Explainable Artificial Intelligence based Cyberbullying Classification System

Udeme E. Udo
Tetfund Centre of Excellence in Computational Intelligence Research
University of Uyo, Uyo, Nigeria

Edward N. Udo
Department of Computer Science
University of Uyo, Uyo, Nigeria

ABSTRACT

The increasing use of social media and digital communication tools has brought many advantages, yet it has also enabled the concerning issue of cyberbullying which involves the use of digital platforms to harass, threaten, dishonour or demean individuals. Techniques to reduce this problem to a certain extent have already been introduced in online systems. However, they possess limitations of usage and lack of explainability. This research presents the development and evaluation of an Explainable Artificial Intelligence (XAI) based cyberbullying classification system intended to detect and interpret harmful content on social media platforms. The system leverages textual and symbolic features from user-generated contents, extracted from social media platform, X (Twitter), to classify posts as bullying or non-bullying. The methodology adopted in this work include data preprocessing (tokenization, stemming, lemmatization, and TF-IDF vectorization), followed by model training using Logistic Regression (LR), Decision Tree (DT), and Multinomial Naive Bayes (MNB) classifiers. To address the class imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was employed, which resulted in improved model fairness and performance. Among the models that were tested, Logistic Regression achieved the highest generalization accuracy of 93.12%. In other to enhance transparency and trust, SHAP (SHapley Additive exPlanations) was integrated, which offered interpretable insights into model predictions and highlights key linguistic features influencing classification results. The system was deployed via a web interface which enables real-time content moderation. While the results demonstrate high accuracy and interpretability, the study also reflects ethical considerations such as contextual misclassification, societal bias, and data privacy.

General Terms

Machine Learning, Classification System

Keywords

Explainable A.I, Cyberbullying, Data Preprocessing, Classifiers, Interpretability, Social Media Platform, SMOTE, SHAP.

1. INTRODUCTION

Artificial intelligence (AI), according to [1], is the ability for computer systems to perform tasks associated with human intellect and behaviour. It leverages data-driven methods inspired by mathematics, cognitive science, physics, to mimic human reasoning, senses, and ways of making decision [2]. This system not only meets but can often exceed the expected requirements set for those tasks. Machine learning (ML) represents a distinct area within artificial intelligence dedicated to creating algorithms and models that empower computers to learn from data and progressively enhance their performance without needing explicit instructions for every individual task. These algorithms include a broad range of computational techniques that allow computers to build predictive models leading to actionable insights [3]. As a core component of AI, machine learning equips systems with the ability to recognize patterns, make informed decisions, and adjust to evolving information, ultimately boosting their capacity to tackle complex problems with flexibility and intelligence. In machine learning, classification systems refer to algorithms or models developed to sort input data into specific, predetermined categories by recognizing patterns learned from training datasets. These systems evaluate the characteristics of the data and allocate labels corresponding to different groups, allowing machines to perform tasks like object recognition, spam filtering, or medical diagnosis. Serving as a key function within machine learning, classification enables AI to effectively process and structure information, supporting precise predictions and well-informed decisions across diverse fields.

In the digital age, the proliferation of social media and online communication platforms has enable instant, ubiquitous connectivity and other numerous benefits, but it has also given rise to negative behaviors and troubling phenomenon such as cyberbullying, racism, sexism and trolling [4]. Cyberbullying, which involves the use of digital platforms and electronic communication technologies (including social media, text messaging and email) to harass, threaten, dishonour or demean individuals, has become a significant concern, particularly among young people [5]. Examples include spreading rumors or outright lies, posting humiliating pictures or videos, sending offensive or derogatory comments, or using fabricated profiles to harass someone [6].

Cyberbullying, as a contemporary phenomenon, impacts not only the victims but also the perpetrators and bystanders. Victims of cyberbullying often experience heightened levels of depression, anxiety, suicidal thoughts and attempts, decreased academic and work performance, and deteriorating physical and mental health [6, 7]. This problem is found in all levels of society, including tertiary education institutions, public and private sectors, thereby affecting individuals regardless of gender, age, or social background [8].

The anonymity and wide reach of the internet exacerbate the impact of such harmful behaviour, leading to severe psychological, emotional, and sometimes even physical consequences for the victims. Bullying is characterized by three key elements: an aggressive intent, repeated actions, and an imbalance of power. It causes physical, mental, or emotional harm to individuals. The negative impact of cyberbullying is more intense than traditional bullying due to the wider online audience and the rapid spread of information on the internet. Perpetrators who engage in and derive satisfaction from cyberbullying may also suffer from mental health issues.

Additionally, bystanders to cyberbullying can develop mental stress and fear as a result of witnessing these events [9]. Cyberbullying can be categorized into many different methods, ranging based on the type and method of harassment or the abuse. These include direct cyberbullying, cyberstalking, flaming, exclusion, impersonation, outing and trickery, cyberbullying by proxy, and catfishing [10].

Traditional methods of combating cyberbullying, such as manual monitoring and reporting, are often insufficient due to the vast amount of content generated online every day. This is where the power of machine learning comes into play. By leveraging advanced algorithms and data analysis techniques, machine learning offers a promising solution to detect and mitigate cyberbullying in real-time. This approach not only enhances the speed and accuracy of identifying harmful content but also provides scalable solutions that can be integrated into various online platforms.

Internet service providers and administrators can create more accurate classifications systems by monitoring the signs of cyberbullying in advance [11].

Explainable AI (XAI) technique gives a transparent and interpretable process of decision making, thereby allowing people to fully understand the working mechanisms and the ways the models reason [12]. Additionally, large datasets can be analyzed using deep learning techniques. The use of machine learning for cyberbullying detection involves several key components. First, natural language processing (NLP) is employed to analyze the text of social media posts, comments, and messages. Natural Language Processing (NLP) and Artificial Intelligence (AI) have been developed as revolutionary technologies in the area of data visualization, which addresses the limitations of traditional methods [13]. NLP techniques enable the machine learning models to understand and interpret human language, identifying patterns and cues associated with abusive behaviour. It analyzes language patterns, syntax, semantics, and context in order to extract meaning [14]. Sentiment analysis, a subset of NLP, plays a crucial role in detecting the emotional tone of the text, helping to distinguish between benign and harmful content. The uses of sentiment analysis include areas like generating market insights from online contents generated by users, predicting virality from linguistic features of articles and newspaper, or identifying top performing individuals by analyzing the language style used in emails [15].

Creating robust datasets is another critical aspect of developing effective cyberbullying detection systems. Dataset is an important component of research and it is used as a core component for training and evaluating models in machine learning in order to detect cyberbullying [16]. These datasets must be representative of the diverse ways in which cyberbullying can manifest, encompassing different languages, cultural contexts, and platforms. Annotating these datasets accurately is also a challenge, as it requires distinguishing between genuine instances of bullying and non-bullying content that might be provocative or controversial [17].

In an effort to create automated techniques for the detection and prevention of cyberbullying, [18] presented an Explainable Multimodal Deep Learning Model for Cyberbullying Detection which included four steps: collection of datasets from different resources, which included images and their captions with binary classes (bullying and non-bullying); application of two techniques, XAI: CNN+GradCam to analyze input images and produce visual explanations, and LSTM+LRP to analyze and interpret input text; employment of two techniques of data fusion (early and late); evaluation of the performance of the EMDL-CBD model based on a set of accuracy metrics. The limitation of their work was the absence of options for multilingual, cross-linguistic, and mix language support.

