Classifying the Severity of Cyberbully from Social Media Comments

Makkala Nokham
Department of Statistics, Chiang Mai University
239 Huaykaew Road, Suthep, Mueang Chiang Mai, Chiang Mai, Thailand 50200

Bandhita Plubin
Department of Statistics, Chiang Mai University
239 Huaykaew Road, Suthep, Mueang Chiang Mai, Chiang Mai, Thailand 50200

Walaithip Bunyatisai
Department of Statistics, Chiang Mai University
239 Huaykaew Road, Suthep, Mueang Chiang Mai, Chiang Mai, Thailand 50200

Thanasak Mouktonglang
Department of Mathematics, Chiang Mai University
239 Huaykaew Road, Suthep, Mueang Chiang Mai, Chiang Mai, Thailand 50200

Suwika Plubin
Department of Statistics, Chiang Mai University
239 Huaykaew Road, Suthep, Mueang Chiang Mai, Chiang Mai, Thailand 50200

ABSTRACT

The proliferating omnipresence of cyberbullying on digital social networks has crystallized into a pressing societal dilemma, exacting substantial emotional and psychological tolls on affected individuals. Traditional methodologies for identifying and combating cyberbullying are hindered by the expansive scope and multifaceted complexity of digital content. This paper explores the utilization of emerging machine learning technologies and sophisticated natural language processing approaches to automate the detection and classification of cyberbullying within social media contexts. Specifically, the study applies Bidirectional Encoder Representations from Transformers (BERT), Naïve Bayes (NB), and Support Vector Machine (SVM) frameworks to systematically classify user-generated comments into non-cyberbullying and distinct tiers of cyberbullying severity, specifically Low, Middle, and High Severity. The dataset consists of 13,204 comments from platforms like Facebook, X (formerly Twitter), and TikTok. The results demonstrate that the SVM model surpasses the performance of its counterparts, achieving a remarkable accuracy of 94% and an F1-Score of 0.95 in binary classification. BERT also demonstrated strong performance, particularly in multi-level severity classification, while NB showed the lowest performance. Stacking also exhibited strong performance, particularly in detecting High Severity Cyberbullying. While NB and BERT performed well, especially in binary classification, they were less consistent in the multi-level severity classification. The findings highlight the effectiveness of SVM for detecting cyberbullying severity, offering valuable insights for future automated moderation and content classification systems.

General Terms

Natural language processing (NLP); text classification; machine learning

Keywords

Cyberbullying detection; BERT; Naïve Bayes; Support Vector Machine; social media analysis

1. INTRODUCTION

In today's world, individuals are increasingly engaged with screens, constantly interacting with phones or computers, and extensively participating in social media and social networks. Technology significantly facilitates human lives, particularly in communication, interaction, and debate. Furthermore, it expands the space for exchanging social opinions more broadly. Social media evidently acts as a platform superseding traditional media such as newspapers or scheduled television programs, which people previously used to share news. Consequently, it can be argued that social media is overcoming the concept of traditional media [1]. The rapid growth of technology, especially social media, has become a powerful tool for individuals to quickly express their opinions or comment on others in the online world. Social media has evolved into a vast medium where cyberbullying has become prevalent.

Cyberbullying can be defined as malicious online behavior involving slander, humiliation, and actions that harm others' mental well-being or personal property. Such actions are frequently observed on social media platforms like Facebook, X, Instagram, TikTok, or in comments on YouTube [2]. Victims of cyberbullying experience intense emotions, including stress, anxiety, sadness, feelings of worthlessness, and anger, which can lead to conflicts at school. Additionally, they experience a constant sense of harassment, even in environments where they should feel safe, such as their homes. This sense of entrapment can sometimes lead to thoughts of self-harm or even suicide [3].

The advancement of communication technology, which enables unrestricted use, has made controlling online bullying increasingly challenging. This includes actions such as threatening harm, impersonating others, and exposing personal secrets. Online behavior has escalated into a growing problem. Various entities, including governments and private sectors, are attempting to address these escalating issues. They have initiated campaigns to raise awareness about the harmful effects of online bullying. Despite efforts to foster protection against issues in online communication, many individuals persist in bullying others on the internet.

Researchers are therefore exploring the use of Machine Learning in conjunction with Natural Language Processing methods to better understand human language [4]. These technologies are being applied in Text Mining to identify the severity of comments involving online bullying. This approach aims to track and monitor online bullying behavior using Bidirectional Encoder Representations from Transformers (BERT), Naïve Bayes (NB), Support Vector Machine (SVM), and Stacking Models.

2. RELATED WORK

In recent years, the proliferation of social media platforms has led to a surge in research focused on detecting cyberbullying and harmful content online. Various machine learning and natural language processing (NLP) techniques have been explored for this purpose. This section reviews some notable studies that have contributed to the development of cyberbullying detection models.

Kit Thananukhun et al. [5] proposed a method for question classification in Thai language chatbot systems using Artificial Neural Networks (ANN) and BERT. Their study focused on improving the accuracy of question-answering systems, a critical task in NLP. The results showed that combining BERT with a Multilayer Perceptron (MLP) achieved the highest accuracy of 92.57%, outperforming other classification models such as SVM and NB. Although the system was not directly related to cyberbullying, their work highlighted the potential of using BERT for text classification tasks.

