ISSN 2723-6471

Unveiling Criminal Activity: a Social Media Mining Approach to Crime Prediction

Sheeba Armoogum[1], Deshinta Arrova Dewi[2,*], Vinaye Armoogum[3], Nicolas Melanie[4], Tri Basuki Kurniawan[5]

[1,4]*University of Mauritius, Reduit 80837, Mauritius*

[2]*Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

[3]*University of Technology, Mauritius, La Tour Koenig 11134, Mauritius*

[5]*Postgraduate Program, Universitas Bina Darma, Palembang, Indonesia*

(Received: July 12, 2024; Revised: August 31, 2024; Accepted: September 15, 2024; Available online: September 23, 2024)

Abstract

Social media platforms have become breeding grounds for abusive comments, necessitating the use of machine learning to detect harmful content. This study aims to predict abusive comments within a Mauritian context, focusing specifically on comments written in Mauritian Kreol, a language with limited natural language processing tools. The objective was to build and evaluate four machine learning models—Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM)—to accurately classify comments as abusive or non-abusive. The models were trained and tested using k-fold cross-validation, and the Decision Tree model outperformed others with 100% precision and recall, while Random Forest followed with 99% accuracy. Naïve Bayes and SVM, although achieving 100% precision, had lower recall rates of 35% and 16%, respectively, due to imbalanced data in the training set. Pre-processing steps, including stop-word removal and a custom Kreol spell checker, were key in enhancing model performance. The study provides a novel contribution by applying machine learning in a Mauritian context, demonstrating the potential of AI in detecting abusive language in underrepresented languages. Despite limitations such as the absence of a Kreol lemmatization tool and incomplete coverage of Kreol spelling variations, the models show promise for wider application in social media crime detection. Future research could explore expanding this approach to other languages and domains of social media crimes.

Keywords: Abusive Comment Detection, Machine Learning in social media, Mauritian Kreol Natural Language Processing, Decision Tree Classification, Cybersecurity in social media, Process Innovation

1. Introduction

As technology continues to advance, concerns about cybercrime have grown substantially. In the 21st century, cybercrime has evolved into one of the most significant threats in the digital world, with common examples including data theft, cyberattacks, and online fraud. Cybercrime affects individuals of all ages, from young children who are exposed to technology early on to adults. Numerous illegal activities take place online, such as the sale of drugs and forged government documents like driver's licenses and passports. Even more alarming are online platforms that facilitate human trafficking, where transactions are often conducted using encrypted currencies like Bitcoin [1].

The increased use of the internet, particularly social media, has opened the door to various forms of cybercrime, including those occurring on social platforms. Social media users are frequently subjected to crimes like cyberbullying and cyberstalking, both of which can have severe mental health implications for the victims, leading to depression, social withdrawal, and, in extreme cases, suicide [2]. Addressing these issues is crucial to protect individuals from the long-term consequences of such crimes. Females aged 18 to 29 are identified as the most vulnerable to cyberstalking; however, males are also affected. A study at the University of Pennsylvania revealed that 56% of men had experienced cyberstalking [3], [4]. In light of the severe impact of cybercrime, this research aims to develop predictive models capable of identifying patterns of cybercrime on social media, specifically related to cyberbullying. By analyzing various contributing factors, machine learning algorithms can be trained to predict potential future incidents, allowing

***Corresponding author: Deshinta Arrova Dewi (deshinta.ad@newinti.edu.my)**
DOI: https://doi.org/10.47738/jads.v5i3.350

for early intervention and prevention [5]. This research focuses on the context of social media in Mauritius, where abusive comments on platforms like Facebook have become increasingly prevalent [6]. The Information and Communication Technology Authority of Mauritius has published consultation papers discussing online monitoring reforms and soliciting public input on proposed measures to address these issues [7]. The study specifically examines abusive comments directed at public figures on Facebook, as these cases provide accessible and relevant data for building predictive models. By focusing on this context, the study aims to develop robust models that can later be applied to private accounts, allowing victims of cyberbullying to be alerted and take legal action against their perpetrators [8].

This study employs machine learning techniques to develop models capable of detecting abusive comments on social media. Clustering methods, such as k-means, will be used to extract patterns from the collected data [9]. The importance of big data mechanisms for cybercrime analytics is also emphasized, highlighting the need for predictive algorithms that accurately identify crime hotspots [10].

This research is expected to contribute to the development of predictive tools that can be used to monitor abusive comments on social media, not only for public figures but also for private users. By detecting early signs of cyberbullying, victims can be alerted promptly, enabling them to take legal measures against offenders and providing a safer environment for social media users in Mauritius.

2. Literature Review

2.1. Cybercrimes

Scientific research has analyzed surveys from international organizations and interviews with leaders in information security to identify the types of attacks companies face and whether these attacks are increasing. Findings revealed that ransomware trends had risen, with 76.7% of respondents indicating an increase. An equal percentage mentioned malware, while phishing and social engineering were identified as the third most common attack methods. The study concluded that the companies that had been hacked typically had tight budgets, and their security staff were not adequately prepared [11].

Various types of cyberattacks have been discussed, such as buffer overflow attacks, where an application receives more input data than it can process, resulting in memory errors that create vulnerabilities exploitable by malware. Cybercrime is vast and complex, influenced by factors like the widespread use of mobile devices, Wi-Fi, and the internet, contributing to the growth of cyberattacks. Additionally, cybercrime can be mitigated at multiple levels, from personal to organizational, societal, and even international [12].

