Decoding Digital Conversations: Emotion, Sentiment, and Cyberbullying Detection on Social Media

Harsimranjeet Singh
Department of AIT (CSE-AIML)
Chandigarh University
Mohali, Punjab, India
harsimranjeets257@gmail.com

Arnav Thakur
Department of AIT (CSE-AIML)
Chandigarh University
Mohali, Punjab, India
arnavthakur2004@gmail.com

Kirti
Department of AIT (CSE-AIML)
Chandigarh University
Mohali, Punjab, India
kirtisharma230819@gmail.com

***Abstract*— Online social media platforms have transformed global communication, fostering connectivity while also introducing challenges such as cyberbullying, emotional distress, and toxic content dissemination. While extensive research has been conducted on cyberbullying detection, sentiment analysis, and emotion classification as independent tasks, their integration into a unified framework remains underexplored. This study presents a comprehensive machine learning approach that simultaneously analyzes sentiment, emotion, and cyberbullying in online interactions. By employing XGBoost and Logistic Regression with TF-IDF-based feature extraction, our methodology effectively captures textual patterns associated with harmful content and emotional expression. Experimental results demonstrate the effectiveness of this approach, reinforcing its potential for enhancing automated content moderation and behavioral analysis in digital environments.**

***Keywords*— *Sentiment Analysis, Emotion Recognition, Cyberbullying Detection, Social Media Monitoring, Automated Content Moderations*—component, formatting, style, styling, insert (**key words**)**

I. INTRODUCTION

Online Social Media Platforms (OSMP) have become an integral part of modern life, enabling people to connect globally and share their ideas, thoughts, and experiences. While these platforms have revolutionized communication, they have also introduced significant challenges, such as cyberbullying, which remains a pervasive issue. Cyberbullying [1] involves using electronic means to harass, intimidate, or harm individuals, often with anonymity and on a large scale. Platforms like Facebook, Twitter, TikTok, and YouTube are frequently exploited for spreading harmful messages, images, or videos, resulting in widespread emotional and social consequences for victims [2].

In parallel, understanding user sentiments and emotions expressed on OSMP is critical for monitoring interactions and identifying potentially harmful behaviors. Sentiment analysis [3] categorizes text into positive, negative, or neutral sentiments, providing an overview of public opinion or individual attitudes. Emotion recognition (ER) [4], on the other hand, delves deeper into the emotional undertones of text, identifying states such as happiness, sadness, anger, and fear. These analyses have proven to be valuable for identifying trends, monitoring well-being, and addressing concerns like emotional distress.

While sentiment analysis, emotion recognition, and cyberbullying detection have been extensively studied as individual tasks, there is a noticeable gap in the research integrating all three into a unified framework. Existing studies often focus on one or two dimensions, neglecting the potential synergy of combining these tasks. For example, understanding sentiment and emotion together can provide context for detecting subtle forms of cyberbullying, while analyzing cyberbullying alongside sentiment can highlight broader emotional patterns among users. Addressing these dimensions collectively can provide a more comprehensive solution for identifying and mitigating harmful behaviour on OSMP.

Recent advancements [5] in machine learning (ML) and deep learning (DL) have greatly enhanced the ability to process and analyze social media data. Techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep belief networks (DBNs) have demonstrated exceptional capabilities in extracting high-level features and capturing complex patterns from textual and multimodal data [6,7]. These methods are particularly suited for large-scale social media datasets, where the data often includes informal language, slang, and diverse forms of expression.

While each of these areas—sentiment analysis, emotion recognition, and cyberbullying detection—has seen individual success, few studies have explored a unified approach combining all three. The majority of existing research focuses on one or two aspects, often neglecting the potential for synergy that could enhance overall effectiveness. This research aims to address this gap by integrating sentiment analysis, emotion detection, and cyberbullying classification into a comprehensive framework, providing a more robust and accurate solution for understanding user behavior on social media platforms.

II. RELATED WORK

The application of machine learning and deep learning techniques in sentiment analysis, emotion detection, and cyber-bullying classification has evolved significantly in recent years, with various studies exploring different models and methodologies to enhance accuracy.

