World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

AI-powered sentiment analysis for classifying harmful content on social media: A case study with ChatGPT Integration

OLADAYO O. AMUSAN [1, *] and AMARACHI M. UDEFI [2]

[1] *Department of Big Data Science and Technology, University of Bradford, England, United Kingdom.*

[2] *Department of Computer Engineering Technology, Grundtvig Polytechnic Oba, Anambra State, Nigeria.*

World Journal of Advanced Research and Reviews, 2024, 24(03), 924–939

Publication history: Received on 28 October 2024; revised on 04 December 2024; accepted on 07 December 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.3.3710

Abstract

Social media platforms have become essential for communication but have also created spaces where harmful content, including cyberbullying, racism, and other abusive behaviors, thrives. This study employs AI-driven sentiment analysis to classify social media posts into three categories: Abusive, Neutral, and Harmless. A dataset of Twitter posts sourced from Kaggle was preprocessed through steps like noise removal, tokenization, and normalization to ensure readiness for analysis. The Sentiment Analysis Model (ChatGPT Integration) was utilized for classification, leveraging its advanced contextual capabilities to effectively analyze linguistic patterns. The model's performance, with an accuracy of 96%, sensitivity of 90%, and precision of 88%, was validated through a confusion matrix analysis, demonstrating its reliability in identifying harmful content. The findings highlight the model's potential as a scalable solution for mitigating online abuse. Future work will focus on addressing challenges such as class imbalance, integrating multilingual datasets, and implementing real-time monitoring to enhance its usability and impact.

Keywords: Sentiment Analysis; ChatGPT Integration; Social Media Content; Cyberbullying; Natural Language Processing (NLP); Preprocessing; Feature Extraction; Classification Model; Performance Metrics.

1. Introduction

Social media platforms have transformed people's interactions, providing communication, expression, and community-building spaces. However, the same platforms have become breeding grounds for harmful behaviors such as cyberbullying, harassment, and abusive language. Cyberbullying, as defined by Nazir and Thabassum (2021), is a specific type of mistreatment characterized by ongoing aggressive and harmful actions directed towards individuals or groups, often to cause emotional distress. The term "cyberbully" is derived from "bully," reflecting their shared intent and similar harmful effects. A dictionary defines a bully as an aggressive individual who perceives themselves as more powerful than others and insults or threatens others in an unhealthy manner. Such behavior has far-reaching consequences, affecting individuals' mental health and contributing to toxic online environments (Perera and Fernando, 2024). Identifying and mitigating harmful content is crucial for maintaining a safe and inclusive digital space.

According to a survey done by the Cyber Crimes Division (CID) of the Law Enforcement Authority, over 1000 instances of cyberbullying were documented in Sri Lanka. Approximately 90% of university students indicated they had encountered cyberbullying, and virtually all survey participants asserted familiarity with an individual who had been subjected to online harassment. 80% of cyberbullying events among Sri Lankans transpired on Facebook. Sixty-five percent of college students have disseminated embarrassing photographs or videos online. Fifteen percent of respondents shared private information online, nine percent propagated falsehoods and misleading information about others, and two percent uploaded abusive comments (Ariyadasa, 2019). Notwithstanding its advantages, social media

* Corresponding author: OLADAYO O. AMUSAN

can adversely affect individuals if it falls into inappropriate hands. The primary challenge is that, because of the huge audience and prolonged visibility, cyberbullying can proliferate swiftly.

Traditional methods, such as keyword filtering and manual moderation, have been deployed to combat abusive content but have significant limitations. Keyword filtering, while straightforward, often misinterprets context, leading to both false positives and false negatives (Litty et al., 2024). Although more accurate, manual moderation is labor-intensive and cannot keep up with the sheer volume of daily content generated (Gongane et al., 2022). These shortcomings highlight the need for automated, context-aware solutions.

Advancements in artificial intelligence (AI) and natural language processing (NLP) have unlocked new opportunities for tackling this challenge. Sentiment analysis, a technique that examines text's emotional tone and sentiment, has shown potential for detecting harmful language by leveraging machine learning models capable of understanding context and linguistic nuances (Hani et al., 2019). Among these, models like ChatGPT excel in analyzing textual data and categorizing it effectively, making them valuable tools for identifying abusive content online (Vanpech et al., 2024).

This study focuses on utilizing AI-powered sentiment analysis to classify social media posts into three categories: Abusive, Neutral, and Harmless. The research leverages a dataset of Twitter posts and employs advanced preprocessing techniques to clean and normalize the data. The sentiment analysis model integrates contextual understanding to accurately distinguish between the severity levels of content. The results aim to provide a scalable and effective solution for managing harmful content, contributing to the creation of safer online environments.

The results of this study demonstrate that integrating ChatGPT into the sentiment analysis workflow provides significant advantages. The model performs robustly distinguishing between categories, effectively reducing the likelihood of false positives and negatives.

2. Review of related literature

2.1. Evolution of Cyberbullying

Cyberbullying, albeit a relatively modern phenomenon, is rooted in traditional bullying. The advent of the internet and the subsequent development of social media platforms have transformed the nature and scope of bullying activities. Initially, internet abuse was confined to emails and chat rooms. Nevertheless, the rise of social networking platforms such as Facebook, Twitter, and Instagram has contributed to the evolution of cyberbullying, rendering it more ubiquitous and menacing. Initial studies on cyberbullying underlined its growth as a serious concern in the late 1990s and early 2000s, emphasizing the particular qualities that set it apart from traditional bullying.

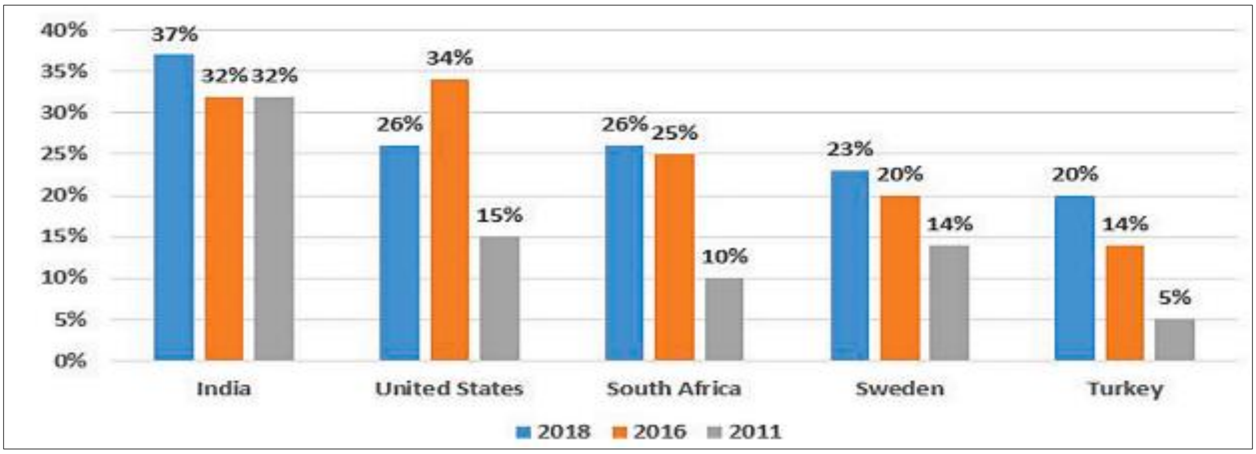


Figure 1 The proportion of children experiencing cyberbullying

Over the past decade, cyberbullying has become a significant issue, particularly impacting children and young people. A recent study by Cook (2020) involved interviews with parents about whether their children had experienced cyberbullying. The percentages of children affected over three different years are illustrated in Figure. 1. The findings reveal that this problem is escalating rapidly, regardless of a country's development level. For example, in Sweden, a nation considered highly developed, the rate of cyberbullying reached a critical level, showing a steady increase from 2011 to 2018.

