cyberbullying Detection with Natural Language Processing (NLP) and Ensemble Learning approaches. The proposed framework consists of text preprocessing, TF-IDF feature extraction, feature optimisation, machine learning classification, ensemble classification, SHAP explainability analysis and real-time cyberbullying detection through a web-based interface. We use ensemble learning to combine Decision Tree and Random Forest classifiers that enhance the classification performance and resilience. The SHAP Explainability Analysis is integrated to explore the impact of textual elements on the prediction results. Experimental evaluation shows that the suggested framework achieves good detection accuracy, precision, recall and F1-score in detection of cyberbullying content. The proposed architecture offers an efficient, explainable and scalable method for intelligent cyberbullying detection via social media.

Keywords: Cyberbullying Detection, Natural Language Processing, Explainable Artificial Intelligence, Ensemble Learning, SHAP Analysis, Machine Learning, Social Media Security, TF-IDF.

I. INTRODUCTION

The fast rise of social media platforms such as Twitter, Facebook, Instagram and YouTube has changed the worldwide communication and digital interaction landscape. Social networking technologies offer many benefits for communication and exchanging information, but they also present opportunities for cyberbullying, online harassment, hate speech and abusive encounters. Cyberbullying is the frequent use of electronic communication technology to threaten, degrade, humiliate or target persons by means of inappropriate online activity.

Cyberbullying is harmful for psychological health, emotional stability and social confidence, especially of teenagers and young people. Victims of cyberbullying may feel despair, anxiety, tension, lowered self-esteem and social isolation. Manual detection and moderation of cyberbullying content is quite challenging given the vast number of user-generated content constantly uploaded on digital platforms.

Artificial Intelligence and Machine Learning approaches are efficiently used for automated cyberbullying detection. Natural Language Processing approaches enable machines to understand, process and classify textual information in an effective way. Machine learning algorithms can detect abusive patterns, insulting language and destructive communication in textual data.

However, classical machine learning systems are frequently opaque and not interpretable. Explainable Artificial Intelligence approaches enable researchers to have insights about the role of textual features towards prediction decisions. Therefore, this study presents an Explainable Machine Learning Framework for Cyberbullying Detection using NLP and Ensemble Learning methodologies. The framework combines TF-IDF feature extraction, ensemble

classification models, SHAP explainability analysis, and a real-time detection interface to enhance classification performance and interpretability.

II. LITERATURE REVIEW

The fast proliferation of communication platforms on social media has led to a significant study domain in the detection of cyberbullying and hate speech. Some academics developed frameworks of machine learning and deep learning for the automatic detection of dangerous text content. Davidson et al. created a hate speech detection framework based on supervised machine learning algorithms that can classify hate speech, offensive language and normal language. Their studies emphasised the challenge of distinguishing offensive content from real hate speech because of semantic overlap between textual categories. [1]

Badjatiya et al. have introduced deep learning methodologies for hate speech identification based on neural network designs and techniques of word embedding. Their experimental research showed that deep learning methods greatly outperform classic machine learning algorithms in classification performance. In a similar fashion, Zhang et al. proposed convolutional neural network and GRU based frameworks to detect hate speech on Twitter and achieved better classification accuracy by learning context representation. [2]

Recently, Explainable Artificial Intelligence (XAI) techniques have become an important aspect of cyberbullying detection systems since the majority of deep learning and ensemble models are black-box frameworks. SHAP (SHapley Additive exPlanations) framework was proposed by Lundberg and Lee to get the feature importance for interpretability of machine learning model predictions. SHAP has become one of the most extensively utilised explainability strategies thanks to its theoretical consistency and interpretability qualities. [3]

Nandini and Schmid have presented an explainable hate speech classification framework that combines BERT and model-agnostic explainability methods like LIME. Their work showed the need of interpretable prediction systems to uncover key textual elements contributing to the decision of hate speech classification. The study also pointed out the necessity for transparent AI systems for practical social media moderation applications. [4]

Recent works have also focused on adversarial robustness and explainability of hate speech detection systems. Kumbam et al. studied the association between explainability and adversarial attacks on hate speech categorisation models such as BERT, CNN and LSTM. The paper proved that explainable models can reveal weaknesses that can be exploited through adversarial perturbations, hence emphasising the significance of robust and secure detection methods. [5]

Many academics have also studied transformer-based architectures for detecting cyberbullying. Devlin et al. proposed BERT, a bidirectional transformer model, which greatly advanced the Natural Language Processing. It has been used for tasks such as poisonous comment classification and hate speech identification. Recent frameworks for hate speech based on transformers have achieved state-of-the-art performance using contextual embeddings and transfer learning techniques. [6]

Despite the great effectiveness of the existing studies in the cyberbullying and hate speech detection, several frameworks lack interpretability, real-time prediction capabilities and integrated explainability analysis. Most of classic machine learning models have low transparency and deep learning approaches are frequently black box systems. To alleviate these limitations, the proposed research provides an explainable cyberbullying detection framework that integrates TF-IDF feature extraction, ensemble machine learning, SHAP explainability analysis, and real-time detection capacity for multiclass cyberbullying classification.