A. Sentiment Analysis

In sentiment analysis, multiple approaches have been proposed to improve classification performance. Studies such as

[8] and [9] employed traditional pre-processing techniques like tokenization, Stopword removal, and stemming, while integrating word embeddings such as Word2Vec and GloVe for feature representation. These studies highlighted the effectiveness of deep learning models, with Bi-LSTM achieving the highest accuracy of 0.893 [8], and BERT outperforming LSTM with an accuracy of 0.924 in [9]. Further, [10] enhanced their approach by incorporating attention mechanisms and spect-based sentiment features, achieving 0.931 accuracy with BERT. Additionally, [11] explored aspect-based sentiment analysis (ABSA) combined with Named Entity Recognition (NER) and POS tagging, demonstrating that BERT outperformed other models with an accuracy of 0.917.

B. Emotion Recognition

Emotion classification has also seen promising results, particularly with transformer-based models. For instance, [12] and [13] used advanced embeddings like BERT and RoBERTa, achieving impressive results of 0.905 and 0.91, respectively. These studies capitalized on pre-processing steps such as NER and emoticon handling, and the use of transformer models for feature extraction. [14] pushed the boundaries of emotion classification, incorporating GPT-based emotion classifiers alongside BERT, resulting in an accuracy of 0.946, the highest among the reviewed studies.

C. Cyberbullying Detection

Cyber-bullying detection, a critical area of research, has benefited from a combination of feature extraction techniques and model innovations. Studies such as [15] and [16] integrated user metadata features with TF-IDF and n-grams, testing models like CNN and LSTM. CNN models were found to yield the best results, with [15] achieving an accuracy of 0.912. On the other hand, [17] demonstrated the efficacy of BERT over traditional machine learning models, achieving 0.94 accuracy in cyber-bullying detection. Further, [18] and [19] also explored deep learning models, with LSTM and SVM models achieving competitive results of 0.94 and 0.91, respectively.

While sentiment analysis, emotion recognition, and cyberbullying detection have been studied extensively, most research treats them as separate tasks. This study bridges that gap by integrating all three into a unified framework, improving the accuracy and robustness of online content analysis.

III. Proposed Methdology

Our proposed approach integrates sentiment analysis, emotion detection, and intent recognition into a unified framework, making it a novel contribution in the field. The methodology follows a structured workflow (Figure 1) consisting of data acquisition, preprocessing, feature extraction, model training & evaluation, and decision making.

The process begins with data acquisition, followed by preprocessing to clean and normalize data. In feature extraction, relevant linguistic and contextual features are identified to enhance model performance. The core novelty lies in training a model that simultaneously processes sentiment, emotion, and intent detection, ensuring a more comprehensive understanding of user input. The trained model is evaluated using standard metrics, and the decision-making phase interprets predictions to provide actionable insights.