2.1.1 Impact of Cyberbullying on Victims

According to a survey (Singhal and Bansal, 2013), social media platforms are significant hubs for cyberbullying, as illustrated in Figure 2. The data reveals that Facebook accounts for the highest proportion of cyberbullying incidents at 70%, followed by Twitter at 27%, and email at 25%. Additionally, the survey highlights the impact of cyberbullying on different groups: 48% of children have experienced bullying-related issues, along with 45% of girls and 30% of boys. Alarmingly, the survey also indicates that approximately 55% of suicide victims were linked to cyberbullying incidents. These have profound psychological, social, and academic effects on victims. It often leads to anxiety, depression, low self-esteem, and feelings of helplessness, with severe cases linked to self-harm and suicidal thoughts. Victims face chronic stress as the harassment invades even their personal spaces. Socially, it can cause withdrawal from peers, weakened support networks, and strained family relationships. Academically, cyberbullying hampers concentration lowers performance, increases absenteeism, and discourages engagement in digital learning environments. The wide-reaching impacts highlight its significant toll on overall well-being.

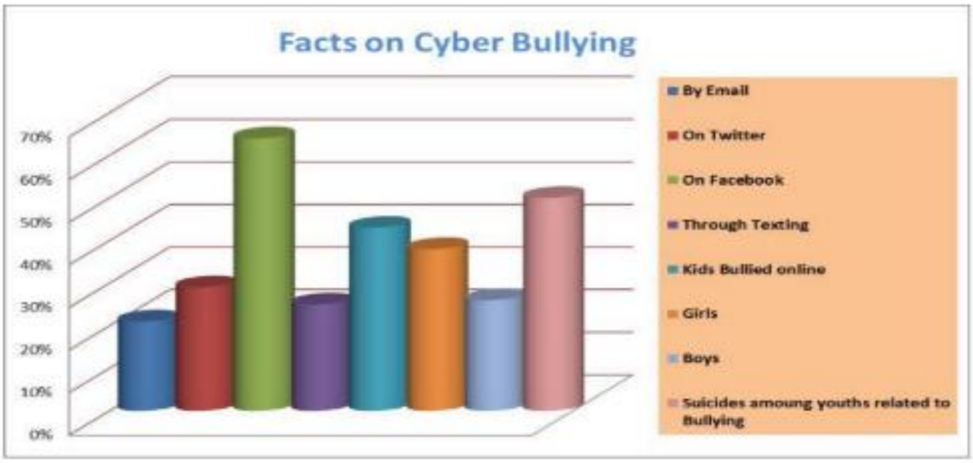


Figure 2 A graph depicting the analysis of cyberbullying**.**

2.1.2 Recognizing Abusive Content

Detecting offensive content faces several language barriers, often leading to classification errors:

- **Humor, Irony, and Sarcasm**: These rhetorical strategies complicate abuse detection by masking harmful intent. Content perceived as humorous or sarcastic can still propagate prejudice and stereotypes, often justified or concealed as satire. Understanding the author's intent and context remains challenging, as such content may unintentionally validate negative stereotypes (Liu et al., 2022).
- **Spelling Variations***:* Social media often features non-standard spellings, such as elongated words ("ohh") or substitutions ("kewl" for "cool"), which reflect cultural identity but also hinder detection. These variations can be adversarial, designed to evade algorithms and create out-of-vocabulary terms that increase classification errors. While text normalization can help, it risks losing meaningful social context. Character or subword-level language modeling offers a more robust approach (Malik et al., 2023).
- **Polysemy**: Words with multiple meanings, including euphemistic or coded phrases, allow covert expression of hate. For instance, benign terms like "Skype" or nominalizations like "the Mexicans" can carry derogatory connotations depending on context. Addressing this requires context-sensitive word representations to distinguish benign from harmful uses (Song et al., 2018).
- **Long-Range Dependencies**: Abusive content often spans multiple sentences or conversation threads, which short-post-focused models struggle to capture. Multi-user dynamics and extended contexts, such as those in Reddit or Wikipedia discussions, require datasets with diverse and longer content for more effective detection (wang et al., 2024).

2.2. Sentiment Analysis

Sentiments are expressions of favorable or negative judgments expressed by language, words, or speech patterns. They embody the emotional tone inside written expressions and serve as a vital signal of human communication (Al-Shabi, 2020). Whether conveyed through written reviews, social media posts, or news stories, sentiments cover a wide spectrum of emotions that impact our perceptions of beliefs and attitudes. Perspectives provided by thought leaders and ordinary persons equally influence the decision-making processes of others, underscoring the important role of sentiment in determining collective and individual choices (Judge et al., 2023). Sentiment analysis, often known as opinion mining, is a natural language processing (NLP) technique that finds sentiment in textual data. It plays a crucial role in understanding user emotions, consumer feedback, and public opinion in today's digital communication era.

2.3. ChatGPT and Contextual NLP Models

ChatGPT, developed by OpenAI, represents a significant advancement in AI-driven text analysis. Built on the GPT architecture, ChatGPT excels in understanding context, detecting sentiment, and generating coherent responses. Its ability to analyze subtle forms of abusive content, such as sarcasm or implicit bias, makes it a valuable tool for cyberbullying detection. Studies leveraging GPT-based models for sentiment analysis have reported significant improvements in classification accuracy compared to traditional methods. For example, Vanpec et al., (2024) demonstrated that GPT-3 models could outperform baseline classifiers in identifying abusive language by leveraging their deep contextual understanding.

2.4. Current Methods for Identifying Cyberbullying and Offensive Content

Sentiment analysis, a crucial component of natural language processing (NLP), has seen significant advancements through the incorporation of machine learning (ML) and artificial intelligence (AI) techniques. Leveraging annotated datasets, ML models have demonstrated proficiency in identifying and categorizing offensive content. Meanwhile, deep neural networks excel in uncovering intricate patterns and contextual nuances across text, images, and videos, particularly in detecting cyberbullying (Al-Qablan, 2023).

2.4.1 Content Filtering and Keyword Analysis

This approach utilizes algorithms to autonomously detect and remove offensive content by analyzing keywords, patterns, and context. Natural Language Processing (NLP) techniques are applied to examine language use and identify abusive patterns.

Ghiassi et al. (2013) introduced a hybrid system that integrates n-gram analysis with dynamic artificial neural networks for sentiment analysis on Twitter data. This method successfully captured nuanced sentiment variations by combining content-based features with neural network techniques.

Dubey et al., (2016) focused on sentiment classification through keyword analysis, highlighting the critical role of keywords in determining sentiment. They demonstrated that combining keyword-based features with machine learning algorithms significantly enhances classification accuracy.