[19] provided a robust framework for multi-class mental health screening by introducing a unified multiclass classification framework for detecting ten distinct mental health and cyberbullying categories from social media data. They conducted a comprehensive evaluation comparing traditional lexical models, hybrid approaches, and several end-to-end fine-tuned transformers, including MentalBERT and a hybrid SHAPLLM explainability framework. However, their Kaggle datasets narrowed broad claims of generalizability, and was limited to English-language text, excluding the rich signals available in multilinguals and multimodal.

Feature engineering, the process of selecting and transforming the most relevant variables from raw data, is essential in enhancing the performance of machine learning models. This is important to capture patterns and properties of relevant data, which enables the model to better understand the relationships between words [20]. Features such as word frequency, user behaviour patterns, and metadata (e.g., time of post, user interactions) can provide valuable insights into detecting cyberbullying.

Continuous model training and updating are crucial for maintaining the efficacy of these systems. It is the basic step where the real learning process occurs, typically by reducing errors through iterative optimization [21]. As online language and behaviours evolve, machine learning models must be regularly updated with new data to adapt to emerging trends and patterns in cyberbullying. This iterative process ensures that the detection systems remain accurate and relevant over time. Explainable AI (XAI) should also be introduced to enhance transparency and understanding for users. Explainability in this context refers to making explicit the details and reasons for a model outcome, and to make its functioning clearer to understand. It seeks to clarify the internal workings of a machine learning model, while aiming to offer clear explanations regarding the methods, procedures and outputs of the model for the users and other stakeholders [22]. Ethical considerations are also paramount when developing and deploying machine learning models for cyberbullying detection. This must be carefully considered to ensure responsible and equitable deployment of the model [23]. Ensuring user privacy, avoiding bias in the models, and providing transparency in how decisions are made are all essential factors. Developers must strike a balance between effective detection and protecting the rights and freedoms of online users [24].

This work therefore employs LR, DT and MNB classifiers to build a cyberbullying classification system using a dataset that contains both textual and symbolic information to aid in cyberbullying detection. SHAP was integrated into the model to offer interpretable insights into the model prediction. The classification system will be deployed via a web interface for real-time moderation.

2. RELATED WORKS

Many works have been done in recent times to identify cyberbullying. This section gives an overview of some previous studies for detecting cyberbullying on social media platforms, focusing on different machine learning techniques.

[25] declared that various research studies conducted by different labs and centers have employed a range of machine and deep learning techniques to identify word phrases and slang. Methods such as k-Nearest Neighbor (kNN), Linear

Regression (LR), Random Forests (RF), Logistic Regression (LogR), Boosting (Bos), Bagging (Bgg), AdaBoost (ADB), Multiple Regression (MR), and Maximum Entropy (MaxE) are being utilized for the detection of cyberbullying on social media. Neural networks, particularly deep learning models, have shown great promise in cyberbullying detection. These models can learn complex patterns and relationships within large datasets, making them well-suited for identifying subtle forms of bullying that might be missed by simpler algorithms. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are among the architectures that have been successfully applied to this problem. CNNs are effective in analyzing the spatial hierarchies in text, while RNNs, especially long short-term memory (LSTM) networks, are adept at capturing temporal dependencies and context in sequences of words. The drawback in the study was insufficient data to train the models for better accuracy, and lack of explainability using any XAI approach.

[26] developed an intelligent, machine learning-driven framework for diabetes prediction, integrating explainable artificial intelligence (XAI) to enhance model transparency and interpretability. Their study utilized a combination of a publicly available dataset, the Pima Indian Diabetes dataset, and a private dataset collected from Rownak Textile Mills Ltd. (RTML) in Dhaka, Bangladesh. A variety of machine learning and ensemble models were evaluated in the development of the diabetes prediction system. The classifiers included Decision Tree, k-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, AdaBoost, XGBoost, Voting Classifier, and Bagging. Among these, XGBoost combined with ADASYN yielded the highest performance metrics and was selected as the final model for deployment. To enable practical, real-time use of the model, the authors developed a web and mobile application interface. The web application frontend was implemented using HTML and CSS, while the backend machine learning model was deployed using Python through the Spyder IDE within the Anaconda environment. For the mobile application, Android Studio was used to design the user interface, and Java was employed as the programming language. These tools provided insight into how individual features influenced model decisions. For a given patient case, the model accurately predicted the diabetes outcome with 80% confidence, primarily due to the patient's glucose level exceeding 140.25 and a history of more than six pregnancies. These interpretable outputs underscore the model’s alignment with established medical knowledge and contribute to its applicability in real-world clinical settings. Their study required additional private data with a larger cohort of patients to get better results, and could not combine machine learning models with fuzzy logic techniques and applying optimization approaches for better performance.

[27] presented a unique framework which integrates sentiment analysis together with machine learning algorithms in order to enhance the detection of cyberbullying on social media. They used publicly available DQE-augmented dataset which is publicly available comprising 47,692 tweets. Text preprocessing was then performed which included lemmatization, remove stop words, and remove special characters and URLs. SMOTE-oversampling was performed on the dataset. Nine ML classifiers were evaluated in their study including Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Gradient Boosting (GB), and Extra Tree (ET) and AdaBoost. The Extra Tree classifier gave the best results with an accuracy of 95.38% and F1 score of 0.95.

[28] introduced a sophisticated model that integrates Recurrent Neural Networks (RNNs) with association rule mining to enhance the detection and interpretation of cyberbullying in textual data. In their model, Sanchez et al. leveraged the power of RNNs to identify and learn from these temporal patterns within conversations. The RNN's ability to process this sequential information allows it to capture such nuances effectively. However, like many deep learning models, RNNs can be difficult to interpret, as they do not naturally provide explanations for their predictions. To address this issue, Sanchez et al. integrated association rule mining into their model. Association rule mining is a technique commonly used in data mining to discover interesting relationships or patterns among variables in large datasets. By applying this technique, the researchers were able to extract human-readable rules from the sequences analyzed by the RNNs. Overall, the combination of RNNs and association rule mining in Sanchez et al.'s model represented a significant advancement in the field of cyberbullying detection. It not only improves the accuracy of detecting complex, context-dependent bullying behaviours but also ensures that the reasoning behind each detection is clear and interpretable. This dual focus on accuracy and explainability makes their approach particularly valuable for real-world applications, where understanding the "why" behind a decision is just as important as the decision itself. However, the study was resource intensive, and required much time to train the model.

[29] advanced the field of cyberbullying detection by proposing a sophisticated, multi-tiered framework capable of classifying and identifying harmful online behavior across a wide spectrum of digital platforms. This architecture integrated cutting-edge machine learning (ML) and natural language processing (NLP) techniques to address the growing complexity of cyberbullying behavior, ensuring that the system remains adaptive to evolving linguistic patterns and the shifting dynamics of online discourse. To support the development of this intelligent system, the researchers curated a comprehensive corpus comprising over 47,000 human-annotated tweets. The authors implemented a robust preprocessing pipeline to prepare the raw tweet data for downstream machine learning tasks. This pipeline encompassed key NLP operations such as tokenization, stopword elimination, and lemmatization or stemming, all aimed at reducing textual noise while preserving semantic integrity. A critical step in the feature engineering process was the use of the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. For the classification task, the study employed BERT (Bidirectional Encoder Representations from Transformers), a deep learning model renowned for its contextual sensitivity and powerful language representation capabilities. The final model yielded a strong performance, achieving an F1 Score of 86% and a matching recall value of 86% on the testing set, signifying a high level of precision and robustness in distinguishing harmful content from benign discourse. Overall, the contribution of Shah et al. represents a meaningful advancement in automated cyberbullying detection systems. By combining well-structured datasets, rigorous preprocessing, and state-of-the-art NLP models, the study offers a scalable and adaptive approach for safeguarding users in increasingly complex digital environments. The work was limited by the inability to consider sentence relationships, lacked the ability to convert emoji to corresponding words or phrases, examine phrases, build a vocabulary of cyberbullying terms with more grammatical details, and used longer words as intensifiers.