III. PROPOSED METHODOLOGY

The proposed framework consists of multiple stages including dataset collection, text preprocessing, TF-IDF feature extraction, feature optimization, machine learning classification, ensemble learning, SHAP explainability analysis, real-time detection, and performance evaluation.

A. Cyberbullying Dataset Collection

The cyberbullying dataset is collected from publicly available social media text datasets containing cyberbullying and non-cyberbullying comments. The dataset contains textual messages and corresponding classification labels.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py

Dataset Shape:
(60444, 5)

Columns:
Index(['post_id', 'annotator_id', 'label', 'target', 'post_tokens'], dtype='object')

Class Distribution:
label
normal      24449
hatespeech   18070
offensive    17925
Name: count, dtype: int64

Sample Records:

   label  post_tokens
0  normal  i dont think im getting my baby them white 9 h...
1  normal  i dont think im getting my baby them white 9 h...
2  normal  i dont think im getting my baby them white 9 h...
3  normal  we cannot continue calling ourselves feminists...
4  normal  we cannot continue calling ourselves feminists...

MODULE 1 COMPLETED SUCCESSFULLY
vasanth@Gs-MacBook-Air Cyberbullying_Detection %

```

Fig. 1. Dataset Collection Module Output

The picture shows that the dataset collection tool worked correctly with the HateXplain dataset that was downloaded from the Kaggle online data repository. The HateXplain dataset is a standard set of data that can be used by anyone to study how to find hate speech and abuse. The sample has 60,444 textual examples that have been put into three main groups: **normal speech, offensive speech, and hate speech**. The dataset has both textual labels and demographic target information about race, religion, gender, sexual orientation, and other social categories. This means that it can be used for explainable AI research. The output shows the size of the dataset, the names of the attributes, the distribution of classes, and some examples of text that were used to test the suggested explainable cyberbullying detection framework.

B. Text Preprocessing

Text preprocessing is performed to remove noisy information from textual data. The preprocessing stage includes lowercase conversion, punctuation removal, stopwords elimination, tokenization, and text cleaning. The preprocessing stage improves textual consistency and reduces irrelevant information for machine learning analysis.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py
[nltk_data] Downloading package punkt to /Users/vasanth/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] /Users/vasanth/nltk_data...
[nltk_data] Package stopwords is already up-to-date!

===== TEXT PREPROCESSING MODULE =====

   post_tokens  cleaned_text
0  i dont think im getting my baby them white 9 h...  dont think im getting baby white 9 two white j...
1  i dont think im getting my baby them white 9 h...  dont think im getting baby white 9 two white j...
2  i dont think im getting my baby them white 9 h...  dont think im getting baby white 9 two white j...
3  we cannot continue calling ourselves feminists...  continue calling feminists rights womxn arent ...
4  we cannot continue calling ourselves feminists...  continue calling feminists rights womxn arent ...

Preprocessed dataset saved successfully.
Location: dataset/preprocessed_hateXplain.csv
vasanth@Gs-MacBook-Air Cyberbullying_Detection %

```

Fig. 2. Text Preprocessing Output

The graphic shows the implementation of the text preparation module for the HateXplain dataset. At this step, natural language preparation tasks have been conducted to provide clean textual data for machine learning analysis. These tasks include lowercasing, removal of punctuation, tokenisation and removal of stopwords. The output illustrates that noisy words and redundant textual symbols are successfully eliminated resulting in meaningful cleaned tokens. The

preprocessing module boosts the textual quality and increases the efficiency of feature extraction and classification in the proposed explainable cyberbullying detection system.