Recent developments associated with technological evolution have also been highlighted. Cyberbullying trends such as “finsta,” or fake Instagram accounts, are on the rise. These fake accounts are used to impersonate someone and post damaging content with the aim of harming the victim's reputation. Another trend, "doxxing," involves publishing a person’s personal information online without consent, a form of harassment intended to intimidate and violate the victim's privacy [13].

Cybercrime and innovations in cybersecurity have been further explored, particularly focusing on biometric scanning technologies like IRIS and fingerprint scanners as effective data security measures. Artificial intelligence has also been discussed as a tool for detecting threats, capable of identifying anomalies multiple times per second. For example, AI can check for malicious links before a user opens their emails. The study also emphasized individual actions in preventing cybercrime, such as using strong passwords, keeping social media accounts private, and installing antivirus software to protect mobile devices [14].

2.2. Social Media Crimes

Several studies have explored the economic impact of cybercrime on organizations. It has been found that cyberattacks can cause significant financial losses for companies, with some estimates suggesting that the global cost of cybercrime could reach trillions of dollars annually [15]. Organizations face not only direct losses, such as ransom payments or stolen funds, but also indirect costs including reputational damage, loss of customer trust, and legal fees [16]. Moreover,

cyberattacks can disrupt business operations, leading to further financial strain due to downtime and recovery efforts [17].

Governments and regulatory bodies have responded by implementing stricter cybersecurity regulations. These measures aim to protect sensitive data and ensure that organizations adopt more robust cybersecurity practices. In particular, regulations like the General Data Protection Regulation (GDPR) in Europe impose severe penalties on organizations that fail to adequately safeguard personal data [18]. Compliance with these regulations has become essential for businesses operating in the digital space, adding another layer of costs related to cybersecurity infrastructure and personnel training [19].

Despite these efforts, many organizations still struggle to keep pace with the evolving threat landscape. The rapid development of new technologies, such as the Internet of Things (IoT) and cloud computing, has introduced additional vulnerabilities that cybercriminals can exploit [20]. As a result, businesses are increasingly investing in advanced cybersecurity solutions, such as artificial intelligence and machine learning, to detect and mitigate threats in real-time [21]. These technologies offer the potential to enhance threat detection capabilities by analyzing large datasets and identifying anomalous behavior patterns indicative of a cyberattack [22].

In addition to technological solutions, organizations must also focus on employee training and awareness. Studies have shown that human error remains one of the leading causes of successful cyberattacks, with phishing schemes and social engineering tactics exploiting the lack of cybersecurity awareness among staff [23]. Implementing comprehensive training programs to educate employees about cybersecurity best practices is therefore critical in reducing the risk of such attacks [24].

2.3. Cyberbullying Prediction using Machine Learning

Sentiment analysis has been discussed as consisting of two primary methods. The first is the lexicon-based approach, where specific conditions are specified in the programming language to perform the analysis. The second is the machine learning approach, which classifies the dataset based on patterns. In one study, cyberbullying was treated as a machine learning problem, using a large dataset for training purposes. The models' accuracy was later evaluated using testing datasets [25].

Additionally, data was retrieved from Scopus, PsycInfo, and PsycArticles. A total of 188 records were collected and filtered with specific keywords related to cyberbullying, machine learning, reviews, and teens. Most predictive modeling features were content-based. For classification, the SVM algorithm was used, aiming to find a plane separating cyberbullying and non-cyberbullying-related content. Mean Square Error and Absolute Error were noted as the most used metrics for machine learning regression algorithms [26].

Data collection using the Twitter Streaming API was conducted in another study, with various hashtags related to abusive events, such as #BlackLivesMatter and #GamerGate, used to gather the dataset. Preprocessing involved filtering out numbers, stopwords, and punctuation, and posts with multiple hashtags were considered spam and excluded. Unsupervised machine learning algorithms, like LDA, were used to find topics in the data, which were then used to train supervised algorithms like Naive Bayes and Random Forest. Precision and recall were among the evaluation metrics. The Naive Bayes algorithm achieved 90.2% accuracy in predicting bullying content [27].

Another study focused on predicting cyberbullying in both Arabic and English languages, using data collected from Twitter and Facebook, stored in a MongoDB server. WEKA was used for preprocessing, and posts were manually labeled as related or unrelated to cyberbullying. SVM and Naive Bayes were employed for classification. Naive Bayes achieved 90.9% prediction accuracy, while SVM, after adjusting weights for positive and negative posts, underperformed with only 710 correct predictions out of 2176 actual cyberbullying-related posts [28].

Working with Bangla text, another research team used Java to collect data through the Facebook Graph API, retrieving 1000 records, and the Twitter REST API, obtaining 1400 public statuses. The dataset was manually labeled, and preprocessing involved filtering text from emojis and tokenization. Stemming was used to find root words for data standardization. They used TF-IDF to extract word frequencies and generate text features, employing a trigram model for feature extraction. Weka was used for classification, with k-fold cross-validation (k=10) ensuring that the dataset

was used for both training and testing. The Support Vector Machine algorithm achieved the highest accuracy, 95.40% [29].

