Deciphering Digital Discourse: Detecting Cyberbullying Patterns in Filipino Tweets Using Machine Learning

January F. Naga[1]* and Rabby Q. Lavilles[2]
[1,2]Department of Information Technology, College of Computer Studies,
MSU-Iligan Institute of Technology
Iligan City, Philippines 9200
Email: [1]january.febro@g.msuiit.edu.ph, [2]rabby.lavilles@msuiit.edu.ph

***Abstract*—The research addresses the escalating challenge of cyberbullying in the Philippines, a concern magnified by widespread social media use. A dataset of 146,661 tweets is analyzed using a pre-trained natural language processing model tailored to detect derogatory Filipino terms. The methodology is designed to preprocess data for clarity and analyze derogatory phrases, using the 23 key terms to indicate cyberbullying. Through quantitative analysis, specific patterns of derogatory term co-occurrence are uncovered. The research specifically focuses on Filipino digital discourse, uncovering patterns of derogatory language usage, which is unique to this context. Combining data mining and machine learning techniques, including Frequent Pattern (FP)-growth for pattern identification, cosine similarity for phrase correlation, and classification technique, the research achieves an accuracy rate of 97.91%. To assess the model's reliability and precision, a 10-fold cross-validation is utilized. Moreover, by examining specific tweets, the analysis highlights the alignment between automated classifications and human judgment. The co-occurrence of derogatory terms, identified through methods like FP-growth and cosine similarity, reveals underlying cyberbullying narratives that are not immediately obvious. This approach validates the high accuracy of the models and emphasizes the importance of a comprehensive framework for detecting cyberbullying in a linguistically and culturally specific context. The findings substantiate the effectiveness of the targeted approach, providing essential insights for developing cyberbullying prevention strategies. Furthermore, the research enriches the literature on digital discourse analysis and online harassment prevention by addressing cyberbullying patterns and behaviors. Importantly, the research offers valuable guidance for policymakers in crafting more effective online safety measures in the Philippines.**

***Index Terms*—Digital Discourse, Cyberbullying Patterns, Social Media, Machine Learning**

I. INTRODUCTION

THE exponential growth of Internet usage in the Philippines has significantly increased online engagement among its population. The country exhibits 76.40 million individuals who actively engage in social media platforms, predominantly aged 18 years and older [1, 2]. The widespread utilization of digital platforms such as Facebook, X (formerly known as Twitter), and TikTok has facilitated significant connectivity among Filipinos. The increase in social media users in the Philippines can be attributed to the country's geographical configuration and a significant population of overseas workers. Filipinos exhibit a noteworthy distinction by attaining the most significant daily social media usage in the Asia-Pacific region, with an average duration of almost four hours [3].

However, this increase in online participation presents various challenges, with cyberbullying being a prominent concern. This form of bullying significantly affects around 70% of young individuals worldwide and is particularly prevalent in the Philippines [4]. The digital domain, while seemingly innocuous, amplifies the negative consequences of bullying behavior, leading to antisocial behaviors, emotional distress, and academic difficulties for those victimized [5–8]. The prevalence of cyberbullying has been facilitated significantly by social media platforms. This trend intensified during the pandemic, with 96% of Internet users aged 16 to 64 actively participating on these platforms as of January 2020 [1].

With its intent to address the urgent need for innovative cyberbullying detection methods, the research leverages data mining and machine learning techniques to analyze X datasets [9–11]. Despite the effectiveness of various machine learning models, such as Support Vector Machines (SVM), Neural Networks, and

Received: Jan. 04, 2024; received in revised form: May 21, 2024; accepted: May 21, 2024; available online: Sep. 09, 2024.
*Corresponding Author

Decision Tree, in identifying cyberbullying instances in English and other languages like Hindi, Roman Urdu language, and Arabic [11–18], there remains a significant gap in research tailored to the Filipino context. Specifically, the complexity of Filipino tweets often encompasses a blend of Filipino and English (Taglish), along with unique slang and derogatory terms, posing challenges for standard cyberbullying detection models.

The research seeks to bridge the research gap by targeting Filipino tweets. The research is uniquely positioned to understand and interpret the mixed language content and slang endemic to Filipino digital communication by leveraging advanced machine learning techniques alongside a pre-trained natural language processing model. The utilized approach aims to identify cyberbullying incidents accurately, contribute valuable insights and methodologies to the field, and address the urgent need for effective cyberbullying detection mechanisms in the Filipino digital space. Furthermore, the research contributes to the broader field of cyberbullying research by highlighting the potential for language-specific models to improve cyberbullying detection across diverse linguistic communities. By developing tailored detection mechanisms that consider social media users' linguistic and cultural diversity worldwide, the research offers a foundational model for future investigations into cyberbullying and online safety measures.

The objectives of the research are:

1) Design a predictive model for classifying tweets based on the presence of Filipino derogatory terms,
2) Employ association rule techniques to discern trends related to these terms,
3) Evaluate the effectiveness of various classification techniques in categorizing tweets within a local database.

A. Cyberbullying

Cyberbullying has become a pervasive issue with the rise of social media, adversely impacting users' mental health and leading to serious consequences such as depression and suicidal tendencies [5–7]. The rapid evolution of Natural Language Processing (NLP) and machine learning offers promising solutions to detect and mitigate cyberbullying by analyzing the vast amounts of textual data generated on social media platforms. Several studies have explored various methodologies for cyberbullying detection, highlighting the role of machine learning algorithms in identifying bullying content. For instance, previous research [12] emphasizes the dominance of aggressive behavior in social media sites and the need for predictive models that incorporate textual and semantic features for accurate cyberbullying detection. Similarly, other studies [13] and [14] explore the efficiency of deep neural networks and pre-trained Bidirectional Encoder Representations from Transformers (BERT) models in identifying cyberbullying content with high accuracy.

The integration of NLP techniques has been crucial in processing and understanding textual data from social media [12, 14, 15]. Techniques, such as tokenization, lemmatization, and vectorization, have been employed to transform raw text into a format that machine learning models can process. Moreover, feature selection methods like Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings have been utilized to extract meaningful features from the text, enhancing the models' ability to distinguish between bullying and non-bullying content.

Table I depicts studies of cyberbullying detection utilizing diverse application of machine learning techniques across various languages and social media platforms. For instance, previous research [6] employs SVM to analyze social networking site data, achieving accuracies of 64% in English and 61% in Dutch, highlighting the linguistic challenges in cyberbullying detection. Similarly, another research [15] utilizes Convolutional Neural Network (CNN) on English social media texts, reaching a 72% accuracy level, which showcases the potential of deep learning in deciphering complex textual patterns. Significantly, a research [16] attains a remarkable 97.11% accuracy using Naive Bayes and SVM on an English Kaggle Dataset, underscoring the effectiveness of combining traditional machine learning algorithms with comprehensive datasets. Research also extends into non-English contexts, focusing on Turkish Twitter data and employing Yapay Sinir Ağı 2 (Turkish for Artificial Neural Network) (YSA2) and a version of BERT that is specifically fine-tuned for the Turkish language, designed to handle the nuances of Turkish text (BERTurk) models to achieve 91% and 77% accuracies, respectively [19, 20]. Similarly, another research [17] demonstrates a high 97.7% accuracy in detecting Persian cyberbullying on Twitter using Deep Neural Network (DNN) and BERT models, emphasizing the efficacy of advanced models in language-specific contexts. Next, other studies [18] and [21] further exemplify the global applicability of these techniques with their successful detection of South African and Vietnamese cyberbullying on X, Facebook, and Instagram, achieving accuracies of 97% and between 86.88% to 90.60%, respectively. Additionally, another research [22] and [23] explores Portuguese and Tunisian Arabic content, with varied

TABLE I
CYBERBULLYING CLASSIFICATION METHODS.

