International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Text Classification on Cybercrime Cases From News Articles Using Supervised Learning

Nor Muhammad Farhan Nor Muhamad Nizam[1], Sofianita Mutalib[1]*, Mohamad Yusof Darus[1], Azlan Ismail[2], Hamam Mokayed[3], and Shuzlina Abdul-Rahman[1]

[1] School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2] Institute of Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[32]Department of Computer Science, Electrical and Space Engineering, Luleå tekniska universitet, Luelå, Sweden

*E-mail address: *sofianita@uitm.edu.my, yusof_darus@uitm.edu.my, azlanismail@uitm.edu.my, hamam.mokayed@ltu.se, shuzlina@uitm.edu.my*

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: The number of cybercrime cases has increased in this country, especially after the pandemic. The nation has created numerous strategic plans, including the introduction of the Malaysia Cyber Security Strategy (MCSS), which sparked a baseline for countering cybercrime. One of the pillars is Enhancing Capacity and Capability Building, Awareness, and Education. To raise awareness effectively, the taxonomy of cybercrime must be easily understandable by the citizens. This project is to study the classification of news postings by applying supervised models that can ease the classification of cybercrime types. Five supervised models with a combination of two feature extractors were examined. The models were experimented with to evaluate their performance using a percentage split of 70:20 and 80:20. Each model is evaluated based on accuracy, F1-measure, and precision. In the experiment, Random Forest with the TF-IDF feature extractor produced the best result. Achieving an impressive accuracy rate of 94.01%, this model stands out for its precision. Naïve Bayes with the Word2vec feature extractor performed the least effectively, with an accuracy rate of 73.48%. This research focused on analyzing textual data by examining word frequency and interpreting topics based on the class labels of Cybercrime Type 1 and Cybercrime Type 2. Each class of cybercrime news uncovered the topic using latent direct allocation, which was interpreted using Chat-GPT. The analysis and the results of the classification model have been effectively visualized in the PowerBI dashboard, enhancing comprehension. To enhance future research, consider adjusting the scope of the data to focus on local Malay news for more targeted insights.

Keywords: Article News, Cybercrime, Machine Learning, Text Classification, Topic Identification**.**

1. INTRODUCTION (10 – 15 PAGES)

Cybercrime refers to illegal activities conducted through the Internet, including fraud and hacking [1]. Examples of cybercrime include fraud, hacking, piracy, child pornography, and online scams. Cybercrime activities have been rapidly increasing since new technology is continuing to grow. Furthermore, 60% of business transactions occur online and require top-notch security. In Malaysia, cybercrime cases have increased by 50% between 2019 and 2021. The Inspector-General of Police attributed this increase to the moderate level of public awareness about cybercrime [2]. In 2022, cybercrime caused a total loss of almost RM 600 million. The Communication and Digital Ministry launched an educational campaign on cybercrime [3]. Addressing cybercrime requires attention not only to technical aspects but also to educational aspects. Being unaware of online threats in cyberspace can have negative consequences. Awareness of online threats lowers the risk of becoming a victim [4, 5]. For example, by educating individuals on how to identify phishing emails and scams, they can better protect themselves from falling victim to cybercrime. Moreover, educating individuals on proper password management techniques can prevent unauthorized access to personal accounts and sensitive information.

The rise of cybercrime cases has become a serious matter. Society has formulated various strategies to address general security threats [6-8], and there is a pressing need for a comprehensive strategic approach to counter the challenges presented by cybercrime in the digital era. One of the strategies implemented by the National Security Council is raising awareness in society. The public needs

E-mail:author's email

to be cyber-aware of online threats and the latest security devices. They need to know about the evolution of cybercrime and the modus operandi of crime [9]. To begin learning about cybercrime, understanding the keywords used in cybercrime is essential. Through this, they will have a clearer understanding of the topic. Omar et al [5] stated that the research they did showed some respondents did not understand cybercrime terms. Secondly, they need to understand the classification of cybercrime based on relevant terms. The classification of cybercrime is important for many reasons. For example, it is used for the identification of cybercrime and to develop countermeasures [10]. Bernama reports that there is a need for a clearer definition of cybersecurity in the jurisdiction since it is evolving [11]. Additionally, classifying cybercrime-related articles into specific types will make it easier for people to get the content of the cybercrime news. Hence, the public needs to have a clear understanding of cybercrime categories derived from textual data. Subsequently, the public should gain a clearer understanding of the topic of cybercrime.

2. RELATED STUDIES

A. *Cybercrime*

Cybercrime is a dynamically evolving subject. Each researcher has their description of understanding this topic. They have attempted to build and use taxonomies. One of the reasons is that the topic of cybersecurity has many variations that have evolved over the past decade [12]. Donalds and Osei-Bryson [10] describe the classification of cybercrime as an important key to the identification and effectiveness of countermeasures. They divide cybercrime into two classes: computer-assisted crimes and computer-focused crimes. Computer-assisted crimes are crimes that start on the internet but are taking on a new form in cyberspace and are related to keywords like fraud, theft, money laundering, sexual harassment, hate speech, and pornography. Meanwhile, computer-focused crimes are crimes that occur in connection with the establishment of the Internet and cannot exist without the Internet and have keywords like hacking, viral attacks, and website defacement. They argue that this categorization is insufficient. Hence, they build an ontology with new classes like objective and attack events. In terms of cyberattacks in the cloud [13], cybercrime can be identified as technical cybercrime, non-technical cybercrime, or hybrid cybercrime. Technical cybercrime focuses on crime on the internet that primarily takes advantage of victims utilizing technical know-how and information technology-age scams such as viruses, spoofing, and keylogger attacks.