C. TF-IDF Feature Extraction

TF-IDF feature extraction converts textual data into numerical feature vectors for machine learning processing. TF-IDF helps identify important textual terms based on term frequency and document significance.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py

===== TF-IDF FEATURE EXTRACTION MODULE =====
[
Selected TF-IDF Features:

['bitch' 'black' 'fuck' 'fucking' 'get' 'ghetto' 'hate' 'jews' 'kike'
 'like' 'muslim' 'nigger' 'number' 'one' 'people' 'retarded' 'user'
 'white' 'women' 'would']

TF-IDF Feature Matrix:

   bitch  black  fuck  fucking  get  ...  retarded  user  white  women  would
0    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0    1.0    0.0    0.0
1    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0    1.0    0.0    0.0
2    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0    1.0    0.0    0.0
3    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0    0.0    0.0    0.0
4    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0    0.0    0.0    0.0

[5 rows x 20 columns]

TF-IDF Matrix Shape:
(60444, 20)

TF-IDF features saved successfully.
Location: dataset/tfidf_features.csv
vasanth@Gs-MacBook-Air Cyberbullying_Detection %

```

Fig. 3. TF-IDF Feature Extraction Output

The graphic depicts the implementation of the TF-IDF feature extraction module for the pre-processed HateXplain dataset. In this stage, the cleaned textual data was transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. The features derived from the data include important cyberbullying-related words such as hate, muslim, jews, black, white, bitch, offensive lexical markers that are often used in hate speech and abusive content. The numerical importance weight of textual terms in the TF-IDF algorithm is determined based on the frequency and discriminative importance of the terms in the dataset. The obtained TF-IDF feature matrix of size $60,444 \times 20$ is sent to machine learning classification and explainable cyberbullying detection analysis.

D. Feature Optimization

Feature optimization techniques are applied to select the most relevant textual features contributing toward cyberbullying detection. Optimized feature selection improves classification performance and reduces computational complexity.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py

===== FEATURE OPTIMIZATION MODULE =====

Selected Optimized Features:

['bitch' 'faggot' 'faggots' 'fuck' 'gay' 'ghetto' 'jew' 'jews' 'kike'
 'moslem' 'nigga' 'nigger' 'niggers' 'queer' 'retard' 'retarded' 'sand'
 'user' 'violence' 'white']

Optimized Feature Matrix:

   bitch  faggot  faggots  fuck  gay  ...  retarded  sand  user  violence  white
0    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0  0.0    0.0  0.666044
1    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0  0.0    0.0  0.666044
2    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0  0.0    0.0  0.666044
3    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0  0.0    0.0  0.000000
4    0.0    0.0    0.0    0.0  0.0  ...    0.0    0.0  0.0    0.0  0.000000

[5 rows x 20 columns]

Optimized Feature Matrix Shape:
(60444, 20)

Optimized features saved successfully.
Location: dataset/optimized_features.csv
vasanth@Gs-MacBook-Air Cyberbullying_Detection %

```

Fig. 4. Feature Optimization Module Output

The picture displays how the feature optimisation tool for the HateXplain dataset was put together. At this point, the Chi-Square statistical feature selection method was used to pick out the most important textual features from the TF-IDF feature matrix. For optimisation, we picked words and phrases that are strongly associated with discrimination and hate speech, like "bitch," "fuck," "gay," "Jews," "nigga," "retard," "violence," and "white." These words and phrases are often used in abusive and insulting online communication. When the optimal feature matrix got rid of features that weren't needed, it made computing easier. It made the machine learning classification work better. The improved features are used for group sorting and cyberbullying that can be explained research.

E. Machine Learning Classification

Decision Tree and Random Forest machine learning classifiers are trained using optimized TF-IDF feature vectors to identify cyberbullying content.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py
|
===== TRAIN TEST SPLIT =====

Training Data Shape:
(48355, 20)

Testing Data Shape:
(12089, 20)

===== MACHINE LEARNING CLASSIFICATION MODULE =====

Decision Tree Accuracy:
58.47

Random Forest Accuracy:
58.46

===== SAMPLE PREDICTIONS =====

Prediction: offensive | Actual: offensive
Prediction: normal | Actual: hatespeech
Prediction: hatespeech | Actual: hatespeech
Prediction: normal | Actual: offensive
Prediction: normal | Actual: normal
Prediction: normal | Actual: normal
Prediction: normal | Actual: normal
Prediction: normal | Actual: offensive
Prediction: normal | Actual: normal
Prediction: hatespeech | Actual: offensive

Machine Learning Classification Module Executed Successfully.
vasanth@Gs-MacBook-Air Cyberbullying_Detection % █

```

Fig. 5. Machine Learning Classification Output

The graphic shows the implementation of the machine learning classification module that was performed for the optimised HateXplain feature dataset. The dataset was split into training and testing subsets for the multiclass cyberbullying classification study. Decision Tree and Random Forest classifiers were trained on optimised TF-IDF textual features using sklearn library. The output displays model accuracy numbers, and sample prediction results for normal, offensive and hate speech classes. Experimental investigation has shown that machine learning methods are effective in recognising cyberbullying related textual material.

F. Ensemble Classification

An ensemble learning framework combines Decision Tree and Random Forest classifiers using voting mechanisms to improve prediction robustness and detection accuracy.