Reference	Data Source	Machine Learning Technique Used	Language	Highest Score (Accuracy)
[6]	Social Networking Sites	Support Vector Machines (SVM)	English,	64%-English, 61%-Dutch
[15]	Social Media Text	Convolutional Neural Network (CNN)	English	72%
[16]	Kaggle Dataset	Naive Bayes, SVM	English	97.11%
[19]	Twitter	Yapay Sinir Ağı 2 (Turkish for Artificial Neural Network) (YSA2)	Turkish	91%
[20]	Twitter	A version of BERT that is specifically fine-tuned for the Turkish language, designed to handle the nuances of Turkish text (BERTurk)	Turkish	77%
[17]	Twitter	Deep Neural Network (DNN), Bidirectional Encoder Representations from Transformers (BERT)	Persian	97.7%
[18]	Twitter	Logistic Regression, Support Vector Machines (SVM), Neural Networks	South African	97%
[21]	Twitter	Bidirectional Encoder Representations from Transformers (BERT)	Vietnamese	86.88%
	Facebook	A variant of BERT that is fine-tuned specifically for the Vietnamese language (PhoBERT)		90.6%
	Instagram			
[22]	Twitter	Support Vector Machines (SVM)	Portuguese	88.36%
		Multilayer Perceptron (MLP)		87.65%
		Logistic Regression		85.19%
		Naive Bayes		76.01%
[23]	Social Media	Naive Bayes	Tunisian	92.9%
		Support Vector Machines (SVM)	Arabic	77.7%

success rates, demonstrating the necessity of adapting detection models to specific linguistic and cultural nuances.

The exploration of machine learning and NLP techniques for cyberbullying detection reveals a promising trajectory towards mitigating the adverse effects of cyberbullying across social media platforms. The integration of sophisticated machine learning algorithms and comprehensive data preprocessing has significantly enhanced the accuracy of cyberbullying detection models [11, 19]. Future research can focus on the implementation of multimodal detection systems that incorporate both textual and visual data, leveraging advancements in deep learning to provide a more holistic approach to cyberbullying detection.

II. RESEARCH METHOD

Quantitative measurement in the research accomplishes understanding the features of cyberbullying texts. Figure 1 is a summary of the research methodology [24]. Using data mining tools and a specific emphasis on Filipino tweets makes it possible to uncover discernible patterns within cyberbullying messages prevalent among Filipinos. The methodology builds upon the established capabilities of RapidMiner for data processing and machine learning, as recognized in the broader research community [25–28]. The approach includes a quantitative evaluation of the frequency and density of offensive terms in tweets. Previous studies [24] and [29] contend that the proportion of "bad" words/terms in a post indicates cyberbullying. Consequently, the research employs a comparable quantitative assessment to detect cyberbullying messages utilizing Filipino tweets as the research focus.

The research centers on tweets in the Filipino language. Consequently, messages penned in other languages that lack derogatory Filipino terms are omitted from the investigation. Following the data retrieval from X, the collected information is stored in a local database. Researchers then employ various data cleaning methods–including tokenization, case transformation, removal of stop words, use of a stemming dictionary, and n-gram techniques–to yield a standardized dataset suitable for analysis. This level may be accomplished using the FP-growth and cosine similarity association criteria. The purpose of these methods is to identify interesting connections between data sets. The established data linkages create frequent occurrences of common patterns within data collection [30].

Following the first stage, the discovered patterns are used to detect cyberbullying tweets. The second stage is accomplished by employing machine learning techniques such as naive Bayes, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest techniques to predict the data class labels. As a result of these studies and the data discoveries, the final output provides a framework for detecting Filipino cyberbullying messages.

While recognizing the advancements and potential of deep learning and transformer-based models in cyberbullying detection, the research leverages associ-

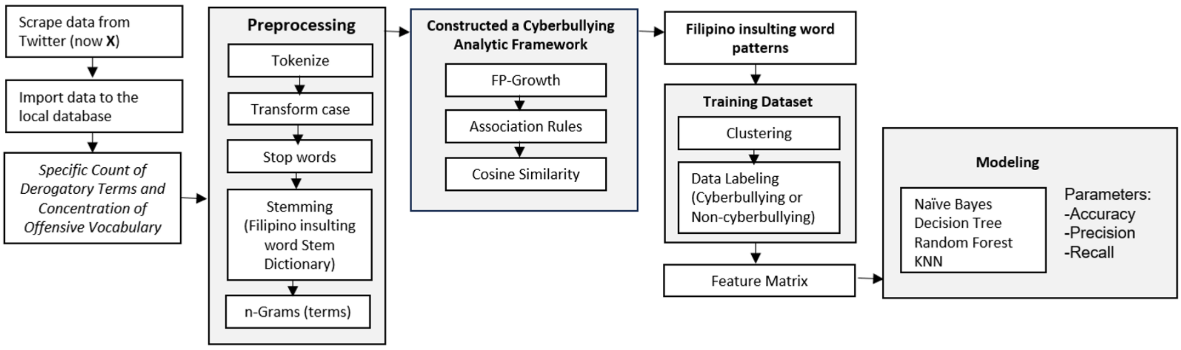
Scrape data from Twitter (now X)
Import data to the local database
Specific Count of Derogatory Terms and Concentration of Offensive Vocabulary
Preprocessing
Tokenize
Transform case
Stop words
Stemming (Filipino insulting word Stem Dictionary)
n-Grams (terms)
Constructed a Cyberbullying Analytic Framework
FP-Growth
Association Rules
Cosine Similarity
Filipino insulting word patterns
Training Dataset
Clustering
Data Labeling (Cyberbullying or Non-cyberbullying)
Feature Matrix
Modeling
Naïve Bayes
Decision Tree
Random Forest
KNN
Parameters:
-Accuracy
-Precision
-Recall

Fig. 1. Framework for cyberbullying detection.

ation rules, Decision Tree, KNN, and Random Forest techniques. These methods are selected based on their compatibility with the dataset’s characteristics and the research goals. Furthermore, the research utilizes the FP-growth algorithm for its efficiency in identifying frequent item sets without the need for candidate generation, cosine similarity for assessing textual instance similarities, and association rules to discover significant patterns of derogatory language use. These methods are chosen for their proven effectiveness in pattern recognition within textual data [31–35], particularly suited to the Filipino language’s distinct linguistic features, which are underrepresented in existing cyberbullying detection literature. The integration of FP-growth, cosine similarity, and association rules with conventional machine learning methods facilitated a nuanced analysis of cyberbullying patterns.

A. Data Collection and Preliminary Analysis

The data collection begins with the utilization of X’s Application Programming Interface (API), meticulously configured to harvest tweets predominantly in the Filipino language. This effort is steered by carefully defining criteria aimed at pinpointing tweets that exhibit markers of cyberbullying. The filters are set to isolate tweets of interest utilizing a blend of key derogatory terms, which are identified through initial research and linguistic analysis. Over a span of four months, from March to June 2022, this approach yielded a dataset of 146,661 tweets.

Choosing X as the focal data source is informed by its significant adoption in the Philippines, serving as a vibrant hub for Filipino online discourse. The platform’s API, known for its sophisticated filtering capabilities, facilitates a targeted collection of tweets, ensuring the dataset’s alignment with the research goals. The dataset’s structure is thoughtfully designed to encapsulate a wealth of information about each tweet, including sequential order, publication time, usernames, retweet metrics, language specifics, tweet text, and geographical details. An illustration of the raw data file garnered from Twitter is provided in Table II, showcasing essential metadata such as date, language, location, and the author’s follower count, along with the tweet content.

The collection process entailes an authentication step to access X’s API, securing a unique identifier that granted us data retrieval capabilities. It includes access to Twitter’s streaming API and two Representational State Transfer (REST) APIs, each offering advantages for exhaustive data collection. Compiled over the specified four-month period in 2022, the dataset is systematically organized into Excel spreadsheets and subsequently exported in CSV format. This process facilitates compatibility with a variety of analytical tools, setting the stage for the subsequent phases of the cyberbullying detection study.

A crucial stage in the data curation process entailes the systematic organization and comprehension of derogatory or insulting phrases commonly used in tweets from the Filipino community. Researchers have started to explore automatic procedures for detecting language patterns signaling harmful content, like previous researchers [6, 24, 36, 37]. Based on the available data, the researchers have identified twenty-three phrases that exhibit frequent association with cyberbullying content for the purpose of context-driven analysis. Consequently, these terms can be categorized as significant derogatory or insulting language markers. The approach employed in the research aligns with prior research conducted [11, 29, 36, 38], wherein the focal point of identifying cyberbullying revolves around offensive language.

