MALAYSIAN JOURNAL OF
CYBERSECURITY AND APPLICATIONS

DEVELOPMENT OF CYBERBULLYING DATASET (CyTED): FLAMING CLASSIFICATION

Nor Izna Mohd Isa [1], ***Madihah Mohd Saudi*** [2*]

[1]Crowe Malaysia Plt,
Level 13, Tower C, Megan Avenue II, 50450 Kuala Lumpur

[2]Cyber Security and Systems (CSS) Research Unit,
Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800 Nilai

Emails: mnorizna@gmail.com[1], madihah@usim.edu.my[2]*

ABSTRACT

Cyberbullying is a pervasive issue with significant psychological impacts, particularly the flaming type, which often occurs on social media like Twitter. Detecting flaming behavior in the Malaysian context is challenging due to the lack of reliable dataset, especially in Malay. This paper addresses this gap by developing a small dataset of Malay and English keywords related to flaming cyberbullying. The objectives for this paper are to extract relevant keywords from Twitter, to develop a flaming classification dataset and evaluate the proposed dataset by applying machine learning algorithms. A total of 3,600 samples (1,800 Malay,1,800 English) are collected using keyword-based searches via the TweetHarvest tool. Data preprocessing, feature extraction, and classification using Logistic Regression, Random Forest, and SVM are conducted with 10-fold cross-validation. Based on the experiment conducted, The Logistic Regression achieved the highest accuracy rate with 94% for Malay and 95% for English keywords. The paper has successfully developed a dataset for flaming classification, which can be used as is basis for developing cyberbullying detection model.

Keywords: Cyberbullying; Flaming; Malay; Dataset; Machine Learning.

1. INTRODUCTION

Cyberbullying is one of the severe social issues where users gradually and persistently misuse social media platforms. It has become a significant concern as most cases lead to psychological distress and negatively impact individuals' well-being. From January 1 to November 15,2013, the Malaysian Communications and Multimedia Commission (MCMC) removed 1,147 pieces of cyberbullying-related content in collaboration with social media platforms [1]. However, traditional detection methods often rely on keyword matching, which fails to capture nuanced forms of cyberbullying. Cyberbullying is widespread, with significant societal impacts. Detecting flaming cyberbullying in Malaysia is crucial for protecting individuals' well-being and mitigating its effects on social media platforms. The constant evolution of language leads to the frequent emergence of new disparaging terms, making it increasingly challenging to maintain comprehensive and up-to-date datasets [2]. Existing methods often lack the nuanced detection capabilities needed to identify subtle variations in cyberbullying behavior, further compounded by the scarcity of datasets, particularly in Malay [3].

To address the identified gaps and improve detection, relevant keywords related to cyberbullying in Malay and English are extracted from Twitter by scraping text comments using the TweetHarvest tool, which was implemented through a Python script. This process has relied on a predefined flaming wordlist consisting of 24 terms for both languages. Flaming is defined as a heated or intense disagreement that occurs on social media platforms, where participants use insulting, offensive, menacing, and inappropriate language. This behavior is often described as an online fight that takes place through emails, instant messaging, or chat rooms, and it shares significant similarities with online harassment. Social media users frequently view flaming as a deliberate attempt to provoke anger or other negative emotions.

Flaming was chosen as the central focus of this research because it is a widespread social issue that causes significant psychological distress and negatively affects individual well-being. This choice is further justified by a critical shortage of reliable, culturally relevant datasets in Malaysia, particularly in the Malay language, which limits the effectiveness of traditional detection methods in identifying nuanced or evolving disparaging terms.

Additionally, flaming often occurs during discussions on sensitive and controversial topic such as religion, race, and politics, where it can escalate into intense disagreements that involve threatening or inappropriate language. By concentrating on this specific category, the research addresses the pressing need for robust, AI-based detection systems capable of maintaining social media integrity and protecting Malaysian users from hostile online interactions that could result in legal defamation.

To address the limitations of keyword-based searches and to capture creative or indirect forms of offensive language, the research implemented advanced preprocessing and machine learning strategies. This included lemmatization to normalize words to their root forms and sophisticated negation handling using "NOT_" prefixes to differentiate contextually offensive terms from constructive ones. Additionally, the models utilized custom vectorizers with n-grams ranging from 1 to 3, which facilitated the detection of nuanced, context-specific patterns rather than relying solely on isolated keywords. The research also optimized detection by fine-tuning classification thresholds between 0.4 and 0.7 and prioritizing the F1 score over simple accuracy. This approach aimed to better manage imbalanced data and reduce false positives. Lastly, a manual validation of 180 raw entries was conducted to compare model predictions against human interpretations, identifying specific areas where the system required refinement to accurately handle subtle or nuanced insults.

Subsequently, a dataset for flaming classification was developed to produce two classification models: label classification (flaming or not flaming detection) and context classification (politics, race, physical insult, general insult, gender, and positive detection). Finally, the effectiveness of the dataset was evaluated using machine learning algorithms, with performance measured through accuracy rate, precision, recall, and F1-score. The scope of this paper is limited to developing a dataset focusing on the Malay and English languages sourced from Twitter through data scraping and preprocessing techniques. This paper targets explicitly flaming classifications, which include categories such as religion, gender, politics, race, physical insults, general insults, and non-flaming, as outlined by Mohdali et al. (2019) [4]. The flaming keywords used in this research are restricted to a predefined list of 24 terms, ensuring a focused approach to this specific cyberbullying category. Furthermore, this paper is geographically limited to Malaysia, reflecting the cultural issues of the region. It does not address other forms of cyberbullying, such as harassment, outing, exclusion, or masquerading, as the primary focus is on flaming classification within the Malaysian context. This paper addresses the gap in cyberbullying detection in Malaysia, focusing on the flaming category by leveraging machine learning techniques to analyze Malay and English social media content. Three key objectives for this paper are to extract flaming-related keywords from social media platform, to develop a small flaming classification dataset, and to evaluate its effectiveness using machine learning algorithms. The findings contribute to building a robust dataset to aid future cyberbullying detection efforts, particularly in Malaysia's multilingual context, where Malay and English are predominantly used. The proposed dataset will support ongoing advancements in AI-based detection systems.

This paper is organized as follows. Section 2 introduces related works, discussing previous studies on cyberbullying detection and classification models. Section 3 describes materials & methodology, detailing the dataset collection, preprocessing techniques, and machine learning models used in this research. Section 4 presents the results and discussion, where the model performance, evaluation metrics, and key findings are analyzed. Finally, Section 5 discusses future work, focusing on potential improvements, dataset expansion, and the integration of deep learning techniques

2. RELATED WORKS

The previous studies focus on various methodologies and studies that have significantly contributed to detecting and classifying cyberbullying in textual data as summarized in Table 1. Notable works include those employing machine learning models such as BERT, Random Forest, Naive Bayes, and SVM for classification tasks. These models effectively handle diverse datasets and provide robust accuracy, precision, recall, and F1-score results.

For instance, Saeid, Kanojia, and Neri (2024) explored machine-learning approaches for decoding cyberbullying on social media, which is closely related to your research's focus on classifying flaming behavior in Malay and English texts [5]. Their study highlights the importance of selecting optimal algorithms and parameters for high detection accuracy. Similarly, Gan, Chua, Jasser, and Wong (2024) proposed a framework categorizing cyberbullying based on intentional dimensions. This aligns with this research's classification of flaming into specific contexts such as gender, Race, politics, and religion [6].