Singh et al., (2020) studied hybrid approaches that mix content analysis with keywords and machine translation to do sentiment analysis across several languages. Their findings indicated that adding keyword-based translations increases sentiment categorization accuracy in varied linguistic contexts

2.4.2 Artificial Intelligence (AI) and Machine Learning

The combination of machine learning (ML) and artificial intelligence (AI) approaches has led to substantial advancements in sentiment analysis, a crucial aspect of natural language processing (NLP) (Udefi et al., 2023). According to Al-Qablan (2023), deep neural networks are excellent at identifying intricate patterns and contexts in text, photos, and videos about cyberbullying, whereas machine learning models trained on annotated datasets are proficient in identifying and categorizing objectionable content.

D'Aniello et al., (2022) provided a comprehensive overview of sentiment analysis, focusing on foundational principles and traditional methodologies, paving the way for incorporating ML and AI approaches. Yadav et al., (2020) shifted from rule-based systems to supervised ML for sentiment classification, using labeled datasets and advanced learning algorithms.

Kim and Jeong (2019) showcased the efficacy of deep learning models, specifically convolutional neural networks (CNNs), in identifying intricate linguistic patterns for sentiment classification. Mahmoud (2018) emphasized the challenges of emotion analysis in concise social media content, highlighting the importance of machine learning (ML) and artificial intelligence (AI) in overcoming these complexities.

Gimenez (2021) explored the application of deep learning methods, such as recurrent neural networks (RNNs), in multilingual sentiment analysis, showcasing the flexibility of ML and AI in handling linguistic diversity. These advancements underline the growing importance of AI-driven techniques in understanding and interpreting sentiment across varied contexts.

2.4.3 Pattern Recognition

Examining user behavior patterns, including post frequency and timing, can aid in identifying potential cases of cyberbullying. This approach focuses on analyzing user interactions to detect aggressive or irregular behaviors that may signify abuse.

Kennedy et al., (2021) provided a foundational perspective on sentiment analysis, focusing on traditional methodologies and introducing key concepts related to pattern recognition. Their study also highlighted the challenges inherent in pattern recognition, emphasizing the need for continued advancements in this area.

Li (2023) demonstrated the application of pattern recognition in classifying anime genres, showcasing its potential for effectively categorizing sentiments. Similarly, Aziz et al., (2017) employed supervised pattern recognition techniques, using labeled datasets and learning algorithms to achieve high accuracy in sentiment classification.

Beigi et al., (2021) pioneered the use of unsupervised pattern recognition, leveraging semantic orientation to classify sentiments, marking a transition from rule-based to automated approaches. Truong et al., (2017) explored the use of convolutional neural networks (CNNs), highlighting their growing role in advancing pattern recognition for sentiment analysis.

2.4.4 Contextual Analysis

Context awareness plays a vital role in distinguishing between harmless interactions and harmful cyberbullying by understanding the context of exchanged messages (Steer, 2022). Semantic analysis, as defined by Goddard (2011), involves examining the meanings behind words and phrases to uncover the intent behind communication.

Ma et al., (2018) presented Sentic Computing, a paradigm for sentiment analysis that incorporates contextual analysis at the idea level and is based on common sense. This technique set the groundwork for interpreting sentiments within a broader contextual framework. Similarly, Wilson et al., (2005) showed how contextual analysis may effectively capture nuanced feelings by proposing a supervised learning technique for detecting contextual polarity at the phrase level.

Ye et al., (2021) emphasized the importance of contextual analysis in their Twitter corpus study, focusing on the unique challenges of analyzing short-form content on social media. Their research highlighted the role of contextual understanding in addressing the nuances specific to platforms like Twitter. Table 1 provides a summary of the reviewed studies.

Despite advancements in sentiment analysis and NLP, several challenges remain. Many models struggle to differentiate between similar sentiments with varying intensity levels, such as distinguishing between neutral and mildly abusive content (Nguyen et al., 2020). Additionally, while models like ChatGPT have shown promise, their reliance on large-scale pretraining datasets may introduce biases or limit their applicability to specific cultural or linguistic contexts. This study addresses these gaps by integrating ChatGPT into a custom classification pipeline tailored to classify social media posts into Abusive, Neutral, and Harmless categories, thereby providing a scalable and context-aware solution for content moderation.

Table 1 A summary of the studies reviewed

Author	Datasets	Methods	Data Mining Task	Compatible Language	Features	Accuracy Result
Chavan and Shylaja (2015)	Kaggle	Logistic Regression & SVM	Classification	English	Textual	LR=73.76% SVM= 77.65%
Chen and Delany (2017)	Online services	SVM	Classification	English	Textual, content-based, & context-based features	SVM= 64%
Zhao and Mao (2016)	Twitter & MySpace	Semantic-enhanced marginalized denoising auto-encoder	Classification	English	Textual & semantics	Twitter = 84.9% MySpace = 89.7%
Rosa et al., (2018)	Me Formspring	Fuzzy Fingerprints	Classification	English	Textual	FFP= 82.9
Van Bruwaene et al., (2020)	VISR dataset	Multi-technique annotation & (SVM, CNN, and XGBoost)	Classification	English	Textual	83.7%, 86.8%, & 86.9% respectively.
Sugandhi et al., (2016)	Twitter	SVM, NB, and KNN	Classification	English	Textual	SVM= 91.31%, NB=87.65%, & KNN= 88.87%.
Bin Abdur Rakib and Soon (2018)	Reddit Formspring, Twitter, & Wikipedia	word2vec skip-gram models with RFCNN, LSTM, BLSTM, & BLSTM	Classification	English	Textual	90%
Al-Ajlan and Ykhlef, (2018)	Twitter	CNN	Classification	English	Textual	CNN = 91%
Banerjee et al., (2019)	Twitter	CNN	Classification	English	Textual	CNN = 93.97
Bozyiğit, et al., (2019)	Twitter	MNB	Classification	Hindi & Marathi	Textual	91.0%
Zhang et al., (2016)	Twitter, Formspring	PCNN	Classification	Turkish	Textual	PCNN = 92.9%
Cheng et al., (2019a)	Instagram, Vine	XBully	Classification	English	Textual, Collaborative	XBully = 90.0%
Cheng et al., (2019)	Twitter	PI-Bully	Classification	English	Textual, User-based	PI-Bully = 80%
Riadi (2017)	Twitter	Naive Bayes	Classification	English	Textual	NB = 86.97%

3. Research methodology

This research is structured into several key stages to ensure a systematic and effective approach. The initial stage involves data collection through the Twitter API, where relevant posts are retrieved to form the dataset. Following this, the preprocessing stage focuses on cleaning and preparing the raw data for analysis by removing noise, normalizing the text, and tokenizing it into structured components. The next stage employs the Sentiment Analysis Model, enhanced by ChatGPT integration, to classify each post's sentiment into one of three categories: Abusive, Neutral, or Harmless. This streamlined process is visually represented in Figure 3, demonstrating the logical flow from raw data acquisition to sentiment classification.