[9] proposed a system for automatic cyberbullying detection

and prevention using supervised machine learning. The system considered key characteristics of cyberbullying, such as the intention to harm, repeated behaviour, and the use of abusive language. Support vector machines and logistic regression were employed to identify cyberbullying and related themes/categories such as race, physical, sexuality, and politics. The method offered a novel theory for the detection of cyberbullying: texting has evolved over time due to changes in context usage, and language. In the dataset that includes tweets, Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression (LR) models were tested along with different Natural Language Processing methods. The accuracy of the system was improved by sentiment analysis, N-gram analysis, and other non-traditional feature extraction methods like Term Frequency-Inverse Document Frequency (TF-IDF) and profanity detection. A limitation of their system was that it could only accurately predict sentences that are correctly spelled words.

[30] introduced a novel framework that combines deep learning and interpretability techniques to enhance the detection and explanation of cyberbullying behaviors in textual data. To achieve this, the authors integrated visual analytics through the LIME (Local Interpretable Model-Agnostic Explanations) algorithm, enabling human-understandable insights into how the model arrived at specific predictions. For empirical evaluation, the study employed the "Cyberbullying Classification" dataset sourced from Kaggle, which contained text samples labeled across various cyberbullying categories including Age, Ethnicity, Gender, Religion, Other types of cyberbullying, and Non-cyberbullying. After extensive preprocessing, including cleaning, sampling, and balancing, the refined dataset was used to train a BERT-based neural network for multi-label classification. The model was deployed and tested within the Google Colab environment, and its predictive capacity was measured using four key metrics: Accuracy (0.956478), Precision (0.963677), Recall (0.956478), and F1 Score (0.960019). These results demonstrated superior performance when benchmarked against existing studies in cyberbullying detection, underscoring the reliability of the model in distinguishing between nuanced forms of abusive language. To enhance interpretability, Krak et al., applied LIME to the trained BERT model, generating three distinct visualization formats that revealed the model's reasoning process: (1) color-coded overlays highlighting influential words; (2) diagrams showing the local importance of individual tokens; and (3) aggregate charts depicting global feature significance across detected categories. This multi-perspective visual analysis added a vital layer of explainability, reinforcing trust in the neural network's outputs and aiding practitioners in understanding how different types of cyberbullying are recognized in textual inputs. The integration of high-performing machine learning with visual interpretability makes this study a meaningful contribution to the development of ethical and transparent AI systems for digital safety. Their model could not work with texts in other languages, and did not conduct experiments with users to assess the impact of visual analytics on human decision-making.

This work, compared to many previous works on cyberbullying detection, integrates explainable AI into the classification process. While earlier systems mostly focused on improving accuracy or classification efficiency, they typically operated as "black boxes", thereby offering little or no understanding into how decisions were made. This work adds SHAP into the framework, allowing users, moderators, and stakeholders to clearly understand which specific linguistic features influenced the classification of a post. The transparency so introduced does not only build trust but also ensures accountability, making the system more ethically aligned with the needs of real-world deployment.

Also, most previous works did not address the imbalance nature of the datasets used, which led to bias predictions and poor generalization. This work uses SMOTE to handle class imbalance in the dataset. The deployment of the model using a web interface for real-time moderation further sets this work apart from other studies on XAI cyberbullying research, as it moves beyond theoretical research into real-life application.

3. METHODOLOGY

The various phases and implementation procedures of the different machine learning algorithms adopted to design the explainable AI based cyberbullying classification system is shown in Figure 1.

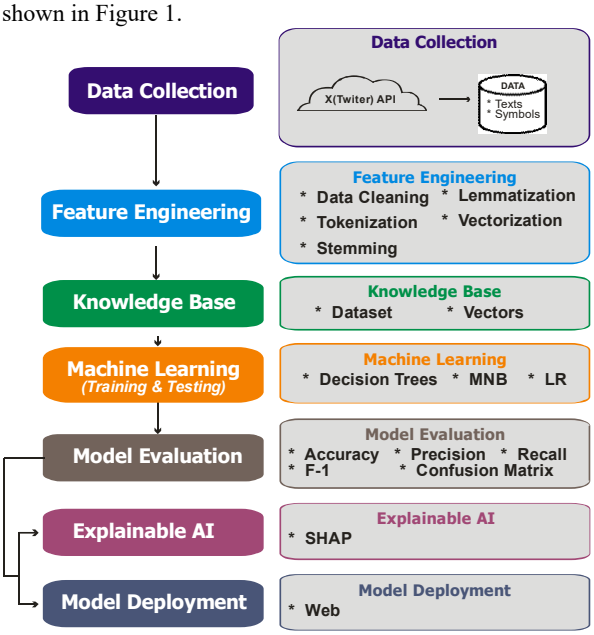


Fig 1: System Flow Diagram of XAI Cyberbullying Classification System

3.1 Data Collection

The datasets used for the study were extracted from the X(Twitter) social media platform for analysis and model evaluation. The data contained text and symbols labelled as bully or not-bully and different types of cyberbullying messages like insults, racism, hate speech, aggression and toxicity as depicted in Figure 2.

"@DGotBricks: What happen to them vixen ent bitches" they got ran and threw to the side like a foothill bit	bully
"@DaeDaviDavie: @white_thunduh im the bitch okay nudes pat 😪😏🐸" wow	bully
"@DevilGrimz: @VigxRArts you're fucking gay, blacklisted hoe" Holding out for #TehGodClan anyway http:/	bully
"@DiamondLoudKush: The fuck be wrong with these bitches?" Nobody knows	bully
"@Dietrich1892: Yall shut up:p" make me bitch	bully
"@Dionalrish: I hate a "I'm pregnant" type of bitch."	bully
"@DoYou_Q: Got bitches in the DM but I don't ever read'em" which is y your top 3	bully
"@DomWorldPeace: Baseball season for the win. #Yankees" This is where the love started	not-bully
"@Dommoneek: Little stupid as bitch I don't fuck with yooooooouuuu.."	bully
"@DreadheadAri: she really asked me that dead ass serious tho, all i could say was "bitch wheet" LOL	bully
"@Dunderball: I'm an early bird and I'm a night owl, so I'm wise and have worms."	not-bully
"@EdgarPixar: Overdosing on heavy drugs doesn't sound bad tonight." I do that pussy shit every day.	not-bully
"@El_Grillo1: Pit Bulls Photographed As Lovely Fairy Tale Creatures http://t.co/Q0Sm89oOLh”	not-bully
"@Ferocious_Ghost: @1stName_Bravo Aw." ...fag, don't tweet "aw" to me lol	bully
"@FloKid88: As long as the Lakers trash from now on, I could careless. And that's real.". CC: @BENBALLER h	not-bully
"@Frosstyy_: @h0rheyd I didn't say anything tho" kiss me then faggot	bully
"@FunnyPicsDepot: this the "I play soccer, cheat on girls, and wear khaki coloured cargos" haircut http://t.c	not-bully
"@G27Status: I could go for a fat ass bitch on my lap" same	bully
"@GEDMelle: 17 missed calls!!!! 😠😡"Das yo P.O bitch twitter finna be screamn #FreeMor	bully
"@GTM_Al: Ya side bitch gotta know it's rules to this shit..anybody ask you my cousin from jersey thinkin b	bully
"@GagaTom1: “@MaleFoot: 3 | Amo los pies http://t.co/4QE1hDkK8i”" fuck yeah	bully
"@GirlThatsVonte: These hoes be thinking Meat won't slap they ass 😴👋" ainna bruh	bully

