Isabela State University Linker:
Journal of Education, Social Sciences, and Allied Health
Volume 2, Issue 1
ISSN (Print): 3082-3692 ISSN (Online): 3082-3706
DOI: https://doi.org/10.65141/jessah.v2i1.n8

Profanity and Sentiment Detection in Filipino Social Media Comments Using Transformer-Based NLP Models

Marc P. Laureta[1], Wendell Alfred Y. Feria[2], Patrick Carl D. Limbag[3], Bienmarc D. Montecillo[4]
College of Informatics and Computing Studies, New Era University, Quezon City, 1107, Philippines [1,2,3,4]
mplaureta@neu.edu.ph; wendellalfred.feria@neu.edu.ph; patrickcarl.limbag@neu.edu.ph; bienmarc.montecillo@neu.edu.ph

RESEARCH ARTICLE INFORMATION

Received: April 09, 2025
Reviewed: May 05, 2025
Accepted: June 17, 2025
Published: June 30, 2025

ABSTRACT

Filipino is considered a low-resource language, which makes it challenging to process due to the limited availability of annotated datasets and linguistic tools. These challenges are further complicated by code switching, regional variations, and the evolving nature of slang in online conversations. To address these issues, the study used a developmental research design and applied three transformer-based models: BERT, DistilBERT, and XLNet. A total of 13,565 Reddit comments were collected using web scraping techniques and the Reddit PRAW API. The dataset underwent preprocessing, including annotation, cleaning, and augmentation. The models were trained and evaluated on their ability to classify profanity into four categories: Non-Profane, Mild, Moderate, and High. Among the models, BERT achieved the highest accuracy of 99.53%, followed by XLNet and DistilBERT. A web application and a Reddit bot were created to demonstrate real-time detection, filtering, and severity-based masking of profane content. Sentiment analysis was also performed to assess the emotional tone and intent behind each comment. The results highlight the system's effectiveness in improving online content moderation through accurate and context-aware detection of profanity and sentiment in Filipino social media posts, and further suggest that handling profanity detection and sentiment analysis as

separate but complementary tasks shows better performance and interpretability.

Keywords: *Profanity detection, NLP, transformer models, Filipino language, sentiment analysis, social media moderation*

Introduction

Profanity, often referred to as swearing or cursing, encompasses the use of offensive, obscene, or taboo language to convey intense emotions such as anger, frustration, or contempt (Roache, 2023). Globally, the spread of profanity has become increasingly concerning in digital spaces, particularly on social media platforms, where it contributes to the proliferation of hate speech, cyberbullying, and toxic discourse (Malmasi & Zampieri, 2019). This trend not only affects users' mental well-being but also poses challenges to online safety, especially for vulnerable populations such as children and adolescents.

In the Southeast Asian context, and particularly in the Philippines, the issue of profanity is further complicated by sociolinguistic diversity. Filipino is a low-resource language characterized by regional linguistic variation and frequent code-switching with English. This type of language presents unique challenges for automatic profanity detection (Raza et al., 2023). Most existing profanity detection models are developed for high-resource languages and rely heavily on lexicon-based approaches. However, research has shown that keyword-based detection methods are inadequate, often failing to identify slang, euphemisms, neologisms, and context-dependent uses of language (Yi et al., 2021). Moreover, these models can misclassify harmless expressions or overlook culturally sensitive offensive terms, leading to both false positives and negatives (Vidgen et al., 2020).

This study addresses the gap in current research by focusing on the detection of profanity in the Filipino language using context-aware natural language processing (NLP) models. Specifically, it aimed to build a machine learning-based system that not only detects profane content but also categorizes its severity (mild, moderate, high) and evaluates the emotional context through sentiment analysis. The system was trained on a curated dataset of Reddit comments, and the best-performing model was deployed in a web application that enables real-time detection of profanity in online discussions.

Theoretically, this research contributes to the broader field of multilingual NLP by advancing profanity detection methods for low-resource languages. It also underscores the importance of context and cultural distinction in developing more equitable and effective moderation tools. By doing so, the study aimed to support safer, more inclusive online environments, particularly in Filipino-speaking digital communities.

Methods

Project Workflow

This chapter focuses on the processes required to proceed and attain the desired result. The project's design and development, testing and operational procedures, project evaluation techniques, and tools are all covered in this chapter.