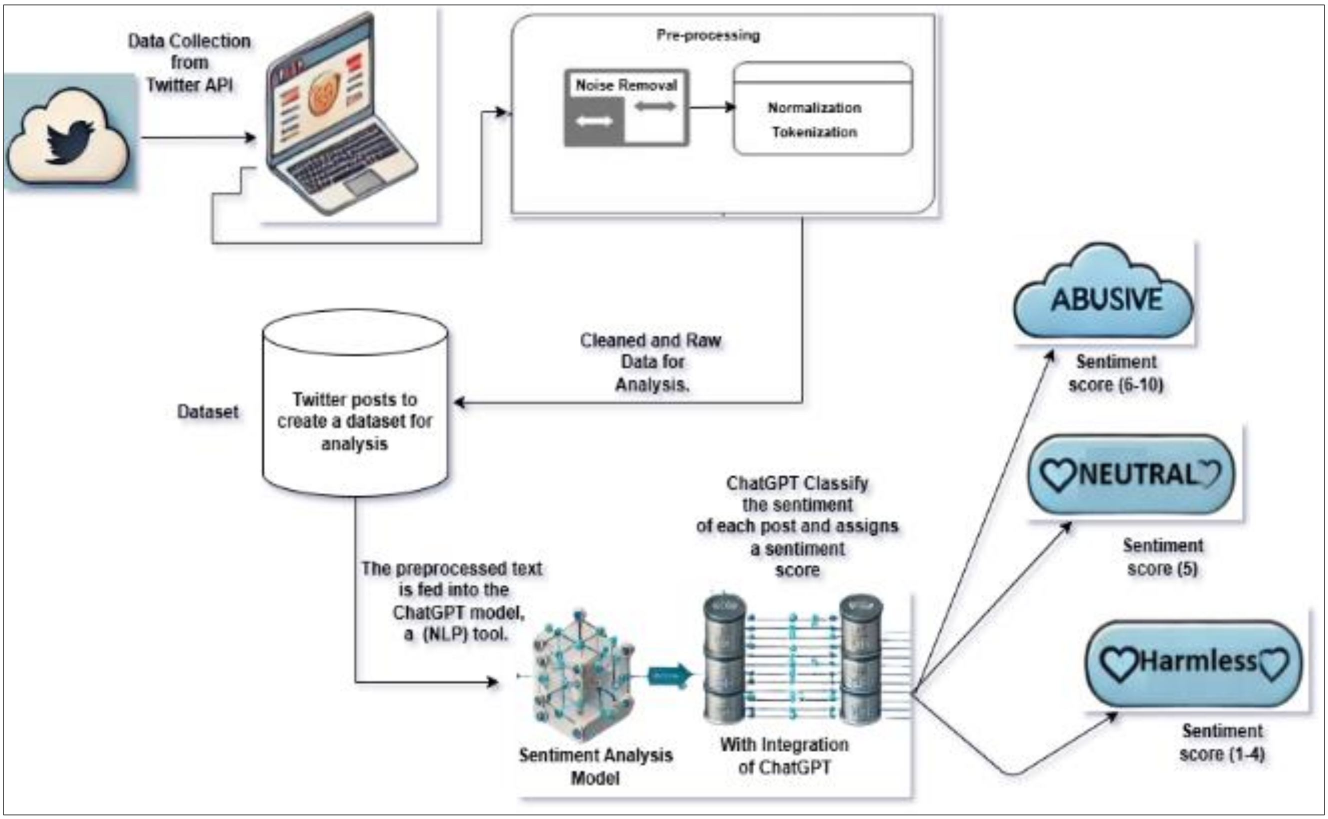

Data Collection from Twitter API
Pre-processing
Noise Removal
Normalization
Tokenization
Cleaned and Raw Data for Analysis.
Dataset
Twitter posts to create a dataset for analysis
The preprocessed text is fed into the ChatGPT model, a (NLP) tool.
ChatGPT Classify the sentiment of each post and assigns a sentiment score
Sentiment Analysis Model
With Integration of ChatGPT
ABUSIVE
Sentiment score (6-10)
NEUTRAL
Sentiment score (5)
Harmless
Sentiment score (1-4)

Figure 3 The developed Model.

3.1. Data Collection

The data collection process for this research involved gathering tweets and comments from social media platforms, focusing on a dataset containing various emotions. This dataset, sourced from Kaggle (2023), includes comments categorized as abusive, harmless, and neutral. The Twitter dataset, deemed most suitable for the study's objectives, is provided in CSV format with five columns: Index, Text, Annotation, Label, and Graded Value. It comprises 505 entries, offering diverse textual data for analysis.

3.2. Preprocessing

Before analysis, the raw data undergoes a series of preprocessing steps to ensure consistency and usability:

- **Noise Removal**: Eliminates unnecessary elements such as URLs, special characters, emojis, and hashtags.
- **Tokenization:** Break down text into individual words or meaningful components for further processing.
- **Normalization:** Standardizes the text by converting it to lowercase and removing stopwords (common words that do not contribute to sentiment).
- **Data Cleaning:** Ensures the dataset is free from duplicates, irrelevant entries, and incomplete records.

3.3. Sentiment Analysis Model (ChatGPT Integration)

This study utilized ChatGPT, a cutting-edge natural language processing (NLP) model, for sentiment classification. The model was fine-tuned to analyze the sentiment of individual posts, leveraging its advanced contextual comprehension to accurately classify sentiments. ChatGPT's capability to interpret subtle nuances, including sarcasm, implicit abuse, and emotional intensity, enables it to provide precise evaluations of content.

Steps in Sentiment Analysis:

- Tokenization is the first step in preprocessing text for transformer models. ChatGPT splits input text into smaller units called tokens (e.g., words, subwords, or characters). These tokens are then converted into numerical indices.

$$For\ an\ input\ sentence\ S = [w_1, w_2,\ .\ .\ ., w_n] \qquad (1)$$

tokenization produces a sequence of tokens $T = [t_1, t_2,\ .\ .\ ., t_m]$ (2)

where $m \geq n$. Each token t_1 is mapped to a unique integer index from a vocabulary V: $t_1 \rightarrow index\ (t_1)$.

- Embedding Layer: Tokens are converted into dense, high-dimensional vectors (embeddings) that represent semantic meaning. If the embedding dimension is d, the embedding matrix E has dimensions $|V| \times d$, where:

$$E = [e_1, e_2,\ .\ .\ ., e_{|V|}] \qquad (3)$$

The input sequence T is transformed into a matrix X:

$$X = [E[t_1], E[t_2],\ .\ .\ ., E[t_m]] \qquad (4)$$

Where:

$E[t_1]$, : The embedding vector for the token t_1.

- Positional Encoding: Transformers lack an inherent sense of sequence order, so positional encoding is added to embeddings. The position p of each token is encoded using sine and cosine functions to generate a unique vector $PE|p|$.

$$PE|p, 2i| = sin\left(\frac{p}{10000^{\frac{2i}{d}}}\right), \qquad (PE|p, 2i+1| = cos\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \qquad (5)$$

The final input to the transformer is:

$$Z = X + PE \qquad (6)$$

- Attention Mechanism: The core of the transformer model is the self-attention mechanism, which computes the importance of each token relative to others in the sequence. This involves three learned matrices: Q (queries), K (keys), and V(values).
  - ✓ Compute the attention scores:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (7)$$

Where:

Q, K, V: Linearly transformed versions of Z and d_k: The dimensionality of keys.

  - ✓ The attention weights highlight the relationships between tokens, allowing the model to focus on relevant parts of the input.
- Feedforward Network: The output of the attention mechanism is passed through a position-wise feedforward network:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \qquad (8)$$

Where:

W_1, W_2: Weight matrices and b_1, b_2: Bias vectors.