Dharani (2024) applied deep learning algorithms to analyze cyberbullying in text data, emphasizing the power of advanced neural network architectures for detecting nuanced patterns in textual content [7]. This is relevant to this research as it demonstrates the potential for deep learning techniques to enhance the performance of flaming detection models. Toktarova, Sultan, and Azhibekova (2024) implemented hybrid deep learning techniques combining convolutional and recurrent models, which could inspire integrating hybrid approaches in dataset evaluation to address complex patterns in multilingual datasets [8]. Lastly, Unnava and Parasana (2024) focused on feature-specific machine-learning approaches for flaming classification, which resonates with this research methodology of using tailored feature extraction [9].

These studies collectively underscore the importance of integrating diverse methodologies, such as feature engineering, algorithm optimization, and advanced classification techniques, to improve the accuracy and scalability of cyberbullying detection systems, as pursued in this research.

Table 1: Related Works

Authors & Year	Model / Technique / Method	Advantages	Disadvantages
Saeid, Kanojia, and Neri (2024) [5]	Machine Learning Models	Optimized for social media data	It may require domain-specific adjustments
Gan, Chua, Jasser, and Wong (2024) [6]	Intentional Dimension Framework	Categories cyberbullying behavior effectively	Limited generalization to other contexts
Dharani (2024) [7]	Deep Learning Algorithms	High accuracy in text analysis	Resource-intensive implementation
Toktarova, Sultan, and Azhibekova (2024) [8]	Hybrid Deep Learning Models (CNN+RNN)	Combines strengths of CNN and RNN	High computational cost
Unnava and Parasana (2024) [9]	Feature-Specific Machine Learning	Tailored feature engineering improves detection	Dataset-specific limitations
Al-Hashedi et al. (2023)[10]	BERT, EDM	High recall and precision in emotion-based detection	Complex implementation and resource-intensive
Wijayanti et al. (2022) [13]	SVM	Effective kernel comparison for classification	Limited to the Indonesian language dataset
Bandeh Ali Talpur & O'Sullivan (2020)[14]	Random Forest, Naive Bayes	Robust in feature-rich datasets, high F1-Score	Struggles with highly imbalanced data
Ali et al. (2017) [15]	Fuzzy Ontology, SVM	High accuracy in web content filtering	Not focused explicitly on cyber bullying

3. MATERIAL & METHODOLOGY

3.1 Data

This research utilized a dataset collected from Twitter using the TweetHarvest tool, where a curated list of 24 flaming-related keywords in Malay and English was used to scrape relevant tweets. Each keyword was limited to 300 tweets, resulting in a raw dataset of 14,400 entries (7,200 per language). After preprocessing and cleaning, the dataset was refined to 3,600 entries (1,800 per language) and stored in CSV format for further analysis. The data collection process was implemented using a Python script that integrated into the TweetHarvest tool via Node

Package Manager (npx). For example, the script set the filename (filename = f'bodoh.csv'), defined the search query (search_keyword = 'bodoh since:2017-04-01 until:2024-11-11 lang:id Malaysia'), limited the number of tweets per keyword (limit = 300), and authenticated using a Twitter API token (twitter_auth_token = '123d5d789733a357bb2a87ef6d707b06178843c5'). The command !npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token} was executed to extract tweets matching the specified criteria. This command stored the collected tweets in a CSV file (-o "{filename}"), defined the search query with keyword, language, date range, and location (-s "{search_keyword}"), retrieved the latest tweets (--tab "LATEST"), enforced the tweet limit (-l {limit}), and authenticated using the provided token (--token {twitter_auth_token}). The dataset was efficiently built by systematically executing this process for all 24 keywords, ensuring comprehensive coverage of flaming-related Twitter discussions in Malay and English.

3.2 Method

The methodology in this research is systematically organized into five (5) primary stages: data collection, preprocessing, labeling, model training, and evaluation. Using Twitter as the source, data was collected using a keyword-based search method. A curated list of 24 flaming-related keywords for Malay and English languages was utilized within a Python script to ensure relevance.

The preprocessing stage involved refining the raw data for subsequent analysis. This included removing unnecessary symbols, punctuation, and URLs, standardizing all text to lowercase for uniformity, and eliminating irrelevant stop words using the Sastrawi library for Malay text. Lemmatization was applied to normalize words to their root forms, ensuring consistency. Additionally, negations, such as "tak" and "tidak," were handled by implementing specific patterns, including the use of "NOT_" prefixes, to represent the context of negated terms accurately. Upon completion of preprocessing, the dataset was reduced to 3,600 entries, equally divided between Malay and English (1,800 entries per language), effectively removing irrelevant and noisy data.

Subsequently, each dataset was labeled into two (2) distinct categories (label classification): flaming versus not flaming, and six contextual categories (context classification) for flaming, including religion, politics, gender, physical insult, general insult, and positive (non-flaming). Each labeled data was used to train three (3) supervised learning models: Logistic Regression, Support Vector Machine (SVM) and Random Forest for both classifications. Using a 10-fold cross-validation approach, the data was transformed into numerical representations with TF-IDF vectorization. Final performance metrics, including accuracy, precision, recall, and F1-score, were computed to evaluate the effectiveness of the models.

Table 2 consists of the hyperparameter values used in this research experiment for prediction models and feature extraction. The sources indicate that the models were initially compared using standard configurations and then the best-performing models (Logistic Regression for label classification and SVM for context classification) were fine-tune. For the final system implementation, the classification threshold was adjusted from the default of 0.5 to an optimal range between 0.4 and 0.7 to maximize the F1-score, which served as the primary metric for model selection due to the imbalanced nature of the dataset.

Table 2: Hyperparameter Values for Prediction Models and Feature Extraction

Category	**Component / Model**	**Hyperparameter**	**Value**
Feature Extraction (TF-IDF)	Custom Vectorizer	Maximum Features (max_features)	2,000
		N-gram Range (ngram_range)	(1, 3)
		Minimum Document Frequency (min_df)	2
		Maximum Document Frequency (max_df)	0.95
Classification Algorithms	Logistic Regression	Maximum Iterations (max_iter)	1,000

	Random Forest	Number of Estimators (n_estimators)	100
	SVM	Kernel	'linear'
		Probability	True
Experimental Setup	K-Fold Cross-Validation	Number of Folds (k)	10
		Shuffle	True
		Random State	100
	Classification Threshold	Fine-tuned Range	0.4 to 0.7

Due to limited and imbalanced datasets, the best models were determined based on the F1 score. After comparing the models, the best-performing model was selected and trained again with fine-tuned parameters to optimize its classification performance. The fine-tuning process utilized custom vectorizers for both languages, specifically configured with optimized parameters such as a maximum feature limit of 2,000 and an n-gram range of 1 to 3, instead of relying on default settings. Additionally, the threshold for classification was fine-tuned beyond default values to enhance the model's accuracy and F1-score. Predictions were aggregated using majority voting across folds, ensuring robust and reliable outcomes.

For testing and validation, pipelines were implemented to process Malay and English inputs, supporting various input types such as single text inputs and CSV files. This research used 180 new raw text entries in a CSV file to test and validate the models for both languages. Predictions were aggregated across 10 folds using a majority voting mechanism to ensure consistent and reliable results. Label and context classifications were displayed as the developed models' detection results. The best-performing models were integrated into the Cyberbullying in Text Detection (CyTED) system, a web platform designed to test the models and validate flaming detection in real-world applications. This structured and rigorous methodology ensured the flaming classification system's reliability, scalability, and adaptability, specifically tailored for a bilingual dataset.

Table 3 summarizes part of the list of flaming keywords for data collection. While Figure 1 represents the flowchart for the CyTED dataset development and Figure 2 and Figure 3 represent the training processes involved in this research. Figure 4 and Figure 5 represent the validation processes.