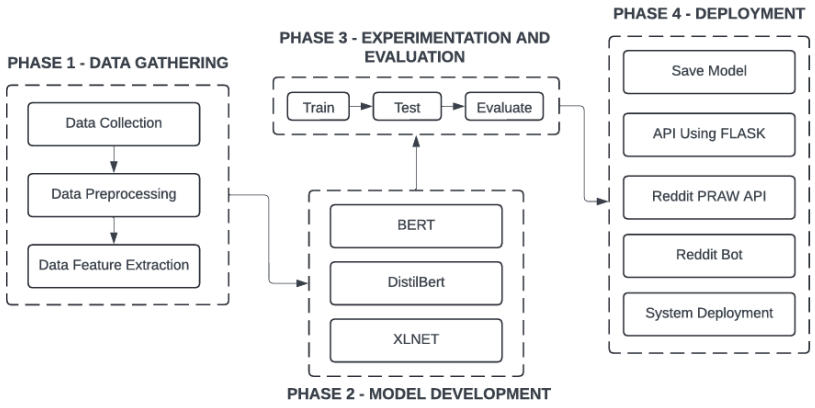

Figure 1. *Project Design of the Study*

The above figure shows the overall workflow of the system, starting from preprocessing the data all the way to deployment. It begins by collecting and preparing Reddit comments through cleaning, labeling, and extracting features. Then, different models like BERT, DistilBERT, and XLNet are trained and tested to find which one performs best. The selected model is then deployed using Flask and the Reddit API to develop a functional web application and bot capable of retrieving Reddit comments and detecting profanity and sentiment in real-time.

Data Collection

The data for this study were gathered by web scraping from the social media platform Reddit. Reddit is widely recognized as a site designed to support open discussions on a broad range of topics, encouraging users from around the world to participate in shared forums for communication and exchange of ideas (Adams, 2022). It was selected due to its large number of active Filipino users and its flexible API, which enables easy access to user-generated content through the Python Reddit API Wrapper (PRAW). PRAW enables efficient access to Reddit's content, allowing for the systematic scraping of relevant posts and comments (Krämer et al., 2024).

The collection period spanned from October to November 2024, during which public comments were scraped from different subreddits such as *r/Philippines, r/Tagalog, r/pinoy, and r/offmychestph*. These subreddits were chosen because they are known to contain informal discussions rich in colloquial expressions and varying levels of profanity, making them suitable for this study's linguistic and contextual analysis.

A total of 13,565 Reddit comments were collected. The comments were further filtered to ensure language consistency and relevance. Various forms and differences of Tagalog profanity were identified using the word list compiled by Esquivel (2022), which served as the foundational reference for detecting explicit and implicit profane expressions.

Data Preprocessing

After collecting various forms of profane words, the researchers proceeded with data preprocessing, which comprised three key components: data cleaning, data annotation, and data augmentation. The data cleaning stage involved removing

duplicate entries, handling missing values, converting all text to lowercase, stripping special characters and punctuation, and removing Tagalog-specific stop words. These steps were essential to produce a clean and standardized dataset, which improves model focus and accuracy in profanity detection. This process is critical in natural language processing (NLP) as it enhances data quality and mitigates bias and noise, contributing to more reliable model performance (Kunilovskaya & Plum, 2021).

The annotation process in the study involved categorizing each Reddit comment based on the frequency of profanity present in the dataset. The researchers created their own framework to classify profanities into three severity levels: mild, moderate, and high. Mild-level profanities were used fewer than 900 times in the dataset; moderate-level profanities appeared between 900 and 1,500 times; and high-severity profanities were used more than 1,500 times. Each comment was labeled according to the highest severity level of profanity it contained.

Table 1. Severity Annotation

Severity Label	Profanity Words
Mild	*Pakshet* | *bwisit* | *pucha* | *kupal* | *yawa*
Moderate	*gago* | *ulol* | *tanga* | *bobo* | *punyeta*
High	*Puta* | *Putangina* | *Tangina* | *pakyu* | *tarantado*

Table 1 shows the profanity words categorized under each annotated severity level, along with their corresponding labels for classification. The researchers utilized the same dataset to manually annotate the sentiment of each sentence. These annotations include labels indicating whether the sentiment expressed is positive, negative, or neutral. This annotated dataset serves as the foundation for training a separate sentiment analysis model, which is integrated into the system alongside the profanity detection model.

Table 2. Sentiment Annotation

Sentiment	Comments
Positive	*"Isa talaga si vico sa iniidolo ko sa lugar na 'yon. Sana po mapost ito. Salamat."*
Neutral	*"gago tawang tawa ko sino may video ng speech nila hahahaha"*
Negative	*"Tangina niyo hugpong mga gunggong ungas"*

Table 2 presents sample Reddit comments from the dataset categorized into three sentiment classes: positive, neutral, and negative. These annotated examples help train the sentiment analysis model to recognize emotional tone, with or without the presence of profanity.

After annotation, the researchers identified a data imbalance in the dataset. To address the issue of class imbalance in the dataset, particularly the underrepresentation of the mild profanity class, data augmentation was conducted using paraphrasing techniques to generate additional diverse samples. As a result, the number of mild instances increased from 1,978 to 4,221. The moderate category grew

from 3,221 to 4,256, and the high category expanded from 4,005 to 4,859. A non-profane category was also introduced, consisting of 5,277 instances to serve as a baseline for comparison. The final dataset used for training and evaluation contained 18,613 Reddit comments, ensuring a more balanced representation across all levels of profanity severity and enhancing the reliability of the classification model.