This non-linearity helps capture complex patterns in the data.

- Sentiment Classification with Integrated Scoring: The classification process includes assigning probabilities to each sentiment category and mapping these probabilities to corresponding scores:
  - ✓ Prediction of Sentiment Scores

The output of the classification layer is a probability distribution over the sentiment categories:

$$P(x|y) = [P(\,Abusive|x), P(\,Neutral|x), P(\,Harmless|x)] \qquad (9)$$

These probabilities represent the model's confidence in each category. The category with the highest probability is selected:

$$\hat{y} = argmax\, P(x|y) \qquad (10)$$

  - ✓ Mapping Sentiment Scores to Categories

The predicted sentiment score $\hat{y}$ is then mapped into one of three categories based on the score range:

    - o Abusive (6–10): Indicates offensive language or cyberbullying.
    - o Neutral (5): Refers to ambiguous or contextually unclear content.
    - o Harmless (1–4): Covers positive or benign interactions.

  - ✓ Mathematical Representation of Classification

If $P(x|y)$ is the probability for a given class, the sentiment score S is derived as:

$$S = \sum_{i=1}^{C} P(y_i|x)\,.\,w_i \qquad (11)$$

Where:

C: Total number of categories (3 in this case)

w_1: Weight assigned to each class ($w_{Abusive}$ =10, $w_{Neutral} = 5, w_{Harmless} = 1$)

  - ✓ Final Sentiment Label Assignment

Based on S, the model assigns the sentiment category:

    - o $S \in [6{,}10] \rightarrow Abusive$
    - o $S = 5 \rightarrow Neutral$

- $S \in [1,4] \rightarrow Harmless$

3.4. Evaluation Metrics

The system's performance was evaluated using the following metrics:

- Accuracy Measures the proportion of correctly classified posts to the total number of posts:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \qquad (12)$$

- Precision indicates the accuracy of the "Abusive" classification:

$$Precision = \frac{True\ Positive}{True\ Positives\ +\ False\ Positives} \qquad (13)$$

- Sensitivity (Recall) measures the model’s ability to identify all "Abusive" posts correctly:

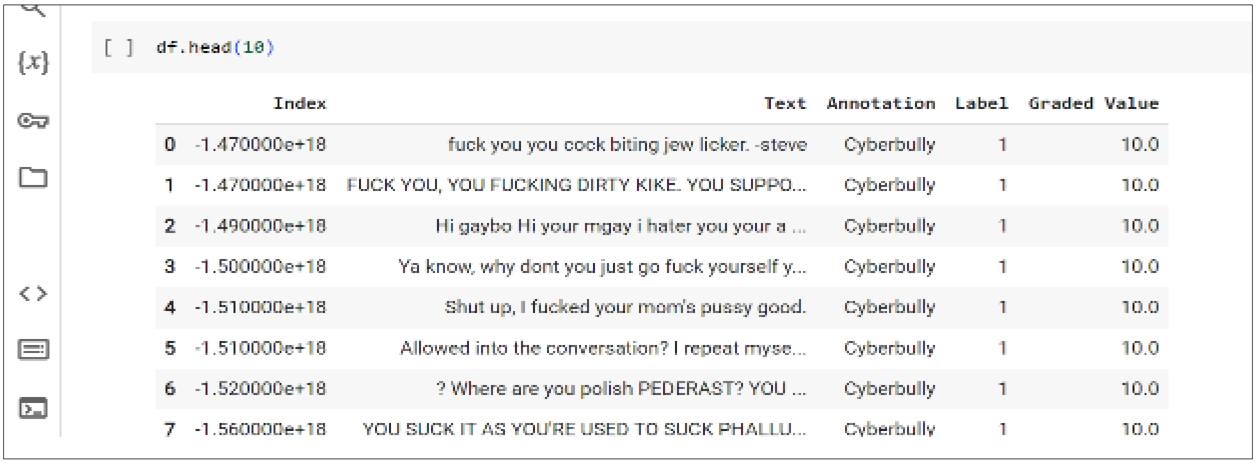
$$Recall = \frac{True\ Positive}{True\ Positives\ +\ False\ Negatives} \qquad (14)$$

- F1 Score combines Precision and Recall into a single metric:

$$Recall = 2.\ \frac{Precision \cdot Recall}{Precision + Recall} \qquad (15)$$

4. Result and Discussion

This section presents the experimental results of the proposed approach. Following the successful preprocessing and selection of relevant datasets, each dataset was evaluated using the language model, and the results were compiled. Scores were assigned to each dataset as described in Equations 1–11. These scores are documented, with the final score for each statement provided in the "Graded Value" column, as illustrated in Figure 4. Additionally, Figure 5 features a pie chart depicting the score distribution based on the nature of the tweets/posts. The analysis reveals that neutral comments accounted for the highest proportion, followed by harmless ones, while abusive content represented the lowest proportion.



```
[ ]  df.head(10)
```

```
          Index                                              Text   Annotation  Label  Graded Value
0  -1.470000e+18           fuck you you cock biting jew licker. -steve   Cyberbully      1          10.0
1  -1.470000e+18  FUCK YOU, YOU FUCKING DIRTY KIKE. YOU SUPPO...   Cyberbully      1          10.0
2  -1.490000e+18            Hi gaybo Hi your mgay i hater you your a ...   Cyberbully      1          10.0
3  -1.500000e+18       Ya know, why dont you just go fuck yourself y...   Cyberbully      1          10.0
4  -1.510000e+18            Shut up, I fucked your mom's pussy good.   Cyberbully      1          10.0
5  -1.510000e+18       Allowed into the conversation? I repeat myse...   Cyberbully      1          10.0
6  -1.520000e+18             ? Where are you polish PEDERAST? YOU ...   Cyberbully      1          10.0
7  -1.560000e+18    YOU SUCK IT AS YOU'RE USED TO SUCK PHALLU...   Cyberbully      1          10.0
```

Figure 4 Tail view of the dataset showing a sample of the Score

Based on the chart conversation in Figure 6, the scores reflecting ChatGPT's evaluation of each dataset are provided. Over half of the datasets were categorized as neutral. Notably, approximately 100 datasets received the most negative score of 10, while around 50 datasets each were assigned scores of 9 and 8. Furthermore, approximately 100 datasets were given a score of 0, indicating no negativity.