Non-technical cybercrime is described as an attack based on the “luring phase,” such as hate speech and cyberbullying. The term "hybrid cyberattack" refers to the skill required to launch and successfully carry out a cybercrime attack on an unwitting victim. The hybrid is a compromise between the technical and non-technical cybercrime attack skill sets that a cybercriminal needs. Examples of hybrid cyberattacks are trojan malware and cyber fraud. The next study is on the cybersecurity topic based on the Capture the Flag challenge. The cybercrime keyword can be described in terms of the technical parts, which are network attacks, web attacks, and deception [12]. In terms of Routine Activity Theory, the victim of cybercrime can be attacked in these two cybercrime categories, which are cyber-dependent crime and cyber-enabling crime. Cyber-dependent crime describes attacks facilitated by internet technologies such as hacking and URL malware, while cyber-enabled crime is considered the integration of internet technologies with traditional crime like routine time spent on SNS. Ahmad et al [14] and Chipa et al [15] researched the keywords and classes of cybercrime based on other research papers. Authors Hijji & Alam describe that cybercrime can be classified based on scams, smishing, extortion, cyberbullying, and cyberstalking [16]. Authors An & Kim classified the crime into two groups. The first group crime article for service contains brute force, phishing, and spamming. The second group is crime articles for products containing exploits, botnets, rootkits, and trojans [17]. More authors concluded the group of cybercrime cases is in different classes and classifying cybercrime is quite challenging as being addressed by many categorizations involved in cybercrime [18].

News articles can be seen as an information-sharing platform on the topic of cybersecurity. They can bring up a situation that is important for the public to know. Information sharing on cybercrime cases is very important for many parties since they want to understand the attack to protect their networks [4, 19]. Information sharing about cybercrime through news articles can be informed on various topics. Meanwhile, the international media focuses on the threat associated with human actors. Previously, we saw that there were different media representations between international and domestic news on a cybersecurity topic. There are also differences in results and similarities in each country. A study from [20] stated that most of the cybercrime reported in the media in Bangladesh is event-driven, and there is less reporting on investigative news. The study also found that there is the least amount of news related to awareness, which we can see is vital for the people in that country. The study found that most of the news is related to hacking, tracking, stealing information, and the security of cyberspace users. Furthermore, research is being made on understanding the news representation of cyber threats in Sweden from 1995 to 2019. Findings show that cybercrime news coverage has evolved, much like Bangladeshi news coverage

B. *Natural Language Processing and Supervised Methods*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

A study by [21-23] used TF-IDF in their research. In a study by [23] on calculating the similarities between documents and descriptions, they used three methods, which are TF-IDF, Universal Sentence Encoder (USE), and Sentence-BERT (SBERT). Among the three compared methodologies, TF-IDF outperforms both USE and SBERT in terms of efficacy, primarily due to its superior management of terminologies. In the evaluation of similarities, TF-IDF achieves the highest scores in recall and F-measure. Another study stated that the combination of TF-IDF with logistic regression has the highest performance in the English dataset. While the combination of TFIDF and Naïve Bayes produces the best performance for the Spanish dataset, in this study, TF-IDF has a problem considering the order of the words. In this study, the NLPs being used are Bag-of-Words (BoW), TF-IDF, Word2Vec, and BERT [22]. Meanwhile, [18] to achieve their research objective, which is to find the relevant term from Twitter, they calculate the highest TF-IDF score. We can see that TF-IDF is still relevant, even though it is a traditional method. But there is a limitation. One study shows that TF-IDF shows that the frequency of a word is low, even though on Twitter the frequency is high. This is probably due to TF-IDF being limited by the size of the vocabulary required to encode important parts of speech. There are also studies using word-embedding techniques. Jáñez-Martino et al [22] stated that Word2vec and BERT are the most significant word embeddings for their research. Since the Word2vec model is based on learning context-independent word representations, whereas BERT is based on learning context-dependent word representations, For the next study, the word embeddings in this study are Word2vec, FastText, and GloVe. The researcher wants to classify building instance types without changing the metadata. Hence, evaluate three-word embeddings to find the best. As a result, Word2vec shows that strong clusters for both classes are uniformly distributed, with no clear central region for the majority class. Comparing all the studies being made, most studies used the traditional method, TF-IDF, since it produced a better performance. While the use of word embedding needs to be based on the usage of extraction.