Fig 2: Extracted Tweets in CSV format

3.2 Data Preprocessing and Feature Engineering

Data preprocessing was performed to ensure that the data is in the best possible form for model training, leading to improved model accuracy, reliability, and efficiency. The preprocessing involved data cleaning, data reduction and quality enhancement and data transformation. Feature engineering crafts and refines input variables from raw data to enhance the effectiveness of machine learning models. By applying domain expertise and analytical methods, it helps reveal useful patterns and reshape the data into a form that supports more accurate and efficient model learning. Figures 3(a) and (b) presents the view of the dataset in a bar and pie charts respectively to reveal its imbalance nature.

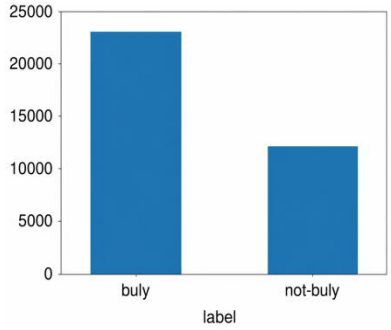


Fig 3(a): Dataset view in Bar Chart

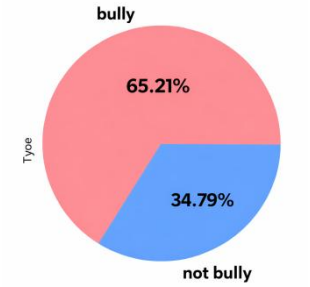


Fig 3(b): Dataset view in Pie Chart

3.2.1 Data Cleaning

Data preprocessing techniques were exploited to eliminate incomplete or corrupt data that could lead to system failures and negatively impact the accuracy of output predictions. Various methods were utilized to detect and remove noisy data, outliers, and other anomalies that could distort the results. By systematically addressing these issues, the overall quality and reliability of the data was enhanced, ensuring that the model's predictions were more accurate and precise.

3.2.2 Tokenization

Tokenization is the process of dividing text into smaller, meaningful units known as tokens, which can include words, sub-words, or characters. This step is essential in natural language processing because it transforms raw text into a structured format that machines can interpret and analyze. By breaking sentences or paragraphs into manageable pieces, tokenization enables downstream tasks such as text classification, sentiment analysis, and machine translation to function effectively. The choice of token type (whether words, characters, or sub-word units) depends on the specific application and model requirements. Figure 4 depicts the untokonized data.

	text	label
0	!!! RT @mayasolovely: As a woman you shouldn't...	not-bully
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	bully
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	bully
3	!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	bully
4	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	bully
...	...	...
35868	RT @Transition: @freebsdgirl just so I'm clear...	not-bully
35869	RT @Leonard_Delaney: @freebsdgirl I know! Holy...	not-bully
35870	FLOSS Weekly, open source projects, and paying...	not-bully
35871	RT @Kasparov63: My WSJ article on Boris Nemtso...	not-bully
35872	"@panelrific: Let's go 🤖🤖🤖🤖🤖🤖🤩"	not-bully

35873 rows × 3 columns

Fig 4: Untokenized Data from X (Twitter)

3.2.3 Stemming

Stemming is a text preprocessing technique in NLP that involves reducing words to their root or base form by removing suffixes and other word endings. The purpose is to group together different variations of a word so they can be treated as a single term during analysis. For example, the words "running," "runner," and "runs" might all be reduced to the stem "run." Unlike lemmatization, stemming does not necessarily produce real words, as it relies on crude rule-based methods that simply trim word endings without considering grammar or context. Despite its simplicity, stemming helps improve search accuracy, reduces vocabulary size, and increases computational efficiency in tasks such as information retrieval and document classification.

3.2.4 Lemmatization

Lemmatization transforms words into their dictionary or base form, known as a lemma, while taking into account the context and grammatical structure of the word. Unlike stemming, which simply chops off word endings, lemmatization uses linguistic rules and vocabulary to ensure that the resulting word is meaningful. For instance, the words "better" and "am" would be converted to their base forms "good" and "be," respectively. This process helps standardize words for more accurate text analysis, particularly in applications like search engines, sentiment analysis, and language modeling, where understanding the correct meaning of a word is essential.

3.2.5 Vectorization

Vectorization is the process of converting the cyberbullying text and symbols into numerical vectors so that machine learning models can interpret and analyze them. This involves representing the words, phrases, and the entire documents as numerical arrays using Term Frequency-Inverse Document Frequency (TF-IDF), enabling the models to detect patterns, make predictions, and understand the language structure.

TF-IDF is a method used in NLP to measure the significance of a word in a particular document compared to its occurrence across a collection of documents. It assigns higher scores to words that are common in one document but uncommon in the broader corpus, allowing models to focus on terms that carry more unique or distinguishing information.

In TF-IDF:

- Term Frequency (TF): Indicates how often a word appears in a document.
- Inverse Document Frequency (IDF): Shows how rear or unique a word is across all documents.

The TF-IDF score for each word is calculated by multiplying these two values as shown in equation (1):

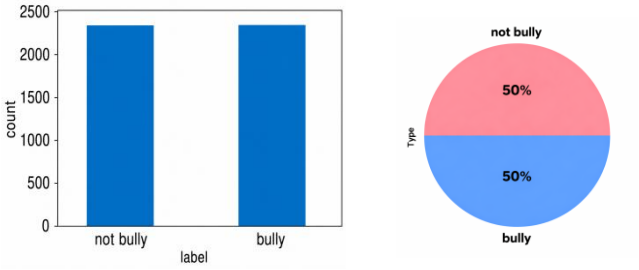
$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

Where:

t = term (word)
d = document
TF(t, d) = (Number of times term t appears in d) ÷ (Total terms in d)
IDF(t) = log(Total number of documents ÷ Number of documents containing t)

3.3 Data Balancing

The dataset before balancing stood at 23,574 (65.72%) of the data labelled as bully, while 12,299 (34.28%) were labelled as not-bully. Given this inequality, data balancing was crucial since imbalanced datasets can lead to biased model that perform well on dominant class but poorly on underrepresented class. SMOTE (Synthetic Minority Oversampling Technique was employed to balance the data because it enhances machine learning performance on imbalanced datasets by generating synthetic samples for the minority class, thereby reducing overfitting compared to simple duplication, preserving the feature space structure, and also improving metrics like F1-score and recall. After the data balancing, bully and not-bully classes stood at 23574 as shown by the bar and pie charts in Figure 5.



a. Bar **b. Pie-chart**

Fig 5: Balanced Dataset after SMOTE

3.4 Knowledgebase

The dataset, along with the identified target outputs and the transformed, vectorized representations of the text, collectively formed the foundational knowledgebase required for training the machine learning classification model. This structured data enabled the model to learn meaningful relationships between input features and expected outcomes, ultimately supporting accurate predictions and consistent performance across different platforms.