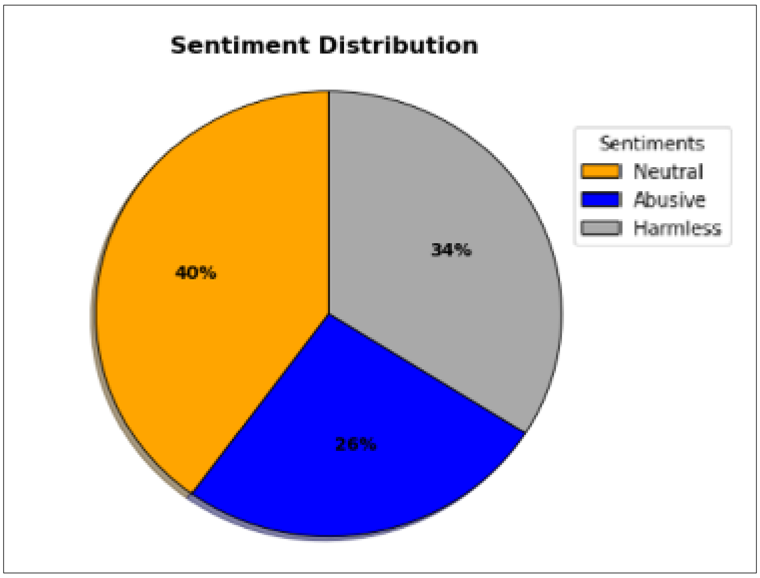


Figure 5 Pie Chart showing the distribution of each class of Graded Value

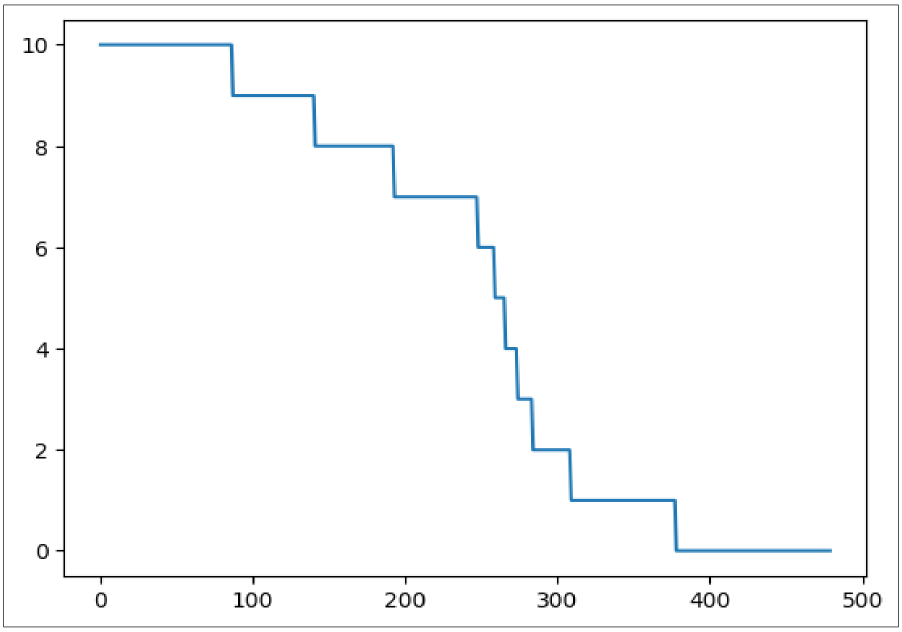


Figure 6 Chart showing the score of each class of dataset

The Word Cloud Python code was employed to visualize commonly used abusive words and expressions associated with cyberbullying as shown in Figure 7. In the word cloud, frequently occurring words are displayed in larger, bolder text. Offensive and vulgar terms such as "fuck," "shit," "asshole," and "bitch" are prominently featured due to their repeated appearance in the analyzed text. The size and prominence of each word in the visualization correspond to its frequency, with words like "fuck" standing out significantly, highlighting both their frequent use and intensity.

The graph in Figure 8 illustrates the performance of the Sentiment Analysis Model (ChatGPT Integration) using three key metrics: Accuracy, Sensitivity, and Precision. These metrics provide valuable insights into the model's effectiveness in classifying social media posts into Abusive, Neutral, and Harmless categories.

The model achieved an impressive accuracy of 96%, indicating that most posts were correctly classified. This high accuracy highlights the model's robustness in distinguishing between abusive, neutral, and harmless content. The strong performance can be attributed to effective preprocessing steps, efficient feature extraction, and the seamless integration of ChatGPT.

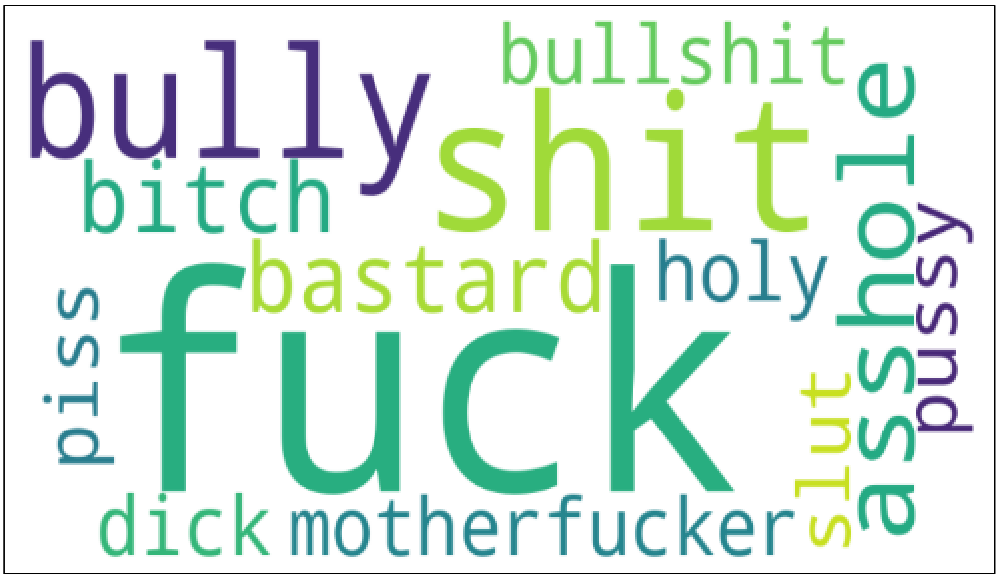


Figure 7 Word cloud showing the frequently used abusive words in the dataset

Sensitivity (90%), also referred to as recall, measures the model's ability to identify all relevant instances of abusive content. A sensitivity score of 90% suggests that the model successfully captures the majority of abusive posts. However, there is still room for improvement, as some subtle cases may be missed. Enhancing sensitivity could involve augmenting the training dataset with more diverse examples or refining the model to better handle ambiguous language and nuanced expressions.

Precision (88%) quantifies the accuracy of positive predictions, reflecting how many posts labeled as abusive were genuinely abusive. An 88% precision score indicates a slightly lower ability to avoid false positives. This metric is critical for ensuring that benign content is not mistakenly classified as abusive, particularly in real-world applications where such misclassifications can have significant consequences.