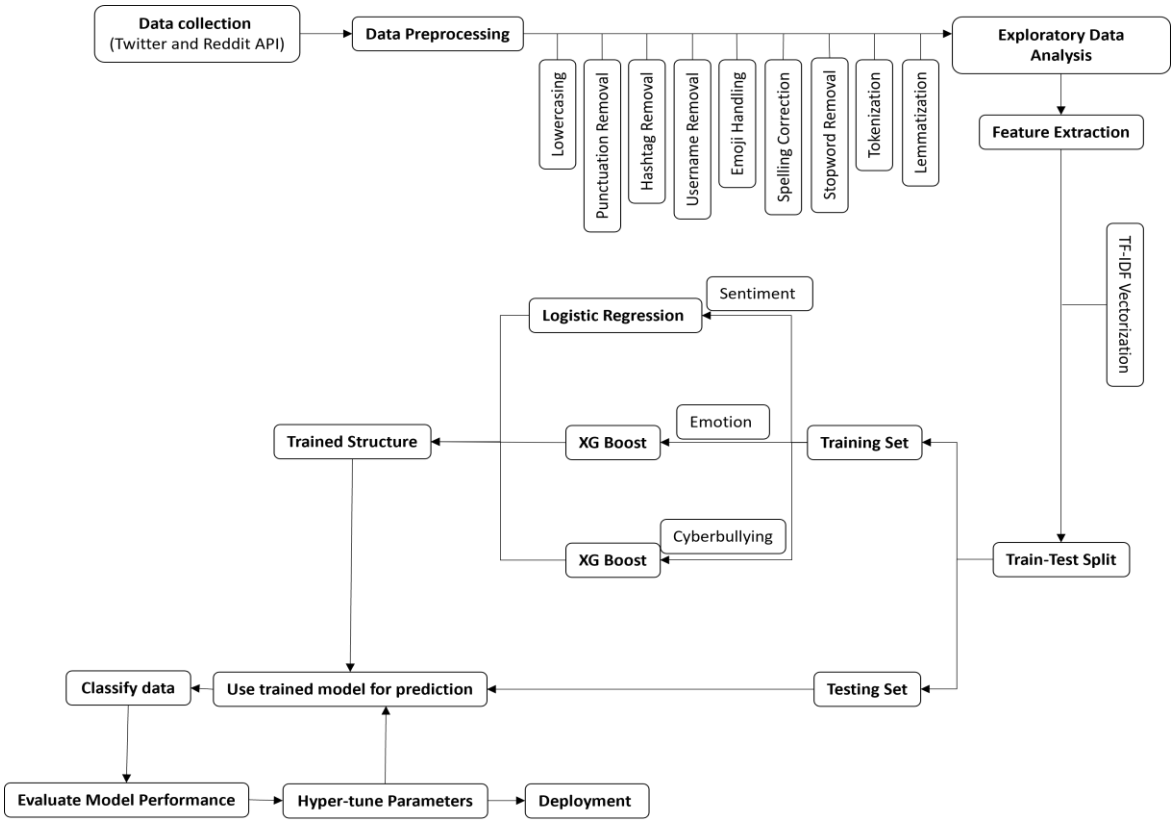

Fig. 1. Workflow

This sections will detail each stage—data preprocessing, feature extraction, model training, and evaluation—highlighting the techniques used to achieve an integrated multi-task learning framework.

A. *Data Collection*

The quality and diversity of the dataset are critical to the effectiveness of sentiment analysis, emotion recognition, and cyberbullying classification. For this research, data was collected from social media platforms using publicly available APIs, specifically from Twitter and Reddit. These platforms were chosen due to their extensive user engagement and the presence of both structured discussions (Reddit) and real-time interactions (Twitter).

Figure 2 presents the normalized distribution of data across platforms, indicating that Twitter contributes the majority share, followed by Reddit. This aligns with previous studies highlighting Twitter as a hotspot for intense public discourse, which often includes emotionally charged content and cyberbullying instances. Reddit, known for its structured discussions and community moderation, provides valuable contrasting data with diverse sentiment expressions.

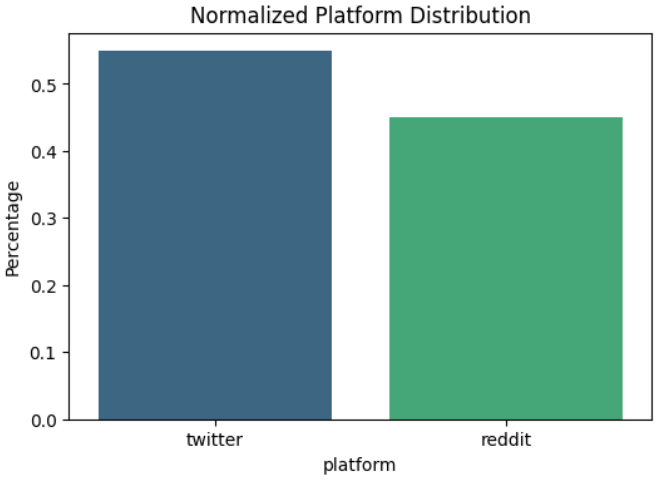

Fig. 2. Normalized Platform Distribution

The collected dataset consists of tweets and Reddit posts that have been manually labeled for sentiment (positive, negative), emotions (e.g., anger, joy, sadness, fear), and cyberbullying content (bullying vs. non-bullying). To ensure data quality, duplicate entries, spam, and irrelevant content were systematically removed. Additionally, careful balancing of sentiment classes, emotional categories, and cyberbullying instances was performed to prevent model bias.

Given the ethical implications of cyberbullying detection, strict anonymization protocols were followed to remove personally identifiable information. Only publicly available data was used, ensuring compliance with ethical guidelines for responsible AI research. These measures contribute to the reliability and fairness of the dataset, supporting its application in real-world moderation strategies.

