Detection of Harassment Toward Women in Twitter During Pandemic Based on Machine Learning

Wan Nor Asyikin Wan Mustapha[1], Norlina Mohd Sabri[2]*,
Nor Azila Awang Abu Bakar[3], Nik Marsyahariani Nik Daud[4], Azilawati Azizan[5]
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu
Kampus Kuala Terengganu, 21080 Kuala Terengganu, Terengganu, Malaysia[1, 2, 3, 4]
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Perak
Kampus Tapah, 35400 Tapah Road, Perak, Malaysia[5]

***Abstract*—Harassment is an offensive behavior, intimidating and could cause discomfort to the victims. In some cases, the harassments could lead to a traumatic experience to the vulnerable victims. Currently, the harassments towards women in social media have become more daring and are rising. The increasing number of the social media users since the Covid-19 pandemic in 2020 might be one of the factor. Due to the problem, this research aims to assist in detecting the harassment sentiments toward women in Twitter. The sentiment analysis is based on a machine learning approach and Support Vector Machine (SVM) has been chosen due its acceptable performance in sentiment classification. The objective of the research is to explore the capability of SVM in the detection of harassments toward women in Twitter. The research methodology covers the data collection using Tweepy, data preprocessing, data labelling using TextBlob, feature extraction using TF-IDF vectorizer and dataset splitting using the Hold-Out method. The algorithm was evaluated using the Confusion Matrix and the ROC analysis. The algorithm was integrated with the Graphical User Interface (GUI) using Streamlit for ease of use. The implementation of the SVM algorithm in detecting the harassments toward women was successful and reliable as it achieved good performance, with 81% accuracy. The recommendations for the SVM model improvement is to train the dataset of other languages and to collect the Twitter data regularly. The performance of SVM would also be compared with other machine learning algorithms for further validations.**

***Keywords*—Harassment; women; detection; twitter; SVM**

I. Introduction

Sentiment analysis is a computational process of identifying and categorizing opinions from a text to extract a particular attitude or belief toward a specific topic. It works by depending on Natural Language Processing (NLP) and implementing a machine-learning algorithm that aims to classify an opinion expressing a positive or negative opinion or sentiment [1]. NLP is a design and implementation of models, systems, and algorithms used for understanding human-like language to solve problems [2]. In the rapid development of social media communities, sentiment analysis has become a hot research topic in NLP [3]. It is crucial to gain deeper insight from public opinion, especially on social media. Many people use the internet to express opinions, thus this situation enables the monitoring of public sentiment for many purposes such as analyzing customers' preferences in businesses and predicting people views or actions on certain important matters. The usage of social media has been increased since the Covid-19 pandemic due to the world wide lock down. People have been more engaged to the internet and the social media has also become the platform for distributing information, besides expressing thoughts.

Harassment covers behaviours of an offensive nature, referring to behaviours that appear to be disturbing, upsetting or threatening. Day by day, cyber harassment toward women is becoming more prevalent and needs to be taken more seriously as it could leave a traumatic impact on an individual. In order to alleviate this problem, many sentiment analysis research have been done to detect the sexual or harassment sentiment on text. The reason sentiment analysis is widely used in analyzing cyber harassment towards women is that it is a reliable and most effective method for analyzing text content. Many harassers become more active anonymously in social media, posting sexist comments and sexist jokes. During the pandemic, these kind of comments could be worst and become increasing since everybody could only meet or socialize over the internet. Sexual harassment can be found in many forms, whether verbally, psychologically, physically, gestural and visually. Exposure to sexual harassment is highly gendered, such that girls are equally or more likely to experience cross-gender harassment, while boys are more likely to experience same-gender harassment [4]. One of the barriers in reporting sexual harassment is victim-blaming, where someone places the responsibility on the victim instead of the person who harmed them. This situation causes difficulties for the victim to come forward and report the problem. This can cause negative impacts on mental health, such as inability to focus, fear, career seatback, and else.