Table III provides an overview of the prevalence and dispersion of these twenty-three terms. Both Google Translate (https://translate.google.com/) and the Cam-

TABLE II
EXAMPLE RAW DATA FILE.

Tweet	Date	Language	Location	Author’s followers
fLSs,, natatawa aq sau :’(:’(:’(:’(:’(bk8 gano’n itsura mo. pangit mo :’((fLSs,, I’m laughing at you :’(:’(:’(:’(:’((I didn’t expect you to look like that. You’re ugly :’()	18/03/2022 08:12			369
Abnormal yung utak mo (Your brain is abnormal)	18/03/2022 04:42	Tagalog	PH	283
Gago ka pala eh dika welcome sa fandom ng atin sama ng ugali mo Todo support kami sa sb19 tas yun ang ganti mo maka bash ka sa bts wagas may pa mongoloid abnormal kapa Kay yoongi Sama ng ugali mo... https://t.co/8lHXmGSRAE (So you’re really an idiot and not welcome in our fandom because of your bad attitude. We fully support SB19 and this is how you repay us by bashing BTS, even calling them mongoloid, you’re also abnormal like Yoongi. Your attitude is terrible... https://t.co/8lHXmGSRAE)	18/03/2022 03:04	Tagalog	PH	
bobo tanga uto uto malandi lahat yan naririning ko Ma ayoko na maging mabait ayoko na maging mabait sa ibang tao (Stupid, dumb, gullible, flirtatious, I hear all that. I don’t want to be nice anymore, I don’t want to be nice to other people.)	20/03/2022 19:56	Tagalog	PH	194
RT @lalacru35021849: *muka pa lang demonyo na* hahaha #GaanoKaKaepal https://t.co/7LchYfPL7P (RT @lalacru35021849: Just by looking at you, you seem like a demon hahaha #HowAnnoyingAreYou https://t.co/7LchYfPL7P)	18/03/2022 07:49	Tagalog	PH	48
RT @thvjjklvrs: *TANGINAMO PUTANGINAMO GAGO KA TARANTADO KA ALAM MO SA LAHAT NG MGA TARANTADO IKAW ANG GAGO SHUTANGINAMO INANG INA KA PUTA K…* (RT @thvjjklvrs: F*** YOU, F*** YOU, YOU’RE AN IDIOT, YOU’RE A FOOL, OUT OF ALL THE FOOLS, YOU’RE THE BIGGEST FOOL. F*** YOU, YOUR MOTHER, YOU WHORE K…)	25/04/2022 07:20			

Note: The table contains example of raw data for various tweets.

TABLE III
INSULTING WORDS AND THEIR FREQUENCY OF OCCURRENCE.

Insulting Words		Occurrences in the Dataset
English	Filipino	
Stupid	*bobo*, *gago*, *estupido*	10002
Bitch	*amputa*, *malandi*	7886
Whore	*puta*, call-girl, prostitute, *pokpok*	7611
Crazy	*baliw*, *siraulo*	5171
Scoundrel	*tanga*, *buhong*	5020
Devil	*demonyo*	5001
Fool	*luka-luka*, *hangal*, *ulol*	4991
Idiot	*tanga*, *idyota*	4860
Ugly	*pangit*	4825
Dog	*aso*	4474
Gay	*bakla*	4391
Blind	*bulag*, *putol*	4352
Fuck	*pakshet*, *kantot*	3828
Satan	*satanas*	3638
Fake	*plastik*, *peke*	3426
Pig	*baboy*	3016
Hick	*probinsyano*, *robinsyana*	1823
Mental	abnormal, *abnox*	1536
Deaf	*bingi*	1329
Monkey	*unggoy*	831
Cursed	*isinumpa*	61
Bastard	*bastardo*	12
Hoe	*asarol*, *asada*	1

bridge Dictionary (https://dictionary.cambridge.org/us/translate/english-filipino/) are employed to verify the translations from English to Filipino. Each term is then cross-referenced with Filipino dictionaries to ensure accuracy.

The dataset presents 88,085 occurrences of insulting phrases, highlighting the considerable importance of these terms in the wider context of cyberbullying in Filipino online conversations. The term with the highest frequency of occurrence is “stupid”, represented in Filipino as “*bobo*”, “*gago*”, and “*estupido*”, with a total count of 10,002 instances. The result is followed by “bitch” (“*amputa*” and “*malandi*”) and “whore” (“*puta*”, “call-girl”, “prostitute”, and “*pokpok*”), with frequencies of 7,886 and 7,611 respectively. These terms indicate a significant emphasis on insults related to intelligence and sexual promiscuity within the dataset.

B. Data Preprocessing

1) Data Cleaning: Although comprehensive, the data obtained from X include additional components such as abbreviations, punctuation marks, and emoticons. Hence, it is imperative to do data cleaning to ensure that the analysis is centered on pertinent material to enhance the data for analytical purposes. Data cleaning improves the adaptability in creating data sources and constructing mining structures. The data cleaning process frequently focuses on specific subsets of data, obviating the necessity for distinct structures for each subset.

Critical data preprocessing steps encompass:

- Tokenization: it splits text into individual words or symbols, aiming for deeper text comprehension.
- Case Transformation: It standardizes text into either lowercase or uppercase.
- Stop Word Removal: It eliminates frequent words, like ‘and’ or ‘the’, which can hinder effective text analysis.
- Stemming or Lemmatization: It simplifies words to their root form (e.g., “liking” becomes “like”).

TABLE IV
SAMPLE OF DATA CLEANING PROCESS: TOKENIZATION, CASE TRANSFORMATION, STOP WORD REMOVAL, STEMMING, AND N-GRAMS.

Process	Before	After
Tokenization	[‘*Bakit iiyak manood lang ng* cooking show? HAHAHA abnormal’] (Why cry while watching a cooking show? HAHAHA that’s abnormal.)	[‘*bakit*’,
Transforming Case	[‘*Bakit iiyak manood lang ng* cooking show? HAHAHA abnormal’]	[*bakit iiyak manood lang ng* cooking show? hahaha abnormal]
Stop Word	[‘*Bakit iiyak manood lang ng*” cooking show? HAHAHA abnormal’]	*bakit iiyak manood lang ng* cooking show abnormal
Stem Dictionary	[‘ok *lang pumangit sa kwento ng iba atleast mas pangit yung nagkwento* HSHAHSHAH’] (Just looking bad in someone else’s story, at least the one telling the story is worse. HSHAHSHAH) [‘*pangit*’, ‘*pumangi’t* (Ugly)	*pangit* (Ugly)
N-Grams (terms)	[‘*napaka sinaunang tao mo tanga*’ (You’re such an ancient person, idiot) Description: Word-level unigram(1-gram)	Token Sequence: 2, Token Value: Two

TABLE V
SAMPLE OF DOCUMENT VECTOR.

Text	*tanga*	*satanas*	*plastik*	*peke*	*puta* (bitch)	*ulol*	*abnoy* (abnormal)	*pokpok*
tanga ulol (stupid/idiot insane)	1	0	0	0	0	1	0	0
satanas (satan)	0	1	0	0	0	0	0	0
plastik ulol tanga (fake or plastic stupid/idiot insane)	1	0	1	0	0	1	0	0
peke tanga (fake stupid/idiot)	1	0	0	1	0	0	0	0
pokpok ulol (whore insane)	0	0	0	0	0	1	0	1

- N-Grams: It forms sequences of a set length of tokens, capturing every subsequent sequence of that size in the dataset.

After cleaning the data, words in every message are converted to their root forms. Table IV illustrates the process of message cleaning using tokenization, case conversion, stop word removal, stemming, and n-grams.