B. *Data Preprocessing*

Raw textual data from social media is inherently noisy, containing extraneous characters, inconsistent formatting, and redundant elements. Effective preprocessing is essential to enhance model efficiency, ensuring more accurate sentiment, emotion, and intent detection. The key preprocessing steps include:

a) Lowercasing

To maintain uniformity and prevent models from distinguishing words based on capitalization, all text is converted to lowercase (e.g., "Happy" → "happy"). While proper nouns and abbreviations (e.g., "USA") may hold significance, a general lowercase transformation is applied, with contextual preservation considered during fine-tuning.

b) Punctuation Handling

Non-informative punctuation, such as commas and special characters (!, @, #, etc.), is removed to streamline text representation. However, punctuation that conveys emotion (e.g., "!!!" indicating excitement or anger) is mapped to corresponding emotion indicators (e.g., "!!!" → "[strong emotion]") to retain sentiment cues.

c) Hashtag and Username Removal

User mentions (@Username) and hashtags (#Topic) are removed as they primarily serve as metadata rather than direct sentiment indicators. Although hashtags may carry sentiment-related information (e.g., "#Happy"), their extraction as distinct features is beyond the scope of this study.

d) Emoji Handling

Emojis play a critical role in conveying sentiment and emotional intensity. Instead of discarding them, they are converted into textual descriptors (e.g., 😊 → "[smiling face]") to preserve sentiment-bearing elements within the text.

e) Spelling Correction

Misspelled words can mislead sentiment classification models. A spell-correction algorithm (e.g., TextBlob, SymSpell) is applied to standardize text while minimizing semantic distortions (e.g., "Ths moovie ws amazng!" → "This movie was amazing!").

f) Stopword Removal

Common stopwords (e.g., "the," "is," "and") are eliminated to reduce dimensionality and enhance feature relevance. However, sentiment-altering words (e.g., "not" in "not happy") are carefully retained to prevent unintended meaning shifts.

g) Tokenization

Text is segmented into individual tokens (e.g., "The movie was fantastic!" → ["The", "movie", "was", "fantastic", "!"]) for structured processing. While word-based tokenization suffices for traditional models, subword tokenization (e.g., WordPiece) is employed for deep learning models to handle out-of-vocabulary terms effectively.

h) Lemmatization

Words are transformed into their base forms while maintaining their grammatical integrity (e.g., "running" → "run", "flies" → "fly"). Unlike stemming, which simply trims suffixes,

lemmatization ensures accurate transformations based on linguistic rules, reducing redundancy in the dataset.

C. *Exploratory Data Analysis (EDA)*

1) *Word Cloud Analysis*

To gain insights into the most frequently occurring words and phrases within the dataset, a word cloud was generated. The word cloud highlights terms commonly used in online discussions, capturing a mix of general sentiment, emotional expressions, and cyberbullying-related language. It includes offensive slurs, derogatory remarks, and abusive terms alongside neutral or positive expressions that may appear in discussions related to cyberbullying. The generated word cloud (Figure 3) prominently features words associated with hate speech, cyberbullying, emotional distress, and victim responses. Examples include offensive slurs ("nigga," "fag," "retard"), words expressing hostility ("kill," "die," "stupid"), and emotionally charged terms ("hurt," "cry," "alone"). Additionally, slang and informal language such as "lol," "lmao," and "wtf" appear, indicating the conversational nature of online interactions. Some words also reflect self-expression and emotional relief, such as "vent," "relieved," and "finally."By analyzing these patterns, the word cloud helps identify key themes and linguistic markers that distinguish cyberbullying-related content from general online discussions. This visualization aids in understanding the frequency and context of specific terms, supporting further analysis in sentiment classification and cyberbullying detection.