The UN Women has reported a global rise in online harassment of women and girls since the onset of the corona virus [5]. The problem is that any people can communicate anonymously and instantaneously in the virtual world, providing the optimal climate for harassment to transpire [6]. The situations become uncontrollable over time, leading to an uncomfortable and dangerous environment especially to the vulnerable victims. Since cyber harassment towards women, especially sexual harassment cases, is proliferating due to a lack of oversight by the authorities, action needs to be taken to exterminate this behavior. Using sentiment analysis is extremely useful in social media monitoring as it allows

gaining an overview of the wider public opinion behind particular topics. Hence, the sentiment analysis to detect harassment on women in social media needs to be initiated to identify harassed and hated messages sent towards women. Sentiment analysis is an ideal technique for analyzing Twitter's tweets to detect the harassment word. As the data in Twitter continues to grow over time, it becomes a suitable platform for sentiment analysis. The factor that makes Twitter a suitable platform for sentiment analysis is that it has varied from regular people, celebrities, politicians, and even companies.

This study is proposing the machine learning based sentiment classification to detect harassments toward women based on Support Vector Machines (SVM). This research is intended to classify tweets into categories, whether the tweet is "Harassment" or "Not Harassment". SVM has been chosen for this research as the algorithm could produce results with excellent accuracy in most studies [7]. The performance of SVM has been promising in the sentiment classification problems. The main objective of this research is to explore the performance of SVM in the sentiment classification of harassments toward women based on Twitter data. The ability of sentiment analysis to distinguish positive and negative sentiments is expected to help the community and also authorities in detecting harassments toward women. Through sentiment analysis, the harassment in social media can be detected and help the authorities monitor people more closely so that harassment cases can be reduced. Thus, it brings a safer community around the world, especially to a woman.

This paper is organized into five main sections which are the Introduction in Section I, Literature Review which contains brief explanation on SVM and the similar works in Section II, Materials and Methods in Section III, Results and Discussion in Section IV and finally Section V concludes the paper.

II. Literature Review

A. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that builds a model by learning from a known class (labelled training data) [8]. SVM is mighty at recognizing a subtle pattern in a complex dataset [9]. SVM works by creating a decision between two classes to predict labels from one or more feature vectors. A boundary separates it called a hyperplane [9]. The hyperplane is oriented so that it is as far away from the closest data points from each class as possible. Support vector refers to the closest point from each of the classes. The key for the SVM algorithm is finding the correct hyperplane position to classify the class [8]. SVM can generate a model based on a small training set while ensuring lower error levels in the test. In a linearly separable class, SVM seeks a hyperplane that separates the two-class vectors, which is a positive and negative class with the greatest margin in linearly separable cases [10]. Then for more complex problems, SVM solves the problem by using additional features such as kernel trick [8]. The techniques purposely convert low dimensional input space into higher dimensional space to solve not separable problems. The data can easily be classified into different classes by using a hyperplane. SVM has many different types of kernel functions; the popular kernel functions are the linear kernel, polynomial kernel, and Radial basis function (RBF) kernel [11]. This research has been implementing the linear kernel due to its suitability with this sentiment classification problem. Support Vector Machine (SVM) is one of the most efficient machine learning algorithms [12]. A study of mobile network prediction by [13] proven that SVM does provide high accuracy in prediction.

B. Similar Works

There are several similar works that studies harassment towards women in social media platforms. The research contains a similar objective and problem but uses a different algorithm. Table I shows similar projects comprising title, project objectives, year, the problem faced, research result and references.

Based on Table I, the first similar work is Understanding the silence of sexual harassment victims through the #WhyIDidntReport movement by [14], where the project aims to identify tweets that disclose a reason for remaining silent after sexual violence. Sexual violence has become more severe across the globe, and many women have been through this experience, urge to the need for research to support and identify the factor of sexual violence. This project used a few algorithms, and SVM outperformed with 92% precision compared to the linear kernel, Random Forest, Naïve Bayes, and Gradient Boosting.

The following project analyses #MeToo hashtagged posts on Twitter by [15]. This project objective is to analyze tweets that reveal patterns and attributes such as emotion and reaction to the hashtag #MeToo. This project is initiated to explore the topics and patterns that lead to sexual harassment. This project uses the approach of modeling aggregate words and generating issues from the particular text.

Analysis of sexual harassment tweet sentiment on Twitter in Indonesia using Naïve Bayes by [16] is the third equivalent work. This research objective is to identify positive and negative sentiment in tweets that lead to sexual harassment within social media. This study is an effort to control and monitor sexual harassment that occurs in social media since the situation leads to a severe impact such as psychological trauma and negatively affects the victims. This project used the Naïve Bayes algorithm and recorded an accuracy of 83 %.