Table 3: List Of Flaming Keywords for Data Collection

Malay	English
1. Anjai/Njir	1. Bastard
2. Babi	2. Big Tits
3. Bangsat	3. Bimbos
4. Barua	4. Bitch
5. Bongok	5. Bull shit
6. Bodo/Bodoh	6. Busty
7. Buto	7. Damn
8. Cibai	8. Dumb
9. Hancing	9. Fake
10. Hanat/Anat	10. Foolish
11. Haprak	11. Fuck
12. Hawau	12. Go die
13. Jahanam/Jahannam	13. Hell
14. Kepam	14. Holy shit

15. Kepala Bana/Pala Bana 16. Kimak 17. Lancau 18. Mak kau hijau/Mak kau ijau 19. Muka Hauk 20. Peghak 21. Puki/Pukimak 22. Sial 23. Walanon 24. Walaun	15. Idiot 16. Loser 17. Motherfucker 18. Selfish 19. Shame on you 20. Shit 21. Stupid 22. Trash 23. Useless 24. Weak

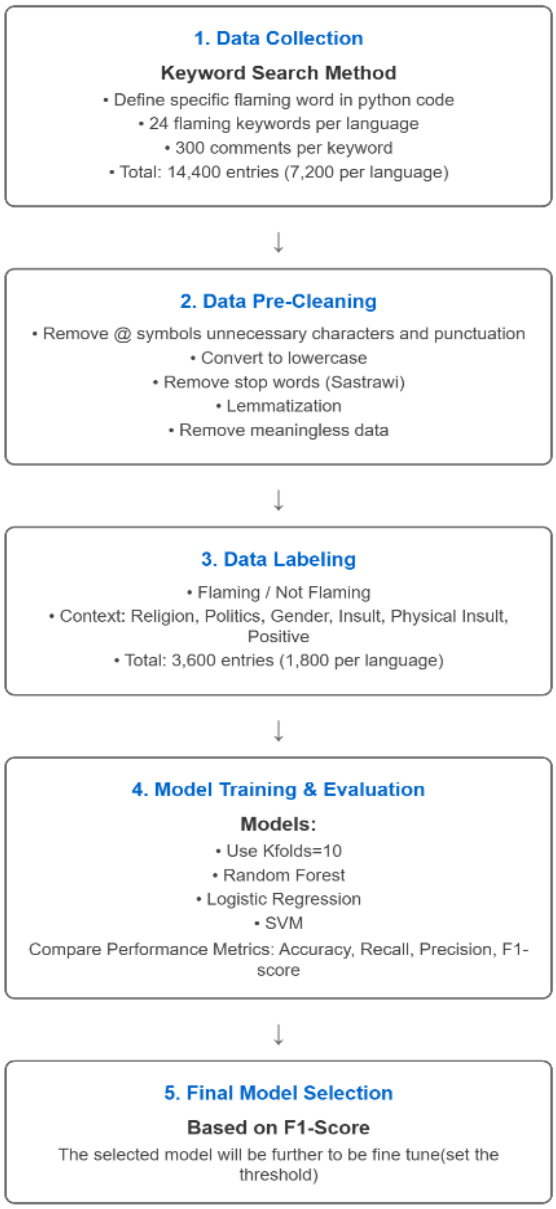


Fig. 1: Workflow for Cyberbullying Dataset Development

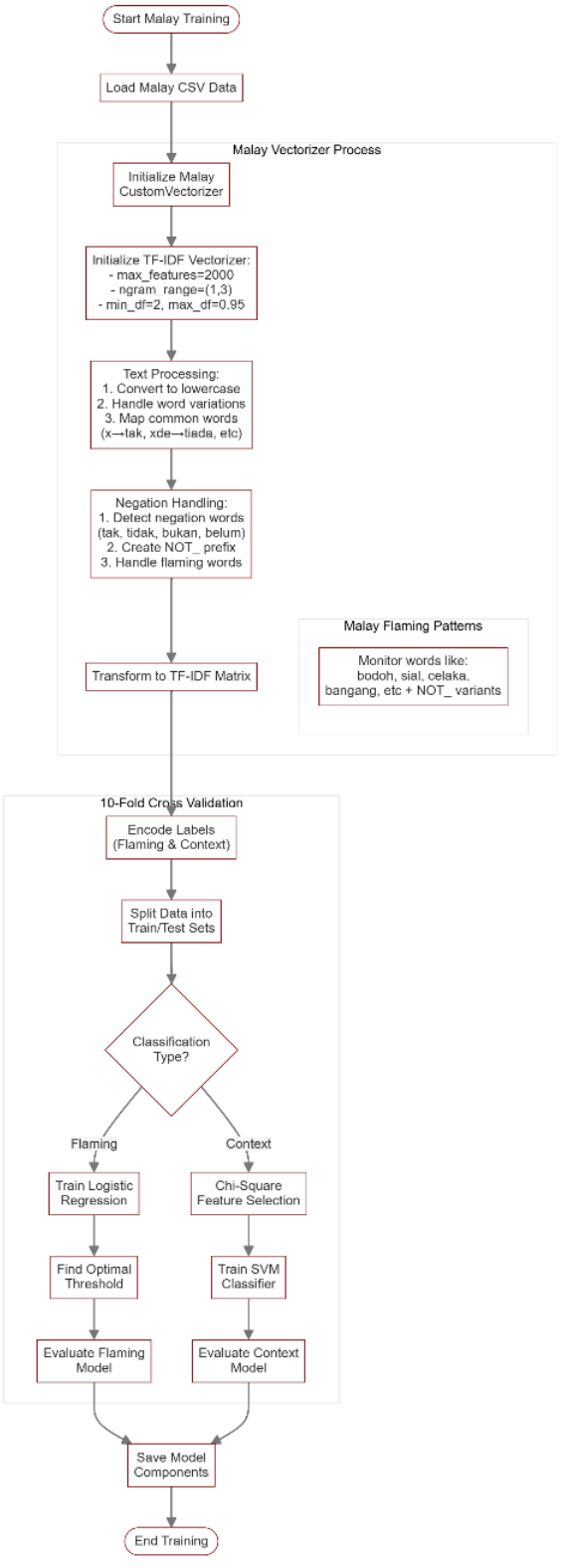


Fig. 2: Flowchart of the Training Process for Selected Model (Malay language)