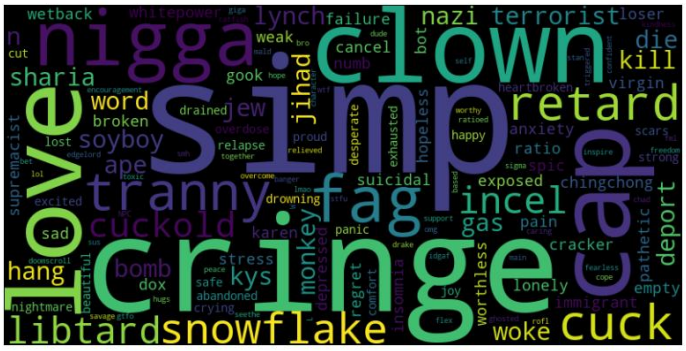

Fig.3 Word Cloud

2) *Cyberbullying Trends*

Cyberbullying distribution (Figure 4) illustrates the proportion of online interactions classified as cyberbullying versus non-cyberbullying. The analysis reveals that approximately 57% of the dataset contains instances of cyberbullying, with abuse and hate speech being the most prevalent categories. This significant presence of harmful interactions highlights the critical need for automated detection systems to enhance online safety and mitigate the negative impact of cyberbullying.

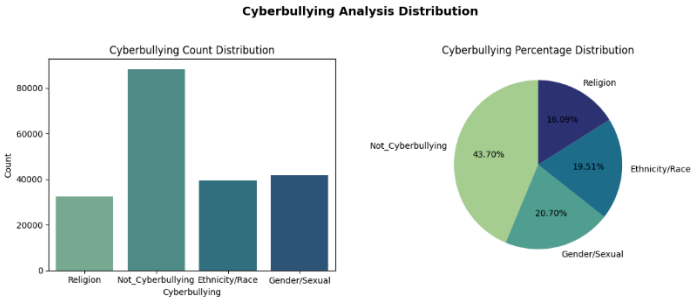

Fig. 4: Distribution of Cyberbullying Classes.

3) *Emotion Distribution*

The emotion analysis (Figure 5) categorizes text-based interactions into six primary emotions: anger, sadness, joy, fear, surprise, and love. The findings indicate that negative emotions (anger, sadness, and fear) collectively constitute approximately 49.79% of the dataset, reflecting a substantial level of distress and anxiety in online interactions. Among these, sadness (18.63%) and anger (14.77%) are the most prevalent, reinforcing prior research that negative emotions, particularly those tied to distress, tend to spread rapidly on social media.Conversely, positive emotions such as love (18.64%) and joy (15.88%) also hold a notable presence, contributing to a more balanced emotional landscape. These insights underscore the necessity for emotion-aware moderation strategies to mitigate the impact of harmful interactions while fostering healthy and constructive online discourse.

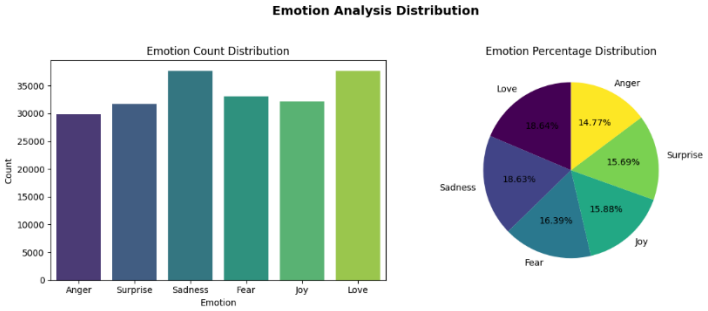

Fig. 5: Distribution of Emotion Classes

4) *Sentiment Analysis*

The sentiment analysis (Figure 6) classifies the dataset into positive and negative sentiments. The results indicate that positive sentiment slightly outweighs negative sentiment, accounting for 54.18% and 45.82%, respectively. While this challenges the assumption that social media is predominantly negative, the substantial proportion of negative sentiment posts suggests that frustration, dissatisfaction, and hostility remain significant concerns in online discussions.

These findings emphasize the importance of sentiment-aware content moderation frameworks to curb potential online toxicity while fostering a more positive and engaging digital environment.