A similar works also study the field of harassment that occurs in social media, which is a study initiated by [17]. The objective is to develop a detection system to detect sexual harassment in text. The problem that arises causes this project's need is the seriousness of the negative and traumatic impacts of cyberbullying. For this project, they used a few different algorithms for the detection. The algorithm that gives the highest accuracy is SGD. Classifier outperformed other algorithms, which are SVC, Multinomial NB, Linear SVC, Decision tree, Random Forest and KNN.

Another research aims to detect hate speech on text data and audio data. This research has implemented SVM, Random Forest (RF), Naive Bayes (NB) and Logistic Regression (LR) to identify bullying and online harassment in the cyberspace. The results have shown that SVM has outperformed other algorithms with the best accuracy of 92.3% [18].

Finally, a research in Bangladesh aims to detect sexual harassments from Bangla texts [19]. The research is motivated by the increasing improper usage of the social media. In this research, CNN-LSTM has outperformed other algorithms. However, SVM has outperformed other machine learning algorithms such as NB, RF, Decision Tree (DT), AdaBoost, SGD, LR and K-Nearest Neighbour.

Based on these similar works, SVM has shown good performance in the research by [14] and [18] with the accuracies of more than 90%. This research has chosen SVM due to its capability and also good performance in solving various other classification problems [20-22]. It is expected that SVM could also produce good results in this sentiment classification problem.

TABLE I. SIMILAR WORKS

No	Title	Objective	Problem	Result	Ref.
1	Understanding the silence of sexual harassment victims through the #WhyIDidntReport movement	To identify tweets that disclose a reason for remaining silent after sexual violence.	Sexual violence has become more severe globally since many women have to go through this experience.	SVM outperform with 92% precision compared to the linear kernel, Random Forest, Naïve Bayes, and Gradient Boosting.	[14]
2	Can women break the glass ceiling?: An analysis of #MeToo hashtagged posts on Twitter	To analyze tweets that reveal patterns and attributes such as emotion and reaction relating to the hashtag #MeToo.	An uprising of online movements on social media (#MeToo) by women worldwide who started to reveal their stories of being sexually harassed along with the #MeToo.	Generates topic form tweets by directly modelling aggregate word co-occurrence.	[15]
3	Analysis of sexual harassment tweet sentiment on Twitter in Indonesia using Naïve Bayes	To identify positive and negative sentiment leading to sexual harassment	The act of sexual harassment has occurred a lot in social media nowadays. This situation causes psychological trauma and negatively affect the victims.	Reaches accuracy of 83%.	[16]
4	Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques	To propose an approach that could be utilized towards developing a detection system to detect sexual harassment in text.	The negative and traumatic impacts of cyberbullying can be severe for society and even lead to absenteeism and suicide.	SGD. Classifier gives the highest accuracy (81%) than SVC, Multinomial NB, Linear SVC, Decision tree, Random Forest, and KNN.	[17]
5	Online Harassment Detection using Machine Learning	To identify bullying and online harassment in cyberspace	Cyberbully causes mental consequences	SVM outperformed RF, NB, LR with 92.3% accuracy	[18]
6	Sexual Harassment Detection using Machine Learning and Deep Learning Techniques for Bangla Text	To detect sexual harassment from Bangla text	Increasing amount of offensive Bangla text in different social media platforms	Deep learning algorithms achieved higher accuracies. However SVM outperform other machine learning algorithms.	[19]

III. MATERIALS AND METHODS

The research was carried out based on six main phases, which were the data collection, data preprocessing, data labelling, feature extraction, classification based on SVM and classifier's performance evaluation. After the performance evaluation, the classifier model was integrated with the graphical user interface to be used by users. The flowchart which briefly demonstrates the system development process is given in Fig. 1.

Based on Fig. 1, the first step is the data collection which involves the scraping of data from Twitter using Tweepy. In this research, a total of 2522 data had been collected and processed. The scraped data or the raw dataset is saved in the CSV file. The raw data is then imported into Google Colaboratory to be preprocessed and transformed into the appropriate form. There are a few steps in the data preprocessing phase which are the duplicate data removal, case conversion, URL removal, punctuation removal, hashtag removal, short word removal, tokenization, stop word removal, POS tag labeling and finally the lemmatization. The following phase is data labeling by using TextBlob. Data will be labelled as positive or negative in this phase.