2) Data Transformation: In the corpus, each term is represented as a vector, with its weight indicating the binary presence of the term: “1” signifies its presence, while “0” denotes its absence. Once this process concludes, Table V displays the sample resulting word vector, with weights determined by the normalized frequency of binary term appearances. Every document is represented as a vector ($d_i = (w_{1i}, w_{2i}, \ldots, w_{ni})$, with each dimension corresponding to a unique term. If a term is present within a document, its corresponding vector value will be non-zero; conversely, if it is absent, the vector value will be zero. The range of terms might encompass both individual words and entire phrases.

III. RESULTS AND DISCUSSION

A. Data Analysis on Filipino Cyberbullying

The primary tools utilized in the analysis are the FP-growth, cosine similarity, and association rules techniques to analyze the patterns of derogatory phrases in Filipino tweets. The FP-growth approach enhances the efficiency of identifying recurring patterns in cyberbullying tweets and eliminating the necessity for intermediary tasks such as candidate generation. The cosine similarity metric elucidates the strength of the correlations among derogatory terms.

By mapping the occurrence frequency of each phrase and converting them into vector representations, the researchers have discovered significant revelations regarding the underlying patterns of these terms. As depicted in Table V and Figs. 2 and 3, certain phrases such as “*tanga ulol*” and “*plastik ulol tanga*”, along with others like “*baboy*”, “*bakla*”, “*abnoy*”, “*bobo*”, and “*gago*”, regularly exhibit co-occurrence patterns, suggesting a potential common narrative or intention within the tweets. The result provides insights into the complex linguistic patterns that characterize cyberbullying in the context of Filipino online communication.

Based upon the association rules methodology, as elucidated previous research [39], the researchers’ ability to ascertain the probability of the concurrent presence of derogatory phrases is facilitated. This methodology, based on conditional statements, reveals connections among seemingly disparate data, where the initial statement represents the detected element, and the subsequent statement emphasizes elements that frequently occur together. By utilizing the recurring patterns of conditional statements and relying on criteria such as support and confidence thresholds, it is

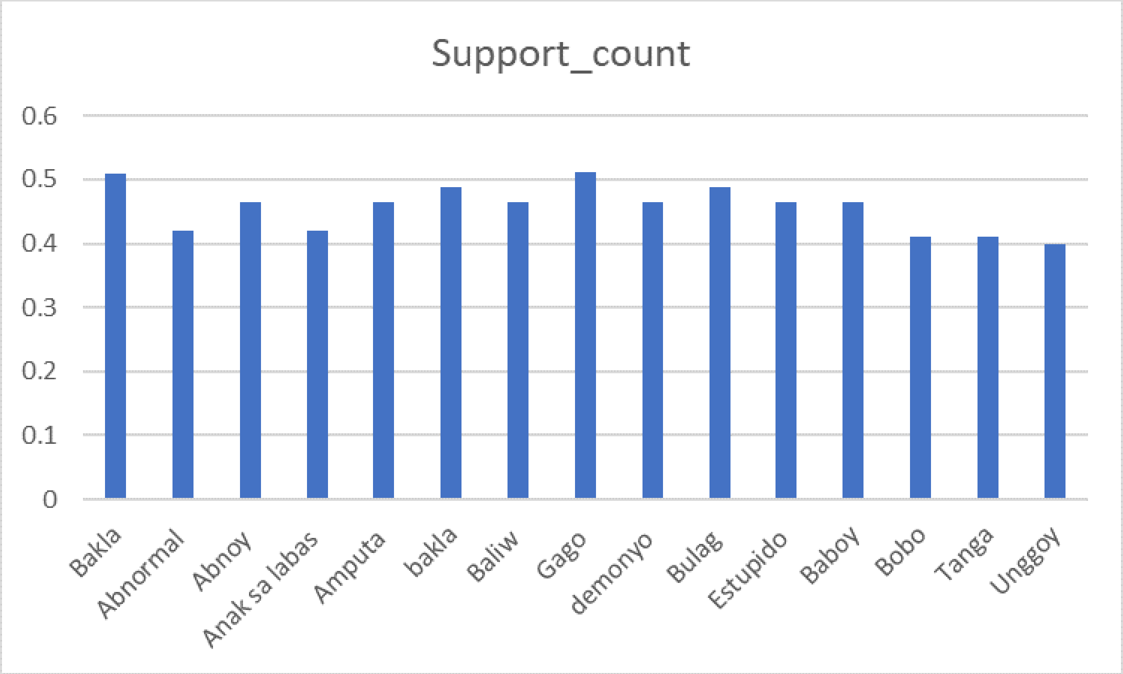

Fig. 2. Derived frequent patterns (Support count).

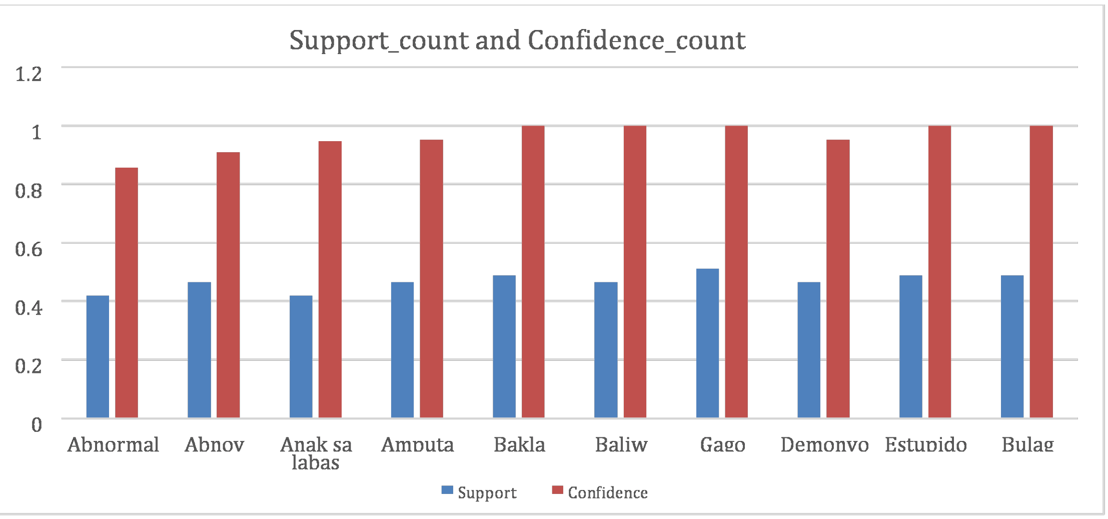

Fig. 3. Derived frequent patterns (Support count and confidence count).

possible to identify the most influential relationships. In essence, it supports quantifies the frequency of item occurrences, whereas confidence evaluates the dependability of conditional claims. For instance, the specific expressions applied during this phase are as follows.

if ((*bobo* + *ulol* + *demonyo* + *baliw* + *abnormal* + *puta* + *malandi*) $>$ 1, “cyberbullying”, “non-cyberbullying”)

The data labeling methodology implemented subsequent to the division process involves the utilization of the K-medoids clustering technique, with a value of k=2 being set to partition the dataset into two separate clusters. The methodology employed is based on the utilization of data labels that are produced from the outcomes of association rules. The K-medoids methodology, renowned for its robustness against outliers, plays a crucial role in the classification procedure. The