Table 3. Dataset After Augmentation

Class	Before	After
Mild	1,978	4,221
Moderate	3,221	4,256
High	4,005	4,859

Table 3 presents the frequency of sentences classified by mild, moderate, and high profanity levels after the data augmentation process, resulting in a total of 18,613, including non-profane.

Feature Engineering

The learning of models is enhanced by generating the key features from text. The use of TF-IDF in this context allows for the differentiation of subtle distinctions in language that may indicate profanity (Hajibabaee et al., 2022). This ranges from breaking the text into word or subword units for better representation to calculating the sentence length, creating term frequency or TF-IDF values, and portraying that the words hold an important value.

Model Development

After completing the dataset, the researchers proceeded to develop the profanity detection models and the sentiment analysis model. This study employed three transformer-based architectures for multi-class profanity classification: BERT, DistilBERT, and XLNet. Each model was fine-tuned to classify Reddit comments into four severity levels: non-profane, mild, moderate, and high.

The BERT model was implemented using the Hugging Face Transformers library. It was selected based on the work of Galinato et al. (2023), who demonstrated BERT's effectiveness in profanity detection for the Filipino language. Their fine-tuned BERT model achieved an accuracy of 86 percent, with F1 scores of 88 percent for the non-abusive class and 83 percent for the abusive class. These results underscore BERT's capability in handling context-rich text classification tasks. In the current study, the pre-trained BERT model was initialized and adapted for the multi-class profanity classification task. BERT utilizes a bidirectional transformer architecture, enabling the model to understand the context by considering both preceding and succeeding tokens in a sentence. The training process employed a labeled dataset and used cross-entropy as the loss function. The associated BERT tokenizer segmented the text into subword units for compatibility. Hyperparameters, including the learning rate, batch size, and the number of epochs, were tuned to optimize model performance.

DistilBERT, a smaller and faster variant of BERT, was also fine-tuned for the same classification task. According to Cruz and Cheng (2019), transformer-based models such as DistilBERT are well-suited for natural language processing tasks in low-resource languages like Filipino. DistilBERT retains approximately 97 percent of BERT's

language understanding capabilities while requiring fewer computational resources due to knowledge distillation during pre-training. In this study, the DistilBERT-base-uncased model from the Hugging Face library was used. The dataset was tokenized using DistilBERT's tokenizer, and training was optimized using the AdamW optimizer, a warm-up learning rate scheduler, and a reduced batch size to enhance computational efficiency.

The XLNet model was also included in the study due to its unique modeling approach. Based on the work of Yang et al. (2019), XLNet addresses some limitations of masked language models like BERT by adopting a permutation-based training mechanism. This approach enables XLNet to capture bidirectional contextual dependencies without masking tokens, improving its capacity to model long-range dependencies in text. In the present study, the XLNet-base-cased model was used, along with XLNet's tokenizer, which is designed to accommodate a wider range of linguistic structures. The same tokenized and preprocessed dataset was used to fine-tune XLNet for the profanity classification task.

Model Training and Testing

The researchers selected the 80:20 data split based on the approach used in the study of Arganosa et al. (2022), where the 80:20 split produced the best performance in terms of accuracy, precision, recall, and F1-score in a similar text classification task involving the Filipino language. The researchers applied the 80:20 dataset partition, having 80% devoted to training and 20% to testing the models. The training set is used to fit the model parameters to the data. The trained model is used to make predictions on the testing data, and the predictions are compared to the actual labels to calculate various performance metrics such as accuracy, precision, recall, and F1-score.

Table 4 shows the splitting of the distribution of the given dataset into sets for training and testing on individual labels. The data is split, whereby 80% for training the models and the remainder 20% for testing as per its functionality. For instance, the word "non-profane" in the training set has recorded 3316 entries, while the test set carries 833. The same 80-20 split is applied across the other labels, mild, moderate, and high, ensuring a balanced and comprehensive dataset for model development and evaluation.

Table 4. Training and Test Split

Label	Training (80%)	Test (20%)
Non-Profane	3316	833
Mild	3374	847
Moderate	3426	830
High	3501	895

The researchers utilized the stratify parameter so that the split preserves the proportion with respect to labels across training and testing. They decided to train the model at five epochs to balance the efficiency of learning with performance since previous research shows that accuracy and F1-score improvements drastically occur within the initial few epochs but tend to level off later on, providing minimal gains beyond five epochs. This decision aligns with research benchmarks, which propose three to five epochs for fine-tuning pre-trained models.

Evaluation and Comparison

The researchers aimed to compare the performance of several transformer-based models for detecting profanity in Tagalog text. Specifically, the researchers compared BERT, DistilBERT, and XLNet to determine which model performs best for the task. To evaluate the performance of each model, the researchers conducted a thorough analysis of key evaluation metrics, including precision, recall, and F1-score, as initially utilized by Hernandez et al. (2021). These metrics were calculated based on the models' predictions on the test datasets, allowing the researchers to assess the accuracy and reliability of each model.

Confusion Matrix

The confusion matrix helps determine the model's performance when it makes a prediction and how accurately it approximates the relationship.