Shivani et al. [6] conducted a study comparing SVM and NB for sentiment analysis using BERT embeddings. The study focused on classifying movie reviews as either positive or negative, leveraging supervised learning techniques. Their results indicated that SVM, with Radial Basis Function (RBF) kernels, outperformed NB, with a 2.5% improvement in accuracy. While their work focused on sentiment analysis, the techniques used for text classification are highly applicable to cyberbullying detection, where sentiment and tone play crucial roles.

Kusumawati et al. [7] compared the performance of NB and SVM in classifying customer service feedback on Twitter. Their study aimed to categorize user feedback into positive or negative sentiment towards the Tokopedia marketplace. The SVM model with a linear kernel outperformed NB with an accuracy of 83.34%. This study demonstrated the potential of SVM for classification tasks and highlighted its application in social media data mining, which is relevant to detecting harmful online content such as cyberbullying.

Venkataramana et al. [8] investigated various machine learning and deep learning models for classifying COVID-19-related content into three categories: positive, negative, and neutral. Their study used datasets from Twitter and applied models such as NB, K-Nearest Neighbors (K-NN), SVM, Decision Tree, Logistic Regression, Long Short-Term Memory (LSTM), and BERT. They found that BERT outperformed other models with an accuracy of 56%. Their findings demonstrate the superiority of BERT in text classification tasks, which is directly relevant to detecting online bullying behavior.

Yi Tian et al. [9] conducted a comparative study on detecting fake news using machine learning and deep learning models. They compared Logistic Regression, Random Forest, SVM, and NB, as well as deep learning models such as Recurrent Neural Network (RNN), LSTM, and BERT. The study concluded that BERT achieved the highest accuracy of 86.76%, demonstrating its potential for detecting misleading or harmful content in online platforms, including cyberbullying.

These studies illustrate the growing interest in applying machine learning and deep learning models, especially BERT, to detect harmful and malicious content online. BERT's superior contextual understanding makes it highly effective in tasks like sentiment analysis, fake news detection, and cyberbullying classification. However, computational complexity and processing time remain challenges when implementing these models at scale.

3. DATASET AND PROPOSED METHODS

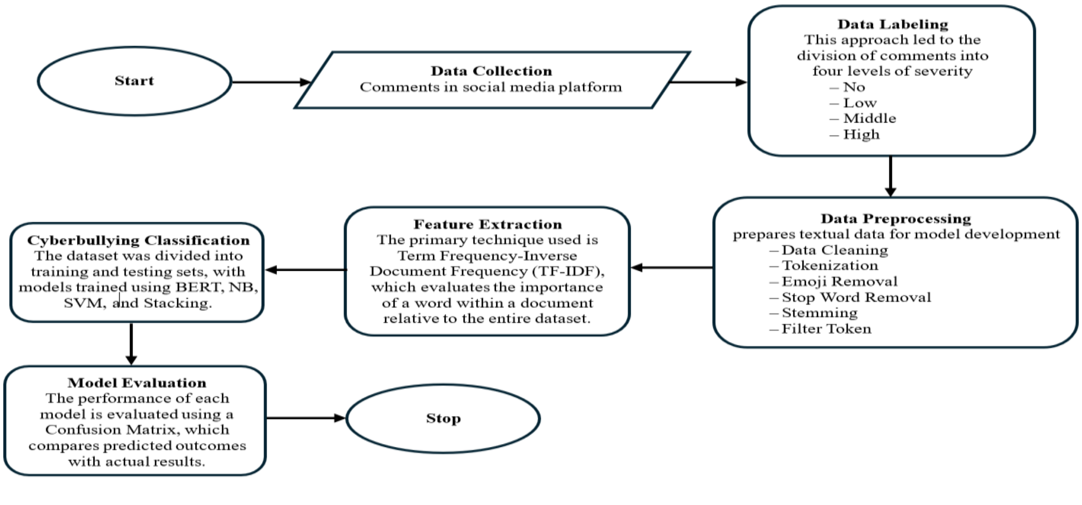
Start
Data Collection
Comments in social media platform
Data Labeling
This approach led to the division of comments into four levels of severity
– No
– Low
– Middle
– High
Data Preprocessing
prepares textual data for model development
– Data Cleaning
– Tokenization
– Emoji Removal
– Stop Word Removal
– Stemming
– Filter Token
Feature Extraction
The primary technique used is Term Frequency-Inverse Document Frequency (TF-IDF), which evaluates the importance of a word within a document relative to the entire dataset.
Cyberbullying Classification
The dataset was divided into training and testing sets, with models trained using BERT, NB, SVM, and Stacking.
Model Evaluation
The performance of each model is evaluated using a Confusion Matrix, which compares predicted outcomes with actual results.
Stop

Fig 1: Main process of analysis

3.1 Data collection

The data collection process involved gathering comments from three major social media platforms: Facebook, X, and TikTok. This process yielded a total of 13,204 messages. The dataset was systematically categorized into 3,906 non-cyberbullying messages and 9,298 cyberbullying messages

3.2 Data labeling

In the absence of definitive tools for assessing cyberbullying severity in social media comments, the study categorized messages by considering their potential impact. These impacts include psychological responses such as irritation, anger, and shame, alongside physical consequences like sleep disturbances and appetite changes. This approach led to the division of comments into four levels of severity:

(1) Level 0 refers to messages that are not categorized as cyberbullying. These messages lack any form of foul language, cursing, slander, or insulting content. The majority of messages classified as Level 0 are either compliments or neutral statements, among other types of non-cyberbullying content. Examples of such messages include:
- “เก่งมากค่ะทำไมใส่ชุดนี้”
  (“You're so good, why are you wearing this?”)
- “การงานราบรื่นพระคุ้มครองครับผม”
  (“May your work be smooth, and may the Buddha protect you.”)
- “สุดในรุ่น”
  (“Ultimate in this version.”)