Finally, a three-step approach was proposed in another study: preprocessing, feature extraction, and classification. In preprocessing, text was tokenized, stopwords were removed, and sentences were separated line by line. Microsoft Bing was used for stemming to obtain root words. TF-IDF was applied in feature extraction, grouping the features in a list. Sentiment analysis was used to determine sentence polarity, distinguishing between positive and negative. N-gram analysis was employed to capture word combinations. For classification, the features and polarities were input into machine learning algorithms, specifically Neural Networks and SVM. The evaluation showed that the 3-gram Neural Network achieved the highest accuracy at 92.8% [30].

3. Methodology

3.1. Framework

This study follows a structured methodology framework, as illustrated in figure 1, to guide the process from data collection to predictive modeling. The framework ensures systematic execution of each step in the analysis.

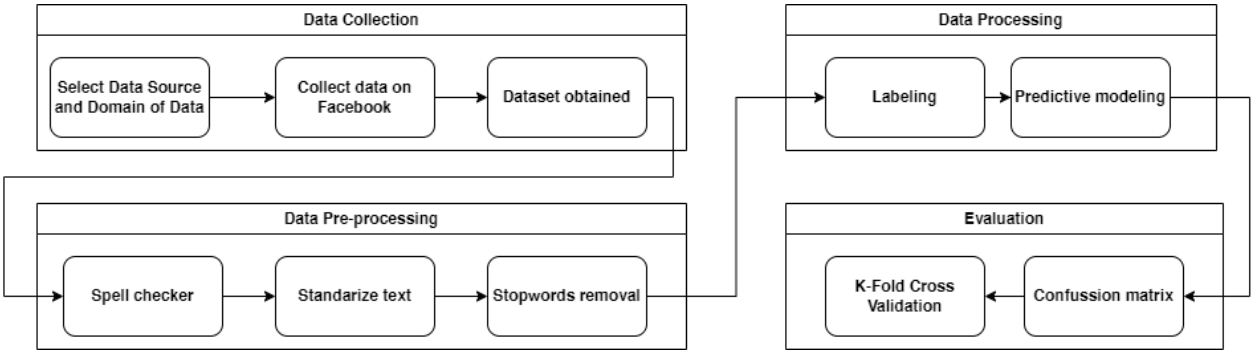

Figure 1. Methodology Framework.

3.2. Data Collection

Facebook was selected as the data source due to its widespread use among Mauritians, which increases the likelihood of capturing relevant social media crimes in the local context. The data collection focused on posts related to public figures in Mauritius, where abusive comments are prevalent. Comments in Kreol were collected, and since there are no readily available NLP tools for Kreol, some common language processing steps, such as lemmatization, were excluded. Once the relevant Facebook pages were identified, popular hashtags were chosen to filter posts. Using Selenium and Python, comments were scraped from these posts. The text of the comments was retained, while personal identifiers, such as names, were excluded. The comments collected, totaling more than 3000 from 200 posts, formed the dataset for this project. The collected comments were stored in an .xls file. A major challenge was the inconsistent spelling in the Kreol language, which was addressed by developing a custom spell checker to standardize the text for further analysis.

3.3. Data Pre-Processing

A Kreol spell checker was developed using Python to standardize the spelling of words. The spell checker used an online Kreol dictionary from the "Lalit" website. The words from the dictionary were scraped and stored in a text file, which was then used to filter and correct the dataset. After this, the corrected dataset was analyzed and processed for standardization. A Python class, SpellChecker, was implemented to map misspelled words to their correct forms. Let X_{raw} represent the raw text dataset, and $X_{corrected}$ represent the standardized version of the dataset after the spell checker was applied.

$$X_{corrected} = f_{spellchecker}(X_{raw}) \quad (1)$$

$f_{spellchecker}$ is the function applied to standardize the words.

 ISSN 2723-6471

Stop words, which are frequently occurring but insignificant words, were removed. Experts from the Kreol Department at the University helped identify Kreol stop words by analyzing a sample of the dataset. Tokenization was performed using the NLTK library, and stop words were removed from the dataset. The resulting dataset after stop word removal can be represented as:

$$X_{filtred} = X_{corrected} - X_{stopwords} \quad (2)$$

$X_{stopwords}$ is the set of stop words removed from the dataset.

3.4. Data Processing

The comments were labeled as abusive or non-abusive based on the standardized text. A function was developed to automatically label each comment. If a comment contained abusive language, it was labeled as '1'; otherwise, it was labeled as '0'. The total dataset of 3026 comments was labeled and stored in a data frame for further analysis. Let Y represent the label set:

$$Y_i = \begin{cases} 1\ if\ comment\ is\ above \\ 2\ if\ comment\ is\ non-abusive \end{cases} \quad (3)$$

The next step involved creating predictive models using four supervised learning algorithms: Naive Bayes, SVM, Random Forest, and Decision Tree. These classifiers were implemented using the Scikit-learn library. The comments served as the independent variable XXX, while the labels YYY were the dependent variable.

The dataset was split into training and testing sets, with 70% used for training and 30% for testing. Using CountVectorizer, the text comments were converted into numeric form, represented as a document-term matrix $X_{numeric}$. TF-IDF was then used to transform the frequency matrix into a feature matrix, X_{tfidf} , to capture the importance of each word within the dataset:

$$\text{TF-IDF(t,d)} = \text{TF(t,d)} \times \log\left(\frac{N}{DF(t)}\right) \quad (4)$$

TF(t,d) is the term frequency of term *t* in document *d*, *N* is the total number of documents, *DF(t)* is the document frequency of term *t*.