Researchers are utilizing various classification models for supervised learning tasks such as classifying cybercrime complaints. Prabhu et al [24] to classify cybercrime complaints. This study utilizes the following classification models: Naive Bayes, K-Nearest Neighbour (KNN), XGBoost, Support Vector Machine (SVM), and Decision Tree. Results indicate that Naïve Bayes and SVM exhibit higher accuracy levels, with SVM achieving 97.4% accuracy. If we compare Naïve Bayes and SVM, SVM scores much better with 97.4%. Mahor et al [25] use Twitter data to detect cybercrime hubs based on Twitter data. The models being used are SVM, Naïve Bayes, and KNN. The results show that linear kernel SVM achieves the highest average precision at 96.57%, while Naïve Bayes demonstrates the lowest precision at 86.51%. An academic paper in the 2019 19th International Symposium on Communications and Information Technologies (ISCIT) titled "Document Classification of Filipino Online Scam Incident Text using Data Mining Techniques" used supervised learning techniques like J48, Naïve Bayes, and Sequential Minimal Optimization (SMO) [26]. The results indicate that J48 achieved the highest accuracy, recall, and f-score, followed by Naïve Bayes. SMO exhibits the lowest accuracy, not exceeding 70%. Comparing these three techniques, J48 is easy to understand, but it took time to build the model. For Naïve Bayes, the researcher agrees that Naïve Bayes is an algorithm that is easy to understand and would be a handful in future cybercrime investigations [26]. Meanwhile, [27] has used several supervised models on textual data. The models are Logistic Regression, Decision Tree, SVM, AdaBoost, Random Forest, Naïve Bayes, Gradient Boosting, and Multilayer Perceptron (MLP). This study shows that decision trees perform the poorest, while SVM emerges as the top classifier for text classification with an accuracy of 84.09%, and logistic regression ranks second with an accuracy of 83.63%. SVM resulted in the highest result due to text data being high-dimensional, which means that SVM can generalize well even in the presence of many functions. In addition, this study states that all models give a better result when performing binary-class classification compared to multi-class classification.

For a single text classifier, Naïve Bayes results with the best accuracy at 0.93, followed by Logistic Regression at 0.81 and SVM at 0.76, while KNN is the worst at 0.72 [28]. In the context of imbalanced data, Naïve Bayes demonstrates superior performance, while KNN yields the least favorable results. The next study, titled "Domain Text Classification Using Machine Learning Models," aims to classify text domains with four algorithms [29]. The techniques being used are SVM, Decision Tree, Random Forest, and Naïve Bayes. As a result, Nave Bayes, Random Forest, and Decision Tree all have accuracy ratings of 76%, 77.38%, and 82.38%, respectively, with SVM having the highest accuracy at 82.38%. [29]. We can see most literature uses Random Forest, Logistic Regression, SVM, and Naïve Bayes. Many studies have been using Naïve Bayes, and the results show it can produce good results, although some of them produced low results. For Naïve Bayes, many works of literature apply this technique

because the algorithm is easy to understand. Logistic regression is one of the simplest and most uncomplicated machine learning models. It focused on predicting the probability for a target class. Some literature shows logistic regression is the second-best classifier, which means it can be considered one of the techniques that can be utilized. Due to its ability to handle multiple dimensions, SVM is one of the traditional methods that is frequently used and produces a good result. Many of the studies in the literature show it gains the highest result for a supervised technique. Lastly, Random Forest. Random Forest operates by having many decision trees formed during the training phase, and at the prediction stage, the class that represents the individual trees' classes is the output. Random Forest demonstrates the potential for extensive utilization as a classifier, although some sources indicate that it may not consistently achieve high accuracy rates.

3. RESEARCH METHODS

In general, there are several phases in establishing the study:

A. *Pre-Modelling Phase*

- Knowledge Acquisition: A research paper with the same domain to identify the problem and technique. The knowledge acquired on the categorization and keywords used in cybercrime cases had been defined. In this research, the classification of cybercrime by Gordon and Ford [30] has been used, which defines the class of cybercrime into two classes, which are Type 1 and Type 2. Type 1 is defined as a crime against the machine, while Type 2 is defined as a crime using the machine. The following Table 1 lists the keywords of cybercrime based on their suitability to be found in the local news article in Malaysia.

TABLE I. TABLE TYPE STYLES

Class	***Keywords***
Type1 Cybercrime	Hacking, Malware, Data Breach
Type2 Cybercrime	Online Gambling, Pornography, Scam

- Data Collection: The data was collected from local English news articles, which are MalayMail, The Sun Daily, and The Star, from January 1, 2018, to October 18, 2023. The data is scraped using the Selenium web driver and the Newspaper3K Python library. The metadata that has been scraped is the title, main text, and publication date.
- Data Preprocessing: Text Cleaning: Remove HTML tags, special characters, numbers, and punctuation marks; tokenization: tokenize the text into tokens using the NLTK library; stopword removal: remove English stopwords using the NLTK stopwords and English stopwords file created by Amir Hosein Sedahati in Kaggle Lemmatization; normalize the word into root form.
- Feature Extraction: Two different feature extractors are being used for the experiment, which are Term Ferquency-Inverse Document Frequency (TF-IDF) and Word2vec TF-IDF. Using TF-IDF for each term allows one to determine the meaning of the terms and helps to understand the key characteristics that distinguish one term from another. This method is a product of term frequency (TF) and reverse document frequency (IDF). The term frequency is defined as the word being more important in a topic of a text if it frequently appears in the text. The Inverse Document Frequency focuses on the frequency with which words appear in multiple texts. If the word has a high frequency across multiple texts, it will show the word is irrelevant. Word2Vec is used to measure the association and similarities between words, which can provide a good insight into the semantic structure of the tokens. Word2Vec has two models, which are the continuous bag-of-words (CBOW) model and the skip-gram m model. In this research, Word2Vec will use the skip-gram m model since it can capture the relationship in the sentence.
- Data Labelling: For each of the data that has been scraped, it will be labeled manually based on the keywords that have been extracted. Which means each row of the data will be assigned a specific label. For Type 1 cybercrime, the keywords will be malware, hacking, and data breaches. While for Type 2 cybercrime, the keywords will be a scam, online gambling, and pornography.