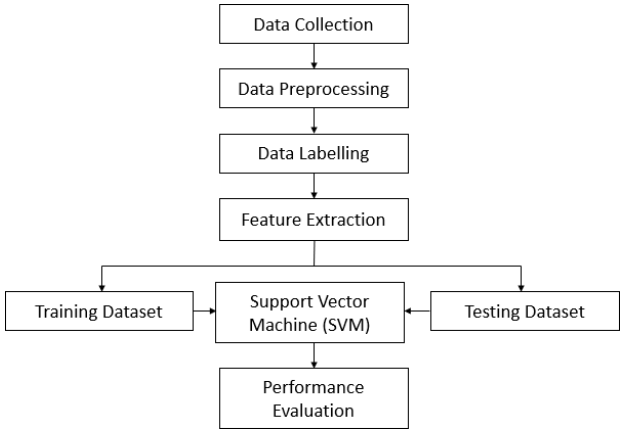

Fig. 1. Flowchart of system development process.

In the feature extraction phase, the dataset will undergo vectorization using TF-IDF. This process enables the dataset to be converted to the format readable by SVM. Then, the dataset will be split into the training and testing dataset using the hold-out method. The dataset will be split into three ratio splits and the best results will be recorded. The performance evaluation phase measures the performance of SVM based on the Confusion Matrix accuracy, recall, precision, F1-Score and also the AUC (Area under the Curve) value. Parameter tuning

is to be done in this phase in order to obtain the best SVM model. Fig. 2 shows the system's architecture which demonstrates the system's overall process starting from the data collection until the SVM model usage by the user. The sentiment classification system will produce the output of "Harassment" or "Not Harassment" based on the user input.

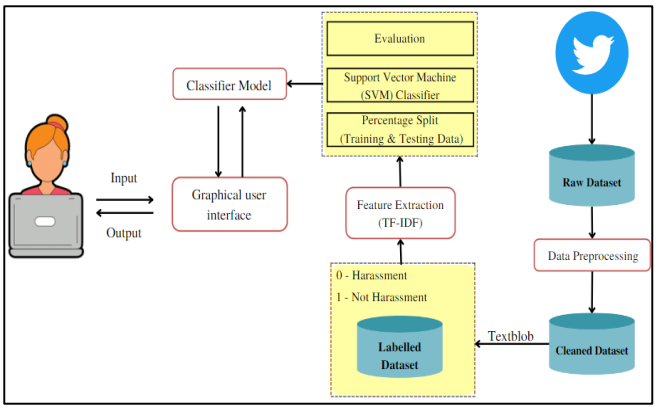

Fig. 2. System architecture design.

A. Data Collection

The data in this research was scrapped using the Twitter Application Programming Interface (API) during the month of March to August 2022. During this time, the world was still in the Covid-19 pandemic. A total of 2522 rows of tweets were scrapped using the Tweepy package. Tweepy is a python library that allows users to access Twitter API. Data collection was done by using eight different keywords, which were #isthisok, #metoo, slut, stupid women, women abuse, women bad, women weak and women whore. These harassment-related keywords were used to get the harassment sentiments as much as possible to train the prototype.

B. Data Pre-processing

The data extracted from Twitter API are unstructured and not uniform. This process is important to ensure reliable and accurate results can be obtained. Reference [23] stated that Pre-processing techniques are used on the target data set to minimize data size and hence boost the system's efficiency. Fig. 3 shows the steps of data preprocessing which include seven stages.

Based on Fig. 3, the first phase is to remove the duplicated data in order to avoid the inconsistent in the prototype developed. Then, the data will undergo the lowercase conversion so that it would be easier to be process. It aids in the maintenance of the consistent flow during NLP activities nd text mining [24]. Afterwards, any URL and links from the dataset will be removed since it is not needed, besides its existence will disturb the efficiency of the prototype developed. This is followed by removing the Twitter handles (@) and punctuation to minimize the data size and increase efficiency.