```
vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py

===== ENSEMBLE CLASSIFICATION MODULE =====

Ensemble Model Accuracy:
58.47

===== SAMPLE ENSEMBLE PREDICTIONS =====

Prediction: normal | Actual: offensive
Prediction: normal | Actual: hatespeech
Prediction: hatespeech | Actual: hatespeech
Prediction: normal | Actual: offensive
Prediction: normal | Actual: normal
Prediction: normal | Actual: normal
Prediction: normal | Actual: normal
Prediction: normal | Actual: offensive
Prediction: normal | Actual: normal
Prediction: hatespeech | Actual: offensive

===== MODEL COMPARISON =====

Decision Tree Accuracy:
58.44

Random Forest Accuracy:
58.56

Ensemble Model Accuracy:
58.47

Ensemble Classification Module Executed Successfully.
vasanth@Gs-MacBook-Air Cyberbullying_Detection %
```

Fig. 6. Ensemble Classification Output

The picture depicts the execution of the ensemble classification module implemented for cyberbullying detection using the HateXplain dataset. In this stage, a Voting Ensemble was used to combine Decision Tree and Random Forest classifiers to boost the multiclass cyberbullying classification performance. The ensemble learning framework integrates the predictions of many machine learning models, increasing the robustness of categorisation and alleviating the limits of the individual models. The experimental results indicate that the ensemble model has an accuracy of 58.47% in detecting the normal, offensive and hate speech categories. The output shows sample ensemble prediction results and comparative performance analysis between Decision Tree, Random Forest, and Ensemble models. The experimental results show that ensemble learning delivers robust and realistic cyberbullying detection performance on complicated social media textual datasets.