(2) Level 1 includes low-severity cyberbullying messages, characterized by psychological effects such as anxiety or unease, which have minimal impact on daily routines. Examples of such messages include:
- “โอ้ยยย ฉาบปูนยี่ห้ออะไรถึงได้ด้านได้ทนขนาดนี้”
  (“Oh my god, what brand of plaster is so thick-skinned and durable?”)
- “สถาบันไม่เกี่ยวเกี่ยวที่สันดานคน”
  (“The institution isn't involved, it's the person's character.”)
- “เรายังสวยกว่าเลย555”
  (“I'm even prettier than you, hahaha.”)

(3) Level 2 refers to moderate cyberbullying messages characterized by harsh words that can disrupt daily life, leading to issues such as loss of self-confidence, stress, or insomnia. Example expressions include 'beast,' 'vile man,' 'mental disorder,' 'irregular,' 'piece of shit,' and 'dregs of society'. Examples of such messages include:
- “มึงแต่งชุดนี้หาผัวดีกว่า ดูจากที่พูดแต่ละคำมึงเก่งกว่าครูอีกไม่ต้องให้ครูสอนหรอกเหมือนมึงจะมาสอนครูสักแล้วนะไอ่หนู”
  (“You should dress like this to find a husband. Judging by every word you say, you're smarter than the teacher. No need for the teacher to teach you; it's like you're about to teach the teacher, little mouse.”)
- “สัตว์นรกส่งมาเกิด”
  (“A hell-spawned animal”)
- “ไปไหนก็ไปไอ้เปรต”
  (“Go wherever, you damn hungry ghost!”)

(4) Level 3 refers to high-severity cyberbullying, where the messages have a critical and devastating impact on the victim, potentially leading to fatal outcomes. Examples of such messages include:
- “ขอให้มันวิบัติตายอย่างทรมาน”
  (“I hope it suffers a disastrous and painful death.”)
- “ขอให้ตายอย่างทรมารตามน้อง ๆที่น่าสงสารไปด้วยเถอะนิสัยชาติชั่วที่สุด 😭 😭 ”
  (“I hope you die a painful death and follow those poor kids. You have the worst character. 😭 😭 ”)
- “มึงสมควรตาย จะตายแบบไหน ก็ควรตาย 😡 😡 😡 ”
  (“You deserve to die. However you die, you should die. 😡 😡 😡 ”)

3.3 Data preprocessing

Data preprocessing prepares textual data for model development, ensuring the resulting models are both efficient and accurate. This process employs various techniques, as detailed below:

3.3.1 Data Cleaning

Text entries containing only numerical values or special characters (e.g., 55555, !!!!, ...) are removed, from the dataset, as they are regarded as lacking semantic value.

3.3.2 Tokenization

Text is segmented into individual words or tokens based on Thai linguistic rules. This step is facilitated by the PyThaiNLP API.

- Input: “ เพื่อนในเฟสเราเต็มเลยค่ะ ”
  (Input: “My Facebook friends list is full”)
- Output: “เพื่อน” // “เฟส” // “เรา” // “เต็ม” // “เลย”
  (Output: “friends” // “Facebook” // “I” // “full” // “very”)

3.3.3 Emoji Removal

Emojis are removed from the text after tokenization.

- Input: “รูปนี้แกงมาก 🤣 🤣 🤣 ”
  (Input: “This picture is so a big prank 🤣 🤣 🤣 ”)
- Output: “รูป” // “นี้” // “แกง” // “มาก”
  (Output: “picture” // “this” // “prank” // “very”)

3.3.4 Stop Word Removal

Common filler words (e.g., “และ (and)”, “หรือ (or)”, “จึง (so)”, “แต่ (but)”, “ที่ (at)”) are eliminated as they do not add meaning.

- Input: “แอนสวยกว่าเยอะค่ะมาก”
  (Input: “Ann is much prettier.”)
- Output: “แอน” // “สวย” // “กว่า” // “เยอะ”
  (Output: “Ann” // “pretty” // “more than” // “much”)

3.3.5 Stemming

Words are reduced to their root forms.

- Input: “ มากกกกกกก ”
  (Input: “veryyyyy”)
- Output: “ มาก ”
  (Output: “very”)

3.3.6 Filter Token

Tokens with fewer than three characters are excluded, as such tokens often lack meaningful content.

- Excluded Tokens: “ มา (come)”, “ รอ (wait)”, “ ไป (go)”

3.4 Feature extraction

Feature extraction transforms textual data into numerical formats, making it suitable for machine learning models. This process involves converting text into numerical vectors, allowing models to analyze data effectively. The primary technique used is Term Frequency-Inverse Document Frequency (TF-IDF), which evaluates the importance of a word within a document relative to the entire dataset.

3.5 Classification method

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

3.5.1 Bidirectional encoder representations from transformers (BERT)

BERT, a neural network-based model developed by Google, is specifically designed to address complex challenges in Natural Language Processing (NLP). By utilizing pre-trained deep learning architectures, BERT interprets the contextual meaning of words by analyzing their interactions with surrounding words within a sentence. This bidirectional approach enables the model to capture nuanced semantic relationships, thereby enhancing the understanding of intricate textual data. BERT generates word embeddings dense semantic vectors that provide a deeper comprehension of language, surpassing the capabilities of traditional models [10].