The next phase is the tokenization and stop word removal. Tokenization is the process of breaking down a raw string into meaningful tokens. The string will be split by referring to non-letter characters such as space, commas, full-stop and other punctuation [25]. The next process is the stop word removal, in which this process eliminates irrelevant words from the dataset in order to provide intelligent patterns or information [25]. After the text is tokenized into words, the word is analyzed one by one using a loop. If that particular word is detected as a stop word, it is removed to save computing time. Natural Language Toolkit (NLTK) can be used to implement the process of stop word removal. NLTK has a pre-built collection of stop words for about 22 languages [26]. The word that matches the word from the NLTK corpus will be removed during the loop. Then, the tokenized word undergoes POS-Tag labelling to label each word to the appropriate part of speech (POS). Part of speech includes nouns, verbs, adverbs, adjectives, pronouns and conjunction. This step differentiates the meaning and weight of every word context. This phase is important because sometimes the same word holds a different importance in different sentences.

The last phase of data preprocessing is the lemmatization. Lemmatization is a process of text normalization in order to maintain uniformity. There are two options for text normalization which are stemming and lemmatization. Stemming is a process that stems the words to its roots by removing the suffixes within the word [26]. While lemmatization is the process of removing inflectional ends from a word and returning the base or dictionary form without changing its meaning [27]. However, lemmatization produces more relevant results compared to stemming [28]. The result generated by lemmatization is a real term in English because it uses corpus to match root forms, which makes it more accurate. Hence, lemmatization is used for text normalization in this research for a better result in classification. Fig. 4 shows the example of the cleaned dataset. During preprocessing phase, the number of data was reduced from 2522 to 1558 after the preprocessing.

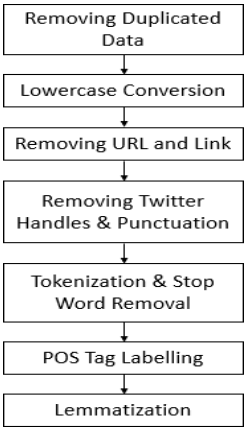

Fig. 3. Pre-processing steps.

index	lemmatized
0	['check', 'april', 'come', 'gender', 'base', 'violence', 'str']
1	['think', 'suppose', 'chair', 'bust']
2	['think', 'suppose', 'chair', 'bust']
3	['check', 'april', 'come', 'gender', 'base', 'violence', 'strategy', 'campaign', 'potentially', 'give', 'planning', 'permission']
4	['watch']
5	['fit', 'campaign', 'amp', '', 'total', 'opposite', 'objectify', 'woman', 'get', 'block']
6	['let', 'test', 'twitteri', 'take', 'microscopic', 'picture', 'cremation', 'remain', 'would', 'love', 'share', 'picture', 'world']
7	['happy', 'see', 'anonobots', 'jump', 'porn', 'good', 'kid', 'trip', 'fail', 'fail', 'racism']
8	['recent', 'crime', 'survey', 'england', 'wale', 'show', 'alarm', 'statistic', 'want', 'speak']
9	['never', 'underestimate', 'handsomely', 'peter', "o'hanorahanrahan"]
10	['comprehensive', 'list', 'trump', "'s", 'sex', 'victim', 'list', 'woman', 'far', 'accuse', 'trump', 'molestation', 'rap']

Fig. 4. Example of cleaned dataset.

C. Data Labelling

The tweets that were extracted from the Twitter do not have corresponding labels [28]. Therefore, the data or tweets need to be labelled before they can be used for training and testing. The APIs that has been used for data labelling was Textblob. Textblob is a Python package for text processing and provides a straightforward API for exploring NLP tasks. Textblob uses Parts-of-Speech (POS) for sentiment labelling [29]. Support Vector Machine (SVM) model performs better when using labels provided by Textblob than other APIs, which are Vivekn, Meaning Cloud and Pattern [30]. Therefore, Textblob was used for data labelling for this research project. The polarity score is as shown in Table II.

This project aims to develop a binary classifier model that only classifies two possible outcomes, whether "Harassment" or "Not Harassment". Therefore, only two class labels are needed to train the classifier model. Due to the reason, sentiment detected as neutral will be removed from the dataset to avoid misdetection. After the neutral sentiment text has been removed, the original number of data has been reduced from 1557 to 1192.