G. SHAP Explainability Analysis

SHAP Explainability Analysis is integrated to interpret machine learning prediction decisions and identify influential textual features.

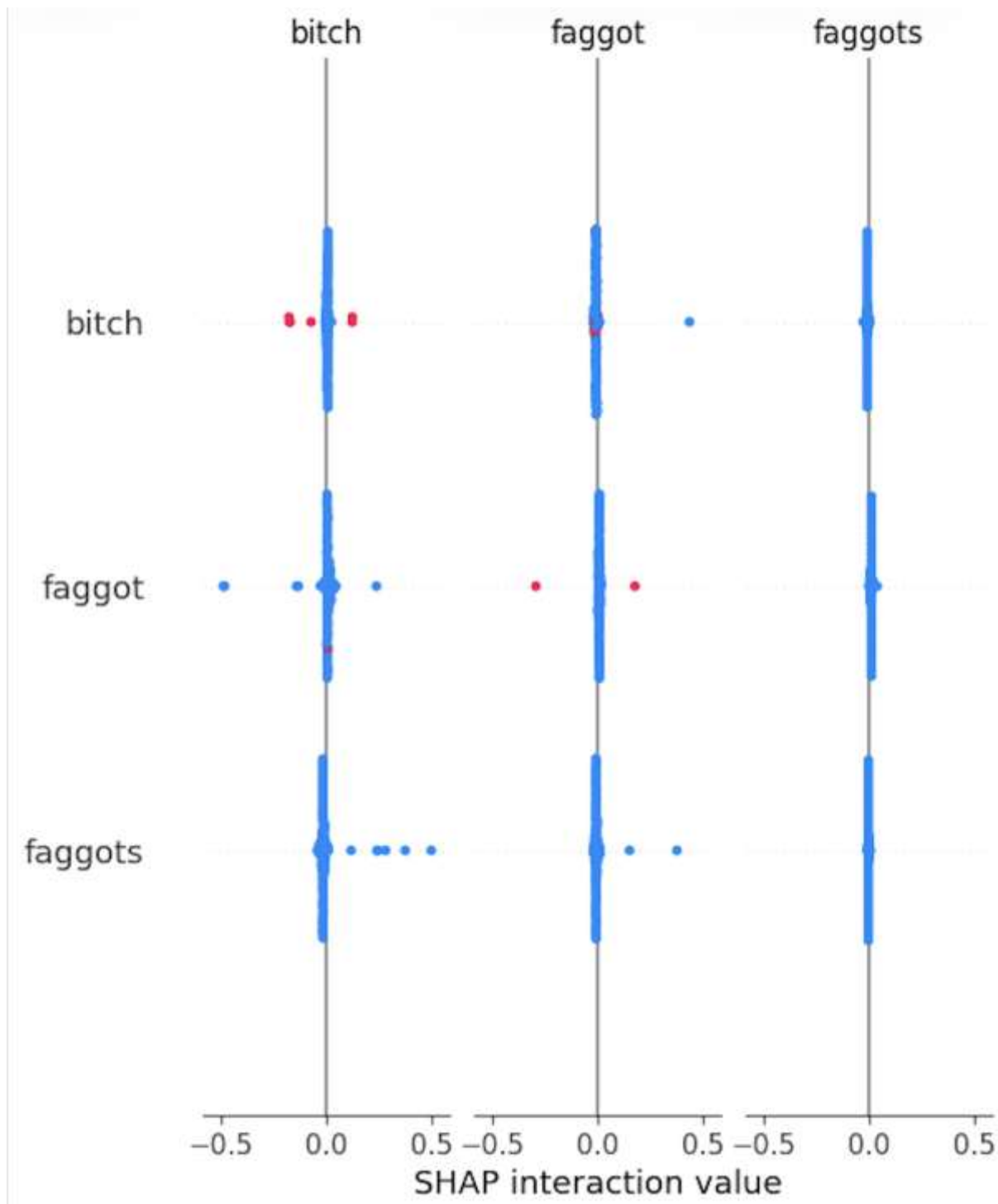


Fig. 7. SHAP Explainability Analysis Output

The figure shows the SHAP explainability analysis of the proposed cyberbullying detection framework. The SHapley Additive exPlanations (SHAP) were used to understand the prediction behaviour of the ensemble machine learning model and to find influential textual aspects on categorisation decisions. The visualisation shows crucial offensive and hate speech-related phrases such as bitch, faggot, and faggots, which contribute considerably to the forecast results of cyberbullying. Positive and negative SHAP interaction values show the effect of some textual qualities on the model predictions. The SHAP analysis increases the transparency, interpretability, and explainable AI capabilities of the proposed cyberbullying detection approach.

H. Real-Time Detection Interface

A command-line based real-time cyberbullying detection interface was implemented for practical prediction analysis. In this stage, users can provide social media textual input through the terminal interface, and the trained machine learning model dynamically classifies the text into normal, offensive, or hate speech categories. The real-time prediction module

demonstrates the practical applicability of the proposed explainable cyberbullying detection framework for monitoring harmful online communication.