3.5.1.1 BERT Model Architecture

BERT is grounded in the Transformer architecture [11], utilizing its encoder portion. The model consists of several layers of bidirectional self-attention mechanisms paired with feed-forward neural networks [11]. The standard BERT base model is characterized by the following parameters:

- L = 12 (the number of layers or Transformer blocks)
- H = 768 (the hidden size)
- A = 12 (the number of self-attention heads)

This configuration yields approximately 110 million parameters for the BERT base model, providing a robust foundation for learning contextual representations of language.

3.5.1.2 Input Representation

BERT's input representation is a combination of three types of embeddings, which together form the final token embedding:

$$E_{token} = E_{token} + E_{segment} + E_{position} \tag{1}$$

where E_{token} represents token embeddings, created using WordPiece embeddings with a vocabulary size of 30,000 tokens, $E_{segment}$ represents segment embeddings, which distinguish between two sentence pairs and $E_{position}$ indicates position embeddings, which encode the position of each token within the input sequence.

3.5.1.3 Self-Attention Mechanism

At the heart of BERT's architecture is the multi-head self-attention mechanism. For each attention $head_i$ the attention is computed as follows:

$$head_i = Attention(Q_i, K_i, V_i) \tag{2}$$

Where Q_i, K_i and V_i are the query, key, and value matrices, respectively. The attention function is defined as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Where d_k denotes the dimension of the key vectors. The outputs of all attention heads are concatenated and linearly transformed to form the multi-head attention:

$$MultiHead(Q,K,V) = concat\left(head_1,...,head_h\right)W^O \tag{4}$$

3.5.1.4 Pre-training Process

BERT undergoes pre-training on a large corpus, including Wikipedia and BookCorpus, using two unsupervised learning tasks:

- Masked Language Model (MLM): 15% of the input tokens are randomly masked, and the model predicts these masked tokens.
- Next Sentence Prediction (NSP): The model predicts whether two given sentences appear consecutively in the original text.

The pre-training loss is computed as the sum of the two task-specific losses:

$$L = L_{MLM} + L_{NSP} \tag{5}$$

Where L_{MLM} is the mean masked language model likelihood, and L_{NSP} is the mean next sentence prediction likelihood.

3.5.1.5 Fine-tuning for Cyberbullying Detection

Following the pre-training phase, BERT is fine-tuned to detect cyberbullying in social media comments using a labeled dataset. The fine-tuning process involves adapting the pre-trained model to the specific task by incorporating a task-specific output layer, defining a suitable loss function, and configuring training hyperparameters to optimize model performance. The first step in fine-tuning involves adding an output layer that is tailored to the task at hand. This output layer computes the final prediction through a softmax function applied to the output of the BERT model, which is weighted by a task-specific parameter matrix W and shifted by a bias term b. Mathematically, this is expressed as:

$$y = soft\max(BERT(x)W^T + b) \tag{6}$$

where x denotes the input sequence, W represents the parameter matrix specific to the task, and b is the bias term. The softmax function ensures that the model outputs a probability distribution over the possible classes for each input sequence.

To guide the learning process, the model is trained using the cross-entropy loss function, which quantifies the difference between the predicted labels and the true labels in the dataset. The objective is to minimize this loss function, which is defined as:

$$L_{CE} = -\sum(y_{true} \times \log(y_{pred})) \tag{7}$$

Here, y_{true} represents the true labels, and y_{pred} refers to the predicted probabilities. This loss function encourages the model to assign higher probabilities to correct predictions and minimize errors during training.

Finally, the fine-tuning process involves configuring several hyperparameters to optimize the training process. The learning rate is adjusted dynamically using a learning rate scheduler to prevent overfitting and facilitate convergence. The batch size is chosen based on available computational resources to ensure training stability. The model is trained for 3 to 5 epochs to ensure generalization to unseen data, and dropout is applied as a regularization technique to prevent overfitting. This configuration ensures that the fine-tuned BERT model effectively learns to detect cyberbullying in social media comments while maintaining the ability to generalize to new, unseen data.

3.5.2 Naïve bayes (NB)

NB is a method of supervised learning used for categorizing data, particularly when the dependent variable is a categorical variable. This model relies on probability theory, evaluating the likelihood of X belonging to each group and classifying X into the group where it has the highest probability of membership [12].

In the classification process, X is assigned to the group with the highest posterior probability. The probability that X belongs to Y, denoted as $P(Y = j \mid X)$, can be computed using Bayes' theorem, known as posterior probability. This can be calculated using the formulas in equations (8) and (9).

$$P(Y = j \mid X) = \frac{P(X \mid Y = j)P(Y = j)}{P(X)} \tag{8}$$

$$P(X) = \sum_{\forall_j} P(X \mid Y = j)P(Y = j) \tag{9}$$

When $P(X \mid Y = j)$ represents the conditional probability of X occurring given that it belongs to group j, $P(Y = j)$ is the prior probability, and j refers to the possible group.

The NB method assumes that all independent variables within each group are conditionally independent of each other. The conditional probability can be calculated using equation (10).

$$P(X \mid Y = j) = P(x_1 \mid Y = j)P(x_2 \mid Y = j) \tag{10}$$

where $X = (x_1, x_2)$.

3.5.3 Support vector machine (SVM)

SVM is a machine learning technique that can be used to solve classification problems by separating data into different groups. This can be achieved using linear equations, both for linearly separable and non-linearly separable data [13].