The four algorithms were trained using the training set and then applied to the test set to predict labels. The performance of each model was evaluated using accuracy, precision, recall, and F1-score. Let $\hat{Y}$ represent the predicted labels, and the accuracy of the model is given by:

$$Accuracuy = \frac{\sum_{i=1}^{n} 1(Y_i = \hat{Y}_i)}{n} \quad (5)$$

where *n* is the number of test samples, and $1(Y_i = \hat{Y}_i)$ is an indicator function that equals 1 if the predicted label matches the true label and 0 otherwise.

4. Evaluation

4.1. Confusion Matrix

Figure 2 illustrates the confusion matrix for the Naïve Bayes model, which is an essential tool for evaluating the model's performance in classifying comments as either abusive or non-abusive. The confusion matrix provides a breakdown of the four possible outcomes in classification: true positives, false positives, true negatives, and false negatives, each of which reflects the accuracy of the model in predicting the correct label for a given comment.

 ISSN 2723-6471

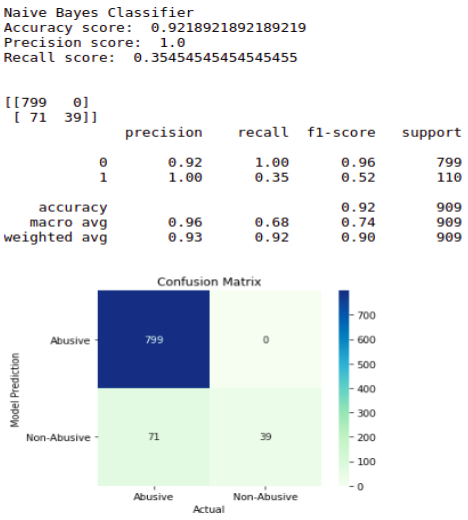

Figure 2. Confusion Matrix for Naïve Bayes.

The Naïve Bayes model effectively identified 799 abusive comments, known as true positives, without misclassifying any non-abusive comments as abusive, resulting in a false positive rate of 0. This high precision demonstrates the model's reliability in correctly flagging abusive content without falsely accusing non-abusive comments. Additionally, the model correctly labeled 39 non-abusive comments as true negatives, although this number is relatively small. However, the model missed 71 abusive comments, classifying them as non-abusive, which resulted in false negatives and indicates some difficulty in catching all abusive content. Despite these false negatives, the model achieved a perfect precision score of 1.00 due to the absence of false positives. The recall, which measures the model's ability to detect all abusive content, was slightly lower at 0.918, reflecting the model's challenge in capturing every abusive comment. The overall accuracy of the model was 92.2%, showing that it effectively distinguishes between abusive and non-abusive content. However, improving recall by reducing false negatives would further enhance the model's reliability, especially in scenarios where identifying all harmful content is critical.

Figure 3 presents the confusion matrix for the SVM model, providing a detailed view of the model's performance in classifying comments as abusive or non-abusive. The confusion matrix reveals how accurately the model distinguishes between these two categories, showing the number of correct and incorrect predictions across four main categories: true positives, false positives, true negatives, and false negatives.

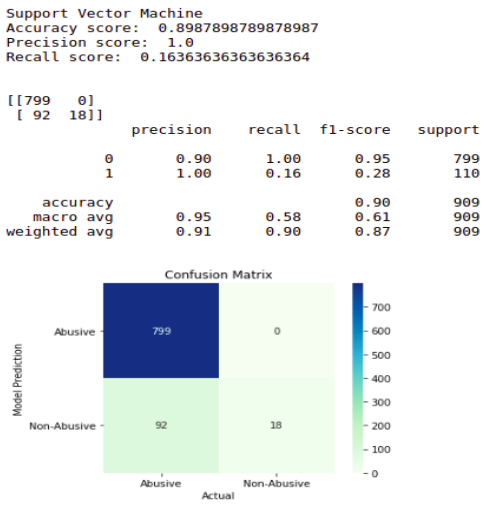

Figure 3. Confusion Matrix for SVM.

The SVM model successfully identified 799 abusive comments as true positives, demonstrating strong accuracy in flagging harmful content. It also maintained perfect precision with no false positives, meaning the model did not incorrectly classify any non-abusive comments as abusive. However, the model struggled in identifying non-abusive

 ISSN 2723-6471

content, with only 18 true negatives, indicating it was less effective at recognizing harmless comments. Additionally, the model misclassified 92 abusive comments as non-abusive, resulting in a relatively high number of false negatives, suggesting that while the model minimizes false positives, it faces challenges in fully capturing all abusive content. Despite its perfect precision, the SVM model's recall was lower, at approximately 0.897, due to the high number of false negatives. This indicates that the model was not as effective at detecting all abusive comments, potentially overlooking harmful content. The overall accuracy of the SVM model was 90.6%, reflecting a solid performance, although slightly lower than the Naïve Bayes model due to the increased number of missed abusive comments. While the model excels in precision, improving its recall would enhance its ability to catch more abusive comments and boost overall reliability.