Accuracy

Accuracy is calculated by dividing the number of correctly classified data instances by the total number of data instances. It reflects how closely the predicted values align with the actual values. The formula for accuracy can be expressed as:

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

Equation 1. *Accuracy*

F1-score

If one wants precision and recalls to be paired, one needs a metric that does so. The F1-score is a statistic that considers both precision and recall and is defined as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 2. *F1-Score*

Precision

Precision, often referred to as the positive predictive value, measures the proportion of correctly identified positive instances out of all instances that were labeled as positive. The formula for calculating precision is expressed as:

Precision = True Positives / (True Positives + False Positives)

Precision = TP / (TP + FP)

Equation 3. *Precision*

Recall

Recall is the proportion of relevant instances that were successfully identified by the model. Commonly known as sensitivity or the true positive rate, it highlights the model's ability to capture all relevant cases.

$$Recall = \frac{TP}{TP + FN}$$

Equation 4. *Recall*

Deployment

To operationalize the profanity detection system and make it accessible for real-world use, the final models were deployed through a web-based application. The system was designed to serve as a content moderation tool capable of detecting and classifying profane language in Filipino social media comments. By providing real-time classification and filtering of offensive language, the deployed system aimed to contribute to safer and more respectful online environments for Filipino users.

Ethical Considerations

When developing a profanity detection system for Filipino social media comments using transformer-based NLP models, three ethical considerations must be carefully addressed. First, data bias was minimized by ensuring the dataset included diverse samples to avoid unfair or inaccurate classifications. Second, the system accounted for cultural and contextual differences in Filipino-English code-switching to avoid misclassifying harmless phrases as profane or missing actual offensive language. Lastly, privacy and data protection were prioritized by anonymizing Reddit comments and ensuring no personally identifiable information was collected. These steps ensured the system remained fair, culturally sensitive, and respectful of user privacy.

Results and Discussion

This study developed and evaluated transformer-based models for detecting and classifying Filipino profane language by severity levels in social media comments.

Classification Performance of Transformer Models

To evaluate the effectiveness of transformer models in classifying the severity of Filipino profanity, three architectures: DistilBERT, BERT, and XLNet, were trained and tested using the same dataset and training configurations. Each model was assessed using precision, recall, F1-score, and accuracy based on its performance across four severity categories: nonprofane, mild, moderate, and high. The comparative results of these models are summarized in Table 5.

Table 5. Classification Report of Transformer Models

Metric	Precision	Recall	F1	Accuracy
DistilBERT	0.9936	0.9934	0.9935	0.9941
BERT	0.9953	0.9953	0.9952	0.9953
XLNet	0.9944	0.9944	0.9944	0.9944

To complement the tabulated metrics, Figure 2 visualizes the confusion matrices of the three models, highlighting the distribution of predicted versus actual labels for each profanity class.

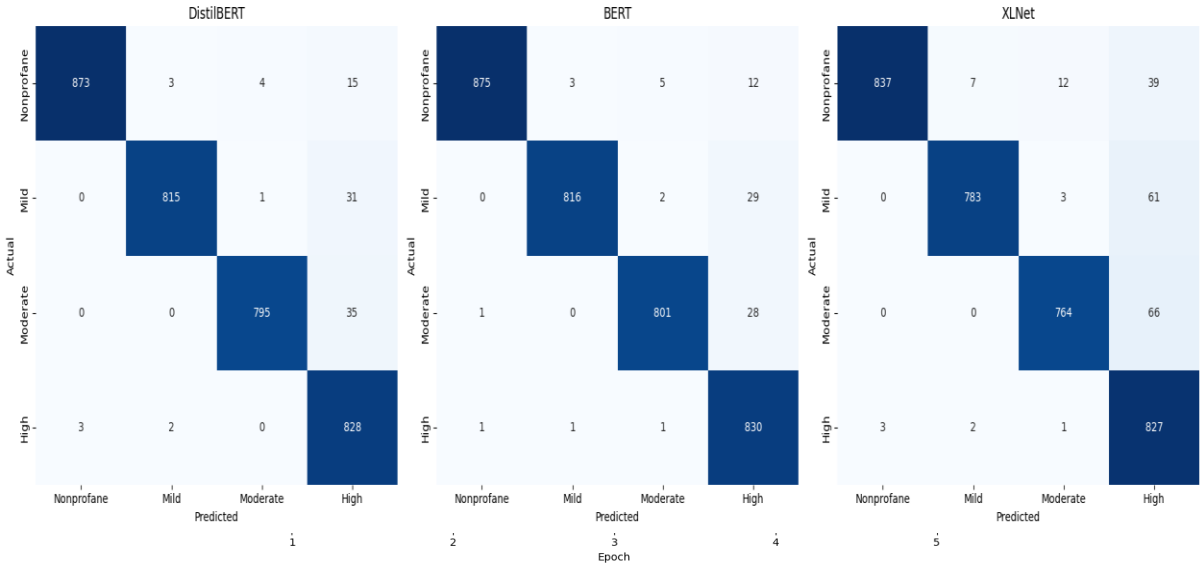

Figure 2. *Transformer Models Confusion Matrix*

BERT showed the most balanced classification performance, with fewer misclassifications across all severity levels. DistilBERT also performed well, particularly in identifying non-profane and high-severity samples, but showed minor confusion between mild and moderate classes. XLNet trailed slightly behind, with more frequent errors in distinguishing non-profane and mild instances, suggesting limitations in its handling of subtle contextual cues.