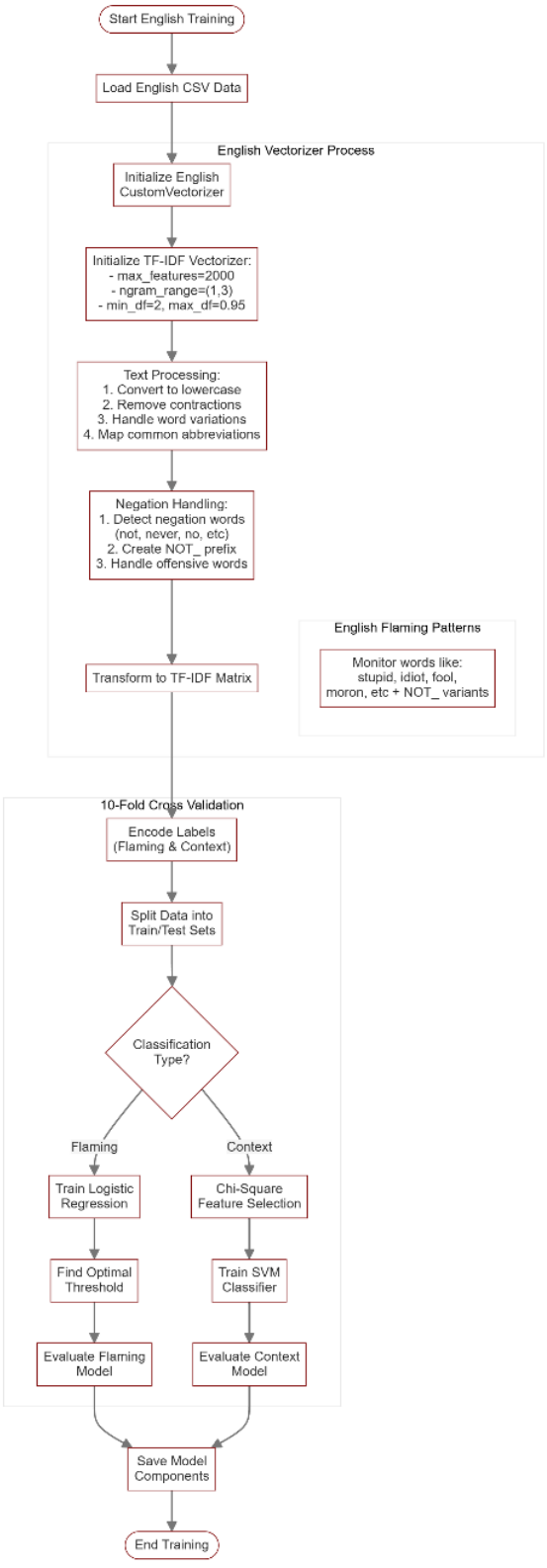


Fig. 3: Flowchart of the Training Process for Selected Model (English language)

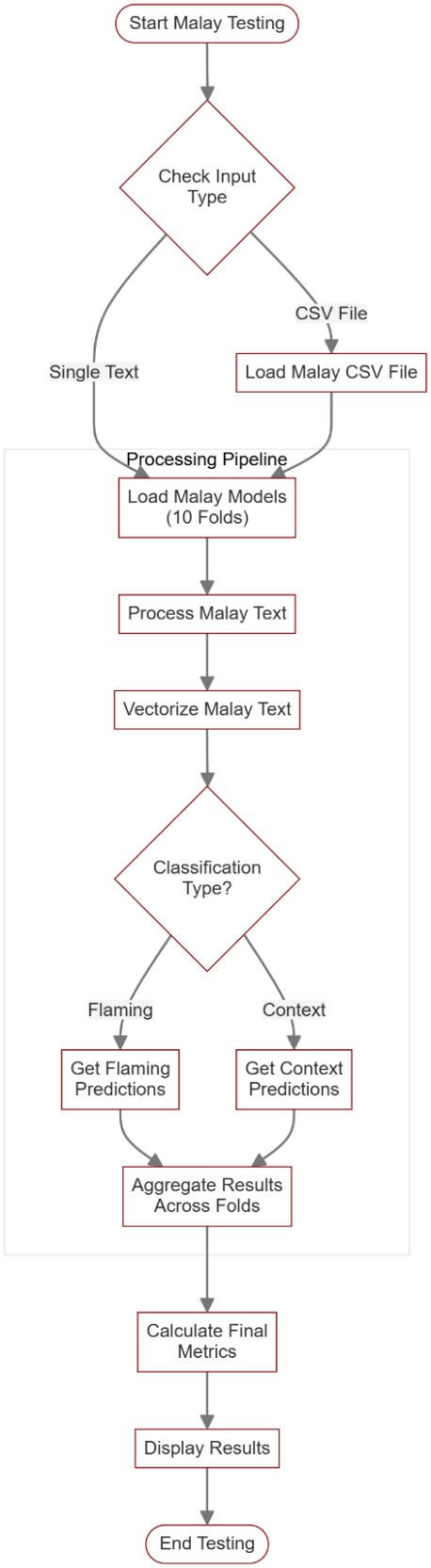


Fig. 4: Validation Process (Malay Language)

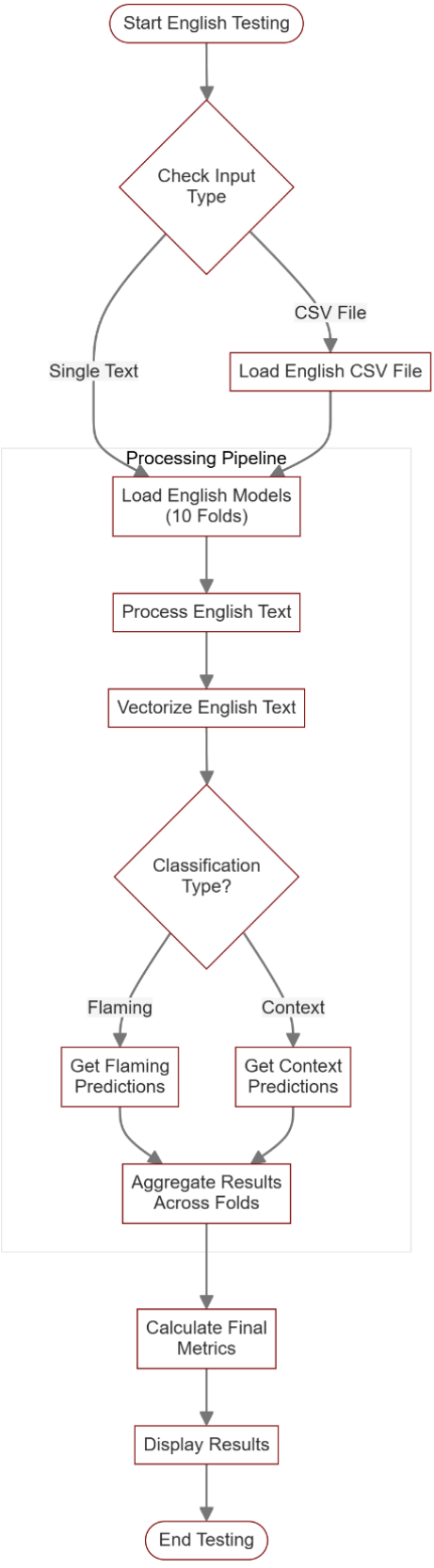


Fig. 5: Validation Process (English Language)

3.3 Equations and Mathematical Expressions

Term Frequency-Inverse Document Frequency (TF-IDF) is used as the keyword extraction model in this research. TF-IDF was a quantitative method to assess the significance of words in documents or a collection of documents called corpus [10]. TF-IDF comprises two main components: Term frequency (TF) and inverse document frequency (IDF).

Definition 1 (See [11]). *TF measured how frequently a term (word) appeared in a document. The assumption was that the more a term appeared in a document, the more important it might be. Eq. 1 is the formula of TF:*

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad (1)$$

Definition 2 (See [11]). *IDF quantified the significance of a term in the context of the entire corpus. If a term was present in numerous documents, distinguishing between them might not have been beneficial. The IDF value would be low if the term was prevalent and appeared in many documents. Conversely, the IDF value would be high if the term was rare. Eq.2 is the formula of IDF:*

$$IDF(t,D) = log\ (\frac{Total\ number\ of\ documents\ in\ corpus\ D}{Number\ of\ documents\ containing\ term\ t}) \quad (2)$$

Definition 3 (See [11]). *TF-IDF was calculated by multiplying a term's Term Frequency (TF) by its Inverse Document Frequency (IDF)(refer Eq. 3).*

$$TF - IDF(t,d,D) = TF(t,d)\ \times\ IDF(\mathrm{t,D}) \quad (3)$$

The model's performance metrics, such as accuracy, precision, recall and F1-score was assessed.