SVM is based on linear classification, which is a supervised learning approach. It works by finding a hyperplane that divides the data into different groups using a linear function positioned between the groups to be separated. The hyperplane equation is provided in equation (11).

$$w \times x + b = 0 \tag{11}$$

When w represents the weight vector and b represents the bias term.

In SVM, the largest margin between the groups is used to determine the optimal hyperplane. The vectors that lie on the boundaries of each group are known as support vectors.

The best separating hyperplane is determined by maximizing the margin on both sides, resulting in new boundaries that define the data groups. The separating hyperplane is the one that touches the closest data points in the feature space. These boundaries are expressed as $(w \times x^{+}) + b \geq 1$ for $y = 1$ and $(w \times x^{-}) + b \leq -1$ for $y = -1$. A wider margin signifies clearer separation between the datasets, making the hyperplane with the widest margin the best. The margin width is given by equation (12), and the values of w and b are calculated from equations (13) and (14), respectively

$$\gamma = \frac{2}{\| w \|} \tag{12}$$

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{13}$$

$$b = -\frac{\max_{y_i = -1}(wx) + \min_{y_i = 1}(wx)}{2} \tag{14}$$

Where γ is the margin and α is the constant coefficient $\alpha_i \geq 0;\ i = 1, 2, 3, ..., N$. After determining w and b, x is classified by substituting it into the equation. A positive y-value assigns x to the first group, a negative y-value to the second, and a zero y-value allows x to be placed in either group.

3.5.4 Stacking model

The stacking model is an ensemble learning technique designed to improve predictive performance by combining multiple machine learning models, such as BERT, NB, and SVM. This technique constructs a meta-model that integrates the strengths of the individual base models to yield more accurate predictions.

In the stacking approach, several base models are independently trained on the same dataset to capture a range of patterns and relationships. The predictions from these base models are then used as input features for a meta-model, which is trained to optimally combine these predictions and generate the final classification output.

This approach offers several advantages, including enhanced accuracy, increased flexibility, and a reduction in overfitting. Stacking is particularly effective in complex tasks such as text classification, where it is essential to capture various linguistic and contextual features. In the context of detecting cyberbullying severity in social media comments, stacking allows for the integration of models like BERT, NB, and SVM, which collectively enhance the overall classification performance.

The stacking procedure consists of training multiple base models, generating meta-features based on their predictions, and subsequently training the meta-model on these meta-features. The final classification is determined by the meta-model, which synthesizes the outputs from the base models to generate a more accurate and reliable prediction [14].

3.6 Model evaluation

The performance of each model is evaluated using a Confusion Matrix, which compares predicted outcomes with actual results. The matrix includes:

Table 1. Confusion Matrix

	Predict		
	Classes	**Yes**	**No**
Actual	**Yes**	TP	FN
	No	FP	TN

- True Positive (TP): Correct predictions of positive outcomes.
- True Negative (TN): Correct predictions of negative outcomes.
- False Positive (FP): Incorrect predictions of positives.
- False Negative (FN): Incorrect predictions of negatives.

(1) Accuracy: Proportion of correct predictions to the total predictions using equation (15).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

(2) Recall (True Positive Rate): Ability to identify all positive cases using equation (16).

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{16}$$

(3) Precision: Proportion of true positives among predicted positives using equation (17).
Words are reduced

$$\Pr ecision = \frac{TP}{TP + FP} \tag{17}$$

(4) F1-Score: Harmonic mean of precision and recall,

balancing both metrics using equation (18).
Tokens

$$F1 - Score = 2 \times \frac{\Pr ecision \times \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \quad (18)$$

4. THE EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the evaluation results of various models used to classify the severity of cyberbullying comments on social media. The analysis encompasses two tasks: first, binary classification (Cyberbullying vs. Non-Cyberbullying), and second, multi-level severity classification (Low, Middle, High severity).

The training arguments for each method were meticulously selected to optimize model performance in classifying cyberbullying severity within social media comments. Table 2 presents a summary of these configurations.

Table 2. Training Arguments for Each Classification Method

Method	TrainingArguments

BERT	learning_rate = 0.00002
	train_batch_size = 32
	eval_batch_size = 32
	train_epochs = 30
	weight_decay = 0.01
	max_length = 512

NB	nb = naive_bayes.MultinomialNB()
	alpha = 1.0

SVM	kernel = linear
	c = 1.0
	degree = 3
	gamma = auto

Stacking	final_estimator = LogisticRegression()
	random_state = 42
	max_length = 512
	stack_method = predict

For the BERT model, a relatively low learning rate of 0.00002 was employed to facilitate fine-tuning of the pre-trained model. The batch sizes for both training and evaluation were set to 32, striking a balance between computational efficiency and model stability. The model was trained for 30 epochs with a weight decay of 0.01 to mitigate overfitting. The maximum sequence length was set to 512 tokens to accommodate longer comments.

NB classifier utilized the MultinomialNB implementation from scikit-learn, with an alpha value of 1.0 for Laplace smoothing. This configuration is well-suited for text classification tasks with discrete features.

SVM was configured with a linear kernel, which is often effective for text classification. The regularization parameter C was set to 1.0, providing a standard trade-off between margin maximization and classification error minimization. While the degree parameter was set to 3, it is not applicable to the linear kernel. The gamma parameter was set to 'auto', allowing the algorithm to automatically determine the appropriate kernel coefficient.