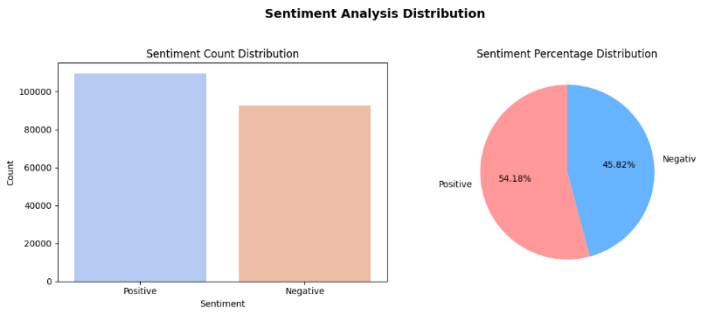

Fig. 6: Distribution of Sentiment Classes

5) Feature Extraction

Machine learning models require numerical representations of textual data. To achieve this, Term Frequency-Inverse Document Frequency (TF-IDF) is employed to quantify the importance of words within a document relative to the entire dataset. TF-IDF effectively captures the relevance of words by reducing the influence of commonly occurring terms while highlighting more distinctive words that contribute to meaning.The TF-IDF formula is given by:

$$TF - IDF = TF(w) \times IDF(w)$$

where:

- TF(w) = Frequency of word *w* in a document
- IDF(w) = Logarithm of inverse document frequency

This approach ensures that important words are assigned higher weights, enhancing the performance of traditional machine learning models in text classification tasks.

D. Train-Test Split

To train the models effectively, the dataset is split into 70% training data and 30% testing data. The training set is used for model learning, while the test set evaluates the model's generalization ability. A stratified split is applied to maintain class distribution across sentiment, emotion, and cyberbullying categories, ensuring balanced learning.

E. Model Training

This research employed two distinct machine learning models, tailored to the specific characteristics of the classification tasks. XGBoost was used for emotion recognition and cyberbullying detection, while Logistic Regression was selected for sentiment analysis. Both models were trained using TF-IDF feature vectors to convert textual data into numerical form, ensuring effective pattern recognition.

a) XGBoost for Emotion Recognition and Cyberbullying Classification

XGBoost, a high-performance gradient boosting algorithm, was chosen for emotion recognition and cyberbullying detection due to its ability to handle imbalanced datasets, manage missing values, and capture complex patterns in textual data. The model was trained using TF-IDF-transformed feature vectors, allowing it to distinguish between various emotional states and identify bullying-related language effectively. XGBoost's interpretability and scalability made it well-suited for these tasks, where nuanced language plays a crucial role.

b) Logistic Regression for Sentiment Analysis

For sentiment classification, Logistic Regression was selected due to its computational efficiency, ease of implementation, and strong performance in binary classification tasks. The input text was transformed using TF-IDF vectorization before being fed into the model. Logistic Regression was particularly advantageous because it provides probabilistic outputs, enabling better interpretability in distinguishing between positive and negative sentiments in social media interactions.

After training the models using TF-IDF feature representation, their performance was evaluated using standard classification metrics. The next section presents the results, highlighting the effectiveness of XGBoost in emotion recognition and cyberbullying detection, as well as Logistic Regression's performance in sentiment analysis.

IV. RESULTS AND EVALUATIONS

This section evaluates the performance of the proposed models for cyberbullying classification, emotion classification, and sentiment analysis. The models are assessed using accuracy, precision, recall, and F1-score, with confusion matrices (Figure 7,8,9) providing further insights into classification effectiveness.

A. Cyberbullying Classification

The cyberbullying detection model achieved an accuracy of 98%, demonstrating high precision across all categories: Ethnicity/Race, Gender/Sexual, Not Cyberbullying, and Religion. The classification report shows consistently high precision, recall, and F1-score values of 0.98, indicating that the model effectively distinguishes between cyberbullying and non-cyberbullying content. The confusion matrix (Figure 7) confirms minimal misclassification, validating the model's robustness in handling various forms of online abuse.