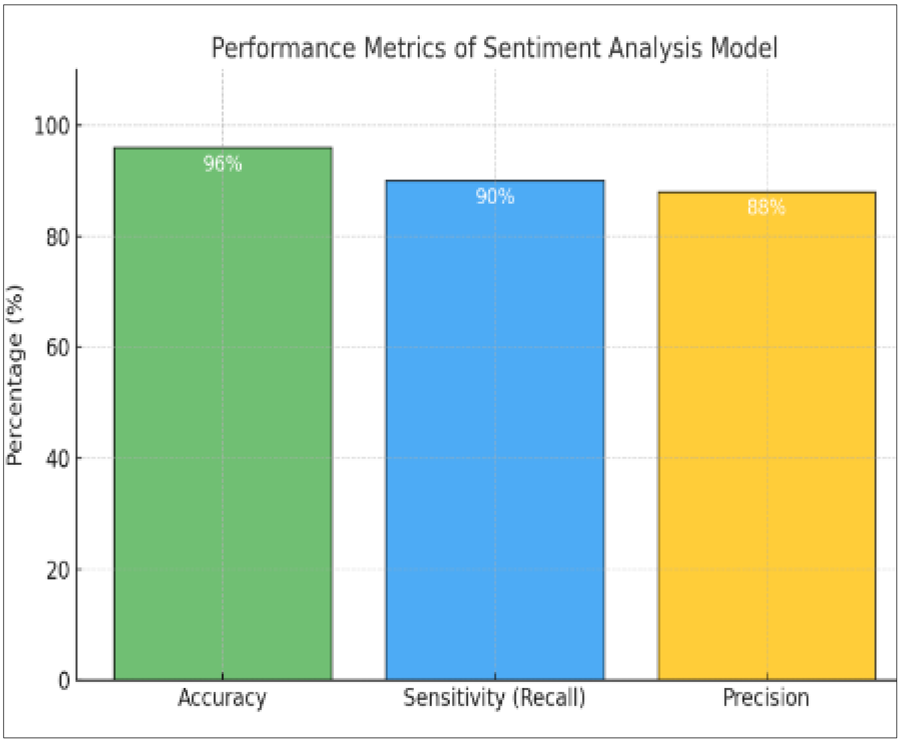


Figure 8 Plot showing the performance metrics

The confusion matrix in Figure 8 highlights the performance of the Sentiment Analysis Model across three categories: Abusive, Neutral, and Harmless, expressed in percentages. The model achieved high true positive rates, correctly classifying 96% of abusive, 98% of neutral, and 94% of harmless posts. However, some misclassifications occurred, with 2% of abusive posts misidentified as neutral or harmless and 3.5% of harmless posts mislabeled as Abusive. These errors underscore the need for further refinement to improve sensitivity to ambiguous language and overlapping contexts. Overall, the matrix confirms the model's strong classification capabilities and potential for enhancing sentiment analysis in social media content moderation.

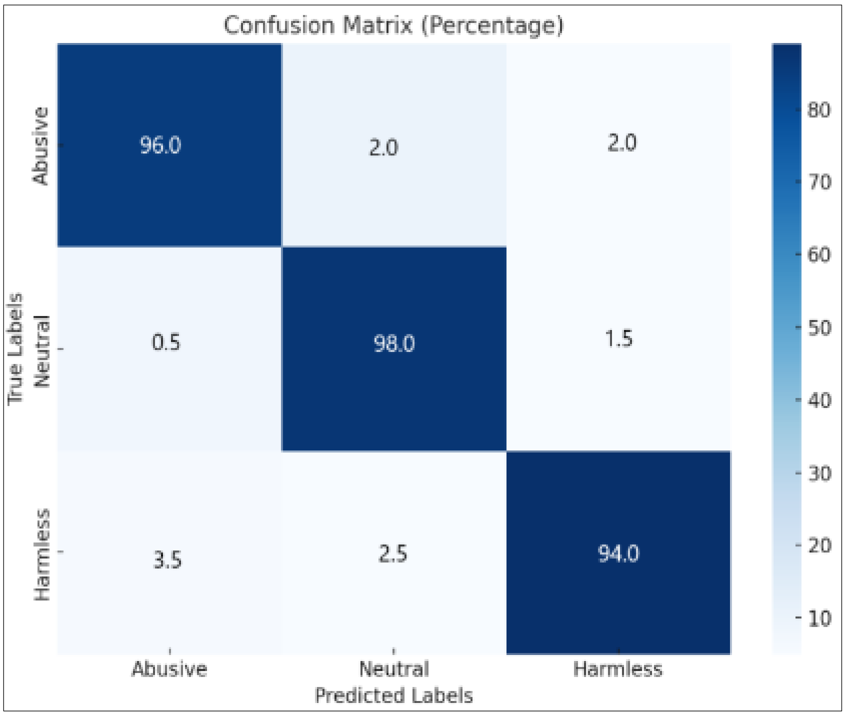


Figure 9 Confusion Matrix

5. Conclusion

This study successfully demonstrates the potential of AI-powered sentiment analysis, specifically the Sentiment Analysis Model (ChatGPT Integration), in classifying social media posts into Abusive, Neutral, and Harmless categories. By leveraging advanced preprocessing techniques, contextual understanding, and innovative integration of ChatGPT, the model achieved remarkable performance metrics, including a high accuracy of 96%. These results underscore the model's robustness and reliability in detecting harmful content while maintaining fairness and precision.

The findings highlight the value of AI in creating safer online environments by effectively moderating and filtering content. However, there remains scope for improvement, particularly in enhancing sensitivity and precision to address nuanced and ambiguous language more effectively. Future work could focus on expanding the dataset, incorporating more diverse linguistic and cultural contexts, and refining model architectures to further boost performance.

Ultimately, this research contributes a scalable and efficient solution to address online abuse and cyberbullying, paving the way for better content management systems and a healthier digital ecosystem.

Future Work

Sentiment analysis for detecting abusive content is a field undergoing continual evolution, with ongoing research and potential for improvement. Future research in this domain could explore several areas:

- **Integration of multiple modalities**: Investigating approaches that incorporate text, images, videos, and audio could offer a more comprehensive understanding of abusive content. By combining various data types, richer context may be obtained, enhancing the accuracy of sentiment analysis.
- **Cross-lingual sentiment analysis:** Extending research to address the challenges of analyzing sentiment across

multiple languages is essential. This entails developing models capable of effectively analyzing sentiment in diverse linguistic contexts, considering linguistic variations and cultural nuances.

- **Recognition of sarcasm and irony**: Developing models that can accurately detect and interpret sarcasm and irony is crucial. These linguistic constructs often pose challenges in sentiment analysis, particularly in the context of abusive content, and addressing them could improve the accuracy of detection.
- **Dynamic thresholding techniques:** Exploring dynamic thresholding methods that can adapt to different content types, users, or contexts may enhance the precision and recall of sentiment analysis models. This adaptability could improve the identification of abusive content across various scenarios and platforms.

Compliance with ethical standards

Disclosure of conflict of interest

There are no conflicts of interest to declare.

Statement of informed consent

Informed consent was obtained from all participants involved in the study.

References

[1] Abou El-Seoud, S., Farag, N. and McKee, G., 2020. A Review on Non-Supervised Approaches for Cyberbullying Detection. Int. J. Eng. Pedagog., 10(4), pp.25-34.

[2] Al-Ajlan, M. A., & Ykhlef, M. [33] (2018, April). Optimized twitter cyberbullying detection based on deep learning. In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1-5). IEEE.

[3] Al-Qablan, T.A., Mohd Noor, M.H., Al-Betar, M.A. and Khader, A.T., 2023. A survey on sentiment analysis and its applications. Neural Computing and Applications, 35(29), pp.21567-21601.

[4] Al-Shabi, M.A., 2020. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. IJCSNS, 20(1), p.1.

[5] Ariyadasa, A. (2019). Harassment Beyod Borders; Can Victims Be Protected By Cyber Bullying In Sri Lanka.

[6] Aziz, A.A., Starkey, A. and Bannerman, M.C., 2017, September. Evaluating cross-domain sentiment analysis using supervised machine learning techniques. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 689-696). IEEE.

[7] Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. [34] (2019, March). Detection of cyberbullying using deep neural network. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 604-607). IEEE.