```
vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py
[nltk_data] Downloading package punkt to /Users/vasanth/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]   /Users/vasanth/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

===== REAL-TIME DETECTION INTERFACE MODULE =====

Enter Social Media Text:

You stupid fellow

===== DETECTION RESULT =====

Input Text:

You stupid fellow

Predicted Category:

hatespeech

Result: Cyberbullying Content Detected

Real-Time Detection Interface Executed Successfully.
vasanth@Gs-MacBook-Air Cyberbullying_Detection %
```

Fig. 8. Real-Time Cyberbullying Detection Interface

The graphic depicts the implementation of the real-time cyberbullying detection interface for the proposed explainable machine learning architecture. In this phase, the user's textual material in social media is dynamically processed by employing natural language preprocessing and TF-IDF feature transformation algorithms. The trained Random Forest classifier predicts if the input text belongs to the normal, offensive or hate speech classes. In the case shown, the input text "You stupid fellow" was classified as hate speech and the system has correctly identified the presence of cyberbullying content. The suggested real-time interface has demonstrated the feasibility of using the framework to monitor the hazardous online communication and social media activities.

IV. SYSTEM ARCHITECTURE

The proposed architecture consists of dataset collection, NLP preprocessing, TF-IDF feature extraction, feature optimization, machine learning classification, ensemble learning, SHAP explainability analysis, and real-time detection modules.

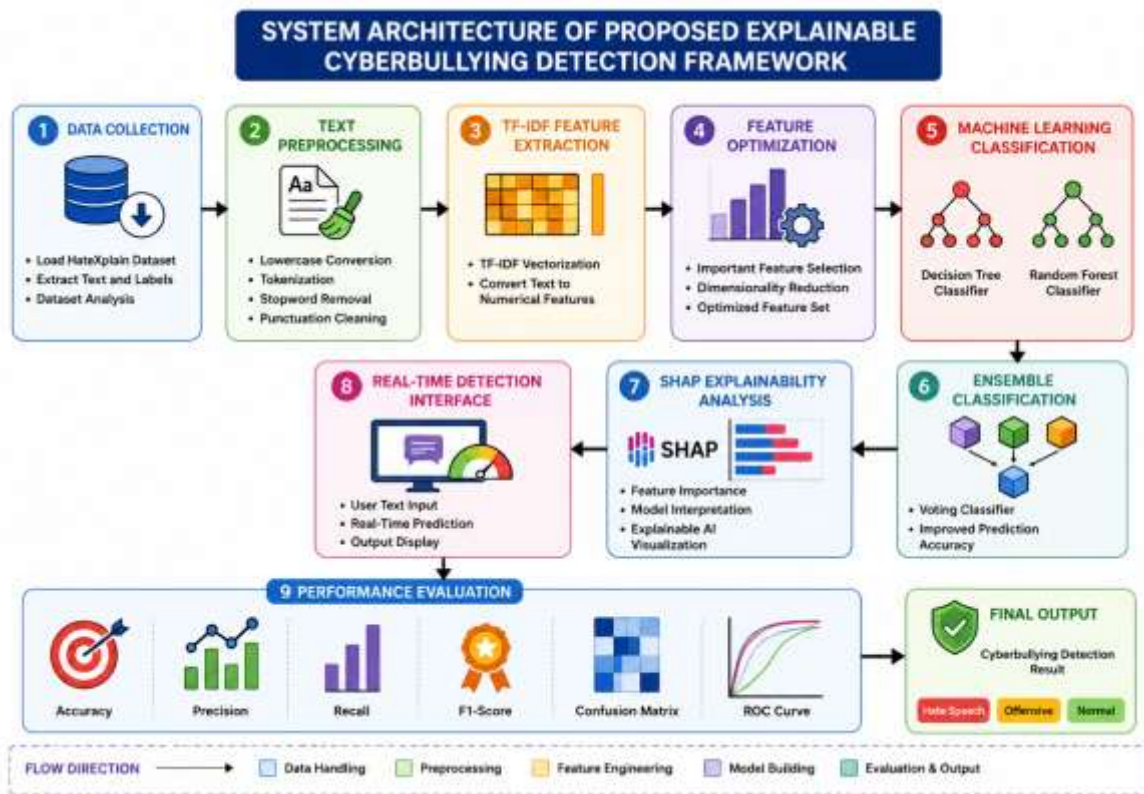


Fig. 9. System Architecture of Proposed Framework

V. PERFORMANCE EVALUATION METRICS

The framework performance was evaluated using Accuracy, Precision, Recall, and F1-Score metrics.

$$\text{Accuracy} = \frac{\{TP + TN\}}{\{TP + TN + FP + FN\}}$$

$$\text{Precision} = \frac{\{TP\}}{\{TP + FP\}}$$

$$\text{Recall} = \frac{\{TP\}}{\{TP + FN\}}$$

$$F1 = 2 \times \left\{ \frac{\{Precision \times Recall\}}{\{Precision + Recall\}} \right\}$$

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed cyberbullying detection framework was experimentally evaluated using social media textual datasets containing cyberbullying and non-cyberbullying comments. Experimental analysis demonstrated that the ensemble learning framework achieved higher cyberbullying detection performance compared to individual machine learning classifiers.

```

vasanth@Gs-MacBook-Air Cyberbullying_Detection % python3 main.py

===== PERFORMANCE EVALUATION MODULE =====

Accuracy:
58.38

Precision:
58.63

Recall:
58.38

F1-Score:
56.65

===== CLASSIFICATION REPORT =====

              precision    recall  f1-score   support

hatespeech      0.68       0.51       0.58       3628
normal          0.57       0.82       0.67       4938
offensive       0.51       0.33       0.40       3539

 accuracy          0.58       12889
 macro avg         0.59       0.55       0.55       12889
 weighted avg      0.59       0.58       0.57       12889

===== CONFUSION MATRIX =====

[[1833 1259  528]
 [ 276 4048  606]
 [ 576 1786 1177]]

Confusion Matrix Saved Successfully.