B. *Model Development*

Model Development: This is the process of fitting a model to the training data. The model learns from the training data, iteratively adjusts its internal parameters, and optimizes them to reduce prediction errors. The data includes an independent variable, the input feature, and a dependent variable, the target labels. This stage aims to adjust hyperparameters to discover the optimal model parameters for accurately predicting class labels in unseen data. Before going into this stage, Balancing the data is crucial in machine learning experiments to prevent bias towards the majority class and improve the model's performance on minority classes. In this experiment, the synthetic minority oversampling technique (SMOTE) was utilized to tackle imbalanced data by generating synthetic samples of the minority class. As a result, both data types in the label will be assigned the same value for consistency. During this stage, the news article and the associated label will be the training data. The model will learn the relationship between the target label and the textual data. In this stage, we will use four classification

models, which are Naïve Bayes, Random Forest, SVM, and Logistic Regression. Naïve Bayes Naïve Bayes is a linear classifier rooted in the Bayesian Theorem. In understanding Naïve Bayes, the fundamental of Bayes' rule needs to be understood, which is the posterior probability. The posterior probability indicates the likelihood that a given instance belongs to a particular class. Random Forest is an ensemble learning. It has a simpler structure than a similar method by applying all the base learners to a decision tree. By building multiple decision trees and combining their predictions, it will increase the accuracy and robustness. SVM Support Vector Machine (SVM) was first introduced in the mid-1990s by Vapink. The fundamental method of SVM involves mapping the original data space, applying a nonlinear transformation to create a high-dimensional feature space, and then identifying the optimal linear classification surface in this transformed space. In other words, it will search for the hyperplane that has the best degree to separate the data points of the two classes of Logistic Regression. Logistic Regression is an inductive learning algorithm that belongs to the group of regression. Logistic regression will focus on describing the relationship between a discrete response variable and explanatory variables. This is different from linear regression, where the response variable is continuous. In this stage, a percentage split of 70:30 and 80:20 has been used to split the data. This stage will give an insight into the strengths and weaknesses of the model, through the evaluation using a few performance measures, which are accuracy, precision, and F1-Score.

4. RESULTS

The results of text classification based on some evaluation metrics during the training and testing process and by using new datasets.

A. Data Analysis and Modeling

The result for the two main phases of the study as below:

- Data Analysis: Three local news stories have had 28,410 data points stolen from them in total. The following is a breakdown of the quantity of data points that were taken from local newspapers: The Sun Daily was at 5102, The Star was at 12243, and MalayMail was at 11065. This whole set of information is being exported to a CSV file. Following pre-processing, there are 5157 total data points for Type 1 cybercrime and 5401 total data points for Type 2 cybercrime. Type 1 is defined as a crime against the machine, while Type 2 is defined as a crime using the machine.
- The Star, MalayMail, and The Sun Daily were the sources of most of the data points, according to the data analysis review. This distribution of data points is important to consider when training and evaluating the model. The breakdown of data points for Type 1 and Type 2 cybercrime will also be crucial in understanding the prevalence of different types of cybercrime in local news articles. This information will guide the training process and help identify patterns and trends in cybercrime reporting.
- Model Evaluation: To determine which model worked best, four classifiers were tested. Every classifier employed a pair of distinct feature extractors. The information was divided into two separate percentages: 80:20 and 70:30, as shown in table II.

TABLE II. ACCURACY

Classifier	**TF-IDF**	**TF-IDF**	**Word2vec**	**Word2vec**
	70:30	***80:20***	***70:30***	***80:20***
Naive Bayes	88.65	88.15	73.71	73.48
SVM	92.35	92.5	79.05	78.53
Logistic Regression	91.76	92.18	78.53	78.53
Random Forest	91.76	92.18	78.53	78.53

In comparing the performance metrics of different machine learning algorithms on the dataset, it is evident that SVM consistently outperforms Naive Bayes, Logistic Regression, and Random Forest in terms of accuracy, F1-score, and precision. The accuracy of SVM ranges from 92.35% to 92.5%, while Naive Bayes, Logistic Regression, and Random Forest have accuracies in the range of 73.71% to 91.76%. Similarly, the F1-scores and precision values of SVM are higher compared to the other algorithms, indicating its superior performance in this particular task. These results suggest that SVM may be the most suitable algorithm for this dataset, given its higher accuracy and precision. Furthermore, when considering the computational efficiency of the algorithms, SVM also stands out as it has a relatively shorter training time compared to Naive Bayes, Logistic Regression, and Random Forest. This makes SVM not only a more accurate choice but also a more efficient one for this specific dataset. Additionally, the robustness of SVM in handling noisy data and outliers further solidifies its position as the top-performing algorithm in this study. The combination of high accuracy, precision, and computational efficiency makes SVM the ideal choice for this particular classification task. This result also supports [29] research, which shows Random Forest is a great classifier to be used,

though much research shows Random Forest is not the best choice since its accuracy is moderate. SVM, which comes in second, shows high accuracy, which is supported by the previous research [24].

- To determine which model worked best, four classifiers were tested. Every classifier Classification of new datasets: An untrained label and the tested models were used to test the algorithm model. The 70:30 % split will serve as the basis for the parameter. There are 2640 unlabeled data points in the test set. This experiment confirms the previous one, which found that Random Forest with TF-IDF feature extractor yields the best result while Naïve Bayes with Word2Vec feature extractor yields the lowest, as shown in Table III.