3.5 Machine Learning Algorithms

Machine learning algorithms used for the XAI based cyberbullying classification system were Multinomial Naïve Bayes (MNB), Decision Trees (DT), and Logistic Regression (LR). After the preprocessing and feature extraction phases, the vectorized text and symbols were inputted to these classifiers in order to evaluate their performance. Model performance metrics, including Accuracy, Precision, Recall, F1-score, as well as Confusion Matrix were computed.

3.5.1 Choice of Machine Learning Algorithms

The classifiers MNB, DT and LR were selected for the natural language processing classification task for the following reasons:

a. Multinomial Naïve Bayes (MNB)

Multinomial Naive Bayes is beneficial for this work because it is fast, interpretable, and effective for text-based data with many features. It offers a good balance of simplicity and performance, making it a go-to choice in various natural language processing tasks. It calculates the class probabilities for a given text using Bayes' rule.

Bayesian networks assume the features or words in a message are conditionally independent in the given class. It calculates the probability of a given message which belongs to a class based on word frequencies. This is represented in equation (2):

$$P(C_k|x) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x)} \quad (2)$$

Where:

- $P(C_k|x)$: The probability of the class C_k given the input x (e.g., message text).
- $P(C_k)$: The prior probability of class C_k (e.g., the probability of a message being cyberbullying or not).
- $P(x_i|C_k)$: The conditional probability of observing feature x_i (word i) given the class C_k.
- $P(x)$: The total probability of the input x, which is used for normalization.

The key part is the product term:

$$\prod_{i=1}^{n} P(x_1|C_k) \quad (3)$$

where the features are treated as conditionally independent given the class, which simplifies the calculation.

b. Decision trees

The decision tree algorithm is selected for NLP applications because of its capacity to process a mix of categorical and numerical data while offering a transparent and interpretable structure. Its hierarchical decision-making process makes it easy to trace how linguistic features contribute to specific outcomes, which is especially valuable in scenarios that demand explainability. Additionally, decision trees perform reliably with limited data, making them a practical choice for systems where clarity and simplicity are prioritized alongside predictive accuracy.

The Decision trees create a model by splitting the data recursively based on feature values, which aims to maximize the homogeneity of the target variable within each split. The equation for the splitting criterion here is based on the Entropy as shown in equation (4):

$$Entropy(D) = -\sum_{k=1}^{k} p_k \log_2(p_k) \quad (4)$$

Where p_k is the probability of class k in dataset D.

c. Logistic Regression

Logistic Regression is chosen for the NLP explainable classification task because it is a straightforward, efficient, and interpretable algorithm that works particularly well for binary classification tasks. Its simplicity, flexibility (with regularization and multiclass support), and probabilistic output makes it a preferred choice for this application. In this algorithm, prediction of the class of a numerical variable is based on its relationship with the label. The algorithm typically computes the class membership probability.

In the context of cyberbully detection, Logistic Regression predicts the probability of a message as bullying or not, using equation (5):

$$P(C = 1|x) = \frac{1}{1+e^{z}} \quad (5)$$

Where $\mathbf{z} = \boldsymbol{w}^{T}\boldsymbol{x} + \mathbf{b}$ -

- $P(C = 1|x)$: The probability that the message belongs to the class (e.g., bullying).

- ***x***: The feature vector (e.g., word counts or TF-IDF scores of words in the message).
- ***w***: The weight vector, which is learned during training.
- ***b***: The bias term.
- ***e***: Euler's number (approximately 2.718).

3.6 Model Evaluation

The performance of the the machine learning models on the explainable AI cyberbullying classification task is assessed.. This involved the use of various metrics such as accuracy, precision, recall, and F1-score to measure the models' correctness, generalization ability, and relevance in understanding and processing natural language.

Accuracy is the proportion of correct predictions, Precision defines how many predicted positives are actually positive, Recall (Sensitivity) showed how many actual positives are correctly predicted, and F1-Score is the harmonic mean of precision and recall. Confusion matrix evaluates the performance of a classification model by displaying the number of correct and incorrect predictions it makes across different classes. It breaks down the results into four categories - true positives, true negatives, false positives, and false negatives- allowing a clear and detailed view of how well the model distinguishes between classes. This structured overview helps identify not just overall accuracy, but also specific areas where the model may be misclassifying data.

3.7 Explainable AI

XAI was used in this study to make the model's decisions transparent, understandable, and trustworthy for users, moderators, and affected individuals. This helped in uncovering the words, phrases, or patterns that influenced the decision, thereby reducing the risk of misclassification, allowing for fairer and more accountable outcomes.

For the explanation, SHAP (SHapley Additive exPlanations) was employed. Shapley values allow the generation of global and local explanation of the ML models. It uses the concept of game theory to identify the contribution of each feature in the prediction. As a post hoc interpretation tool, SHAP introduces the interpretability without compromising the performance of the model. The scope of our interpretation is global as it explains how the entire model behaves across all inputs.

3.8 Model Deployment

The trained machine learning model was deployed within a Flask framework, allowing it to be used effectively in real-world scenarios. Flask, known for its simplicity and flexibility, serves as a suitable platform for building APIs that can host and deliver machine learning predictions seamlessly to end users or external systems.

4. RESULTS AND DISCUSSION

This section outlines the experimental outcomes of the Explainable Artificial Intelligence (XAI)-driven cyberbullying classification system.

4.1 Dataset Structure

The extracted dataset from the X (Twitter) API contains 35873 rows of tweets labelled as bully or not-bully, in 2 columns captioned 'text' and 'label' as indicated in Figure 6. The text column contains the raw tweet content as string objects, while the label column denotes the classification outcome. At compilation, the dataset consumed about 560.6KB of memory, and the tweets were classed as object data type.

	text	label
0	!!! RT @mayasolovely: As a woman you shouldn't...	not-bully
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	bully
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	bully
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	bully
4	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	bully
...	...	...
35868	RT @Transition: @freebsdgirl just so I'm clear...	not-bully
35869	RT @Leonard_Delaney: @freebsdgirl I know! Holy...	not-bully
35870	FLOSS Weekly, open source projects, and paying...	not-bully
35871	RT @Kasparov63: My WSJ article on Boris Nemtso...	not-bully
35872	"@panelrific: Let's go 🙂🙂🙂🙂🙂🙂😀"	not-bully