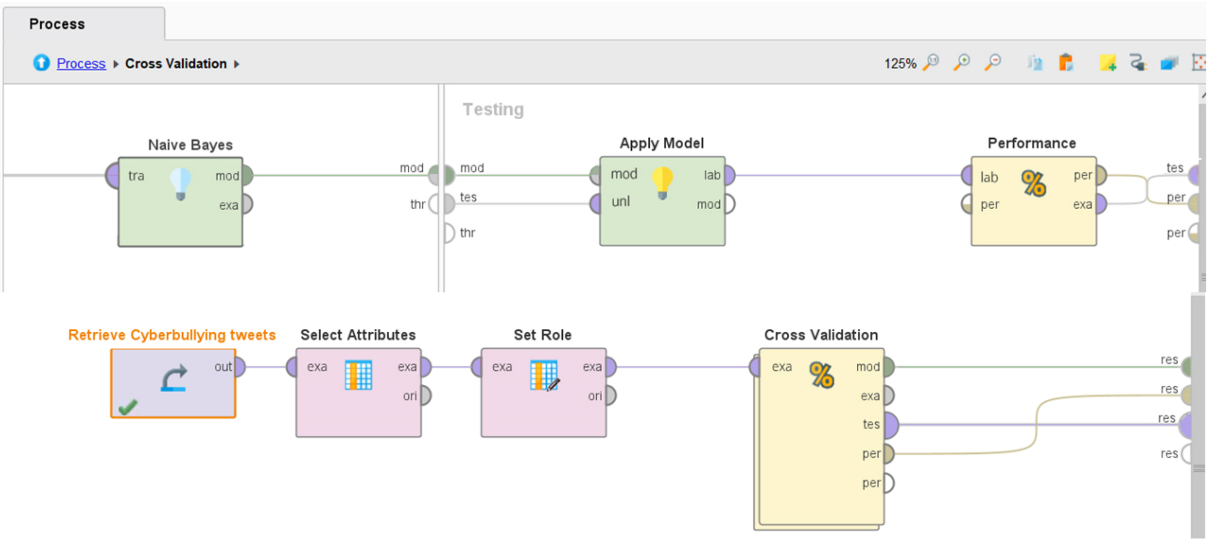

Fig. 4. Naive Bayes workflow.

research employs the Cluster Distance Performance metric to evaluate the effectiveness of centroid-based clustering, particularly with the implementation of K-medoids. This analysis assesses the performance of the model based on cluster centroids, specifically the “Average within cluster distance” and the “Davies-Bouldin index”. By means of this approach, the average distance between the centroid and all instances belonging to the two clusters is computed, offering valuable insights about the performance of the clustering algorithm.

Next, each term undergoes a thorough cross-referencing process with dictionaries, thereby aligning the overall sentiment derived from word translations to guarantee the accuracy of the labels. A comprehensive examination is conducted to address discrepancies in labeling, resulting in establishing a conclusive label for every tweet. The utilization of this composite technique serves to mitigate potential biases, building a strong foundation for the subsequent training of the machine learning algorithms.

One of the main objectives of the research is to investigate and confirm the practicality of automated cyberbullying detection methods specifically designed for Filipino tweets. Given the vast volume of data generated on social media platforms daily, the scalability of human annotation poses significant challenges, making it impractical for real-time or large-scale applications. Although the researchers recognize the critical contribution of human annotation in establishing accurate labels and improving the precision of automated systems, the research aims to create and evaluate an automated detection framework that can effectively function on a large scale. The utilization of FP-growth, cosine similarity, and association rules are utilized for their ability to handle large datasets efficiently. These methods allow the rehearses to systematically create labels by identifying specific patterns and associations that indicate cyberbullying. They provide a foundation for automated detection, which can be improved and validated through targeted human annotation in future iterations of the research.

B. *Machine Learning*

The research begins with collecting data using X’s API, focusing on tweets in the Filipino language. This initial phase involves leveraging RapidMiner for data cleaning, setting a solid foundation for subsequent model training by concentrating on relevant textual features. Clustering to organize the data based on message similarities is allied and then transformed the cleaned text into word vectors to separate cyberbullying from non-cyberbullying content. This crucial step in pattern identification utilizes FP-growth for mining frequent item sets and cosine similarity for evaluating textual similarities, aiding in the detection of meaningful derogatory language patterns. Following this, Naive Bayes, Decision Tree, KNN, and Random Forest algorithms for cyberbullying detection are implemented. The effectiveness of each model is evaluated using a cross-validation strategy, with performance metrics particularly highlighting the precision of the Decision Tree and Random Forest algorithms.

Naive Bayes (Fig. 4) is chosen due to its probabilistic approach, suited for analyzing the textual nature of the dataset. The Naive Bayes classifier is based on

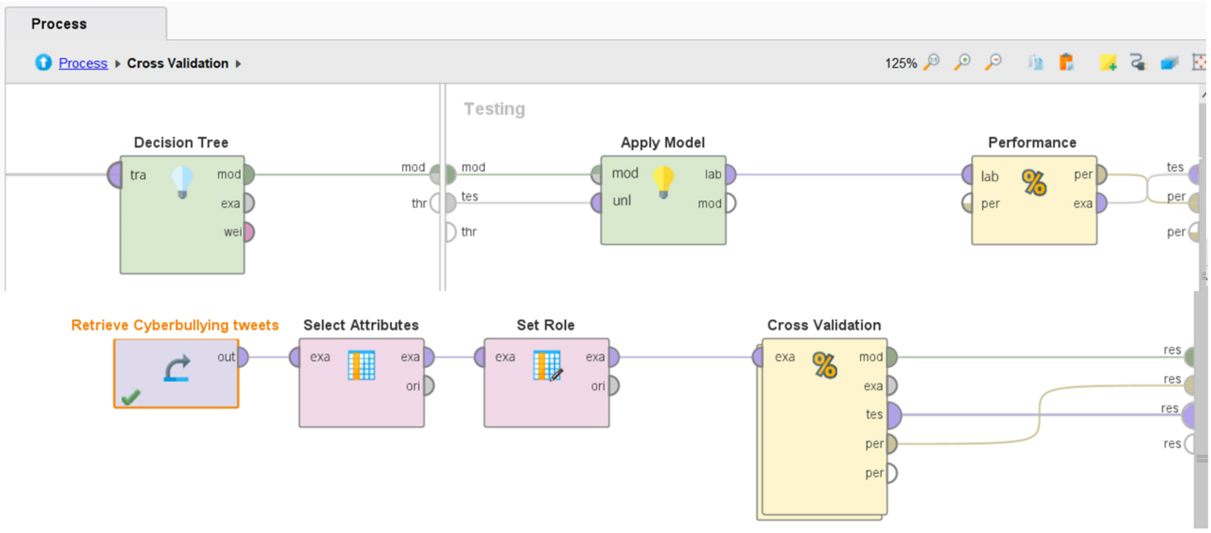

Fig. 5. Decision Tree workflow.

applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable [40]. Equation (1) shows the formula of Naive Bayes. It shows $P(C_k \mid x)$ as the posterior probability of class C_k given predictor x, $P(C_k)$ as the prior probability of class, $P(x \mid C_k)$ as the likelihood, which is the probability of predictor given class, and $P(x)$ as the prior probability of predictor.

$$P(C_k \mid x) = \frac{P(C_k)\, P(x \mid C_k)}{P(x)}. \tag{1}$$

Decision Tree (Fig. 5) provides a straightforward model for identifying key phrases linked to cyberbullying. Decision tree use the concept of information gain, which is based on the reduction in entropy or impurity in the dataset after a dataset is split on an attribute [33]. The Information Gain is calculated using Eq. (2). It consists of $G(T, a)$ as the information gain of an attribute a for the total set T, Entropy(T) as the entropy of the entire set. The second term calculates the entropy for each subset T_v that results from splitting T by attribute a, weighted by the size of subset v relative to T.

$$IG(T, a) = \text{Entropy}(T) - \sum_{v \in \text{Values}(f)} \frac{|T_v|}{|T|} \text{Entropy}(T_v). \tag{2}$$

Random Forest (Fig. 6) improves the Decision Tree’s performance through ensemble learning, enhancing accuracy and reducing the risk of overfitting by averaging predictions from multiple trees [34]. Then, KNN (Fig. 7) relies on the similarity between data points for classification, grouping similar messages together to effectively identify cyberbullying content. KNN classifies a sample based on the majority class among its k nearest neighbors [35]. Equation. (3) shows $d(x, y)$ as the distance between points x and y, each having n dimensions and x_i and y_i are the values of the i-th dimension of points x and y, respectively.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{3}$$

Each model’s performance is evaluated using a confusion matrix and other metrics, showcasing the models’ varied effectiveness in cyberbullying detection within the dataset [41]. Accuracy is the proportion of accurate predictions to the total number of predictions made. Furthermore, the evaluation includes the assessment of precision, which indicates the accuracy of predictions, and recall. It measures the model’s ability to identify relevant instances correctly [30]. The primary objective is to optimize these models, ensuring that they closely correspond with the training data and enhance their predictive capabilities for practical use cases. The researchers utilize the 10-fold cross-validation technique to assess the model’s reliability and precision. The dataset is partitioned in a strategic manner for the purposes of training and evaluation. Specifically, 70% of the dataset is allocated for training, 15% for validation, and the remaining 15% for testing. The utilization of this distribution allows for a