For the Stacking model, Logistic Regression was chosen as the final estimator due to its effectiveness in combining predictions from diverse base models. A random state of 42 was set to ensure reproducibility of results. The maximum sequence length was maintained at 512 tokens for consistency with the BERT model. The stacking method was set to 'predict', utilizing the predicted class labels from base models for the final classification.

4.1 Binary classification divides the comments into two groups

- Non-Cyberbullying: 1,600 comments
- Cyberbullying: 1,600 comments

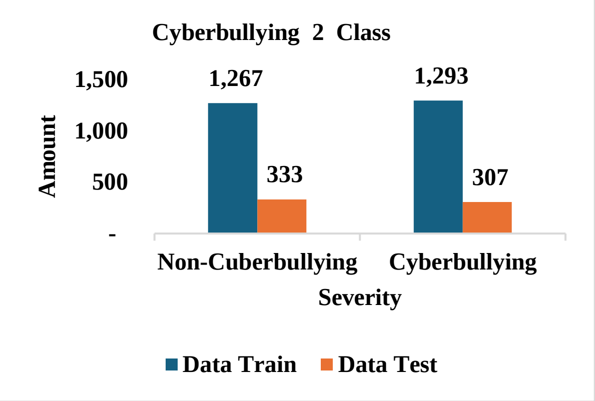

Fig 2: Distribution of Training and Testing Data for Cyberbullying Classification

The dataset was divided into training and testing sets, with models trained using BERT, NB, SVM, and Stacking. The confusion matrix was used to evaluate each model's performance based on key metrics such as Precision, Recall, F1-Score, and Accuracy.

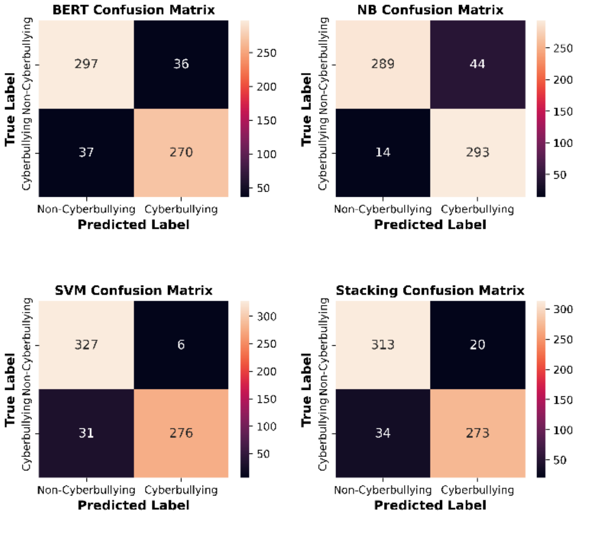

Fig 3: Confusion Matrices of BERT, NB, SVM, and Stacking Models for Binary Classification

From Fig 3, the confusion matrices reveal that the SVM model achieved superior performance compared to the other models. SVM recorded the lowest number of false positives and false negatives, resulting in the highest accuracy and F1-Score. The Stacking model also demonstrated strong performance, effectively balancing predictions across both Non-Cyberbullying and Cyberbullying classes, making it a competitive alternative to SVM. Although the NB model excelled in Cyberbullying detection due to its high true positive rate, its overall accuracy was limited by a higher number of false negatives in Non-Cyberbullying classification.

Conversely, the BERT model exhibited consistent performance but was outperformed by the other models due to comparatively higher error rates.

The performance comparison presented in Table 3 demonstrates that SVM model outperformed all other models across evaluation metrics, particularly excelling in both F1-Score and Accuracy. SVM achieved an Accuracy of 94%, supported by high F1-Scores of 0.95 for Non-Cyberbullying classification and 0.94 for Cyberbullying classification, reflecting its superior capability in accurately handling both classes. The Stacking model exhibited strong performance with an Accuracy of 92% and balanced F1-Scores of 0.92 for Non-Cyberbullying and 0.91 for Cyberbullying, positioning it as a competitive alternative to SVM NB also performed well with an Accuracy of 91%, showing high precision in Non-Cyberbullying detection (0.95) but slightly lower recall values for both classes, impacting its overall F1-Scores. Meanwhile, BERT demonstrated moderate performance, achieving an Accuracy of 89%, with an F1-Score of 0.89 for Non-Cyberbullying and 0.88 for Cyberbullying. These results reinforce the efficacy of the SVM model, highlighting its ability to provide the most reliable and precise classification in binary tasks.

Table 3. Performance Comparison of Models for Binary Classification

Model	Severity	Precision	Recall	F1-Score	Accuracy
BERT	**Non-Cyberbullying**	0.89	0.89	0.89	0.89
	Cyberbullying	0.88	0.88	0.88	
NB	**Non-Cyberbullying**	0.95	0.87	0.91	0.91
	Cyberbullying	0.87	0.95	0.91	
SVM	**Non-Cyberbullying**	0.91	0.98	0.95	0.94
	Cyberbullying	0.98	0.90	0.94	
Stacking	**Non-Cyberbullying**	0.9	0.94	0.92	0.92
	Cyberbullying	0.93	0.89	0.91	

4.2 Multi-Level Classification (Severity Levels)

This section reviews the findings from the multi-level classification task. So, comments are categorized into four severity levels. Some comments are just non-cyberbullying, while others are categorized as Low Severity Cyberbullying. Additionally, there are comments that fall under Middle Severity Cyberbullying, and lastly, there is High Severity Cyberbullying. To ensure fairness, each level consisted of 1,800 comments. The confusion matrix was employed to calculate evaluation metrics such as Precision, Recall, F1-Score, and Accuracy.