Definition 4 (See [11]). *Accuracy was a metric that quantifies the ratio of accurate predictions made by a classification model to the total number of predictions. The calculation divided the sum of true positive and true negative predictions by the total number of predictions generated by the model.*

$$Accuracy = (\frac{true\ positives\ +\ true\ negatives}{total\ prediction}) \quad (4)$$

Definition 5 (See [11]). *Precision was the ratio of accurately predicted positive cases to the model's total number of positive predictions.*

$$Precision = true\ positive/(true\ positives + false\ positives) \quad (5)$$

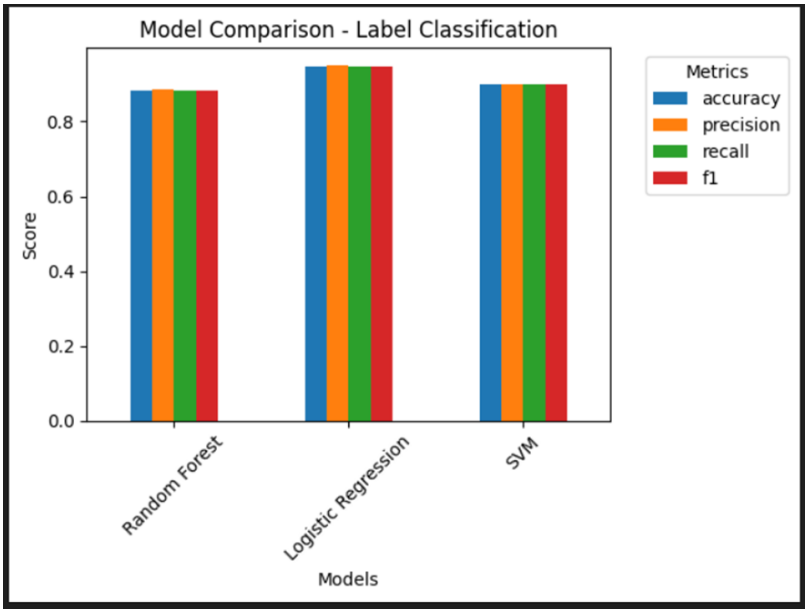
Definition 6 (See [11]). *Recall was a metric that quantifies the accuracy of positive predictions by dividing the number of true positive predictions by the total number of actual positive instances in the test data.*

$$Recall = true\ positive/(true\ positives + false\ negatives) \quad (6)$$

Definition 7 (See [11]). *The F1-score was calculated as the harmonic mean of precision and recall, offering a consolidated value that considers both metrics.*

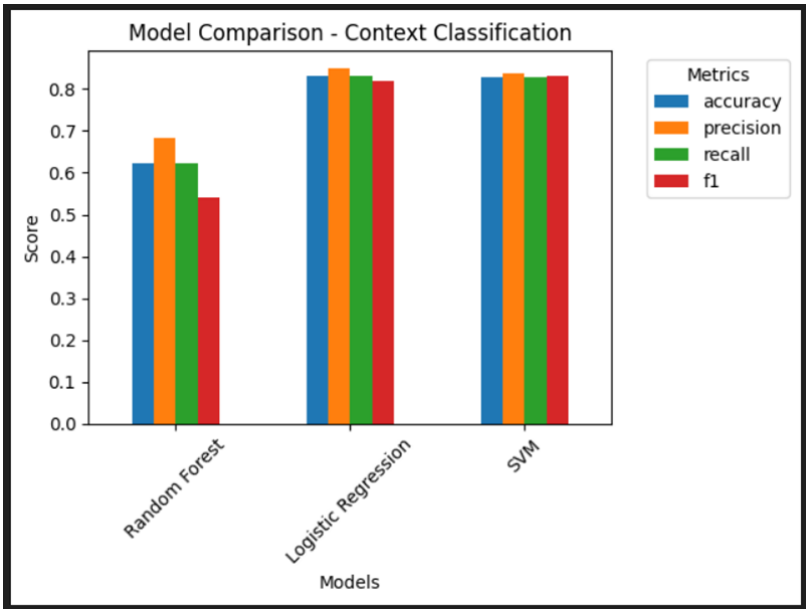
$$F1 - score = 2 * precision * recall/(precision + recall) \quad (7)$$

The results of performance metrics for Definition [4] – [7] are given in Figure 6 - Figure 9 as follow.



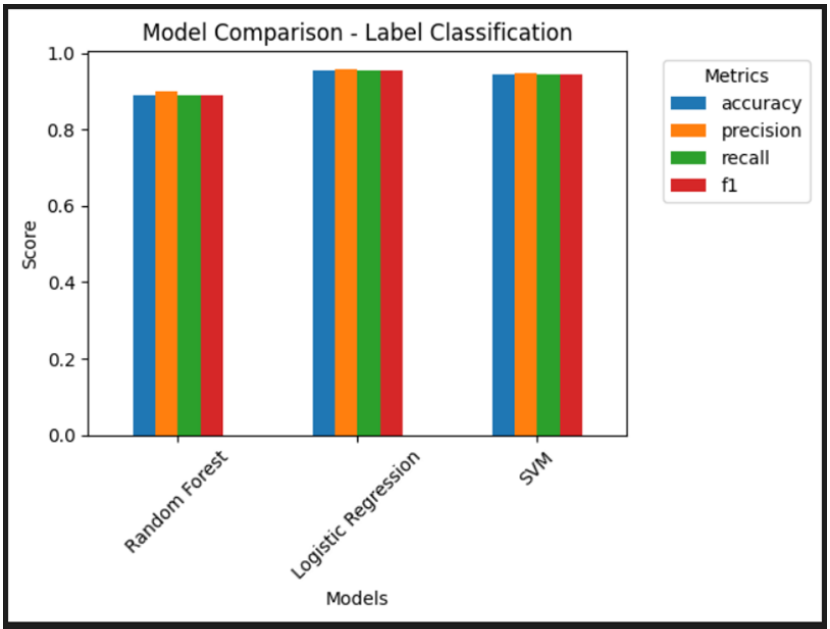

```
Detailed results for Label Classification:
                    accuracy  precision  recall      f1
Random Forest         0.8839     0.8875  0.8839  0.8839
Logistic Regression   0.9483     0.9494  0.9483  0.9484
SVM                   0.8989     0.9004  0.8989  0.8989
```

Fig. 6: Computational Results of Performance Metrics for Label Classification in Malay




```
Detailed results for Context Classification:
                    accuracy  precision  recall      f1
Random Forest         0.6217     0.6829  0.6217  0.5395
Logistic Regression   0.8317     0.8484  0.8317  0.8197
SVM                   0.8294     0.8371  0.8294  0.8303
```