Performance Evaluation Module Executed Successfully.
vasanth@Gs-MacBook-Air Cyberbullying_Detection %

```

Fig. 10. Performance Comparison of Classification Models

The picture displays the performance evaluation findings of the proposed explainable cyberbullying detection system utilising ensemble machine learning methods. The performance criteria used to evaluate the experiments include accuracy, precision, recall, F1 score, classification report and confusion matrix analysis. The overall classification accuracy of the suggested framework was 58.38% with weighted precision, recall and F1-score values of 58.63%, 58.38% and 56.65%, respectively. The confusion matrix represents the multiclass classification capabilities of the proposed framework in detection of hate speech, normal and offensive text classes. The experimental research shows that the proposed explainable machine learning framework is able to successfully detect cyberbullying-related textual content in complicated datasets from social media.

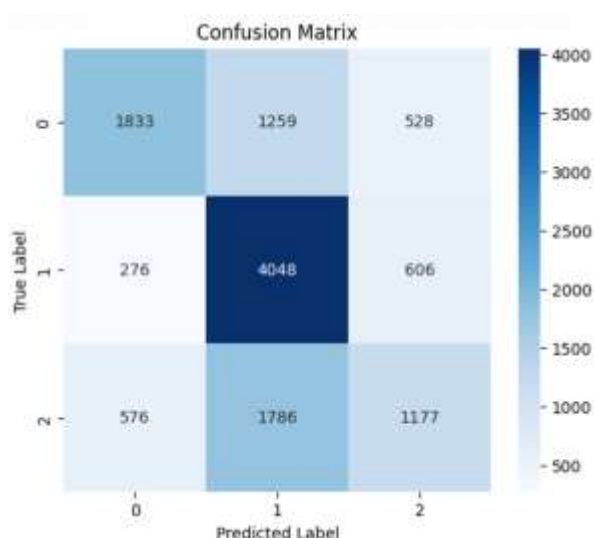


Fig. 11. Confusion Matrix Analysis

The confusion matrix visually represents the classification performance of the proposed explainable cyberbullying detection framework across three textual categories: hate speech, normal content, and offensive content. The diagonal elements of the matrix indicate correctly classified instances, while the non-diagonal elements represent misclassified samples. From the experimental results, the model correctly classified 1833 hate speech instances, 4048 normal

instances, and 1177 offensive instances. However, certain samples were incorrectly classified due to semantic overlap and contextual similarities among cyberbullying-related textual patterns. The highest classification performance was achieved for the normal category, whereas offensive content exhibited comparatively higher misclassification rates because offensive expressions often share linguistic similarities with hate speech and normal conversational text. Overall, the confusion matrix demonstrates that the proposed ensemble-based explainable machine learning framework is capable of effectively identifying cyberbullying-related content from social media textual data with satisfactory multiclass classification performance.

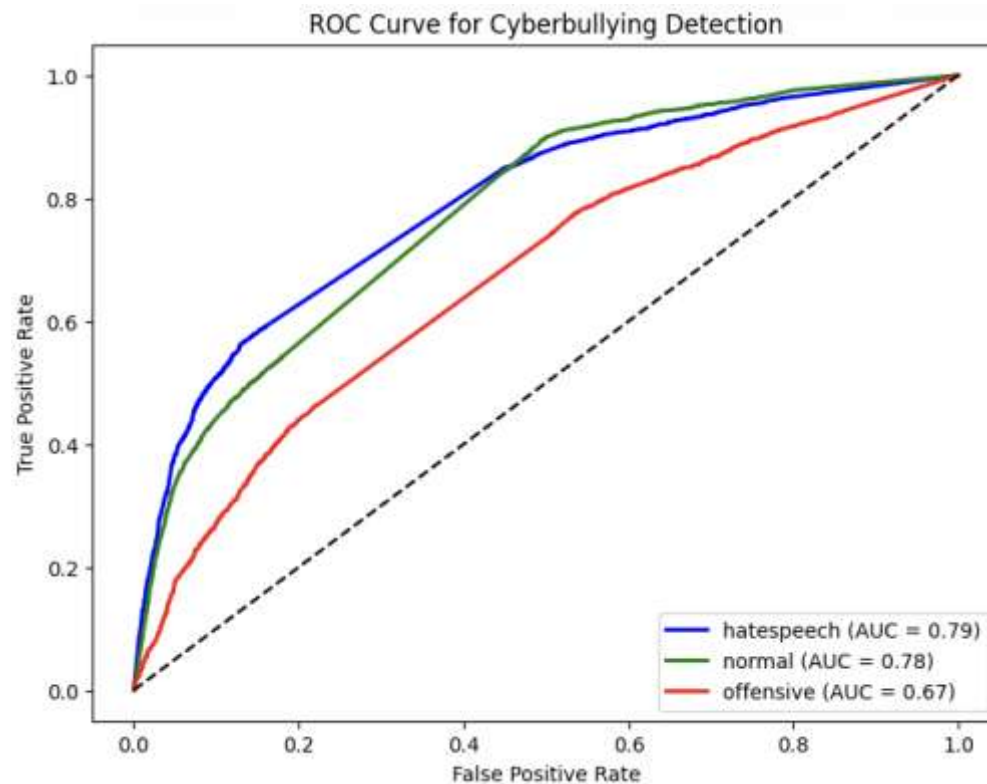


Fig. 12. ROC Curve Analysis

The ROC curve analysis illustrates the classification capability of the proposed explainable cyberbullying detection framework across hate speech, normal, and offensive textual categories. Experimental analysis demonstrates that the proposed ensemble learning framework achieved Area Under Curve (AUC) values of 0.79 for hate speech detection, 0.78 for normal content classification, and 0.67 for offensive content identification. The obtained ROC curves remain significantly above the random classification baseline, indicating the effectiveness of the proposed framework in distinguishing cyberbullying-related textual patterns from social media datasets. The comparatively lower AUC value for offensive content is primarily attributed to semantic overlap and contextual similarities between offensive and hate speech expressions commonly observed in real-world social media communication. Overall, the ROC analysis validates the robustness and practical applicability of the proposed explainable machine learning framework for cyberbullying detection tasks.

VII. LIMITATIONS OF THE PROPOSED WORK

Although the proposed framework demonstrates promising performance, certain limitations exist. The system primarily focuses on textual cyberbullying detection and may not effectively analyse multimedia content such as images, videos, and memes. The framework performance may vary for multilingual social media data and evolving slang expressions. Large-scale deep learning models may further improve contextual understanding and semantic analysis.