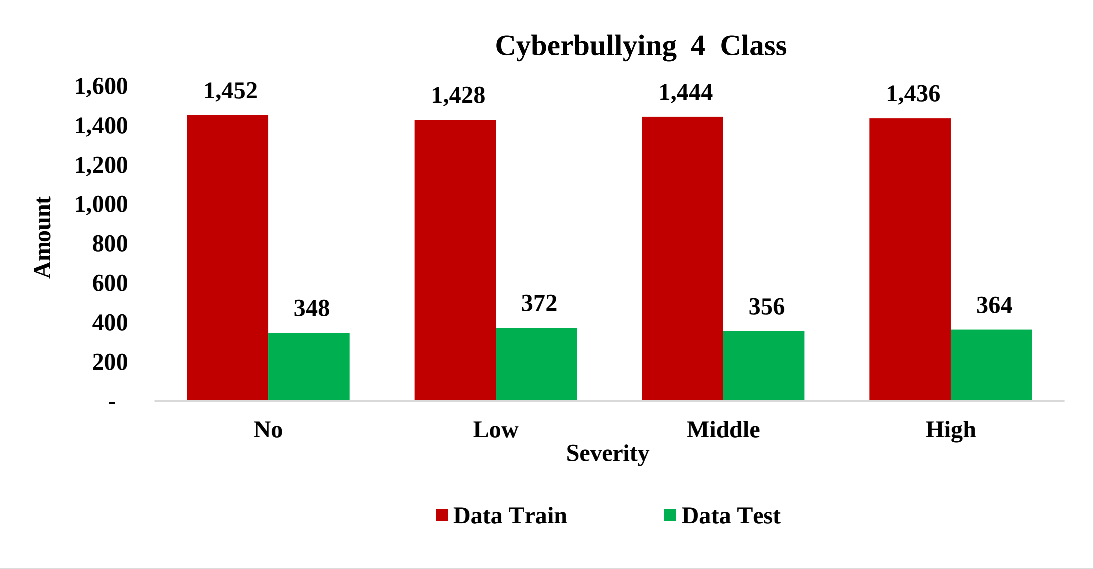

Fig 4: Distribution of Training and Testing Data for Cyberbullying Classification

The confusion matrix analysis in fig 5 reveals that SVM outperformed other models, consistently achieving the highest accuracy and F1-Score across all severity levels due to low false positive and false negative rates, making it the most robust and balanced model for multi-level classification. Stacking demonstrated strong performance, particularly in High and Middle Severity Cyberbullying, although it faced marginal challenges in balancing predictions across classes. BERT showed strengths in High Severity classification, with competitive results in Low Severity, but struggled with distinguishing adjacent severity levels such as Low and Middle due to higher false negatives and positives. NB excelled in Low Severity detection, achieving a high true positive rate, but its performance was hindered by significant misclassification rates in Middle and High Severity levels, making it less effective than the other models.

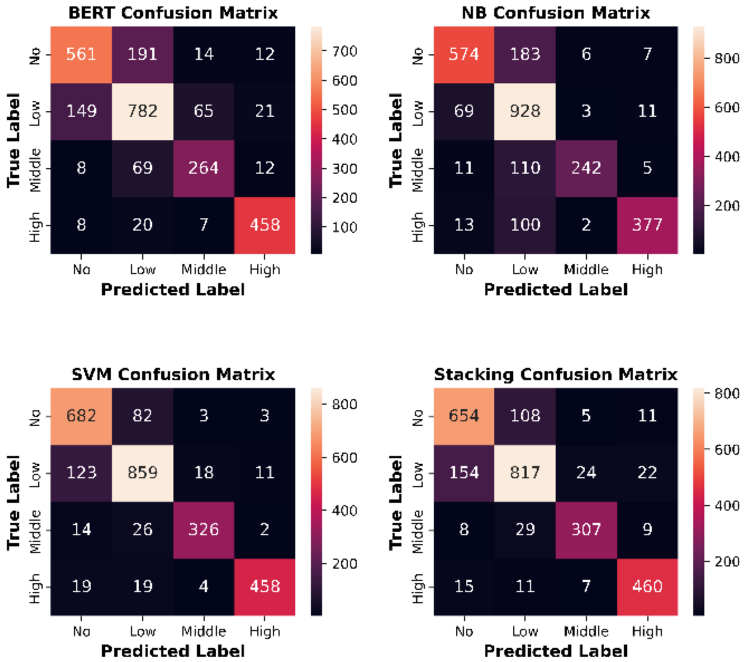

Fig 5: Confusion Matrices of BERT, NB, SVM, and Stacking Models for Multi-Level Classification

The results showed that the SVM model outperformed all other models in both F1-Score and Accuracy. The Stacking model also demonstrated strong performance, particularly yielding high results in the Middle and High Severity classes.