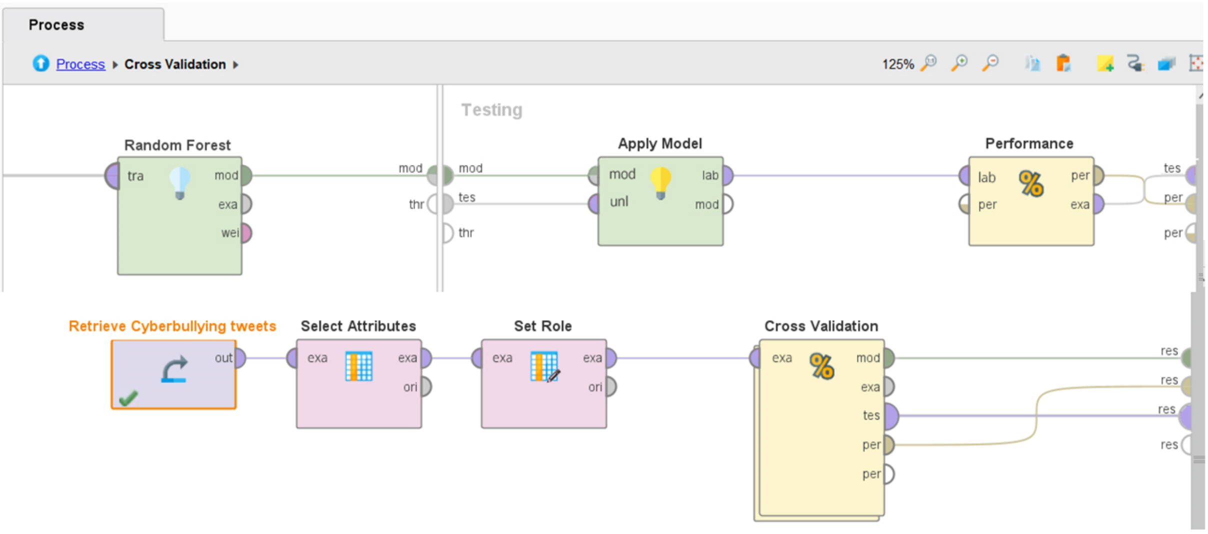

Fig. 6. Random Forest workflow.

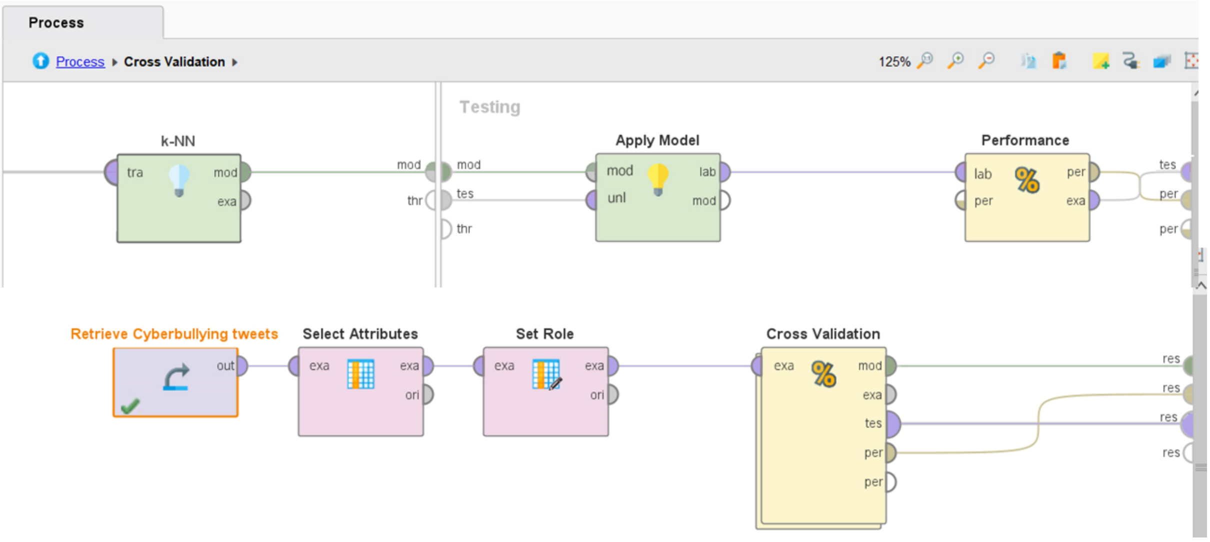

Fig. 7. K-Nearest Neighbors (KNN) workflow.

comprehensive evaluation of the model’s performance across several subsets of data.

Incorporating FP-growth, cosine similarity, and association rules, the research takes a comprehensive approach to analyze cyberbullying patterns. The FP-growth algorithm is particularly efficient in identifying frequent item sets without needing candidate generation, while cosine similarity helps to assess the similarity between textual instances. Association rules are used to find significant patterns of derogatory language use, informing the labeling process. Combining traditional machine learning methods with FP-growth, cosine similarity, and association rules allows for an in-depth analysis of cyberbullying patterns in Filipino tweets, efficiently identifying cyberbullying content and contributing to the broader understanding of cyberbullying detection methodologies.

C. Qualitative Analysis of Cyberbullying Instances

A qualitative analysis of selected tweets classified as instances of cyberbullying provides deeper insights into the models’ detection capabilities. This analysis focuses on the contextual interpretation of derogatory terms and the subtleties of cyberbullying in the digital discourse among Filipinos. By examining specific tweets, the researchers highlight the alignment between automated classifications and human

TABLE VI
THE RESULTS OF PERFORMANCE MEASURE.

Parameter	Naive Bayes	Decision Tree	Random Forest	K-Nearest Neighbors (KNN)
Accuracy	93.33	97.91	97.91	97.33
Precision	96.33	96.79	96.79	96.00
Recall	89.11	96.52	96.50	97.67

judgment, underscoring the models’ effectiveness in navigating the complexities of cyberbullying detection within a linguistically and culturally specific context.

Example 1: “*Gago ka pala eh, dika* welcome *sa* fandom *ng atin, sama ng ugali mo*” (Translation: “You’re an idiot, you’re not welcome in our fandom, you have a bad attitude.”).

The first example is flagged as cyberbullying due to the use of “*Gago*” (idiot), a derogatory term prevalent in the dataset associated with cyberbullying behavior. The aggressive tone, direct address, and exclusion from a community (“not welcome in our fandom”) exemplify cyberbullying tactics aimed at isolating and demeaning the recipient. The models accurately identify this instance, demonstrating sensitivity to both explicit language and implied social aggression.

Example 2: “RT @username: *muka pa lang demonyo na hahaha* #GaanoKaKaepal” (Translation: “Just by the face, already a demon haha #HowShameless”).

Retweets with derogatory comments, as seen in the second example, pose a challenge due to the mixed signals of potential humor (“hahaha”) juxtaposed with insult (“demon”). The hashtag “#HowShameless” further amplifies the negative sentiment. Despite the complexities, the analysis framework effectively discerns the underlying intent of public shaming, validating the model’s capability to interpret nuanced cyberbullying cues within social interactions on X (formerly Twitter).

Example 3: “*bobo tanga uto uto malandi lahat yan naririning ko Ma ayoko na maging mabait ayoko na maging mabait sa ibang tao*” (Translation: “Stupid, fool, gullible, flirt, I hear all that, Ma, I don’t want to be nice, I don’t want to be nice to other people.”).

The third example presents a complex case where derogatory terms are self-directed or recounted from others (“I hear all that”), reflecting the individual’s internalization of bullying and its impact on their behavior (“I don’t want to be nice to other people”). The nuanced understanding of self-directed negativity and its correlation with cyberbullying instances underscores the model’s depth of analysis, recognizing the broader spectrum of cyberbullying’s psychological effects.