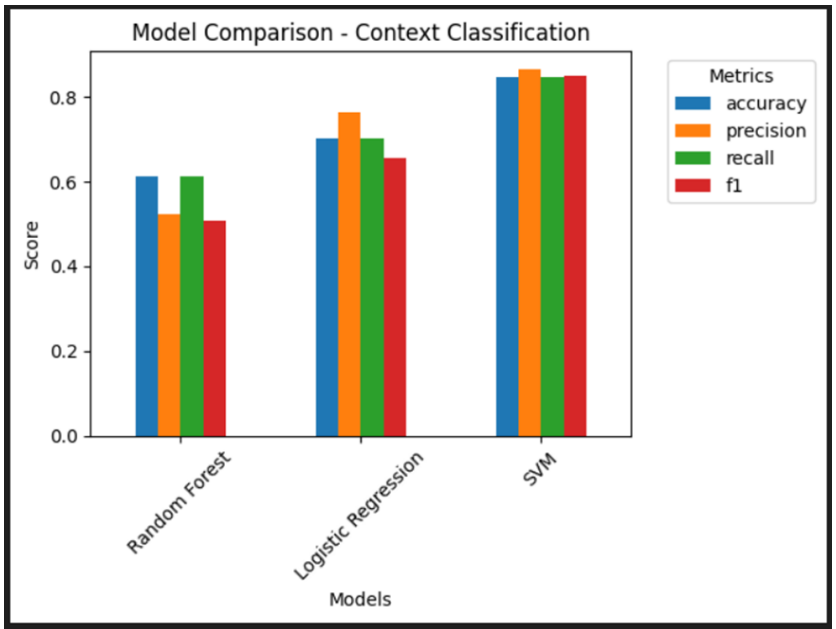
Fig. 7: Computational Results of Performance Metrics for Context Classification in Malay



Detailed results for Label Classification:

	accuracy	precision	recall	f1
Random Forest	0.8883	0.9007	0.8883	0.8883
Logistic Regression	0.9550	0.9562	0.9550	0.9549
SVM	0.9433	0.9460	0.9433	0.9434

Fig. 8: Computational Results of Performance Metrics for Label Classification in English



Detailed results for Context Classification:

	accuracy	precision	recall	f1
Random Forest	0.6117	0.5236	0.6117	0.5078
Logistic Regression	0.7033	0.7624	0.7033	0.6564
SVM	0.8467	0.8642	0.8467	0.8484

Fig. 9: Computational Results of Performance Metrics for Context Classification in English

Figure 6 until Figure 9 present the computational results of performance metrics on accuracy, precision, recall, and F1-score. These metrics are used to evaluate the classification of flaming behaviour across the bilingual Malay and English datasets. Figures 6 and 8 focus on label classification to distinguish flaming from non-flaming content, where Logistic Regression (LR) proved most effective, yielding high accuracies of 94–95% for Malay and 95–96% for English. While Figures 7 and 9 evaluate context classification across seven categories (religion, politics, race, gender, physical insults, general insults, and positive), with Support Vector Machine (SVM) identified as the optimal model due to its superior F1-scores of 83% for Malay and 85% for English. The researchers prioritized the F1-score over simple accuracy to account for the imbalanced nature of the dataset, which contains six sub-categories for flaming but only one for positive content, thereby ensuring a more reliable detection system that balances precision and recall while avoiding false alarms in complex linguistic scenarios.

4. RESULTS AND DISCUSSION

4.1 Result

Table 4: Summary Of the Best Models Based on Accuracy

Accuracy	Malay	English
Best Model of Label Classification	LR(95%)	LR(96%)
Best Model of Context Classification	LR(83%)	SVM(0.85%)

Table 5: Summary of The Best Models based on F1-Score

F1-Score	Malay	English
Used Model of Label Classification	LR(95%)	LR(95%)
Used Model of Context Classification	SVM(83%)	SVM(85%)

Table 6: Evaluation Results for Flaming Detection and Context Classification on Malay Texts

Category	Model Prediction	Actual Data (Human Understanding)	Difference (Absolute)	Difference (Percentage)
Label Classification				
Flaming	89	103	-14	-7.78%
Not Flaming	91	77	14	7.78%
Context Classification				
Positive	109	105	4	2.23%
General Insult	54	68	-14	-7.78%
Physical Insult	0	0	0	0
Race	9	3	6	3.33%
Religion	5	2	3	1.67%

Politics	1	1	0	0
Gender	2	1	1	0.5%
Total text analyzed	180	180	0	0

Table 7: Evaluation Results for Flaming Detection and Context Classification on English Texts

Category	Model Prediction	Actual Data (Human Understanding)	Difference (Absolute)	Difference (Percentage)
Label Classification				
Flaming	91	97	-6	-3.33%
Not Flaming	89	83	6	3.33%
Context Classification				
Positive	107	105	4	1.11%
General Insult	60	68	-14	-4.45%
Physical Insult	2	1	1	0.55%
Race	7	5	2	1.11%
Religion	4	3	1	0.55%
Politics	1	1	0	0
Gender	3	2	1	0.56%
Total text Analyzed	180	180	0	0

4.2 Findings

Table 4 summarizes the best-performing models for label and context classification across the Malay and English datasets. Logistic Regression (LR) achieved the highest performance for label classification for both languages. In context classification, LR was the best-performing model for Malay, while Support Vector Machine (SVM) outperformed other models for English. While, Table 5 presents the best models based on F1 score, where LR was selected for label classification in both languages, and SVM was chosen for context classification due to its higher F1 score compared to LR, despite LR achieving a higher accuracy.

The evaluation of 180 Malay text entries assessed the flaming detection and context classification models, as shown in Table 6. In flaming detection, the model correctly identified 89 out of 103 non-flaming entries (86.41%) and 91 out of 77 flaming entries, showing a +7.78 percentage point overprediction. For context classification, it performed well in the positive category (96.33%) but underperformed in insult detection (-7.78 percentage points). Race and religion showed slight overpredictions (+3.33 and +1.67 percentage points), while Politics remained accurate, and physical insult was not detected. The model demonstrated moderate accuracy but required improved handling of nuanced insults and reducing overpredictions.

The evaluation of 180 English text entries assessed the flaming detection and context classification models, as shown in Table 7. In flaming detection, the model correctly identified 89 out of 83 flaming instances but showed

a +3.33 percentage point overprediction, indicating a tendency to classify more texts as flaming than validated by human understanding. It also accurately classified 91 out of 97 non-flaming instances (93.81%), with a -3.33 percentage point difference, suggesting a slight underestimation of non-flaming content. For context classification, the model performed well in the positive category (98.13%), but showed underperformance in insult detection (-4.45 percentage points). Race and religion had slight overpredictions (+1.11 and +0.55 percentage points, respectively), while Politics, Gender, and Physical Insult had minimal discrepancies. The model demonstrated strong performance but required refinement in handling false positives and improving nuanced classification.

4.3 Explanatory Text

The selection of models for the CyTED website was based on the F1 score, as it provides a balanced evaluation in imbalanced datasets where one class is underrepresented. Since flaming classification involved six categories while non-flaming had only one, accuracy alone was not a reliable metric. While precision and recall measure correctness and detection ability, optimizing one often compromises the other. The F1 score mitigates this trade-off, ensuring consistent performance. Given the complexity of context classification, where overlapping language patterns increase misclassification risks, prioritizing the F1 score ensures a more reliable detection system for flaming comments.

The testing and validation of the model measured performance using true positive (TP), true negative (TN), and false positive (FP) cases. True positives referred to correctly detected flaming text, while true negatives indicated accurate non-flaming classifications. False positives occurred when non-flaming text was misclassified as flaming, contributing to overprediction. The F1 score balanced precision and recall, particularly in context classification, where overlapping language patterns increased misclassification risks. The findings confirmed the model's effectiveness in detecting flaming content, highlighting areas for refinement in detecting insults and minimizing overpredictions to improve accuracy and reduce false positives.

4.4 Discussion

Table 4 summarizes the best-performing models for each classification task (Label and Context) across the Malay and English datasets. Logistic regression (LR) performs best for label classification in both languages. In context classification, LR was the best model for Malay, and SVM led for English. Based on this initial analysis, the best-performing models would be integrated into the CyTED (Cyberbullying in Text Detection) website, allowing users to test the models with their own text inputs. However, in selecting the optimal model for integration into the CyTED website, the F1 score was prioritized over other metrics such as accuracy, precision, and recall. The F1 score, being the harmonic mean of precision and recall, provided a balanced measure of a model's performance, especially in cases of imbalanced datasets where one class may be underrepresented. The dataset was described as imbalanced in terms of the flaming context itself as flaming had 6 categories while not flaming had only 1 category.