Training Stability and Model Behavior

Given BERT's superior classification performance in detecting profanity severity, this section focuses on analyzing its training stability and learning behavior. Monitoring training and validation loss across epochs provides insight into how effectively the model generalized to unseen data.

As shown in Figure 3, the BERT-based profanity detection model demonstrated a dynamic yet ultimately stable training behavior across five epochs. Initially, the training loss was relatively high at epoch 1 but dropped sharply by epoch 2, indicating rapid learning in the early phase. A temporary spike in training loss occurred at epoch 3, likely due to learning rate fluctuations or batch variability, but this was quickly corrected in subsequent epochs, where the loss again decreased to near zero by epoch 5. Validation loss showed a more gradual decline over the first three epochs, reaching its lowest value at epoch 3. However, a slight increase was observed in epochs 4 and 5, while still remaining relatively low and stable. Notably, the gap between training and validation loss remained modest, suggesting minimal overfitting despite the fluctuations.

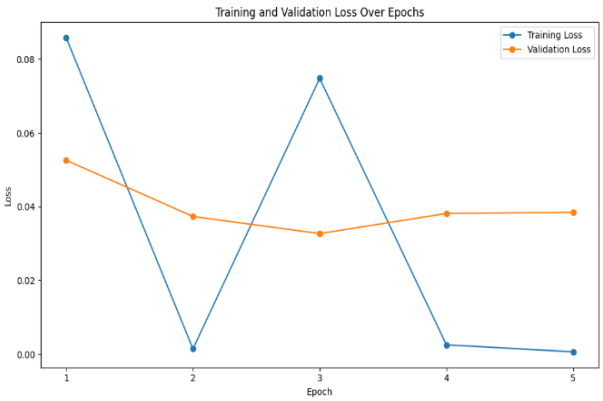

Figure 3. *Loss Curve for BERT Profanity Detection Model*

This pattern indicates that the model effectively generalized to unseen data while continuing to refine its internal representations. The stable convergence by the final epochs supports the robustness of the model, making it well-suited for deployment in real-world applications such as automated moderation of online discourse.

Extension to Sentiment Analysis Using BERT

Building on BERT's strong performance in the profanity classification task, the model was also applied to sentiment analysis involving Filipino social media comments. The rationale for this decision stems from BERT's consistent precision, recall, and F1 scores, which indicate its capacity to understand distinct contextual relationships in multilingual and code-switched environments. Given these characteristics, the same model architecture and training methodology were adopted for the sentiment classification task.

The sentiment analysis model was trained for seven epochs. This epoch count was selected based on earlier trends observed during profanity detection, where performance metrics plateaued around the fifth to seventh epoch. Validation accuracy and loss were monitored continuously to evaluate the model's learning behavior and to avoid overfitting. Checkpoints were saved throughout the training to preserve the best-performing version of the model.

The model demonstrated stable convergence during training. Both training and validation loss curves decreased steadily, with minimal divergence between them, signifying generalization rather than memorization. The smoothness of the loss curve also reflects consistent learning throughout the epochs.

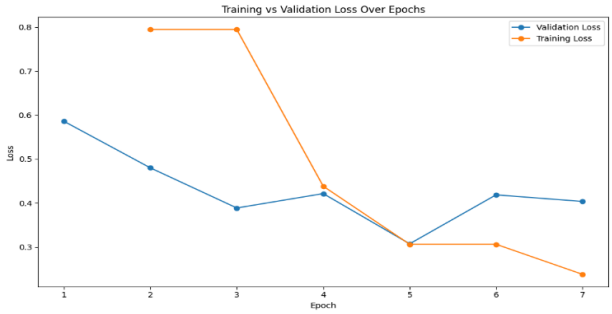

Figure 4. *Loss Curve for BERT Sentiment Analysis Model*

Figure 4 presents the training and validation loss over seven epochs. The model initially exhibited higher loss values, but both metrics steadily declined as training progressed. By epoch 5, the losses reached their lowest points, indicating effective learning. Although a slight uptick in validation loss occurred afterward, the gap between training and validation remained small, suggesting minimal overfitting and strong generalization. These results highlight the model's capacity to extract meaningful patterns from the dataset, affirming its suitability for sentiment classification in real-world moderation systems.