35873 rows × 2 columns

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35873 entries, 0 to 35872
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    35873 non-null  object
 1   label   35873 non-null  object
dtypes: object(2)
memory usage: 560.6+ KB
```

Fig 6: Compiled Dataset after Extraction

4.2 Text Tokenization

The texts in the tweets were divided into smaller, semantically meaningful units known as tokens in 1.0982 seconds using RegexpTokenizer (Regular expression-based Tokenizer) class from the NLTK (Natural Language Toolkit) library in Python. This step is important in natural language processing as it transforms unstructured text into analyzable components such as symbols, words, and phrases. Also the RegexpTokenizer employed here uses regular expressions to define precise rules for splitting text, thereby allowing for greater control over how tokens are extracted while ensuring that punctuation, hashtags, mentions, and other special characters common in social media text are handled correctly. Figure 7 shows the tokenized version of the dataset.

Fig 7: Tokenized Data with Regular expression-based Tokenizer

4.3 Word Stemming and Lemmatization

Words in the datasets were reduced to their root or base form using stemming and lemmatization. In stemming, heuristic rules are applied to truncate words to their stem often producing slightly crude forms, while lemmatization relies on linguistic knowledge bases to return the proper dictionary form or lemma of a word, taking into account its part of speech. For example words such as "Weekly, open source projects, and paying" were

stemmed to "week, open, sourc, project, and, pay", which was further text-processed to "week open sourc project and pay the bill". Figure 8 shows a sample of the dataset as stemmed and processed.

	text	label	text_tokenized	text_stemmed	text_processed
0	!!! RT @mayasolovely: As a woman you shouldn't...	not-bully	[RT, mayasolovely, As, a, woman, you, shouldn...	[rt, mayasolov, as, a, woman, you, shouldn, t...	rt mayasolov as a woman you shouldn t complain...
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	bully	[RT, mleew, boy, dats, cold, tyga, dwn, bad, f...	[rt, mleew, boy, dat, cold, tyga, dwn, bad, fo...	rt mleew boy dat cold tyga dwn bad for cuffin ...
2	!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	bully	[RT, UrKindOfBrand, Dawg, RT, sbaby, life, You...	[rt, urkindofbrand, dawg, rt, sbabi, life, you...	rt urkindofbrand dawg rt sbabi life you ever f...
3	!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	bully	[RT, C, G, Anderson, viva, based, she, look, l...	[rt, c, g, anderson, viva, base, she, look, li...	rt c g anderson viva base she look like a trann
4	!!!!!!!!!! RT @ShenikaRoberts: The shit you...	bully	[RT, ShenikaRoberts, The, shit, you, hear, abo...	[rt, shenikarobert, the, shit, you, hear, abou...	rt shenikarobert the shit you hear about me mi...
...	...	...	...	...	...
35868	RT @Transition: @freebsdgirl just so I'm clear...	not-bully	[RT, Transition, freebsdgirl, just, so, I, m, ...	[rt, transit, freebsdgirl, just, so, i, m, cle...	rt transit freebsdgirl just so i m clear you h...
35869	RT @Leonard_Delaney: @freebsdgirl I know! Holy...	not-bully	[RT, Leonard, Delaney, freebsdgirl, I, know, H...	[rt, leonard, delaney, freebsdgirl, i, know, h...	rt leonard delaney freebsdgirl i know holi cru...
35870	FLOSS Weekly, open source projects, and paying...	not-bully	[FLOSS, Weekly, open, source, projects, and, p...	[floss, week, open, sourc, project, and, pay, ...	floss week open sourc project and pay the bill...
35871	RT @Kasparov63: My WSJ article on Boris Nemtso...	not-bully	[RT, Kasparov, My, WSJ, article, on, Boris, Ne...	[rt, kasparov, my, wsj, articl, on, bori, nemt...	rt kasparov my wsj articl on bori nemtsov s mu...
35872	"@panelrific: Let's go 👌👌👌👌👌👌"	not-bully	[panelrific, Let, s, go]	[panelrif, let, s, go]	panelrif let s go

35873 rows × 5 columns

Fig 8: Processed Dataset after Stemming and Lemmatization

4.4 WordCLOUD

WordCloud, a feature in Natural Language Processing, was built to display the most frequent words in the cyberbully tweets, with the size of each word indicating how often the words appear in the tweets. Figures 9a and 9b show the wordClouds for bully and not-bully class in the dataset.

a. Bully class **b. Not-bully class**

Fig 9: NLP wordCloud for Bully and Not-bully Class

It is observed from Figure 9 that some words, such as 'bitch' and 'rt' seem to appear prominently in both bully and not-bully classes. This observation highlights the importance of further semantic analysis and the need for employment of explainability tools like SHAP to resolve such overlaps.

4.5 Count Vectorization

TF-IDF vectorization was employed to generate the numerical vectors version of the text for machine learning models' interpretation and analysis. Here, Term Frequency (TF) measures how often a word appears in a document relative to the total number of words in that document. In Inverse Document Frequency (IDF), terms that are unique or less frequent across the entire corpus are assigned higher weight, and common words that appear in many documents are down-weighted. The technique effectively captured term importance across the corpus, and generated the numerical vectors that enabled the models to distinguish between relevant and irrelevant features. These vectors were represented in an array form as indicated in Figure 10.

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

Fig 10: Vectorized Text with TF-IDF

4.6 Model Building

Three different models (Decision Tree, Multinomial Naive Bayes, and Logistic Regression) were built using the balanced dataset, which was split into 80% for training and 20% for testing. Figure 11 displays the classification report of the Accuracy, Precision, Recall, and F1-score for the different models.

```
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

       bully       0.97      0.89      0.93      4707
   not-bully       0.90      0.97      0.93      4723

    accuracy                           0.93      9430
   macro avg       0.93      0.93      0.93      9430
weighted avg       0.93      0.93      0.93      9430

Accuracy score on train dataset :  0.9476377326475423
Accuracy score on test dataset :  0.9311770943796395


Classification Report for Decision Tree:
              precision    recall  f1-score   support

       bully       0.94      0.92      0.93      4707
   not-bully       0.92      0.94      0.93      4723

    accuracy                           0.93      9430
   macro avg       0.93      0.93      0.93      9430
weighted avg       0.93      0.93      0.93      9430

Accuracy score on train dataset :  0.9981971472506496
Accuracy score on test dataset :  0.9299045599151644


Classification Report for Multinomial Naive Bayes:
              precision    recall  f1-score   support

       bully       0.92      0.90      0.91      4707
   not-bully       0.91      0.93      0.92      4723

    accuracy                           0.91      9430
   macro avg       0.91      0.91      0.91      9430
weighted avg       0.91      0.91      0.91      9430