Table 4. Performance Comparison of Models for Multi-Level Severity Classification Comparison

Model	Severity	Precision	Recall	F1-Score	Accuracy
BERT	**Non**	0.77	0.72	0.75	0.78
	Low	0.74	0.77	0.75	
	Middle	0.75	0.75	0.75	
	High	0.91	0.93	0.92	
NB	**Non**	0.86	0.75	0.80	0.80
	Low	0.70	0.92	0.80	
	Middle	0.96	0.66	0.78	
	High	0.94	0.77	0.85	
SVM	**Non**	0.82	0.89	0.85	0.88
	Low	0.87	9.85	0.86	
	Middle	0.93	0.89	0.91	
	High	0.97	0.93	0.95	
Stacking	**Non**	0.79	0.84	0.81	0.85
	Low	0.85	0.80	0.82	
	Middle	0.90	0.87	0.88	
	High	0.92	0.93	0.92	

The results presented in Table 4 highlight that SVM model achieved superior performance across all metrics, particularly excelling in the High Severity and Middle Severity categories, with the highest F1-Scores and an overall Accuracy of 88%. SVM demonstrated robust prediction capabilities, achieving F1-Scores of 0.95 for High Severity and 0.91 for Middle Severity, complemented by consistently high precision and recall values across all severity levels. The Stacking model also delivered strong performance, demonstrating its effectiveness in classifying Low Severity and High Severity categories, with overall Accuracy reaching 85% and F1-Scores of 0.82 and 0.92 for the respective severity levels.

In comparison, BERT model showed dependable results in High Severity classification, achieving an F1-Score of 0.92, but faced challenges in other categories such as Non-Cyberbullying, where its F1-Score dropped to 0.75. NB excelled in precision for specific levels, such as Middle Severity (0.96), but its lower recall values, particularly for High Severity and Middle Severity categories, resulted in overall F1-Scores of 0.85 and 0.78, respectively.

These results underscore the exceptional performance of SVM as the most reliable model for multi-level severity classification, particularly in correctly identifying Medium and High Severity cyberbullying levels. The Stacking model emerged as a competitive alternative, while BERT and NB demonstrated strengths in specific severity levels but fell short in overall consistency.

5. CONCLUSIONS

This study demonstrates the effectiveness of applying machine learning models and natural language processing methods to classify cyberbullying comments based on their severity on social media platforms. The research employed models such as Bidirectional Encoder Representations from Transformers (BERT), Naïve Bayes (NB), Support Vector Machine (SVM), and a Stacking Model to evaluate their performance in both binary and multi-class classification tasks.

The results indicate that SVM outperformed other models, achieving the highest accuracy and F1-scores in both tasks. For binary classification, the SVM achieved an accuracy of 94% and an F1-score of 0.95, showcasing its efficacy in accurately classifying comments into non-cyberbullying and cyberbullying categories. In multi-level severity classification, the SVM model achieved an overall accuracy of 88% and an F1-score of 0.95 for high-severity cyberbullying cases, demonstrating its robustness and reliability in identifying varying levels of harm in online comments.

The study emphasizes the importance of advanced text mining techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and contextual embeddings, in enhancing the ability of models to understand and classify complex human language. The integration of models like BERT and SVM offers a significant advantage in identifying online harassment patterns, enabling real-time monitoring and mitigation of cyberbullying incidents.

Future work could focus on enhancing the generalizability and scalability of the models by expanding the dataset to include comments from additional social media platforms, languages, and cultural contexts. Exploring newer and more advanced Natural Language Processing (NLP) models, such as transformer-based architectures or hybrid ensemble methods, may further improve classification accuracy. Additionally, deploying these models in real-world applications, such as automated content moderation systems, online support systems, and cyberbullying intervention tools, can greatly contribute to fostering safer and more inclusive digital environments. These advancements would not only improve cyberbullying detection but also serve as a foundation for broader applications in harmful content detection and online safety.

6. REFERENCES

[1] Daopradub, P. (2017). The influence of social media on communication and interaction. Research article.

[2] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM-11).

[3] Hee, W. L., Tan, S. P., & Tay, A. A. (2015). The psychological effects of cyberbullying on adolescents. Journal of Youth and Adolescence, 44(2), 235–246. https://doi.org/10.1007/s10964-014-0117-0

[4] Big Data Thailand. (2022). Machine learning and natural language processing for analyzing online behavior. Report.

[5] Thananukhun, K., et al. (2023). Question classification for Thai conversational chatbots using artificial neural networks (ANN) and multilingual BERT. In Proceedings of the International Conference on Natural Language Processing and Computational Linguistics.

[6] Shivani, S., et al. (2022). Comparison of SVM and Naïve Bayes for sentiment classification using BERT embeddings. International Journal of Computer Applications, 179(10), 44–52. https://doi.org/10.5120/ijca2022921443

[7] Kusumawati, S., et al. (2019). Comparing the performance of Naïve Bayes and SVM for Tokopedia's Twitter service classification. Indonesian Journal of Computer Science, 11(2), 101–109. https://doi.org/10.11591/ijcs.v11i2.22679

[8] Venkataramana, V., et al. (2022). COVID-19 sentiment classification using BERT and other machine learning models. International Journal of Data Science and Machine Learning, 6(3), 213–225. https://doi.org/10.11648/j.ijdsm.20220603.15

[9] Tian, Y., et al. (2021). Fake news detection using machine learning and deep learning models. IEEE Transactions on Information Forensics and Security, 17(1), 204–215. https://doi.org/10.1109/TIFS.2021.3081124

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (NeurIPS 2017) (pp. 5998-6008).

[12] Hassan, S., Rafi, M., & Shaikh, M. S. (2011). Comparing SVM and naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment. In Proceedings of the 2011 IEEE 14th International Multitopic Conference (pp. 31–34). IEEE. https://doi.org/10.1109/INMIC.2011.6143977

[13] Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and Their Applications, 18–28. https://doi.org/10.1109/5254.708791

[14] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1.