Interpretation and Implications

BERT outperformed other models due to its bidirectional encoder architecture, which captures both left and right context, making it well-suited for handling the informal, code-switched, and complex nature of Filipino social media comments. Its attention mechanism helps distinguish profanity severity by understanding context, allowing the same word to be interpreted differently depending on usage. In real-world applications, a BERT-based system enhances content moderation by reducing false positives and negatives and enabling severity-based responses like warnings or comment masking. This supports the development of ethical, context-aware moderation systems in multilingual settings like the Philippines.

Model Real-Time Testing

To evaluate the usability and effectiveness of the trained BERT-based profanity detection model in live environments, the researchers conducted a real-time testing phase. This aimed to assess the model's ability to detect profane language in actual user inputs from real-world contexts, particularly in social media and online discussions.

As part of this testing, a Reddit bot was developed and deployed. The bot automatically scanned Filipino-English code-switched comments in Reddit threads, detecting and filtering profanity in real time. This bot served as a demonstration of how the model could be used to support content moderation in dynamic and user-driven platforms.

Figure 5 illustrates the profanity detection bot in action, actively classifying and filtering profane language from Reddit comments. The bot implements a severity-based censorship mechanism: mild profanities are masked by replacing a single character (e.g., Kupal becomes Kup*l), moderate profanities are obscured with two characters (e.g., Gago becomes G*g*), and high-severity profanities are masked entirely except for the first letter (e.g., Tangina becomes T******). This graduated filtering strategy offers a practical solution that balances content moderation with the preservation of readability in user-generated text.

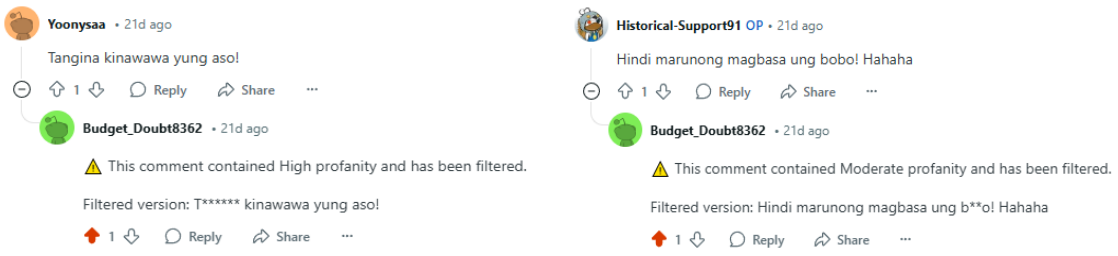

Figure 5. *Subreddit Filtered Comments*

Following the successful implementation of real-time profanity detection using a Reddit bot, the researchers proceeded with the deployment of a web-based profanity detection application.

System Deployment

The profanity detection system was deployed as a web application built using the Flask framework. This application allowed users to input Reddit post URLs, which were then processed by the model to detect profanity and assess both the sentiment and intensity of the comments.

In addition to profanity detection, the system computed sentiment scores and magnitude scores to quantify the strength of the sentiment. This dual analysis enabled moderators to better understand not only whether a comment was profane, but also the emotional tone behind it and the critical context in content moderation decisions.

Figure 6 presents the homepage interface. Users are provided with a simple and accessible input field to paste Reddit URLs. The design prioritizes usability, ensuring users of all technical backgrounds can operate the system with ease.

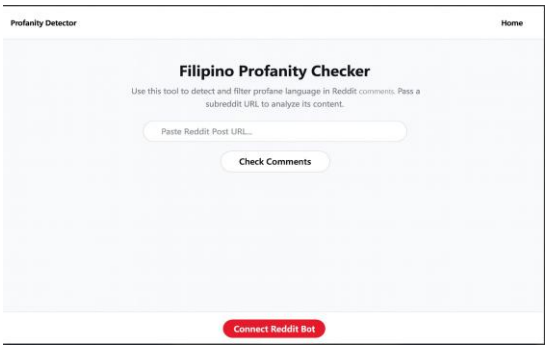

Figure 6. *System Home Page*

Results Visualization and Insights

Once a Reddit thread is submitted, the system retrieves and analyzes the associated post and comments. Results are presented with visual summaries and detailed tables for deeper interpretation. Figure 7 presents the system's results interface, which displays the analyzed Reddit post title and content, along with a summary of the total number of comments and the subset identified as profane. To enhance the interpretability of the analysis, the system incorporates three key visualizations.