TABLE III. ACCURACY

Classifier	Feature Extractor	Accuracy (%)
Naïve Bayes	TF-IDF	88.65
	Word2vec	62.39
SVM	TF-IDF	93.18
	Word2vec	73.11
Logistic Regression	TF-IDF	92.54
	Word2vec	72.88
Random Forest	TF-IDF	94.58
	Word2vec	74.45

This demonstrates that the choice of feature extractor has a significant impact on the accuracy of the classifiers. Random Forest with TF-IDF feature extractor outperforms the other combinations, while Naïve Bayes with Word2vec feature extractor performs the worst. This information will be crucial in determining the best approach for future experiments and model testing. Additionally, further analysis will be conducted to understand the reasons behind these varying results and to optimize the performance of the classifiers.

B. Data Analysis and Modeling

The result for the two main phases of the study as below:

- To gain a deeper understanding of the results obtained from the model validation evaluation, exploratory data analysis will be conducted. This will involve creating visualizations such as word clouds to identify patterns and trends within the data. By analyzing the frequency of words and phrases within the text data, we can gain insights into the key features that are driving the performance of the classifiers. This information will be instrumental in making informed decisions for future experiments and model optimization. The word cloud of Cybercrime Type1 news displays prominent words like data, company, attack, hacker, and security, as shown in Fig.1. These words are commonly linked to Type 1 cybercrime news. Additionally, words such as China and Facebook indicate local Type 1 cybercrime news related to these entities.

Figure 1. Example of a figure caption. (figure caption)

Continuing to analyze the word cloud, we can see that the frequency of words such as data, company, attack, hacker, and security indicates the common themes and topics discussed in Type 1 cybercrime news. The presence of words like China and Facebook suggests that there may be specific incidents or events involving these entities that are of interest in the cybercrime landscape. By examining these patterns and trends, we can better understand the current state of cyber threats and potentially identify areas for further research and investigation. This insight can be valuable in developing strategies to prevent and mitigate cyber-attacks in the future. In cybercrime type 2 news, the presence of words like police, victim, and scam indicates their association with this type of cybercrime, concerning Fig. 2. The Word Cloud also shows the words' bank account, Malaysia, 'datuk', and woman are frequently mentioned in cybercrime type 2 news. The Word Cloud raised questions and prompted a deeper analysis of the frequent mention of specific words in both cybercrime type 1 and type 2 news.

Figure 2. Example of a figure caption. (figure caption)

C. Topic Modeling

Identifying hidden themes in the text is a method used for uncovering underlying subjects. In this study, we utilized topic modeling to identify topics related to types 1 and 2. The topic modeling is conducted using a method called latent direct allocation (LDA). The number of topics with the highest coherent score indicates the optimal number of topics present in the text. Through this analysis,

we hope to gain a better understanding of the common themes and issues present in both types of cybercrime news. By utilizing topic modeling and LDA, we aim to uncover the underlying topics that are prevalent in these news articles and shed light on the key factors driving cybercrime in today's digital landscape. This research will provide valuable insights into the motivations and tactics of cybercriminals, ultimately helping to inform strategies for the prevention and mitigation of cyber threats. For example, by analyzing a dataset of cybercrime news articles using topic modeling, researchers may discover prevalent themes such as phishing scams, ransomware attacks, and data breaches. By identifying these common topics, they can better understand the emerging trends and patterns in cybercrime activity. This information can then be used to develop targeted interventions and strategies to combat cyber threats effectively.

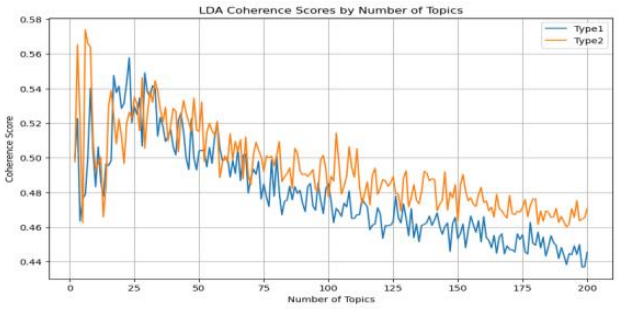

Figure 3. Example of a figure caption. (figure caption)

Based on Figure 3, 200 topics have been tested for topic discovery in the study. For cybercrime type 1 news, the highest coherent score of 23 is associated with the discovery of topics. However, the results also reveal the presence of many overlapping topics, which can affect the outcome. Therefore, 8 topics are determined to be the most suitable number for analysis. It has the highest coherent score with the least amount of topic overlap. In the case of cybercrime type 2, 6 topics are considered the most appropriate for exploration. It scores the highest and has the least overlap with other topics. For topic interpretation, it is essential to utilize the language model (LM), specifically a generative-pretrained transformer (GPT), known for its ability to generate text based on extensive training data. The pre-trained in the word refers to the model having been trained using a huge dataset. The GPT has been largely used for natural language processing applications. Referring to the study conducted by [31], interpretation can be effectively performed using ChatGPT, which has demonstrated satisfactory results. Topic interpretation: Cybercrime Type 1 news For Cybercrime Type 1 news, topic 3 shows the most prevalent topic, while topic 6 is the least prevalent. Topic 3 in the Cybercrime Type 1 news category seems to be the most prevalent, indicating that it is a significant focus within this particular area of cybercrime. On the other hand, topic 6 appears to be the least prevalent, suggesting that it may not be as commonly discussed or reported on about Cybercrime Type 1 incidents. This information can help researchers and analysts better understand the trends and patterns within this specific subset of cybercrime, allowing for more targeted and informed responses to emerging threats.