The co-occurrence of certain derogatory terms within tweets, as detected through FP-growth and cosine similarity, often reveals underlying cyberbullying narratives not immediately apparent through superficial analysis. Utilizing the recurring patterns of conditional statements and relying on criteria, such as support and confidence thresholds and association rules methodology, make it possible to identify the most influential relationships among the data. These insights provide evidence of the approach, validating the high accuracy rates achieved and underscoring the importance of a comprehensive methodological framework for detecting cyberbullying. The results can be seen in Table VI.

Table VI presents a comparative assessment of four distinct machine learning algorithms: Naive Bayes, Decision Tree, Random Forest, and KNN. These algorithms are evaluated using three performance metrics: accuracy, precision, and recall.

The Naive Bayes method achieves an accuracy of 93.33%. Although the value is respectable, it is lower than other models. However, it is worth noting that both the Decision Tree and Random Forest algorithms have exceptional categorization capabilities, as evidenced by their impressive accuracy rate of 97.91%. Then, KNN algorithm achieves an accuracy of 97.33%, which is somewhat lower than the top-performing models but still demonstrates strong performance.

Then, Naive Bayes exhibits a precision of 96.33%, denoting that its positive predictions are accurate. Both the Decision Tree and Random Forest models exhibit a precision of 96.79%. The results are slightly superior to Naive Bayes. It indicates a slightly enhanced level of dependability in positive predictions. The precision of the KNN model is 96.00%. Although the value is slightly lower than the precision of the other models, it is still considered high.

Next, Naive Bayes exhibits a recall rate of 89.11%. It is significantly inferior to other models. This result indicates its reduced ability to accurately identify all pertinent positive instances. Meanwhile, the Decision Tree and Random Forest models demonstrate impressive recall rates of 96.52% and 96.50% respectively, suggesting their exceptional ability to accurately identify favorable cases. Then, KNN attains a recall rate of 97.67%, rendering it the most efficient in catching all pertinent instances compared to other models.

Overall, the Decision Tree and Random Forest models are the most effective, particularly in terms of accuracy and precision. Although KNN lags slightly behind in accuracy and precision, it outperforms in recall, demonstrating its ability to detect all relevant positive cases. Similarly, Naive Bayes is a reliable model. However, it falls short in terms of recall, which can restrict its applicability in situations where it is vital to identify as many relevant cases as possible.

D. Discussion

In response to the evolving landscape of cyberbullying detection, the study leverages a combination of FP-growth, cosine similarity, association rules techniques, and classification approaches such as Naive Bayes, Decision Tree, and KNN. This diverse methodological framework acknowledges the extensive body of research that underscores the efficacy of various methodologies in this domain. Reflecting on the methodology, the researchers recognize the untapped potential of incorporating deep learning and transformer-based models in future iterations of the research. Such models, known for their ability to understand complex language patterns, can enhance the accuracy of cyberbullying detection, especially as datasets grow in size and complexity. For instance, previous research [42] explores the use of a hybrid model that combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (BiLSTM) networks (CNN-BiLSTM) deep learning detection model. It is capable of identifying cyberbullying content in tweets posted in English, Hindi, and Hinglish, achieving significant success in automatic cyberbullying detection with an accuracy rate of 95%. Similarly, another research [43] utilizes the Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) approach on an English Twitter dataset. They lay the groundwork for these advanced analytical approaches, inviting further exploration into their applicability in diverse linguistic and cultural contexts.

However, adopting these sophisticated models is often curtailed by the need for significant computational resources and exhaustive training datasets, posing challenges for their immediate application in large-scale, real-time scenarios. The chosen methodology, grounded in data mining techniques and conventional classification algorithms, is specifically tailored for its versatility and efficiency, allowing researchers to tackle the vast datasets characteristic of social media. Thus, it offers a scalable framework for detecting cyberbullying within the Filipino context. The proven efficacy of the selected methods in text classification and pattern recognition [31, 33, 35, 44], particularly for tasks demanding nuance content differentiation, positions the approach as aptly suited for the intricate nature of cyberbullying across different linguistic and cultural landscapes [45–47].

Acknowledging the critical role of human annotation in validating and refining automated labeling processes, the researchers envisage incorporating this element in future iterations of the research, aiming to enhance the accuracy and reliability of the cyberbullying detection framework. This methodological strategy reflects a conscientious effort to balance comprehensive and real-time analysis with the pursuit of accuracy, thereby contributing valuable insights into detecting cyberbullying dynamics across social media platforms and linguistic contexts.

Cyberbullying stands as a formidable challenge, especially prevalent on widely used social media platforms like X (formerly Twitter), where it employs verbal aggressions and the vast reach of digital media to inflict profound psychological distress. The methodologies used align with those of seminal works in the field [6, 24, 36, 37] emphasizing the identification of offensive language patterns as key indicators of cyberbullying. The research seeks to uncover instances of cyberbullying among Filipino tweets by meticulously analyzing the prevalence and nuances of derogatory language, thereby shedding light on the behaviors indicative of cyberbullying.

With a reported user base of more than 10.50 million individuals in the Philippines [48], X demonstrates its ability to facilitate positive and harmful interactions. The frequent emergence of derogatory terms in the Filipino language, as indicated in Table II, underscores the association between such language usage and cyberbullying incidents. The aforementioned tendencies, which have also been documented in prior research, contribute to further substantiating the association between the use of offensive language and the occurrence of cyberbullying.

Comparing the results with these studies, it is observed that while the patterns of derogatory language may vary, the underlying behaviors and impacts of cyberbullying are consistent across different cultures and languages. For instance, previous research [6] utilizes SVM to analyze cyberbullying in English and Dutch social media text, achieving 64% and 61% accuracy, respectively. Similarly, another research [17] uses deep learning techniques on Persian tweets, achieving a high accuracy of 97.7%. In the Turkish context, previous research [49] employs a combination of traditional and deep learning methods, achieving an accuracy of 91% with the YSA2 model. Additionally, previous research [18] demonstrates the effectiveness of a semi-supervised learning technique for detecting abusive language in South African social media, achieving a notable accuracy of 97%. The success of these advanced models in different linguistic contexts underscores the need for language-specific approaches in cyberbullying detection. Their findings emphasize the importance of adapting detection models to each region's unique linguistic and cultural nuances. This reinforces the importance of developing culturally and linguistically tailored cyberbullying detection models.

IV. Conclusion

The research findings reveal significant insights into the patterns of derogatory term usage, offering a nuanced understanding of cyberbullying dynamics unique to the Filipino context. This targeted analysis provides a foundation for developing more effective cyberbullying prevention strategies tailored to Filipino online communities' linguistic and cultural nuances. Combining data mining and machine learning techniques, including FP-growth for pattern identification, cosine similarity for phrase correlation, and Decision Tree and Random Forest algorithms for classification, the research achieves an accuracy rate of 97.91%.

Moreover, the research contributes to the broader field of cyberbullying detection by showcasing the potential of language-specific models. This approach enhances detection accuracy across diverse linguistic communities and underscores the importance of contextual sensitivity in analyzing digital discourse. The research also marks a significant step forward in understanding and combating cyberbullying by offering valuable insights for policymakers, educators, and technology developers. It emphasizes the critical role of technological innovation in fostering safer online environments, ensuring that digital spaces remain conduits for positive social interaction despite the challenges posed by cyberbullying.

The research offers valuable insights into detecting cyberbullying patterns in Filipino tweets, but several limitations must be acknowledged. The research relies solely on data from X or formerly Twitter, which may not represent the full range of online interactions. The findings are specific to Filipino users and may not apply to other platforms or demographics. Language evolves rapidly, and the dataset captures only a snapshot in time, necessitating ongoing updates for relevance and accuracy. Moreover, the analysis identifies patterns of derogatory language but does not explore the underlying psychological, social, and cultural factors contributing to these behaviors. Furthermore, the machine learning models used may overlook subtle forms of cyberbullying that do not involve explicit derogatory language. Advanced models like deep learning and transformer-based approaches can provide more nuanced detection capabilities. Recognizing these limitations allows future research to build on the findings and address these gaps.