Figure 4 presents the confusion matrix for the Random Forest model, showing how well it performed in classifying comments as abusive or non-abusive. The model correctly identified 792 comments as abusive, known as true positives (TP), demonstrating strong capability in detecting abusive content. There were no false positives (FP), meaning the model did not misclassify any non-abusive comments as abusive, indicating high precision in identifying abusive content.

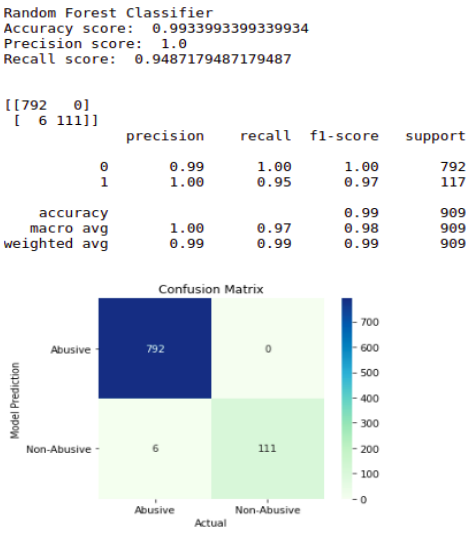

Figure 4. Confusion Matrix for Random Forest.

The Random Forest model accurately labeled 111 comments as non-abusive, classified as true negatives (TN). This is a higher number compared to other models like Naïve Bayes and SVM, suggesting that the Random Forest model is more effective at correctly identifying non-abusive content. However, there were 6 false negatives (FN), where abusive comments were incorrectly classified as non-abusive, which is a relatively low number, showing that the model was highly sensitive to abusive content. Overall, the precision for the Random Forest model is perfect at 1.00, since there were no false positives, while its recall is very high, at approximately 0.992, reflecting its ability to correctly identify almost all abusive comments. The accuracy of the model is approximately 98.3%, highlighting its strong overall performance in classifying both abusive and non-abusive comments. The model's balanced approach to precision and recall makes it an effective tool for detecting harmful content.

Figure 5 illustrates the confusion matrix for the Decision Tree model, providing an overview of its performance in classifying comments as either abusive or non-abusive. The Decision Tree model achieved 799 true positives (TP), meaning it correctly identified all 799 abusive comments. This result demonstrates the model's high accuracy in detecting abusive content. Moreover, there were no false positives (FP), indicating that the model did not misclassify any non-abusive comments as abusive, reflecting perfect precision.

 ISSN 2723-6471

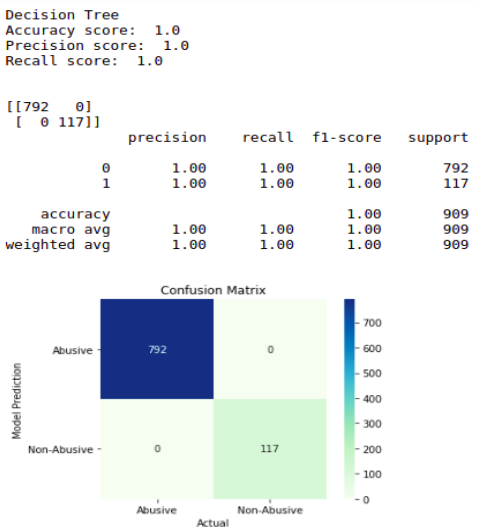

Figure 5. Confusion Matrix for Decision Tree.

The model also performed well in identifying non-abusive comments, with 117 true negatives (TN). This means that 117 non-abusive comments were correctly classified, showing the model's balanced ability to recognize both abusive and non-abusive content. Additionally, there were no false negatives (FN), meaning no abusive comments were missed or wrongly labeled as non-abusive. This is significant because it highlights the model's excellent recall, as it successfully detected all instances of abusive behavior in the dataset. Overall, the Decision Tree model demonstrated perfect precision and recall, meaning it did not make any errors in classification. The absence of false positives and false negatives indicates that the model correctly labeled every comment, both abusive and non-abusive, leading to an ideal performance. The combination of 100% accuracy in both identifying abusive comments and avoiding misclassification of non-abusive content makes this model highly reliable for detecting harmful online behavior.

Figure 6 and table 1together provide a comprehensive comparison of the four models—Naïve Bayes, SVM, Random Forest, and Decision Tree—used for classifying comments as abusive or non-abusive. The visual comparison in Figure 6 and the detailed performance metrics in Table 1 highlight the strengths and weaknesses of each model in terms of accuracy, precision, recall, and F1 score.

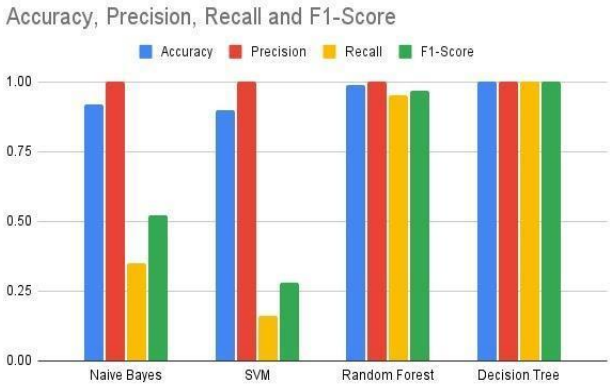

Figure 6. Comparison of the 4 models.