- Topic Interpretation Cybercrime Type1 News

Based on Table 4, Topic 3 in the Cybercrime Type 1 news category seems to be the most prevalent, indicating that it is a significant focus within this particular area of cybercrime. On the other hand, topic 6 appears to be the least prevalent, suggesting that it may not be as commonly discussed or reported on about Cybercrime Type 1 incidents. This information can help researchers and analysts better understand the trends and patterns within this specific subset of cybercrime, allowing for more targeted and informed responses to emerging threats.

- Topic Interpretation Cybercrime Type2 News

For Cybercrime Type2 news, topic 1 shows the most prevalent topic while topic 6 is the least prevalent, as shown in Table 5. For example, a recent ransomware attack targeted a major hospital, encrypting patient records and demanding a large sum of money for decryption. This resulted in delays in patient care and potential breaches of sensitive information. In another instance, a DDoS attack on a popular e-commerce website caused it to crash during a major sale, resulting in significant financial losses for the company.

TABLE IV. CYBERCRIME TYPE1 NEWS TOPIC IDENTIFICATION

Topic	Title	Keywords	Interpretation
1	Cybersecurity Attack	'attack', 'security',' hacker', 'cyber', 'agency', 'cybersecurity', 'network', 'company', 'ransomware', 'hacking	Discuss on cyber threats and efforts to secure the network by agencies
2	Data Breach and Personal Security	'data',' breach', 'personal', 'security', 'online', 'customer', 'service', 'cyber', 'malaysia', 'digital'	Discussion on incident where personal information had been breach. The word Malaysia indicates the topic focus on the region of Malaysia.

3	China's Technology and Markets	'china', 'market', 'cent',' technology', 'trade',' chinese', 'global', 'company', 'data', 'share'	Discuss on the China role in technology market which has probability include data and information sharing
4	Legal Enforcement	'court', 'police', 'phone', 'woman', 'law', 'hacking', 'authority', 'lawyer', 'charge', 'surveillance'	Discussion on legal action in the context of cybercrime made by law enforcement
5	Device Security and Malware	'device', 'user', 'security', 'malware', 'apple', 'password', 'account', 'email' ,'software', 'apps'	Discusses on the protection of electronic devices like smartphones from any threat of malware
6	Social Media and Data Privacy	'facebook', 'data', 'user', 'twitter', 'account', 'privacy', 'company', 'platform', 'social medium', 'firm'	Discuss on how the social medium company handling social medium user's data
7	Cybersecurity in Political Interference	'trump', 'election', 'russia', 'russian', 'campaign', 'intelligence', 'president', 'committee', 'moscow', 'putin'	Discuss on the topic of foreign interference and cyber threats in a country's election.
8	Cryptocurrency and Security	'uber', 'cryptocurrency', 'hacker', 'security', 'bitcoin', 'money', 'crypto', 'company', 'data', 'exchange',	Discuss on security challenge on companies that involve in cryptocurrency exchanges.

TABLE V. CYBERCRIME TYPE2 NEWS TOPIC IDENTIFICATION

Topic	Title	Keywords	Interpretation
1	Online frauds	'victim', 'rm', 'police', 'scam', 'account', 'bank', 'money', 'suspect', 'public', 'online'	Discuss on individuals that been victim of various online frauds with the law enforcement attempt to mingle this issue
2	Financial transaction on business	'online', 'financial', 'cent', 'business', 'customer', 'market', 'game', 'industry', 'money', 'sex'	Discussion on monetary aspect of business in online activities.
3	Inappropriate content for children	'child', 'online', 'content', 'law', 'user', 'platform', 'internet', 'pornography', 'sexual', 'facebook	Discuss on laws and measure on protecting children from inappropriate content.
4	Legal action against sexual offences	'child', 'court', 'woman', 'charge', 'sexual', 'pornography', 'sex', 'victim', 'kelly', 'abuse'	Discussion on law enforcement for individual that commit sexual content especially those that involve with children
5	Malaysian and international scam	'malaysian', 'victim', 'scam', 'police', 'malaysia', 'job', 'chinese', 'cambodia', 'authority', 'syndicate'	Discuss on Malaysian that been victim of international scam. This involved the movement of cross-border criminals.
6	Police operation against online gambling	police', 'gambling', 'online', 'investigation', 'syndicate', 'activity', 'arrested', 'raid', 'illegal', 'operation'	Discussion on police efforts on combating illegal online gambling

5. CONCLUSIONS

Based on the research results presented in this article, the following main conclusions can be drawn.

- Strength and limitation: This study's strength lies in determining the best classification model for cybercrime terminology. The classification of cybercrime can give news media an advantage in classifying cybercrime topics in their news. This research also indicates that Random Forest remains a viable classifier, despite not always yielding the optimal results. The limitations of this research are limited to two types of cybercrime terminology. As cybercrime classification is expected to evolve, further research is necessary to classify emerging terminologies. Another limitation is the data collected. English news on cybercrime is typically less prevalent in publication compared to Malay news. Furthermore, the data utilized for dashboard visualization is not updated in real time.
- Future works: A few suggestions are in order regarding this project. The first suggestion is to broaden the study of Malay because this nation has seen many Malay news articles on cybercrime published. The second suggestion is to include the most recent cybercrime taxonomy in the classification of cybercrime. The taxonomy of cybercrime might change. Such as, the five categories are the most recent additions to the Council of Europe's (COE) Convention on Cybercrime's classification scheme; nonetheless, it will be extremely difficult to locate news items that discuss these five categories. This research may always be expanded to include the analysis of social media texts like threads and Twitter, and it can be updated. Furthermore, the victim of

cybercrime may benefit from additional textual data analysis.