Future directions for the research include exploring multimodal detection systems that integrate textual and visual data. Such advancements can leverage deep learning techniques. It offers a comprehensive approach to cyberbullying detection that accounts for the evolving nature of online communication.

Acknowledgement

The research was supported by a grant from MSU-Iligan Institute of Technology SO 02105-2022 under the Futures Thinking Project.

Author Contribution

Writing—original draft, J. N. and R. L.; Methodology, J. N.; Analysis and modelling, J. N. and R. L.; Analysis result review, J. N. All authors have read and agreed to the published version of the manuscript.

References

[1] K. D. Peña, "PH social media craze: 77% of Filipinos more engaging online than in real life," 2023. [Online]. Available: https://shorturl.at/Co5xS

[2] S. Kemp, "Digital 2023: The Philippines," 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-philippines

[3] Statista, "Social media in the Philippines - Statistics & facts," 2024. [Online]. Available: https://www.statista.com/topics/6759/social-media-usage-in-the-philippines/#topicOverview

[4] UNICEF, "Online bullying remains prevalent in the Philippines, other countries," 2019. [Online]. Available: https://shorturl.at/Gawld

[5] K. H. Chan, Tommy, C. M. K. Cheung, and Z. W. Y. Lee, "Cyberbullying on social networking sites: A literature review and future research directions," vol. 58, no. 2, pp. 1–16, 2021.

[6] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of

cyberbullying in social media text," *PloS ONE*, vol. 13, no. 10, pp. 1–22, 2018.

[7] J. Li, Y. Wu, and T. Hesketh, "Internet use and cyberbullying: Impacts on psychosocial and psychosomatic wellbeing among Chinese adolescents," *Computers in Human Behavior*, vol. 138, pp. 1–10, 2023.

[8] R. Garett, L. R. Lord, and S. D. Young, "Associations between social media and cyberbullying: A review of the literature," *Mhealth*, vol. 2, pp. 1–7, 2016.

[9] R. Lokeshkumar, O. A. Mishra, and S. Kalra, "Social media data analysis to predict mental state of users using machine learning techniques," *Journal of Education and Health Promotion*, vol. 10, pp. 1–23, 2021.

[10] C. Zachlod, O. Samuel, A. Ochsner, and S. Werthmüller, "Analytics of social media data – State of characteristics and application," *Journal of Business Research*, vol. 144, pp. 1064–1076, 2022.

[11] A. Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data," *Journal of Big Data*, vol. 8, pp. 1–20, 2021.

[12] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[13] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. Coimbatore, India: IEEE, 2019, pp. 604–607.

[14] Y. Yadav, P. Bajaj, R. K. Gupta, and R. Sinha, "A comparative study of deep learning methods for hate speech and offensive language detection in textual data," in *2021 IEEE 18th India Council International Conference (INDICON)*. Guwahati, India: IEEE, 2021, pp. 1–6.

[15] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *2016 23rd International conference on pattern recognition (ICPR)*. Cancun, Mexico: IEEE, 2016, pp. 432–437.

[16] Noviantho, S. M. Isa, and L. Ashianti, "Cyberbullying classification using text mining," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*. Semarang, Indonesia: IEEE, 2017, pp. 241–246.

[17] M. Dehghani, D. T. Dehkordy, and M. Bahrani, "Abusive words detection in Persian tweets using machine learning and deep learning techniques," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*. Tehran, Islamic Republic of Iran: IEEE, 2021, pp. 1–5.

[18] O. Oriola and E. Kotzé, "Improved semi-supervised learning technique for automatic detection of South African abusive language on Twitter," *South African Computer Journal*, vol. 32, no. 2, pp. 56–79, 2020.

[19] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Systems with Applications*, vol. 179, 2021.

[20] F. Beyhan, B. Arık, I. Arin, A. Terzioglu, B. Yanikoglu, and R. Yeniterzi, "A Turkish hate speech dataset and detection system," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 4177–4185.

[21] S. T. Luu, K. Van Nguyen, and N. L. T. Nguyen, "Impacts of transformer-based language models and imbalanced data for hate speech detection on Vietnamese social media texts," *Research Square Platform*, 2022.

[22] A. Silva and N. Roman, "Hate speech detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression," in *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2020, pp. 1–12, sociedade Brasileira de Computação - SBC.

[23] H. Haddad, H. Mulki, and A. Oueslati, "T-HSAB: A Tunisian Hate Speech and Abusive Dataset," in *International Conference on Arabic Language Processing*, vol. 1108, 2019, pp. 251–263.

[24] H. Margono, X. Yi, and G. K. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," in *Proceedings of the Thirty-Seventh Australasian Computer Science Conference*, vol. 147, 2014, pp. 115–124.

[25] P. Ristoski, C. Bizer, and H. Paulheim, "Mining the web of linked data with RapidMiner," *Journal of Web Semantics*, vol. 35, pp. 142–151, 2015.

[26] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 4968–4982, 2022.

[27] M. Z. Naser, "Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering

problems," *Journal of Infrastructure Intelligence and Resilience*, vol. 2, no. 1, pp. 1–15, 2023.

[28] E. D. Madyatmadja, D. J. M. Sembiring, S. M. B. P. Angin, D. Ferdy, and J. F. Andry, "Big data in educational institutions using RapidMiner to predict learning effectiveness," *Journal of Computer Science*, vol. 17, no. 4, pp. 403–413, 2021.

[29] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science*, vol. 181, pp. 605–611, 2021.

[30] J. Han, M. Kamber, and J. Pei, "Chapter 6: Mining frequent patterns, associations, and correlations: Basic concepts and methods," in *Data mining: Concepts and techniques*. Morgan Kaufmann, 2011, pp. 243–278.

[31] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the 20th International Conference on Very Large Data Bases, (VLDB'94)*. Santiago de Chile, Chile: Morgan Kaufmann, 1994.

[32] A. A. Amer and H. I. Abdalla, "A set theory based similarity measure for text clustering and classification," *Journal of Big Data*, vol. 7, pp. 1–43, 2020.

[33] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[35] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[36] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine Learning and Applications and Workshops*. Honolulu, USA: IEEE, 2011, pp. 241–244.

[37] B. A. Talpur and D. O'Sullivan, "Cyberbullying severity detection: A machine learning approach," *PLoS ONE*, vol. 15, no. 10, pp. 1–19, 2020.

[38] M. Alzaqebah, G. M. Jaradat, D. Nassan, R. Alnasser, M. K. Alsmadi, I. Almarashdeh, S. Jawarneh, M. Alwohaibi, N. A. Al-Mulla, N. Alshehab, and S. Alkhushayni, "Cyberbullying detection framework for short and imbalanced Arabic datasets," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, pp. 1–11, 2023.

[39] J. Han, M. Kamber, and J. Pei, "Chapter 9: Classification: Advanced methods," in *Data mining: Concepts and techniques*. Morgan Kaufmann, 2012, pp. 393–442.

[40] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*, 1995, pp. 338–345.

[41] J. Han, M. Kamber, and J. Pei, "Chapter 8: Classification: Basic concepts," in *Data mining: Concepts and techniques*. Morgan Kaufmann, 2012, pp. 327–391.

[42] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application to detect cyberbullying using machine learning and deep learning techniques," *SN Computer Science*, vol. 3, pp. 1–13, 2022.

[43] B. Ogunleye and B. Dharmaraj, "The use of a large language model for cyberbullying detection," *Analytics*, vol. 2, no. 3, pp. 694–707, 2023.

[44] Q. Huang, V. K. Singh, and P. K. Atrey, "On cyberbullying incidents and underlying online social relationships," *Journal of Computational Social Science*, vol. 1, pp. 241–260, 2018.

[45] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48.

[46] S. R. Safavian and D. Landgrebe, "A survey of Decision Tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[47] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[48] Statista, "Leading countries based on number of X (formerly Twitter) users as of January 2024 (in million)," 2024. [Online]. Available: https://shorturl.at/rSdj7

[49] A. Bozyiğit, S. Utku, and E. Nasiboğlu, "Cyberbullying detection by using artificial neural network models," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*. Samsun, Turkey: IEEE, 2019, pp. 520–524.