TABLE II. RANGE OF POLARITY

Polarity	*Sentiment Class*
More than 0	Not Harassment (positive)
Less than 0	Harassment (negative)
0	Neutral

D. Feature Extraction

Before the model development, text or word must be converted into the numerical representation which could be understood by SVM to be processed. In this phase, the text words will be converted into numerical vectors. The feature extraction technique that will be implemented is the TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a numerical measure meant to indicate the importance of a word to a document in a certain corpus [31]. A study by [22] shows SVM that uses TF-IDF as a feature extraction technique generates a good and reliable classification result. There are 3 calculations in order to get the numerical presentation of each word which are Term Frequency (TF), Inverse Document Frequency (IDF) and TF-IDF score. Firstly, the Term Frequency (TF) value is calculated using (1). TF represents the frequency of a word in a tweet. The number "n" indicates the number of times that phrase appears in the document.

$$TF = n / \text{Number of term in the document} \quad (1)$$

Then, Inverse Document Frequency (IDF) reflects the frequency on which a term appears in a corpus of tweets. Eq. (2) is used to calculate the IDF value.

$$IDF = \text{No. of documents} / \text{No. of document with term 't'} \quad (2)$$

The last one is to calculate the number TF-IDF score for each word using Eq. (3). Words with a higher score are considered more significant, whereas those with a lower score are considered less essential.

$$TFIDF = TF \times IDF \quad (3)$$

For each new word inserted to the corpus, it will loop to find the word variable in the corpus and calculate their vector value. If the word variable is not found, the TF-IDF vectorizer will recalculate the corpus to find the value of vector for that word. This process continues until all line of data in that dataset is vectorized. The number of column or word variables produces in this phase was 3982.

E. Dataset Splitting

After the data has been vectorized using the TF-IDF, the dataset has been split into two datasets which are the training and the testing dataset. The training dataset was used to train the SVM classifier model, while the testing dataset was used to evaluate the performance of the classifier. The method used to split the dataset was the hold-out method. The hold out method is a method that permanently splits the data into certain percentages. This approach was chosen because it was efficient and easier to implement. In this research, the split ratio was set to 3 splits, which were 90:10, 80:20 and 70:30. 10.

F. SVM Implementation

In this research, after the dataset has been trained and tested, the codes has been deployed as the SVM classifier model. Fig. 5 shows the phases of the SVM implementation in this research. The following sections will explain the process of implementation of SVM.

1) Applying SMOTE to training dataset: SMOTE (Synthetic Minority Oversampling Technique) is a process of handling imbalance dataset by duplicating the minimum data. The labelled data used for training were not balanced, where the dataset contains 498 negative data and 455 positive data. In this case, the positive data was duplicated until it reaches the same value as negative data which was 498. If the imbalanced data is not handled, it can cause bias situation where the classifier always go to the origin, which means the classifier cannot perform the classification efficiently.

Creating Kernel Matrix: Kernel matrix is a set of mathematical functions that is used in SVM. It takes the dataset as training dataset and transforms the data into required form to be learnedby SVM algorithm.

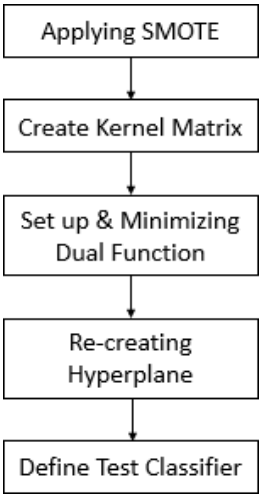

Fig. 5. SVM implementation phases.

2) Set up and minimizing dual function: Setting up and minimizing dual function are done by applying cvxpy package. This is to express complex optimization problem from previous stage into a readable form, so that the solver is able to translate the problem into support vector. In this phase, the value support vector is calculated, higher value indicates the point is more important as a support vector. Support vector is a point used to represent the data. Then it is classified to positive or negative using default hyperplane. This model development used linear SVM and utilized Eq. (4) to classify the weight of support vector.

$$k(X1, X2) = X1 \times X2 \quad (4)$$

3) Re-creating hyperplane: This section is used to recreate the hyperplane or boundaries to effectively classify the support vector. If this phase is ignored, the default hyperplane that has been created in previous phase will be used for classification which is not effective.

4) Define test classifier: In this phase, the classifier will classify each data in the dataset and the results of classification are compared with the original labels from Textblob. The accuracy of the system is determined by comparing the total of correct calculation.