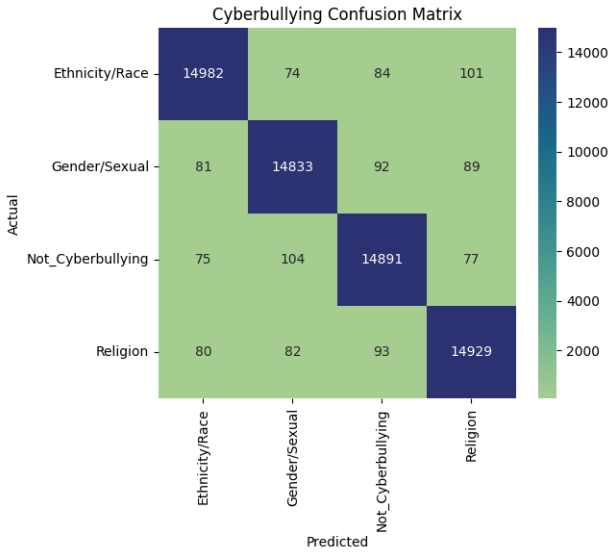

Fig.7. Cyberbullying Confusion Matrix

B. *Emotion Classification*

The emotion classification model achieved an accuracy of 90%, effectively categorizing emotions into Anger, Fear, Joy, Love, Sadness, and Surprise. The precision, recall, and F1-score for all emotions range between 0.90 and 0.91, demonstrating strong reliability in detecting emotional nuances. The confusion matrix (Figure 8) suggests a well-balanced classification, though minor misclassifications exist in closely related emotions (e.g., Sadness and Fear). This highlights the complexity of emotional expression in textual data.

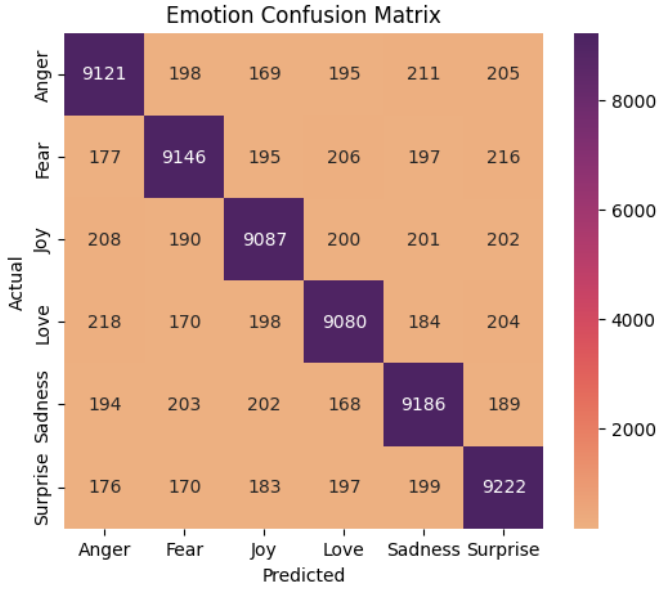

Fig.8. Emotion Confusion Matrix

C. Sentiment Analysis

The sentiment analysis model performed exceptionally well, achieving an accuracy of 96% in distinguishing between positive and negative sentiments. The classification report shows a precision, recall, and F1-score of 0.96 for both classes, confirming the model's ability to identify sentiment-driven expressions with high confidence. The confusion matrix (Figure 9) indicates balanced classification, with minimal misclassification errors.

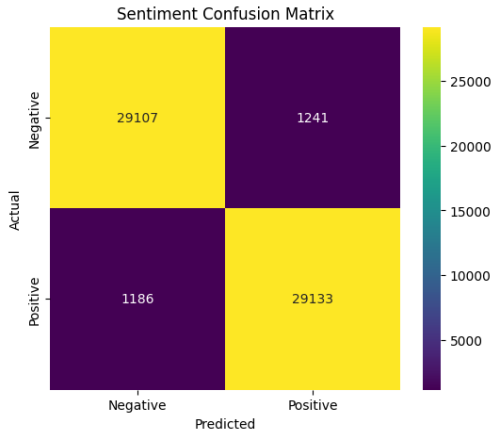

Fig.9. Sentiment Confusion Matrix

The evaluation results confirm that all models exhibit strong predictive capabilities and high classification accuracy across their respective tasks. The cyberbullying detection model demonstrated exceptional performance, while the emotion and sentiment analysis models effectively captured emotional and sentiment-based nuances. While the models show robustness and reliability, slight misclassifications in emotion detection suggest that contextual ambiguities could be further addressed in future improvements.

These findings validate the effectiveness of the proposed approach, making it a promising solution for online content moderation and behavioral analysis in real-world applications.

V. Conclusion

This study proposed a comprehensive machine learning-based approach that integrates cyberbullying classification, emotion classification, and sentiment analysis, rather than addressing these aspects in isolation. By combining these tasks, our approach enables a holistic understanding of online interactions, improving the accuracy and context-awareness of automated content moderation systems.