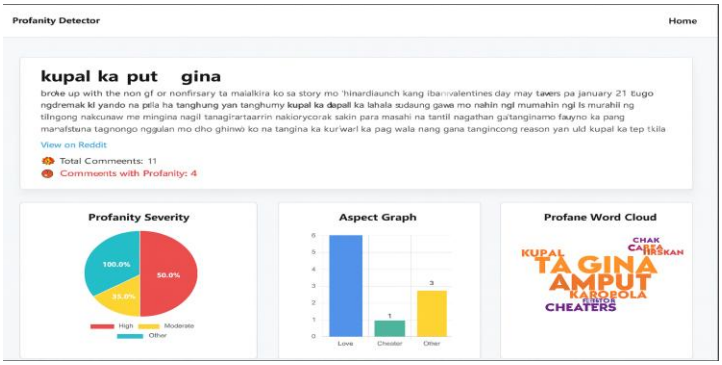

Figure 7. *System Results Page Graph and Word Cloud Visualization*

The visualizations include a Profanity Severity Chart showing the distribution of comments by severity (mild, moderate, high), an Aspect Graph categorizing profane comments by topics like politics, gender, or religion, and a Profane Word Cloud highlighting the most frequently detected profane terms. Together, they provide clear insights into the severity, context, and common usage of offensive language for effective moderation. To enhance the interpretability of detected profanities, the system includes an aspect classification feature. It uses keyword-based matching to categorize the target or theme of a comment (e.g., gender, politics). This allows moderation efforts to be more context-aware.

Table 6. Aspect Keywords

Aspect	Keywords
Politics	gobyerno | presidente | senador | duterte | marcos
Gender	babae | lalaki | LGBT | bakla | tomboy
Religion	Diyos | pari | INC | iglesia | catholic
User	ikaw | mo | g*g* ka | t*ng* ka | drivers
Class	mahirap | mayaman | etilista | masa | burgis
Entertainment	kapamilya | tv5 | gma | abs

Table 7 shows keyword matching technique enables the system to automatically determine the subject or context of a comment alongside profanity classification. By associating specific keywords with predefined aspect categories, the system can identify the underlying topic being discussed. This aspect-based classification enhances the system's ability to interpret the intent and context of profane content, offering a deeper understanding of user behavior and discourse in Filipino social media. It also supports more distinct and context-aware moderation by considering not just the severity of the language used, but also the topic in which it appears.

This method provides actionable insights into the type of profanity used, not just its intensity, also enhancing moderation tools with topic-specific targeting.

Profanity Detection Output

Next are the detection results, which are shown in Figure 8, showcasing the system's ability to present profane content alongside contextual emotional analysis in a structured format. Figure 8 presents the Detection Results Table generated by the profanity detection system. Each row corresponds to a comment retrieved from a Reddit post, which was analyzed to determine the presence and severity of profane language.

Detection Results

Comment	Severity	Sentiment Score	Magnitude Score
naaawa ako sa mga halaman na nagppakahirap magphotosynthesis para lang maproduce yung oxygen ng mga cheater na walang ibang ginawa sa mundo kundi kalibugan	*Non-Profane*	-1.0229	26.5953
screenshot mo tong post mo imyday mo na public tapos tag mo sya baka makita ni bago unless aware syang niligawan sya na kayo pa pero joke lang to ha mamaya gawin mo nga	*Non-Profane*	0.2269	7.713
p******** niyong mga c******* tas kayo pa pavictim a***** magsama sama kayo sa impyerno mga h**** kayo	*High*	-0.9555	16.2427
play traitor by olivia rodrigo btw cheating na po yan talaga kasi kahit kayo pa may ineentertain na	*Non-Profane*	0.0001	0.0022
ohhhh t****** ng mga ganyan hahahaha same thing that happened to me 5 days pa lang kaming break may chikinini na sa bagong peofile pic a***** yun pala may bago na nakipagbreak lang para masabing hindi siya nagcheat t****** ng mga ganyan hahahaha	*High*	-1.0611	45.6292
message the new girl na prolly nagoverlap kayo	*Non-Profane*	-0.031	0.2483
baka inunahan ka lang tutal sa laki ng mga iniipon mong galit imposibleng hindi mo sya palitan eventually pag nakabwelo ka na mindset ba mindset	*Non-Profane*	0.19	4.7508
ku*al talaga sabi nya pa last chat ko na to sayo kasi may iba na palang chinachat bw*set	*Mild*	-1.0721	19.2987
takot sa consequence si g**o ayaw matawag na manloloko hahahahah	*Moderate*	-1.0161	10.1611

Figure 8. *Detection Results of Retrieved Data from Reddit*

The table is divided into four columns: filtered comments, severity, sentiment score, and magnitude score. The filtered comments display text where profanity is masked based on its severity; mild profanities are partially censored, moderate ones have additional letters censored, and high-severity profanities are heavily redacted, leaving only the first letter visible. The severity column indicates whether the comment is non-profane, mild, moderate, or high in terms of profanity.

This categorization allows for a better understanding of the intensity of offensive content. The sentiment score reflects the emotional tone of each comment, with negative values indicating negative sentiment and positive values showing positive sentiment. The magnitude score measures the overall emotional strength of the comment, regardless of whether it is positive or negative. This detection result highlights the system's ability not only to identify and classify profane content but also to analyze the emotional context of each statement. The combination of severity level and sentiment analysis provides a deeper understanding of the nature and impact of user comments in online discussions.