Text classification is the first step in educating the public about the definition of cybercrime, according to the scientific significance of this study. The research's result will educate and enlighten the public on news pertaining to cybercrime. Combating cybercrime is an arduous task. Since combating cybercrime is a shared responsibility, educating the public is the first step in combatting crime. The examination of cybercrime-related materials by public bodies can serve as a strategic planning tool for the factors and subjects associated with cybercrime. Therefore, Malaysia's internet will be safer with a well-thought-out strategic plan to combat cybercrime. Media publishers will find it easier to post news based on the classification of cybercrime, which will make it easier for readers to access it.

ACKNOWLEDGEMENT

The authors would like to express the gratitude to College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

REFERENCES

[1] K. Phillips, J. C. Davidson, R. R. Farr, C. Burkhardt, S. Caneppele, and M. P. Aiken, *Conceptualizing Cybercrime: Definitions, Typologies and Taxonomies*. Forensic Sciences, vol. 2, no. 2, pp. 379–398, Apr. 2022, doi: 10.3390/forensicsci2020028.

[2] F. Zolkepli and R. Vethasalam, *Scammers' adaptability has led to 50% increase in cybercrime over past two years, says IGP*. The Star, Sep. 26, 2022. https://www.thestar.com.my/news/nation/2022/09/26/scammers039-adaptability-has-led-to-50-increase-in-cybercrime-over-past-two-years-says-igp

[3] Bernama, *Almost RM600 million lost to cyber crime in 2022*. NST Online, Jan. 14, 2023. https://www.nst.com.my/news/nation/2023/01/870171/almost-rm600-million-lost-cyber-crime-2022

[4] M. A. Pitchan, S. Z. Omar, and A. H. Ahmad Ghazali, Amalan Keselamatan Siber Pengguna Internet terhadap Buli Siber. Pornografi, E-Mel Phishing dan Pembelian dalam Talian (Cyber Security Practice Among Internet Users Towards Cyberbullying, Pornography, Phishing Email and Online Shopping. Jurnal Komunikasi: Malaysian Journal of Communication, vol. 35, no. 3, pp. 212–227, Sep. 2019, doi: 10.17576/jkmjc-2019-3503-13.

[5] S. Z. Omar, K. Kovalan, and J. Bolong, *Effect of Age on Information Security Awareness Level among Young Internet Users in Malaysia*. International Journal of Academic Research in Business and Social Sciences, vol. 11, no. 19, Dec. 2021, doi: 10.6007/ijarbss/v11-i19/11733.

[6] Mokayed, H., Quan, T.Z., Alkhaled, L. and Sivakumar, V., *Real-time human detection and counting system using deep learning computer vision techniques*. Artificial Intelligence and Applications, Vol. 1, No. 4, pp. 221-229, doi:10.47852/bonviewAIA220239

[7] Khalid, M., Yusof, R. and Mokayed, H., *Fusion of multi-classifiers for online signature verification using fuzzy logic inference*. International Journal of Innovative Computing, 2011, 7(5), pp.2709-2726, doi: 10.1023/IJIC/122343/

[8] Mokayed, H., Clark, T., Alkhaled, L., Marashli, M.A. and Chai, H.Y, *On restricted computational systems, real-time multi-tracking and object recognition tasks are possible*. IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2022, pp. 1523-1528, doi: 11.2083/IEEE/11811.

[9] A. Palassis, C. P. Speelman, and J. A. Pooley, *An Exploration of the Psychological Impact of Hacking Victimization*, SAGE Open, vol. 11, no. 4, p. 215824402110615, Oct. 2021, doi: 10.1177/21582440211061556.

[10] C. Donalds and K.-M. Osei-Bryson, *Toward a cybercrime classification ontology: A knowledge-based approach*, Computers in Human Behavior, vol. 92, pp. 403–418, Mar. 2019, doi: 10.1016/j.chb.2018.11.039.

[11] Bernama, *Perlu definisi jelas 'keselamatan siber' dalam penggubalan rang undang-undang*, *Astro Awani*, Jun. 23, 2023. https://www.astroawani.com/berita-malaysia/perlu-definisi-jelas-keselamatan-siber-dalam-penggubalan-rang-undangundang-424260 (accessed Mar. 14, 2024).

[12] V. Švábenský, P. Čeleda, J. Vykopal, and S. Brišáková, *Cybersecurity knowledge and skills taught in capture the flag challenges*, Computers & Security, vol. 102, p. 102154, Mar. 2021, doi: 10.1016/j.cose.2020.102154.

[13] H. S. Brar and G. Kumar, *Cybercrimes: A Proposed Taxonomy and Challenges*, Journal of Computer Networks and Communications, vol. 2018, pp. 1–11, 2018, doi: 10.1155/2018/1798659.

[14] R. Ahmad and R. Thurasamy, *A Systematic Literature Review of Routine Activity Theory's Applicability in Cybercrimes*, Journal of Cyber Security and Mobility, Jun. 2022, **Published**, doi: 10.13052/jcsm2245-1439.1133.

[15] I. H. Chipa, J. Gamboa-Cruzado, and J. R. Villacorta, *Mobile Applications for Cybercrime Prevention: A Comprehensive Systematic Review*, International Journal of Advanced Computer Science and Applications, vol. 13, no. 10, 2022, doi: 10.14569/ijacsa.2022.0131010.