[8] Beigi, O.M. and Moattar, M.H., 2021. Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. Knowledge-Based Systems, 213, p.106423.

[9] Bin Abdur Rakib, T., & Soon, L. K. [30] (2018). Using the reddit corpus for cyberbully detection. In Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part I 10 (pp. 180-189). Springer International Publishing.

[10] Bozyiğit, A., Utku, S., & Nasiboğlu, E. [37] (2019, September). Cyberbullying detection by using artificial neural network models. In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 520-524). IEEE.

[11] Bozyiğit, A., Utku, S., & Nasibov, E. [2] (2021). Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, 115001.

[12] Chavan, V. S., & Shylaja, S. S. [22] (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2354-2358). IEEE.

[13] Chen, H., Mckeever, S., & Delany, S. J. [23] (2017, August). Presenting a labelled dataset for real-time detection of abusive user posts. In Proceedings of the international conference on web intelligence (pp. 884-890).

[14] Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019a, January). Xbully: Cyberbullying detection within a multi-modal context. In Proceedings of the twelfth acm international conference on web search and data mining (pp. 339-347).

[15] Cheng, L., Li, J., Silva, Y., Hall, D., & Liu, H. (2019, August). PI-bully: Personalized cyberbullying detection with peer influence. In The 28th International Joint Conference on Artificial Intelligence (IJCAI).

[16] Cook, S. (2020). Cyberbullying facts and statistics for 2020. https://www.comparitech. com/internet-providers/cyberbullying-statistics/ [last accessed 05 July 2024]

[17] D'Aniello, 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.

[18] Dubey, G., Rana, A. and Ranjan, J., 2016. A research study of sentiment analysis and various techniques of sentiment classification. International Journal of Data Analysis Techniques and Strategies, 8(2), pp.122-142.

[19] Ghiassi, M., Skinner, J. and Zimbra, D., 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16), pp.6266-6282.

[20] Giménez Fayos, M.T., 2021. Natural language processing using deep learning in social media (Doctoral dissertation, Universitat Politècnica de València).

[21] Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. Social Network Analysis and Mining, 12(1), 129.

[22] Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(5), 703-707.

[23] Judge, M., Kashima, Y., Steg, L. and Dietz, T., 2023. Environmental Decision-Making in Times of Polarization. Annual Review of Environment and Resources, 48.

[24] Kennedy, B., Ashokkumar, A., Boyd, R. L., & Dehghani, M. (2021). Text analysis for psychology: Methods, principles, and practices.

[25] Kim, H. and Jeong, Y.S., 2019. Sentiment classification using convolutional neural networks. Applied Sciences, 9(11), p.2347.

[26] Li, P., 2023, November. Predicting Emotions from Twitter Posts: A Comparative Study of Machine Learning Methods. In Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023) (Vol. 108, p. 122). Springer Nature.

[27] Litty, A., Jahin, Z., & Jesan, Z. (2024). Detecting and Preventing Cyberbullying on Social Media Platforms Using Deep Learning Techniques. EasyChair Prepr.

[28] Liu, C., Liu, Z., & Yuan, G. (2022). Longitudinal associations between cyberbullying victimization, mindfulness, depression, and anxiety: a mediation analysis. Journal of Aggression, Maltreatment & Trauma, 31(1), 121-132.

[29] Ma, Y., Peng, H. and Cambria, E., 2018, April. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[30] Mahmoud Shehabat, A., 2018. Beyond Twitter Revolutions: The Impact of Digital Media Logistics on Terror Networks of Communication in Iraq and Syria from 2014 to 2016.

[31] Malik, Z. U. A., & Ayaz, M. (2023). Analyzing the Use of Language in Cyber Threats, Propaganda, and Communication: A Case Study of Pakistan. Research Mosaic, 3(2), 08-16.

[32] Nazir, T. and Thabassum, L., 2021. Cyberbullying: Definition, types, effects, related factors and precautions to be taken during COVID-19 pandemic. The International Journal of Indian Psychology.

[33] Perera, A., & Fernando, P. (2024). Cyberbullying detection system on social media using supervised machine learning. Procedia Computer Science, 239, 506-516.

[34] Riadi, I. [16] (2017). Detection of cyberbullying on social media using data mining techniques. International Journal of Computer Science and Information Security (IJCSIS), 15(3).

[35] Rosa, H., Carvalho, J. P., Calado, P., Martins, B., Ribeiro, R., & Coheur, L. (2018, July). Using fuzzy fingerprints for cyberbullying detection in social networks. In 2018 IEEE international conference on fuzzy systems (FUZZ-IEEE) (pp. 1-7). IEEE.

[36] Singh, N.K., Tomar, D.S. and Sangaiah, A.K., 2020. Sentiment analysis: a review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing, 11, pp.97-117.

[37] Singhal, P., & Bansal, A. (2013). Improved textual cyberbullying detection using data mining. International journal of Information and Computation technology, 3(6), 569-575.

[38] Song, J., & Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. Computers in Human Behavior, 78, 273-282.

[39] Sugandhi, R., Pande, A., Agrawal, A., & Bhagat, H. [29] (2016). Automatic monitoring and prevention of cyberbullying. International Journal of Computer Applications, 8(8), 17-19.

[40] Truong, Q.T. and Lauw, H.W., 2017, October. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In Proceedings of the 25th ACM International Conference on Multimedia (pp. 1274-1282).

[41] Udefi, A. M., Aina, S., Lawal, A. R., & Oluwarantie, A. I. (2023). An Analysis of Bias in Facial Image Processing: A Review of Datasets. International Journal of Advanced Computer Science and Applications, 14(5).

[42] Van Bruwaene, D., Huang, Q., & Inkpen, D. [25] (2020). A multi-platform dataset for detecting cyberbullying in social media. Language Resources and Evaluation, 54(4), 851-874.

[43] Vanpech, P., Peerabenjakul, K., Suriwong, N., & Fugkeaw, S. (2024, February). Detecting Cyberbullying on Social Networks Using Language Learning Model. In 2024 16th International Conference on Knowledge and Smart Technology (KST) (pp. 161-166). IEEE.

[44] Vanpech, P., Peerabenjakul, K., Suriwong, N., & Fugkeaw, S. (2024, February). Detecting Cyberbullying on Social Networks Using Language Learning Model. In 2024 16th International Conference on Knowledge and Smart Technology (KST) (pp. 161-166). IEEE.

[45] Wang, C., Gao, T., Cheng, X., & Li, B. (2024). Social media use, cyber victimization, and adjustment during COVID-19 virtual learning: A short-term longitudinal study among Chinese middle school students. School Psychology, 39(2), 176.

[46] Wilson, T., Wiebe, J. and Hoffmann, P., 2005, October. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of human language technology conference and conference on empirical methods in natural language processing (pp. 347-354).

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

[48] Ye, X., Dai, H., Dong, L.A. and Wang, X., 2021. Multi-view ensemble learning method for microblog sentiment classification. Expert Systems with Applications, 166, p.113987.

[49] Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., ... & Dillon, E. (2016, December). Cyberbullying detection with a pronunciation based convolutional neural network. In 2016 15th IEEE international conference on machine learning and applications (ICMLA) (pp. 740-745). IEEE.

[50] Zhao, R., & Mao, K. [27] (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Transactions on Affective Computing, 8(3), 328-339.