While accuracy indicated the overall correctness of the model, it can be misleading in scenarios with class imbalance. For instance, in a dataset where non-flaming comments vastly outnumbered flaming ones, a model predicting all comments as non-flaming would achieve high accuracy but failed to identify actual flaming comments. Precision and recall individually focused on the correctness of positive predictions and the ability to identify all positive instances, respectively.However, optimizing one often led to a compromise in the other. The F1 score addressed this trade-off by considering both metrics simultaneously, ensuring a more comprehensive evaluation of the model's effectiveness (W&B, 2024) [12]. Consequently, the model selected for deployment on the CyTED website might differ from those identified as the best model based solely on accuracy or other metrics. By emphasizing the F1 score, the chosen model ensures a balanced performance, effectively identifying both flaming and non-flaming comments, thereby enhancing the reliability and user experience of the CyTED platform.

Table 5 depicted the best models based on F1 score.For label classification, LR was chosen to be used in the CyTED web system for both languages. For context classification, SVM was selected as the best model due to its higher F1 score compared to LR even LR's accuracy was higher. A higher F1 score indicated that SVM achieved a better balance between precision and recall, which was crucial for accurately identifying various context categories within flaming comments.Having many context categories made classification more challenging because each context had unique language patterns and tone. This variety meant the model must learn subtle differences between, for example, a comment on Race versus one on politics, even if both have similar words or phrases. With many contexts, there was a higher chance the model might confuse one for another, leading to

misclassifications. For example, words used in political discussions might overlap with those in other contexts, making it harder for the model to decide the correct label. The F1 score became even more valuable here, as it ensured the model's balance in identifying the true instances of each context while avoiding too many false alarms.

The testing and validation of Malay models were conducted on a random subset of 180 Malay text entries to assess the effectiveness of the flaming detection and context classification models for the Malay language. In flaming detection, the model correctly classified 89 out of 103 entries (86.41%) as not flaming, showing a slight underperformance of -7.78 percentage points compared to human evaluation. This represented true negative (TN) cases, where the model successfully identified non-flaming text. The model correctly identified 91 out of 77 entries for flaming text, with an overprediction of +7.78 percentage points, representing true positive (TP) cases where the model accurately detected flaming text. However, the overprediction suggested that the model tended to classify more entries as flaming than had been validated by human evaluation. A false positive (FP) occurred when the model incorrectly classified a non-flaming comment as flaming, contributing to the overprediction rate. Overall, the flaming detection showed moderate accuracy but required adjustments to improve the balance between identifying flaming and non-flaming text, particularly considering the nuanced expressions in Malay.

In context classification, the model performed well in the positive category, correctly identifying 109 entries compared to 105 actual ones (96.33%), resulting in a +2.23 percentage point difference. For insult detection, the model identified 54 out of 68 entries (79.41%), but exhibited a -7.78 percentage point difference, highlighting challenges in accurately detecting insults in Malay texts. The model slightly overpredicted categories like Race and Religion, with +3.33 percentage points and +1.67 percentage points, respectively, indicating that it was sensitive to text patterns commonly associated with these contexts. Minor discrepancies were observed for gender (+0.55 percentage points), while Politics remained consistent with no variation, and physical insult was not detected in the analyzed subset.

This random sample evaluation of Malay text demonstrated the model's reasonable effectiveness in detecting flaming and classifying contexts, particularly for positive texts. However, it revealed areas for improvement, such as handling nuanced insults and reducing overpredictions in categories like Race and Religion. The analysis of true positive (TP), true negative (TN), and false positive (FP) cases highlighted how the model had performed correctly identifying flaming and non-flaming text while also exposing limitations in false classifications. These findings emphasized the model's potential for application in Malay, pointing to specific areas where refinements were needed to enhance its accuracy and reliability.

The testing and model validation were also conducted on a random subset of 180 English text entries to assess the effectiveness of the flaming detection and context classification models, with a particular emphasis on flaming detection. For flaming text, the model successfully identified 89 out of 83 instances but showed a +3.33 percentage point overprediction. This represented true positive (TP) cases, where the model correctly detected flaming text, but also indicated a tendency to classify slightly more texts as flaming than verified by human understanding. The flaming detection model demonstrated strong overall performance, correctly identifying most entries in both categories. However, the slight overprediction suggested that the model was sensitive to keywords or phrases that might not always indicate flaming in their specific context. The model also correctly classified 91 out of 97 instances (93.81%) of non-flaming text, with a -3.33 percentage point difference compared to human evaluation. This represented true negative (TN) cases, where the model accurately identified non-flaming content. However, the slight underestimation of non-flaming texts indicated some misclassification. A false positive (FP) occurred when the model incorrectly classified a non-flaming comment as flaming, contributing to the observed overprediction.

In context classification, the model accurately identified positive texts, correctly classifying 107 out of 105 instances (98.13%), with a +1.11 percentage point difference. For insults, the model identified 60 out of 68 instances (88.24%), with a -4.45 percentage point difference, suggesting difficulty in capturing subtle nuances of insults. Race and Religion categories showed slight overpredictions, with Race (7 vs. 5 instances, +1.11 percentage points) and religion (4 vs. 3 instances, +0.55 percentage points). Other categories, such as Politics, Gender, and Physical Insult, had minimal discrepancies, with only small overpredictions.

The flaming detection model proved to be highly effective, correctly classifying over 85% of the entries in both flaming and non-flaming categories, showcasing its reliability in identifying harmful content. However, the slight overprediction of flaming cases highlighted areas for improvement, particularly in distinguishing false positives from true flaming incidents. The identification of true positive (TP), true negative (TN), and false positive (FP) cases provided insights into the model's strengths and weaknesses. True positives referred to correctly identified

flaming text, while true negatives indicated accurate classification of non-flaming content. However, false positives, where non-flaming comments were misclassified as flaming, revealed the model's sensitivity to certain terms. This evaluation confirmed the model's capability in handling English texts while identifying areas for further refinements to enhance its accuracy and robustness in real-world scenarios.

Given the bilingual nature of the dataset, this research successfully addresses both Malay and English. This research addresses the distinct cultural and linguistic nuances of its bilingual dataset by using language-specific curated keywords and tailored preprocessing pipelines for both Malay and English. For the Malay dataset, this research incorporates regional slang and culturally specific insults. It also employs the Sastrawi library to effectively manage stop-word removal and handle regional negations. For the English dataset, nuances are processed using the Natural Language Toolkit (NLTK), which provides specialized tokenization, stop-word removal, and lemmatization. To capture subtle variations and indirect language, the models implement negation handling with "NOT_" prefixes and utilize custom TF-IDF vectorizers with an n-gram range of 1 to 3. This approach allows the system to learn multi-word, context-specific patterns rather than relying solely on isolated terms. By adopting this structured methodology, the research ensures that each language is processed with sensitivity to its unique patterns. It utilizes independent pipelines and optimized algorithms, such as Logistic Regression and Support Vector Machines (SVM), to maintain robust classification accuracy across a diverse multilingual social media landscape.