G. Performance Evaluation

Performance evaluation is necessary to see whether the system is fulfilling the objectives and shows an effective classification. The performance metrics used in this research are the Classification Report, Confusion Matrix and the ROC curves. Classification report and Confusion Matrix show the value of accuracy, recall, precision and F1-score. Eq. (5) is used to calculate the classifier's accuracy, Eq. (6) calculates the precision and Eq. (7) calculates the recall value. TP represents the True Positive, TN is the True Negative, FP represents the False Positive and FN is the False Negative.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

Eq. (8) calculates the F1-score, which is the value of the weighted average of precision and recall and it is almost the same as accuracy.

$$\text{F1-Score} = 2((\text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision})) \quad (8)$$

ROC curves is use to illustrate and evaluate the effectiveness of a classification algorithm. Rather than providing only values, it provides graphical representation of the classifier's performance [32]. The performance of classifier is measured by observing the value of AUC (area under ROC curve). The greater the AUC number, the better the classifier's performance (between 0.5-1.0).

IV. Results and Discussion

This section reports the evaluation results which are the Classification Report, Confusion Matrix and the ROC curve. Before the training and testing phase, the data had been balanced using the SMOTE techique. Fig. 6 shows the accuracy before applying SMOTE, while Fig. 7 shows the accuracy of the model after applying SMOTE on the dataset. Based on the figures, there is quite a significant difference in the accuracy values between the model that uses SMOTE and the one which does not. The model that used SMOTE has a higher accuracy and could produce more reliable classification

```
print("Accuracy Model  : ",accuracy)

Accuracy Model  :  0.7857142857142857
```

Fig. 6. Accuracy before SMOTE implementation.

```
print("Accuracy Model  : ",accuracy)

Accuracy Model  :  0.8075313807531381
```

Fig. 7. Accuracy after SMOTE implementation.

In the training and testing phase, there were three split ratios that had been used. The accuracy results from the different percentage splits are shown in the Table III. From the table, the dataset of 80:20 split produces the highest accuracy of 80.75%. Therefore, the percentage split of 80:20 is applied throughout the model development. For splitting of the 1192 data using the 80:20 ratio, 950 data is used to train the classifier while the other 224 data is used to evaluate the model performance.

TABLE III. Split Ratio Results

Percentagesplit	Accuracy	Precision	Recall	F1-Score
90:10	78.33%	77.36%	74.55%	75.93%
80:20	80.75%	78.94%	80.36%	79.64%
70:30	79.05%	78.48%	78.03%	78.26%

A. Classification Report

The Classsification Report was used to report the accuracy, recall, precision and F1-score. The classification was calculated using the Scikit Learn module available on Python. Fig. 8 shows the output of classification report of the model constructed. The accuracy achieved by the classifier is 81%, which is considered as good and acceptable. The precision measures the exactness of the classifier and the value is 79%. This value denotes less false positive obtained by the module. The recall value measured is 80%, which represents less false negative in the classification. Less false negative and the less false positive denote that the system can accurately classify the input given by the user. The value of F1-Score obtained by user is 80%, which indicates the harmonic balance between the recall and precision values. These values can be concluded as good, reliable and had achieved the classifier's objective in solving the sentiment classification problem.

```
Classification report :
              precision    recall  f1-score   support

          -1       0.82      0.81      0.82       127
           1       0.79      0.80      0.80       112

    accuracy                           0.81       239
   macro avg       0.81      0.81      0.81       239
weighted avg       0.81      0.81      0.81       239
```

Fig. 8. Classification report.

B. Confusion Matrix

Fig. 9 shows the Confusion Matrix which represents the value of True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) predicted by the classifier. Confusion matrix is the calculation of the number of right and wrong predictions, which are then summarized with the count values and divided by class. By looking at the confusion matrix plot, it gives an insight of the error made by the classifier model and the type of error being made. Based on Fig. 9, 90 is the value of TP, which indicates the number of not harassing tweets which is predicted correctly, while value 22 represents FP, which is supposedly to be non-harassment tweet but incorrectly predicted as the harassment tweet. The value of 24 represents FN which is the number of harassment tweets that are incorrectly labelled as non-harassment. Lastly, the TN value indicates 103 harassment tweets which are predicted correctly by the classifier. From the results, it can be observed that the classifier has made 46 incorrect predictions out of the 239 predictions made. The classifier only made incorrect predictions for about 19%. Hence, the model still can be concluded as good and reliable as the false prediction percentage is still small.