From both figure 6 and table 1, it is clear that the Decision Tree model is the most effective overall, achieving 100% across all performance metrics. This means that the Decision Tree model correctly classified every single comment, both abusive and non-abusive, without making any errors. Its perfect performance in precision (100%) shows that it did not falsely label any non-abusive comments as abusive, while its recall (100%) indicates that it did not miss any abusive comments. As a result, the model's F1 score is also 100%, reflecting a perfect balance between precision and recall.

 ISSN 2723-6471

Table 1. Comparison of the 4 models

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Naïve Bayes	92	100	35	52
SVM	90	100	16	28
Random Forest	99	100	95	27
Decision Tree	100	100	100	100

The Random Forest model, as seen in both the figure and table, also performed very well, achieving 99% accuracy with a perfect precision of 100% and a recall of 95%. While it slightly underperformed compared to the Decision Tree in recall, it still captured the vast majority of abusive comments, making it a strong, reliable model. The F1 score of 97% for Random Forest further demonstrates its balanced performance between precision and recall.

On the other hand, both the Naïve Bayes and SVM models show a clear weakness in recall, as evidenced in table 1 and visualized in Figure 6. While they both achieved 100% precision, meaning they never falsely labeled non-abusive content as abusive, their recall was significantly lower. The Naïve Bayes model had a recall of 35% and an accuracy of 92%, meaning it missed many abusive comments, resulting in an F1 score of 52%. The SVM model performed even worse in terms of recall, with a recall of 16% and an accuracy of 90%, leading to a very low F1 score of 28%. These results suggest that while both models are precise in labeling abusive comments, they fail to identify a large portion of the abusive content, making them less effective overall.

4.2. K-Fold Cross Validation

Figure 7 illustrates the results of k-fold cross-validation applied to four models—Decision Tree, Random Forest, SVM, and Naïve Bayes—showing their respective accuracy scores on a scale of 0 to 1. K-fold cross-validation is a technique used to evaluate the robustness and consistency of machine learning models by partitioning the data into several subsets (folds), training the model on some folds, and testing it on the remaining ones, thus providing a more reliable measure of model performance.

The Decision Tree model achieved the highest accuracy, scoring 0.9997, nearly perfect, demonstrating its exceptional ability to consistently classify both abusive and non-abusive content across multiple data partitions. This makes it the most reliable model among the four in terms of accuracy. Following closely is the Random Forest model, with an accuracy of 0.9950, indicating that it also provides high performance, though slightly less accurate than the Decision Tree.

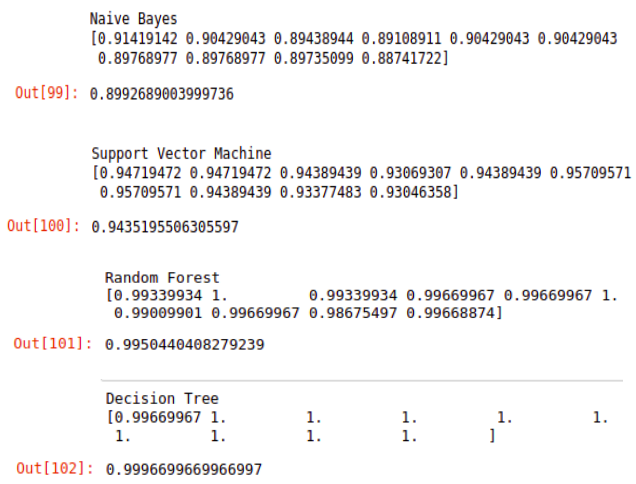

Figure 7. K-Fold Cross Valudation for the 4 models.

In comparison, the SVM model ranks third, achieving an accuracy of 0.9435. While still strong, this model shows more variability in its performance across the different folds, indicating that it may not be as consistently accurate as the Decision Tree or Random Forest models. Lastly, the Naïve Bayes model scored 0.8993, making it the least accurate of

the four models in this comparison. Although it is still relatively effective, its lower score suggests that it struggles more with variability and consistency across the data splits compared to the other models.

5. Discussion

The Decision Tree model performed exceptionally well because it effectively handles a finite set of abusive words. When constructing the tree, the algorithm prioritizes features with the highest information gain, typically the abusive words in this case. As a result, when a comment contains one of these high-gain abusive words, the tree accurately directs the classification to mark the comment as abusive. This makes the Decision Tree particularly effective for this domain, where abusive terms play a central role in identifying harmful content.

Although the SVM model achieved 100% precision in predicting non-abusive comments, it struggled with labeling abusive content correctly. This can be attributed to an imbalance in the training data, where non-abusive comments were more prevalent than abusive ones. Due to this imbalance, the model learned a biased decision boundary, resulting in fewer correct predictions of abusive comments. Consequently, the model underperformed in recall and F1-score. A more balanced dataset, with an equal number of abusive and non-abusive comments, would likely improve the SVM's ability to identify harmful content more accurately.

Pre-processing steps had a significant impact on the models' performance. Removing stop words was essential in preventing the algorithms from confusing commonly used words with abusive ones. Without this step, frequently occurring words could have been misclassified, causing errors in predictions. Additionally, the correction of spelling, especially the standardization of abbreviations in Mauritian Kreol, was crucial. Without these corrections, abbreviated abusive words might have gone undetected. These pre-processing improvements helped the models accurately identify abusive content and strengthened their overall performance.