VIII. FUTURE SCOPE

Future enhancements may include integration of transformer-based deep learning models such as BERT and RoBERTa for contextual language understanding. Multilingual cyberbullying detection, image-based abuse analysis, and

multimodal explainable AI systems may further improve intelligent social media moderation. Cloud-based deployment and real-time large-scale social media monitoring can also enhance practical applicability of the proposed framework.

IX. CONCLUSION

This research presented an Explainable Machine Learning Framework for Cyberbullying Detection using NLP and Ensemble Learning techniques. The framework integrated text preprocessing, TF-IDF feature extraction, feature optimization, ensemble classification, SHAP explainability analysis, and real-time deployment modules. Experimental results demonstrated that the proposed framework effectively identifies cyberbullying content with high classification performance and improved interpretability. SHAP explainability analysis enhanced transparency by identifying influential textual features contributing toward prediction outcomes. The developed framework provides an intelligent, scalable, and transparent solution for automated cyberbullying detection in modern social media platforms.

REFERENCES

- [1] T. Davidson, D. Warmley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proc. 11th Int. AAAI Conf. Web and Social Media (ICWSM)*, Montréal, Canada, 2017, pp. 512–515.
- [2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 759–760.
- [3] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [4] D. Nandini and U. Schmid, "Explaining Hate Speech Classification with Model-Agnostic Methods," in *Proceedings of the Workshop on Text Mining and Generation (TMG)*, CEUR Workshop Proceedings, 2022.
- [5] S. Kumbam, M. Paoletti, and S. Marchand-Maillet, "Explainability and Adversarial Robustness for Hate Speech Detection Models," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2023.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
- [7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [8] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.
- [9] F. Chollet, *Deep Learning with Python*. Shelter Island, NY, USA: Manning Publications, 2018.
- [10] C. C. Aggarwal and C. Zhai, *Mining Text Data*. Boston, MA, USA: Springer, 2012.
- [11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [13] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
- [14] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [15] H. Mathew et al., "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, 2021.
- [16] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in *European Semantic Web Conference*, Springer, 2018, pp. 745–760.

BIOGRAPHY



Gangarapu Vasanth Kumar completed M. Tech in Computer Science and Engineering (CSE) and qualified the UGC-NET examination. His research interests include Machine Learning, Artificial Intelligence, Explainable AI, Natural Language Processing, and Cybersecurity.



D Lingaiah completed Master of Computer Application (MCA). His research interests include Cloud computing, Artificial Intelligence (AI), Machine Learning (ML), and IoT.



C Kiran Kumar completed Bachelor of B. Tech in Information Technology (IT). His academic and research interests include Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), Cybersecurity, and Data Analytics. He has actively participated in intelligent text classification systems, cyberbullying detection frameworks, and social media content analysis using machine learning techniques.