Conclusion and Future Works

This study developed a profanity detection system designed for the Filipino code-switching context, incorporating separate models for profanity classification and sentiment analysis. By using transformer-based architectures, particularly BERT, the system effectively categorized profane content into four severity levels and assessed the sentiment behind user comments. Key contributions include the creation of a manually annotated dataset, the use of sentiment analysis to support contextual understanding, and the deployment of real-time moderation tools through a Reddit bot and web interface.

The system provides practical support for content moderation by enabling more informed decisions based on both the severity and emotional tone of user content. This is especially useful in managing online discussions where Filipino and English are used interchangeably, a setting not well-supported by most existing moderation tools.

Future work can improve the system by broadening the dataset to include content from other platforms such as Facebook, Twitter, and YouTube. Expanding to

audio-based detection could also allow the system to handle spoken content in real-time, which would be valuable for moderating video and voice-based communication.

References

[1] Adams, N. (2022). 'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of COVID-19. *International Journal of Social Research Methodology*. https://doi.org/10.1080/13645579.2022.2111816

[2] Arganosa, S., Marasigan, R., Villanueva, J., Wenceslao, K., & Ponay, C. (2022). *Hate speech in Filipino election-related tweets: A sentiment analysis using convolutional neural networks*. Proceedings of the 2022 3rd International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 451–456.

[3] Cruz, J. C. B., & Cheng, C. (2019). *Evaluating language model fine-tuning techniques for low-resource languages*. arXiv. https://arxiv.org/abs/1907.00409

[4] Esquivel, O. J. (2022). *A sociolinguistic analysis of Tagalog profanities through variables: Age, sex, and context*. ResearchGate. https://www.researchgate.net/publication/383431040

[5] Galinato, V., Amores, L., Magsino, G. B., & Sumawang, D. R. (2023). *Context-based profanity detection and censorship using Bidirectional Encoder Representations from Transformers (BERT)*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4341604

[6] Hajibabaee, P., Malekzadeh, M., Ahmadi, M., Heidari, M., Esmaeilzadeh, A., Abdolazimi, R., & Jones, J. H. (2022). Offensive language detection on social media based on text classification. In *2022 Computing and Communication Workshop and Conference (CCWC)* (pp. 92–98). IEEE. https://doi.org/10.1109/CCWC54503.2022.9720804

[7] Hernandez Urbano Jr, R., Uy Ajero, J., Legaspi Angeles, A., Hacar Quintos, M. N., Regalado Imperial, J. M., & Llabanes Rodriguez, R. (2021, August). *A BERT-based hate speech classifier from transcribed online short-form videos*. In *Proceedings of the 2021 5th International Conference on e-Society, e-Education and e-Technology (ICSET)* (pp. 186–192). https://doi.org/10.1145/3485768.3485806

[8] Krämer, S., Saxena, S., & Pundir, A. S. (2024). *Revolutionizing sentiment analysis: Accelerated data science approaches for Reddit submissions*. In *2024 IEEE INDISCON* (pp. 1–6). IEEE. https://doi.org/10.1109/indiscon62179.2024.10744244

[9] Kunilovskaya, M., & Plum, A. (2021). Text preprocessing and its implications in a digital humanities project. In *Proceedings of RANLP 2021 Student Research Workshop* (pp. 85–93). INCOMA Ltd. https://aclanthology.org/2021.ranlp-srw.13/

[10] Raza, M. O., Mahoto, N. A., Hamdi, M., Reshan, M. S. A., Rajab, A., & Shaikh, A. (2023). Detection of offensive terms in resource-poor language using machine learning algorithms. *PeerJ Computer Science, 9*, e1524. https://doi.org/10.7717/peerj-cs.1524

[11] Roache, R. (2023). What is swearing? In *For fck's sake: Why swearing is shocking, rude, and fun* (online ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190665067.003.0002

[12] Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics, 17*(1), 66-78. https://doi.org/10.1080/19331681.2019.1702607

[13] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding*. arXiv. https://arxiv.org/abs/1906.08237

[14] Yi, M., Lim, M., Ko, H., & Shin, J. (2021). Method of profanity detection using word embedding and LSTM. *Mobile Information Systems, 2021*, 6654029. https://doi.org/10.1155/2021/6654029

[15] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)* (pp. 1415–1420). Association for Computational Linguistics. https://aclanthology.org/N19-1144/

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgement

The authors would like to express their sincere gratitude to Prof. Marc P. Laureta for his expert guidance and valuable feedback throughout the development of this study. His insights were instrumental in refining the methodology and strengthening the overall quality of the work. The authors also acknowledge the support of their peers and colleagues, whose constructive suggestions and encouragement contributed meaningfully to the completion of this research. They further extend their appreciation to Reddit as the source of the publicly available user-generated content used for dataset construction.

Finally, they recognize the broader academic and institutional environment that enabled the successful execution of this study.

Artificial Intelligence (AI) Declaration Statement

In this study, AI tools such as ChatGPT and Gemini were utilized to assist with idea refinement and code development. These tools supported the researchers in model implementation. All AI-generated content was thoroughly reviewed, edited, and verified by the authors to ensure accuracy, clarity, and alignment with the study's objectives.