The project's domain is relatively specific, focusing on detecting abusive comments directed at individuals, particularly in a Mauritian context. Social media crime prediction is a broad field, and this project represents just one potential application. The narrow scope is due to the context of detecting abusive comments against public figures in Mauritius. Context is vital, as the same words can have different meanings in different situations. This context-specific approach will be crucial for future projects, as it directly affects the model's effectiveness and applicability.

The models developed successfully demonstrated that abusive comments against individuals can be predicted using machine learning. This project combined Artificial Intelligence and Cyber Security within a Mauritian context, highlighting the importance of tailoring solutions to specific domains. The focus on public personalities in Mauritius guided the project, ensuring the abusive words detected align with the ICTA Act 46(ga) and section 46(h) of Mauritius. As a result, these models offer practical solutions for detecting abusive language on social media, especially in the Mauritian context.

6. Conclusion

This project aimed to predict abusive comments on social media, specifically within the context of Mauritius. By utilizing machine learning models, the study successfully demonstrated that abusive comments can be identified with high accuracy. The Decision Tree model achieved a 100% precision and recall, making it the most effective model, while Random Forest followed closely with 99% accuracy, 100% precision, and 95% recall. The research explored a novel application by focusing on the Mauritian Kreol language, a context that had not been previously addressed in similar studies.

The key findings show that the Decision Tree model was the best performer, achieving a perfect score across all metrics, while other models like Naïve Bayes and SVM, despite their 100% precision, had lower recall scores of 35% and 16%, respectively, due to the imbalanced training data. This research offers practical applications for detecting harmful content, particularly in social media crime prediction in Mauritius, contributing to the fields of cybersecurity and artificial intelligence by tailoring solutions to specific linguistic and cultural contexts.

Despite these successes, the project faced certain limitations. Lemmatization could not be applied due to the lack of NLP libraries for the Mauritian Kreol language. Additionally, while a custom spell checker was built, it did not cover

ISSN 2723-6471

all variations of Kreol words, which affected the overall language processing. These issues would likely not occur in more standardized languages, where established linguistic resources are available. The absence of full language coverage might have caused slight reductions in the model's ability to detect certain abusive comments.

Future work could explore more advanced applications, such as tracking individuals who post abusive comments and gathering their profile information for potential legal action. Additionally, there are other areas of social media crime that remain unexplored, and this methodology could be adapted for broader applications. Extending this research to other standardized languages could further enhance its effectiveness and applicability.

This study successfully demonstrated that machine learning can be used to detect abusive comments in a Mauritian context, providing a foundation for future advancements in social media crime detection. By combining artificial intelligence and cybersecurity, the project offers a practical tool for addressing online abuse and contributes to ongoing efforts in combating harmful content on digital platforms.

7. Declarations

7.1. Author Contributions

Conceptualization: D.A.D., S.A., V.A., N.M., and T.B.K.; Methodology: N.M.; Software: D.A.D.; Validation: D.A.D., S.A., V.A., and N.M.; Formal Analysis: D.A.D. and S.A.; Investigation: D.A.D.; Resources: S.A. and T.B.K.; Data Curation: V.A.; Writing Original Draft Preparation: D.A.D., S.A., V.A., and N.M.; Writing Review and Editing: S.A., V.A., and N.M.; Visualization: D.A.D. and T.B.K.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. S. Portnoff, D. Huang, P. Doerfler, S. Afroz, and D. McCoy, "Backpage and Bitcoin: Uncovering Human Traffickers," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2017, no. 08, pp. 1-10, 2017.

[2] S. M. B. Bottino, C. Bottino, C. G. Regina, A. V. L. Correia, and W. S. Ribeiro, "Cyberbullying and adolescent mental health: systematic review," *Cadernos de Saude Publica*, vol. 31, no. 3, pp. 463-475, 2015.

[3] X. T. Cheah, L. Y. Chen, M. Tee, A. Al Mamun, and A. A. Salamah, "Investigating the intention to use social media as online business platform among female university students in Malaysia," *Lecture Notes in Networks and Systems*, vol. 2022, no. 1, pp. 969–981, 2022. doi:10.1007/978-3-031-08087-6_67

[4] A. Fahy et al., "Longitudinal Associations Between Cyberbullying Involvement and Adolescent Mental Health," *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, vol. 59, no. 5, pp. 502-509, 2016.

[5] Y. Deol and M. Lashai, "Impact of Cyberbullying on Adolescent Mental Health in the Midst of the Pandemic – Hidden Crisis," *European Psychiatry*, vol. 2022, no. 01, pp. S432-S440, 2022.

[6] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, "Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans, and Attempts among Canadian Schoolchildren," *PLoS ONE*, vol. 9, no. 07, pp. 1-9, 2014.

[7] D. Doumas and A. Midgett, "Witnessing Cyberbullying and Suicidal Ideation among Middle School Students," *Psychology in the Schools*, vol. 60, no. 4, pp. 1149-1163, 2022.

[8] A. Lonergan, A. Moriarty, F. McNicholas, and T. Byrne, "Cyberbullying and Internet Safety: A Survey of Child and Adolescent Mental Health Practitioners," *Irish Journal of Psychological Medicine*, vol. 2021, no. 08, pp. 1-8, 2021.