Accuracy score on train dataset :  0.9463386181663928
Accuracy score on test dataset :  0.9146341463414634
```

Fig 11: Classification Report for the Machine Learning Models

From the classification results, Logistic Regression gave accuracy score of 94.76% on training and 93.12% on testing, with a macro Average of 93% and weighted average of 93%. Decision Tree yielded 99.82% accuracy on training and 92.99% on testing, while offering 93% macro average and 93% weighted average. For the Multinomial Naive Bayes model, 94.63% was gotten as the training accuracy and 91.46% as test accuracy, as well as 91% macro average and 91% weighted average.

All models demonstrated optimal performance, with Logistic Regression slightly outperforming others in terms of generalization. It is worthy to note that Logistic Regression exhibited robust generalization with balanced precision and recall, which makes it suitable for real-world deployment. The

Decision Tree showed a near-perfect training accuracy which indicates potential overfitting, though its test accuracy remained competitive. Multinomial Naive Bayes gave consistent performance but slightly varied in precision and recall, suggesting its limitations in handling complex feature interactions.

The confusion matrices for the binary class classification of the models are shown in Figure 12 which further confirmed the models' ability to distinguish between bully and not-bully tweets, with minimal false positives and negatives.

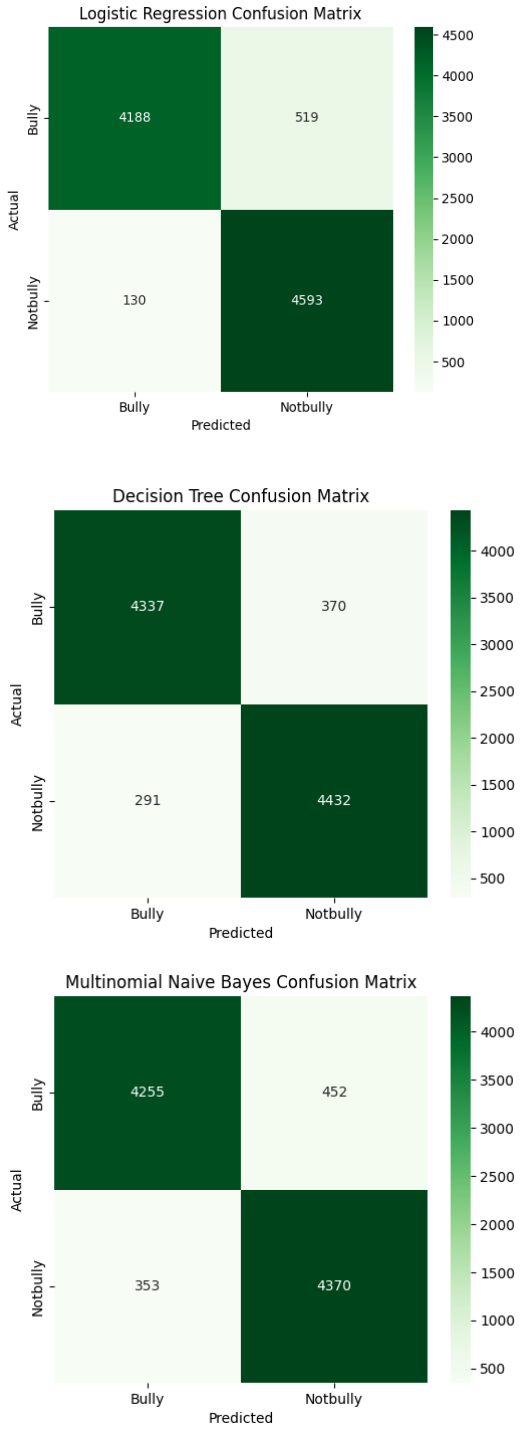


Fig 12: Confusion Matrices of the Models

4.7 SHAP Summary Plot

SHAP summary plot for the models visualizes the influence of the different words on the output of the models by listing the most influential features (words) for the classification. It aggregates feature importance in the dataset thereby showing both the magnitude and direction of contribution made by each word. Words with strong relationships to bullying behavior appeared prominently, therefore pushing the model's predictions toward the bully class. Furthermore, words such as "bitch", "fuck", "hoe", "black", "wear", "pussi", were discovered to have high impact on the models' predictions. This process is crucial since it not only validates the model's learning process but also gives actionable insights into the linguistic markers of online harassment. Figure 13 displays the SHAP summary plot showing the overall feature impact while Figure 14 shows the SHAP individual prediction explanation.

SHAP explainability plots revealed the linguistic patterns that heavily influenced classification outcomes. This transparency is important for ethical AI deployment, especially in sensitive domains like cyberbullying detection. This allows the stakeholders such as platform moderators, developers, and users, to understand why certain content is flagged, fostering trust and accountability. Moreover, the SHAP's individual prediction explanations give case-by-case interpretability, which enables manual review and potential model refinement. This feature is particularly valuable in borderline cases where context may alter the perceived offensiveness of a term. In all, SHAP explainability plots serve as a bridge between algorithmic efficiency and ethical AI deployment which empowers stakeholders to balance automation with human oversight. This ensures that models does not only perform well statistically but also operate responsibly in real-world applications.

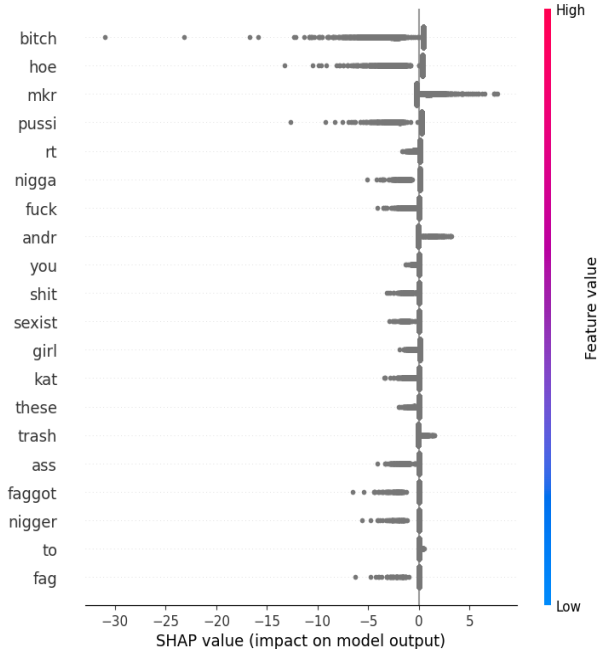


Fig 13: SHAP Summary Plot

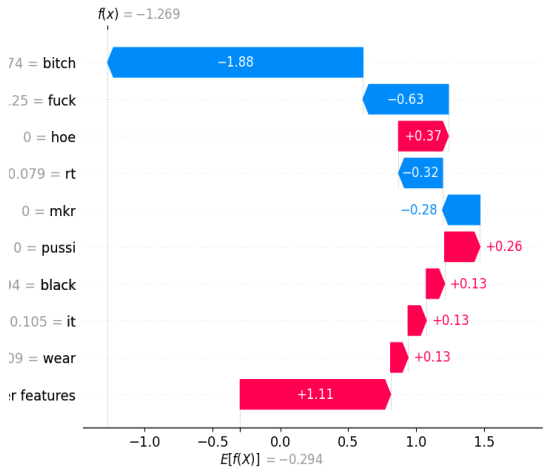


Fig 14: SHAP Individual Prediction Explanation

5. CONCLUSION

This study presented the design, development, and evaluation of an Explainable Artificial Intelligence (XAI) based cyberbullying classification system which aimed at detecting harmful content on social media platforms. The system was developed to address the growing problem of online harassment by using machine learning models that not only classify content as bullying or non-bullying but also provide transparent explanations for their decisions. The study started with the collection of a labeled dataset comprising tweets containing both textual and symbolic elements. A comprehensive preprocessing methods were implemented, including tokenization, stemming, lemmatization, and stopword removal, which was followed by vectorization using Term Frequency-Inverse Document Frequency (TF-IDF). To handle the issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, which resulted in a balanced dataset that improved the fairness and performance of the models. Three machine learning algorithms which include Logistic Regression, Decision Tree, and Multinomial Naive Bayes, were trained and evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Among these models, Logistic Regression demonstrated the best generalization capability with a test accuracy of 93.12%, making it the most preferred for deployment. The Decision Tree model showed signs of overfitting despite high training accuracy, while the Naive Bayes model performed consistently but was less effective in handling complex feature interactions. A key contribution of this work is the integration of SHAP (SHapley Additive exPlanations) to enhance model interpretability. SHAP provided explanations into the influence of specific features or words on classification outcomes, enabling stakeholders to understand and trust the decisions of the system. This level of transparency is crucial for ethical AI deployment, especially in sensitive applications like cyberbullying detection. Furthermore, the system was successfully deployed through a web interface, allowing real-time interaction and content moderation. However, the research also acknowledged ethical challenges which include the risk of misclassifying contextually patterned language, potential biases in feature importance, and concerns around data privacy and user consent.