The models demonstrated high accuracy and reliability, with cyberbullying detection achieving 98% accuracy, emotion classification 90%, and sentiment analysis 96%. These results validate the effectiveness of machine learning techniques in identifying harmful behavior, emotional states, and sentiment expressions, ensuring a more nuanced and insightful analysis.

While the proposed approach is effective, further enhancements can improve its robustness and adaptability. Future work can focus on:

- Expanding training datasets to cover a more diverse range of linguistic and cultural contexts.
- Integrating contextual and semantic understanding to reduce misclassifications, potentially incorporating BERT-based models while maintaining computational efficiency.
- Developing real-time detection mechanisms that dynamically adapt to evolving online interactions.
- Enhancing interpretability to ensure AI-driven moderation decisions are transparent and explainable.

By addressing these areas, the proposed approach can evolve into a more adaptive, intelligent, and effective solution for safeguarding digital platforms, improving user experience, and mitigating online abuse in real time.

References

[1] Balakrisnan, V. and Kaity, M., 2023. Cyberbullying detection and machine learning: a systematic literature review. *Artificial Intelligence Review*, *56*(Suppl 1), pp.1375-1416.

[2] Shakeel, N. and Dwivedi, R.K., 2022. A survey on detection of cyberbullying in social media using machine learning techniques. In *Intelligent Communication Technologies and Virtual Mobile Networks: Proceedings of ICICV 2022* (pp. 323-340). Singapore: Springer Nature Singapore.

[3] Mahalakshmi, L. and Anbalagan, E., 2024, August. National Language Processing for Sentiment Analysis in Social Media-A Comprehensive Review. In *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)* (Vol. 1, pp. 504-508). IEEE.

[4] Al Maruf, A., Khanam, F., Haque, M.M., Jiyad, Z.M., Mridha, F. and Aung, Z., 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*.

[5] Hassan MM, Alam MGR, Uddin MZ, Huda S, Almogren A, Fortino G (2019) Human emotion recognition using deep belief network architecture. Inf Fusion 51:10–18

[6] Prakash PR, Anuradha D, Iqbal J, Galety MG, Singh R, Nee lakandan S (2023) A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification. J Control Decis 10(1):54–63

[7] Yadav SP, Zaidi S, Mishra A, Yadav V (2022) Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Arch Comput Method Eng 29(3):1753–1770

[8] Dang, N.C., Moreno-García, M.N., & De la Prieta, F. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics, 9(3), 483.

[9] Yadav, A. and Vishwakarma, D.K., 2020. Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), pp.4335-4385.

[10] Dang, C.N., Moreno-García, M.N. and De la Prieta, F., 2021. Hybrid deep learning models for sentiment analysis. Complexity, 2021(1), p.9986920.

[11] D'Ani+A9:G19ello, G., Gaeta, M. and La Rocca, I., 2022. KnowMIS-ABSA: An overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. Artificial Intelligence Review, 55(7), pp.5543-5574.

[12] Graterol, W., Diaz-Amado, J., Cardinale, Y., Dongo, I., Lopes-Silva, E. and Santos-Libarino, C., 2021. Emotion detection for social robots based on NLP transformers and an emotion ontology. Sensors, 21(4), p.1322.

[13] Nandwani, P. and Verma, R., 2021. A review on sentiment analysis and emotion detection from text. Social network analysis and mining, 11(1), p.81.

[14] Plaza-del-Arco, F.M., Curry, A., Curry, A.C. and Hovy, D., 2024. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. arXiv preprint arXiv:2403.01222.

[15] Mitra, S., Tasnim, T., Islam, M.A.R., Khan, N.I. and Majib, M.S., 2021, December. A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms. In 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

[16] Muneer, A. and Fati, S.M., 2020. A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. Future Internet, 12(11), p.187.

[17] Salawu, S., 2021. Detection and Prevention of Cyberbullying on Social Media (Doctoral dissertation, Aston University).

[18] Perera, A. and Fernando, P., 2021. Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, pp.605-611.

[19] Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E.D., Stringhini, G., Vakali, A. and Kourtellis, N., 2019. Detecting cyberbullying and cyberaggression in social media. ACM Transactions on the Web (TWEB), 13(3), pp.1-51.