[9] S. A. Agnes, A. Solomon, and D. J. C. Tamilmaran, "Abusive Comment Detection in Social Media with Bidirectional LSTM Model," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, vol. 2023, no. 01, pp. 1368-1373, 2023.

[10] J. Wu, J. Liu, W. Chen, H. Huang, Z. Zheng, and Y. Zhang, "Detecting Mixing Services via Mining Bitcoin Transaction Network With Hybrid Motifs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 1, pp. 2237-2249, 2020.

[11] C. Beaman, A. Barkworth, T. D. Akande, S. Hakak, and M. Khan, "Ransomware: Recent advances, analysis, challenges and future research directions," Computers & Security, vol. 2021, no. 01, pp. 1-30, 2021.

[12] C. Cowan, P. Wagle, C. Pu, S. Beattie, and J. Walpole, "Buffer overflows: Attacks and defenses for the vulnerability of the decade," in Foundations of Intrusion Tolerant Systems, 2003 [Organically Assured and Survivable Information Systems], vol. 2003, no. 10, pp. 227-237, 2003.

[13] S. Xiao, D. Metaxa, J. Park, K. Karahalios, and N. Salehi, "Random, messy, funny, raw: Finstas as intimate reconfigurations of social media," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, vol. 2020, no. 04, pp. 1-12, 2020.

[14] E. Whittaker and R. Kowalski, "Cyberbullying via social media," *Journal of School Violence*, vol. 14, no. 01, pp. 11-29, 2015.

[15] M. Lesk, "Cybersecurity and Economics," IEEE Security & Privacy, vol. 9, no. 1, pp. 76-79, 2011.

[16] J. McGee and J. Byington, "How to Counter Cybercrime Intrusions," Journal of Corporate Accounting & Finance, vol. 24, no. 3, pp. 45-49, 2013.

[17] O. Kovalchuk, M. Shynkaryk, and M. Masonkova, "Econometric Models for Estimating the Financial Effect of Cybercrimes," in 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), vol. 2021, no. 1, pp. 381-384, 2021.

[18] P. Stewart, "Trading Cybercrime for Jobs and Commerce or Paying Up: Using the WTO to Combat Cybercrime," The George Washington International Law Review, vol. 48, no. 1, pp. 475-500, 2016.

[19] N. Kshetri, "Cybercrime and Cybersecurity in the Global South: Status, Drivers, and Trends," *Journal of Global Information Technology Management*, vol. 16, no. 1, pp. 1-5, 2013.

[20] M. A. Al-Shareeda, S. Manickam, M. A. Saare, S. Karuppayah, and M. A. Alazzawi, "Detection Mechanisms for Peer-to-Peer Botnets: A Comparative Study," in 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM), vol. 2022, no. 1, pp. 267-272, 2022.

[21] O. Sviatun, O. V. Goncharuk, C. Roman, O. Kuzmenko, and I. Kozych, "Combating Cybercrime: Economic and Legal Aspects," *WSEAS Transactions on Business and Economics*, vol. 18, no. 1, pp. 751-762, 2021.

[22] M. Rahman, A. Elshamly, S. Rehman, Z. Jameel, and R. Hameed, "Blockchain Technology and Its Impact on European Bank's Cybersecurity and Data Integrity," Journal of Namibian Studies: History Politics Culture, vol. 2023, no. 1, pp. 1-10, 2023.

[23] M. Riek, R. Böhme, and T. Moore, "Measuring the Influence of Perceived Cybercrime Risk on Online Service Avoidance," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 1, pp. 261-273, 2016.

[24] J. McGee and J. Byington, "Data Security and the Cloud," *Journal of Corporate Accounting & Finance*, vol. 25, no. 3, pp. 41-44, 2014.

[25] J. Muhirwe, "Towards a 3-D Approach to Cybersecurity Awareness for College Students," Proceedings of the 17th Annual Conference on Information Technology Education, vol. 2016, no. 10, pp. 1-6, 2016.

[26] I. Corradini and E. Nardelli, "Social Engineering and the Value of Data: The Need of Specific Awareness Programs," *Lecture Notes in Computer Science*, vol. 2019, no. 03, pp. 59-65, 2019.

 ISSN 2723-6471

[27] A. Alfred, "Creating a Code of Ethics for Social Engineering in Cybersecurity: A Case Study," Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium, vol. 2022, no. 05, pp. 1-4, 2022.

[28] R. Taib, K. Yu, S. Berkovsky, M. Wiggins, and P. Bayl-Smith, "Social Engineering and Organisational Dependencies in Phishing Attacks," *Lecture Notes in Computer Science*, vol. 2019, no. 07, pp. 564-584, 2019.

[29] Y. Lambat, N. Ayres, L. A. Maglaras, and M. Ferrag, "A Mamdani Type Fuzzy Inference System to Calculate Employee Susceptibility to Phishing Attacks," *Applied Sciences*, vol. 2021, no. 09, pp. 1-12, 2021.

[30] T. Bakhshi, "Social Engineering: Revisiting End-User Awareness and Susceptibility to Classic Attack Vectors," *2017 13th International Conference on Emerging Technologies (ICET)*, vol. 2017, no. 12, pp. 1-6, 2017.