The research emphasizes the importance of protecting individual well-being by developing the CyTED system, which aims to reduce the severe psychological and societal impacts of flaming behavior on social media platforms like Twitter. To address ethical concerns regarding data acquisition, this research used secure, authenticated access through the Twitter API and the TweetHarvest tool. This approach ensured compliance with the platform's technical requirements and avoided any account restrictions. The methodology used in this research focuses on identifying hostile language patterns in highly sensitive and controversial contexts, which is intended to maintain the integrity of the platform rather than targeting specific user identities.

5. CONCLUSION

This research has successfully developed a flaming classification dataset with 3,600 entries (1,800 per language) by extracting relevant cyberbullying keywords from Twitter using the Tweet-Harvest tool. The dataset underwent extensive cleaning and preprocessing, including tokenization, stop-word removal, and text normalization, ensuring its quality for machine learning.

The dataset was trained and tested using Logistic Regression, Random Forest, and SVM with 10-fold cross-validation. The best models were selected based on the F1-score to address dataset imbalance. Logistic Regression achieved 95% accuracy for Malay and 96% for English in label classification, while SVM performed best for context classification with 83% F1-score for Malay and 85% for English. The dataset was limited in size and imbalanced, affecting model generalizability. The model's sensitivity to specific keywords led to occasional false positives, where non-flaming content was misclassified as flaming. Additionally, contextual misclassification occurred due to the nuanced nature of offensive language, particularly across different cultural and linguistic contexts. The research demonstrated the effectiveness of machine learning in detecting cyberbullying, with the best models integrated into the CyTED system for real-world testing. By prioritizing the F1 score over accuracy, the model ensured a balanced and reliable evaluation metric, making it suitable for automated moderation systems in social media platforms, online communities, and digital safety tools.

Future research should expand dataset diversity by incorporating more languages, dialects, and text variations. Increasing keyword coverage and regularly updating offensive terms will enhance flaming detection. Improving annotation quality through multi-layer validation can ensure higher accuracy. Additionally, deep learning models (e.g., BERT) should be explored to improve contextual understanding and reduce false classifications. Furthermore, future research in flaming detection should prioritise expanding dataset diversity to include multilingual, dialectal, and low-resource language variations, while continuously updating keyword coverage to reflect evolving offensive expressions across social media contexts. In parallel, improving annotation quality through data-centric approaches, multi-layer validation, and statistical quality estimation can significantly enhance dataset reliability and model performance. Furthermore, the integration of advanced deep learning architectures, particularly transformer-based models such as BERT, has demonstrated strong capability in capturing contextual nuances and reducing false classifications, especially when combined with hybrid or ensemble techniques. Collectively, these directions highlight a shift toward more adaptive, inclusive, and context-

aware detection frameworks that align with the complexity of modern cyber-enabled communication environments [16-23].

REFERENCES

[1] Shahira, N. (2023, November 27). 1,147 kandungan buli siber diturunkan di media sosial – KKD. Buletin TV3. https://www.buletintv3.my/nasional/1147-kandungan-buli-siber-diturunkan-di-media-sosial-kkd/

[2] Garnsey, M. R., & Fisher, I. E. (2008). Appearance of New Terms in Accounting Language: A Preliminary Examination of Accounting Pronouncements and Financial Statements. Journal of Emerging Technologies in Accounting, 5(1), 17–36. https://doi.org/10.2308/jeta.2008.5.1.17

[3] Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Zainal, Z. (2019). Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 10(1). https://doi.org/10.14569/ijacsa.2019.0100146

[4] Mohdali, R., Zakaria, W. N. W., Ali, N. a. M., & Salam, Z. A. (2019). Qualitative investigation of sensitive topics in tax compliance study in Malaysia. International Journal of Academic Research in Business and Social Sciences, 9(3). https://doi.org/10.6007/ijarbss/v9-i3/5798

[5] Saeid, A., Kanojia, D. and Neri, F., 2024, June. Decoding Cyberbullying on Social Media: A Machine Learning Exploration. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 425-428). IEEE.

[6] Gan, M.F., Chua, H.N., Jasser, M.B. and Wong, R.T., 2024, June. Categorization of Cyberbullying based on Intentional Dimension. In 2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) (pp. 285-290). IEEE

[7] Dharani, M., 2024. An Analysis of Cyberbullying in Text Data using Deep Learning Algorithms. Communications on Applied Nonlinear Analysis, 31(3s), pp.61-73.

[8] Toktarova, A., Sultan, D. and Azhibekova, Z., 2024, May. Review of Machine Learning Models in Cyberbullying Detection Problem. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 233-238). IEEE.

[9] Unnava, S. and Parasana, S.R., 2024. A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. Engineering, Technology & Applied Science Research, 14(4), pp.15607-15613

[10] Al-Hashedi, M., Soon, L., Goh, H., Lim, A. H. L., & Siew, E. (2023). Cyberbullying detection based on emotion. IEEE Access, 11, 53907–53918. https://doi.org/10.1109/access.2023.3280556

[11] Hrushikesh T, & Kavya Sree Koneti. (2023). Sentiment analysis of cyberbullying. Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. Retrieved from https://www.diva-portal.org/smash/get/diva2:1779496/FULLTEXT02

[12] W&B. (2024, November 16). Weights & Biases. W&B. https://wandb.ai/mostafaibrahim17/ml-articles/reports/An-Introduction-to-the-F1-Score-in-Machine-Learning--Vmlldzo2OTY0Mzg1

[13] Wijayanti, S. S., Utami, E., & Yaqin, A. (2022). Comparison of Kernels on Support Vector Machine (SVM) Methods for Analysis of Cyberbullying. In 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). https://doi.org/10.1109/icitisee57756.2022.10057761

[14] Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. PLoS ONE, 15(10), e0240924. https://doi.org/10.1371/journal.pone.0240924

[15] Ali, F., Khan, P., Riaz, K., Kwak, D., Abuhmed, T., Park, D., & Kwak, K. S. (2017). A fuzzy ontology and SVM–Based web content classification system. IEEE Access, 5, 25781–25797. https://doi.org/10.1109/access.2017.2768564

[16] Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*.

[17] Gupta, A., Cheung, J., Meng, P., & O'Brien, S. (2025). EnDive: A cross-dialect benchmark for fairness and performance in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

[18] Mridha, M. F., Wadud, M. A. H., Hamid, M. A., & Alamri, A. (2021). L-Boost: Identifying offensive texts from social media posts in Bengali. *IEEE Access, 9*, 149801–149811. https://doi.org/10.1109/ACCESS.2021.3123456

[19] Alshahrani, E. S., Aksoy, M. S., & Emam, A. (2025). Detection of hate speech and offensive language in Arabic text: A systematic literature review. *Applied Computational Intelligence and Soft Computing, 2025*, Article 8891234. https://doi.org/10.1155/2025/8891234

[20] Nou, S., Lee, J.-S., Ohyama, N., & Obi, T. (2024). The improvement of ground truth annotation in public datasets for human detection. *Machine Vision and Applications, 35*, 45. https://doi.org/10.1007/s00138-024-01456-7

[21] Klie, J.-C., Haladjian, J., Kirchner, M., & Nair, R. (2024). On efficient and statistical quality estimation for data annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

[22] Azroumahli, C., Elyounoussi, Y., & Badir, H. (2024). BERT for Arabic NLP applications: Pretraining and finetuning MSA and Arabic dialects. *Communications in Computer and Information Science, 1910*, 123–135. https://doi.org/10.1007/978-3-031-56789-0_10

[23] Dhiman, D., Asha, V., Devi, M., & Gurav, U. (2025). Hybrid sentiment analysis model combining BERT and gradient boosting for enhanced text classification. In *Proceedings of the 3rd International Conference on Data Science and Information System (ICDSIS 2025)*.