The research demonstrated that effective preprocessing techniques, such as the tokenization, stemming, lemmatization, and TF-IDF vectorization, practically enhanced the quality of input data, enabling the models to learn meaningful patterns. The application of SMOTE to normalize class imbalance further improved the fairness and reliability of the models, as evidenced by balanced precision, recall, and F1-scores. Among the evaluated machine learning models, Logistic Regression emerged as the most suitable for deployment, which offers a strong balance between generalization and interpretability. The Decision Tree model which was highly accurate on training data, showed signs of overfitting, and the Naive Bayes model, though consistent, only struggled with nuanced feature interactions. The introduction of SHAP explainability plots added an important layer of interpretability, allowing stakeholders to understand the rationale behind each classification. This is particularly valuable in sensitive domains like cyberbullying detection, where context and intent are often ambiguous. However, the system also highlighted the challenges of contextual misclassification and potential biases in feature importance, which underscores the need for continuous monitoring and ethical oversight.

6. REFERENCES

[1] Martin, F., Zhuang, M., & Schaefer, D. (2024). Systematic review of research on artificial intelligence in K-12 education (2017–2022). Computers and Education: Artificial Intelligence, 6, 100195.

[2] Gupta, A. K., Kumar, A., & Kumar, B. (2025). Advancing sustainable food packaging: Integrating machine learning, deep learning, and artificial intelligence. Trends in Food Science & Technology, 163, 105148.

[3] Sharma, A., Lysenko, A., Jia, S., Boroevich, K. A., & Tsunoda, T. (2024). Advances in AI and machine learning for predictive medicine. Journal of Human Genetics, 69(10), 487-497.

[4] Saifullah, K., Khan, M., Jamal, S. and Sarker, I (2024). Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 11, 1–12.

[5] Prama, T. T., Amrin, J. F., Anwar, M. M., & Sarker, I. H. (2025). AI Enabled User-Specific Cyberbullying Severity Detection with Explainability. *arXiv preprint arXiv:2503.10650*.

[6] Aronna, S. R., Zumma, T., Prodhan, R., Zohora, F., Sakib, N. and Tahmiduzzaman, K. (2023). A Study of Cyber Bullying Classification Using Social Media and Textual Analysis Based on Machine Learning Approaches. In Proceeding of 14th ICCCNT IEEE Conference, IIT, Delhi, India.

[7] Garzia-Mendez, S. and Arriba-Perez, F. (2025). Promoting Security and Trust on Social Networks: Explainable Cyberbullying Detection using Large Language Models in a Stream-based Machine Learning Framework. arXiv:2505.03746v1 [cs.SI].

[8] Armas, D. G. A., Toapanta, S. M. T., Díaz, E. Z. G., Guerrero, J. L. J., Arellano, M. R. M., & Hifóng, M. M. B. (2025). Influence of Social Media and Artificial Intelligence on Cyberbullying for Decision-Making with Legal or Judicial Foundations in Ecuador. Journal of Internet Services in Information Security 15(1), 32-50.

[9] Perera, A., & Fernando, P. (2024). Cyberbullying

detection system on social media using supervised machine learning. Procedia Computer Science, 239, 506-516.

[10] Unnava, S., & Parasana, S. R. (2024). A study of cyberbullying detection and classification techniques: A machine learning approach. *Engineering, Technology & Applied Science Research*, *14*(4), 15607-15613.

[11] Ambareen, K., & Meenakshi Sundaram, S. (2023). A Survey of Cyberbullying Detection and Performance: Its Impact in Social Media using Artificial Intelligence. *SN Computer Science*, *4*(6), 859.

[12] Xu, Q., Feng, Z., Gong, C., Wu, X., Zhao, H., Ye, Z., ... & Wei, C. (2024). Applications of explainable AI in natural language processing. Global Academic Frontiers, 2(3), 51-64.

[13] Uddin, M. K. S. (2024). A review of utilizing natural language processing and AI for advanced data visualization in real-time analytics. Global Mainstream Journal, 1(4), 10-62304.

[14] Sarella, P. N. K., & Mangam, V. T. (2024). AI-driven natural language processing in healthcare: transforming patient-provider communication. Indian Journal of Pharmacy Practice, 17(1), 21-26.

[15] Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative AI. Customer Needs and Solutions, 11(1), 3.

[16] Fashakh, A. M., Çevik, M., Aydoğan, Ş. K., & Ibrahim, A. A. (2025). Detection cyberbullying using AI and sentiment analysis to examine psychological impacts on vulnerable groups. Egyptian Informatics Journal, 32, 100856.

[17] Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 29(3), 1839-1852.

[18] Abood, M. M., & Al-Bayati, M. A. (2024). Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). Journal Port Science Research, 7(3), 268 – 280

[19] Ajayi, E., Kachweka, M., Deku, M., & Aiken, E. (2025). A Machine Learning Approach for Detection of Mental Health Conditions and Cyberbullying from Social Media. arXiv preprint arXiv:2511.20001.

[20] Wulandari, W., Makmur, H., Surianto, D. F., Risal, A. A. N., Budiarti, N. A. E., Zain, S. G., & Wahid, A. (2025). Semantic Feature Engineering with LSA-SVM for Cyberbullying Comment Classification on instagram, Informatica, 49(15), 165 – 178.

[21] Zhao, Z. (2025).0 Let Network Decide What to Learn: Symbolic music understanding model based on large-scale adversarial pre-training. In Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA. ACM, New York, NY, USA,

[22] Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. ACM Computing Surveys, 57(2), 1-44.

[23] Tilala, M. H., Chenchala, P. K., Choppadandi, A., Kaur, J., Naguri, S., Saoji, R., ... & Tilala, M. (2024). Ethical considerations in the use of artificial intelligence and machine learning in health care: a comprehensive review. Cureus, 16(6), e62443. doi: 10.7759/cureus.62443

[24] Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: the promise and limitations of explainable machine learning in cardiology. Canadian Journal of Cardiology, 38(2), 204-213.

[25] Ige, T., & Adewale, S. (2022). AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. arXiv preprint arXiv:2207.11897.

[26] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1-10.

[27] Almufareh, M. F., Jhanjhi, N. Z., Humayun, M., Alwakid, G. N., Javed, D., & Almuayqil, S. N. (2025). +Integrating sentiment analysis with machine learning for cyberbullying detection on social media. IEEE Access, 13, 78348-78359.

[28] Sanchez, R., Hernández, P., & Fernandez, E. (2023). Integrating RNNs with rule-based explanations for temporal pattern recognition in cyberbullying. Applied Intelligence, 53(4), 3562-3578.

[29] Shah, V., Sinha, A., Navalkar, N., Gupta, S., Gonsalves, P., & Malik, A. (2023). ML and Natural Language Processing: Cyberbullying Detection System for Safer and Culturally Adaptive Digital Communities. Journal of Smart Internet of Things, 2, 193-205.

[30] Krak, I., Sobko, O., Molchanova, M., Tymofiiev, I., Mazurets, O., & Barmak, O. (2024). Method for Neural Network Cyberbullying Detection in Text Content with Visual Analytic. In CS&SE@ SW, 298-309.