[16] M. Hijji and G. Alam, A Multivocal Literature Review on Growing Social Engineering Based Cyber-Attacks/Threats During the COVID-19 Pandemic: Challenges and Prospective Solutions, IEEE Access, vol. 9, pp. 7152–7169, 2021, doi: 10.1109/access.2020.3048839.

[17] J. An and H.-W. Kim, *A Data Analytics Approach to the Cybercrime Underground Economy*, IEEE Access, vol. 6, pp. 26636–26652, 2018, doi: 10.1109/access.2018.2831667.

[18] M. G. Almatar, H. S. Alazmi, L. Li, and E. A. Fox, *Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait*, ISPRS International Journal of Geo-Information, vol. 9, no. 12, p. 702, Nov. 2020, doi: 10.3390/ijgi9120702.

[19] [19] K. Toeppe, H. Yan, and S. K. W. Chu, *Diversity, Divergence, Dialogue*, Springer Nature, 2021. [Online]. Available: http://books.google.ie/books?id=wcIkEAAAQBAJ&pg=PR4&dq=978-3-030-71305-8&hl=&cd=1&source=gbs_api

[20] R. Shikder, U. Talukder, and O. Islas, *A Study on Cybersecurity News Coverage in Bangladeshi Newspapers*, Razón y Palabra, vol. 25, no. 112, Jan. 2022, doi: 10.26807/rp.v25i112.1911.

[21] M. G. Almatar, H. S. Alazmi, L. Li, and E. A. Fox, *Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait*, ISPRS International Journal of Geo-Information, vol. 9, no. 12, p. 702, Nov. 2020, doi: 10.3390/ijgi9120702.

[22] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, *Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach*, Applied Soft Computing, vol. 139, p. 110226, May 2023, doi: 10.1016/j.asoc.2023.110226.

[23] K. Kanakogi et al., Tracing CVE Vulnerability Information to CAPEC Attack Patterns Using Natural Language Processing Techniques, Information, vol. 12, no. 8, p. 298, Jul. 2021, doi: 10.3390/info12080298.

[24] A. V. Prabhu, M. Jefiya, J. D. Joseph, T. Sunny, and C. M. Abraham, *Cyber Complaint Automation System*, Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Jan, 2023, doi: 10.1109/accthpa57160.2023.10083354.

[25] V. Mahor, R. Rawat, S. Telang, B. Garg, D. Mukhopadhyay and P. Palimkar, *Machine Learning based Detection of Cyber Crime Hub Analysis using Twitter Data*, IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-5, doi: 10.1109/GUCON50781.2021.9573736.

[26] E. B. B. Palad, M. S. Tangkeko, L. A. K. Magpantay, and G. L. Sipin, *Document Classification of Filipino Online Scam Incident Text using Data Mining Techniques*, International Symposium on Communications and Information Technologies (ISCIT), Sep. 2019, doi: 10.1109/iscit.2019.8905242.

[27] B.-M. Hsu, *Comparison of Supervised Classification Models on Textual Data*, Mathematics, vol. 8, no. 5, p. 851, May 2020, doi: 10.3390/math8050851.

[28] J. Ahmed and M. Ahmed, *Online news classification using machine learning techniques*, IIUM Engineering Journal, vol. 22, no. 2, pp. 210–225, Jul. 2021, doi: 10.31436/iiumej.v22i2.1662.

[29] A. V. S. Siva Rama Rao, D. Ganga Bhavani, J. Gopi Krishna, B. Swapna, and K. Rama Sai Varma, *Domain Text Classification Using Machine Learning Models*, Lecture Notes in Networks and Systems, pp. 573–582, 2022, doi: 10.1007/978-981-16-7657-4_46.

[30] S. Gordon and R. Ford, *On the definition and classification of cybercrime*, Journal in Computer Virology, vol. 2, no. 1, pp. 13–20, Jul. 2006, doi: 10.1007/s11416-006-0015-z.

[31] M. Mansurova, *Topic Modelling using ChatGPT API - Towards Data Science*, Medium, Oct. 05, 2023. https://towardsdatascience.com/topic-modelling-using-chatgpt-api-8775b0891d16

Nor Muhammad Farhan Nor Muhamad Nizam: Bachelor's Degreee in Information System (Hons.) Intelligent System Engineering from Universiti Teknologi MARA …

Sofianita Mutalib is currently Associate Professor at the School of Computing, College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Shah Alam. …

Mohamad Yusof Darus: Mohamad Yusof Darus holds the position of Associate Professor at the School of Computing, College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Shah Alam. His current research primarily centers on Cybersecurity, with a specific focus on areas such as SQL Injection Attacks, DDoS Attacks, and related topics. His scholarly contributions have been featured in Scopus-indexed journals.

Azlan Ismail is currently Associate Professor at the IBDAAI, Universiti Teknologi MARA (UiTM) Shah Alam. …

Hamam is working as a senior lecturer with Prof. Marcus Liwicki in the EISLAB Machine Learning at Luleå tekniska universitet (Luleå University of Technology), Sweden. His main research interests are in the field of vehicle intelligence system (VIS) and AI for education. He was previously working as a Senior staff researcher at MIMOS, Kuala Lumpur, Malaysia. His research interests include machine learning, natural language processing, image processing and system design in real-time environment… …

Shuzlina Abdul Rahman is currently Associate Professor at the School of Computing, College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Shah Alam. …