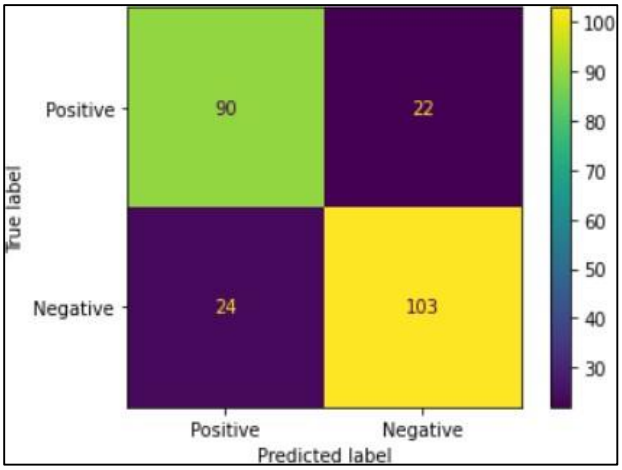

Fig. 9. Confusion matrix.

C. ROC Curve

Fig. 10 shows the ROC curve for the detection of the harassment tweets. The AUC value obtained for this classifier is 0.81, which is near to 1. As the AUC value is approaching value 1, it is proven that the SVM classifier model obtained is reliable and could classify the data correctly. Thus, the 0.81 AUC value demonstrates that the classifier model obtained is capable in distinguishing the classes of "Harassment" and "Not Harassment" in this research.

D. Discussion

Based on the evaluation results obtain in the evaluation phase, the SVM classifier has been able to detect the harassment tweets towards women in this research with good and reliable performance. The finding during the labelling of data has shown that there were more negative (Harassment) tweets than the positive ones. This shows that something has to be done to curb more harassments toward women over the Twitter. Perhaps Twitter itself could block any contents that has bad sentiments towards women. This is crucial as more people are using Twitter nowadays since the pandemic. Fig. 11 shows the word cloud for the negative words that have been analyzed. The word cloud is the visual presentation of the words used in the tweets. The bigger words such as 'stupid' and 'woman' show that the words were the most frequently used in the tweets.

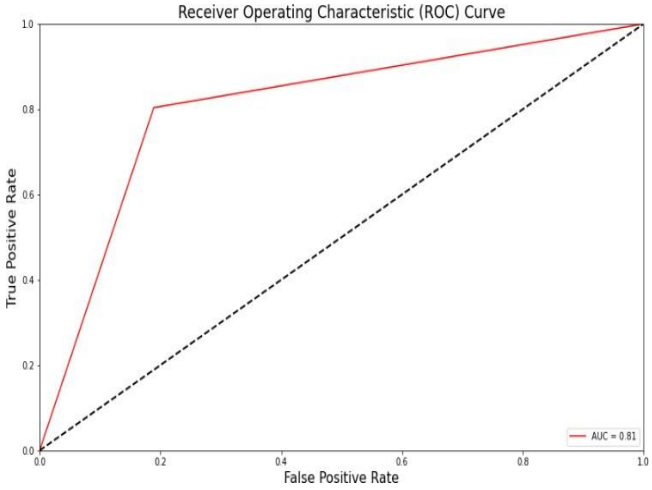

Fig. 10. ROC curve.

Fig. 11. Word cloud for negative words.

V. Conclusion

This research has explored the capability of SVM in the sentiment classification of harrassments toward women based on Twitter data. The accuracy achieved by the SVM classifier was 81%, which indicates the good and acceptable performance of the model. This research could contribute to the community as it help to detect harassments toward women through Twitter. By detecting harassments that occurs within the social media, law enforcement could be made and actions could be taken against the harassers. Apart from legal action, this classifier would also help in detecting and taking down the harassment posts. This could increase the quality of the social media content and make a safer space for women on social media. The limitation associated with this classifier is that it can only analyze English Tweets. It would be better if the classifier could also detect harassments in other languages as

harassments occur in various languages in the world. In future works, besides collecting and training dataset with other languages such the Malay language, data should also be scraped regularly. This is to collect more relevant and latest data in order to produce a better performance machine learning model. The results of SVM classifier model could also later be compared with other machine learning techniques such as the Naive Bayes and the deep learning algorithms.

ACKNOWLEDGMENT

Special gratitude goes to Universiti Teknologi MARA Cawangan Terengganu for the continuous support given towards the advancement of research in the university.

REFERENCES

[1] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," Journal of King Saud University - Engineering Sciences, vol. 30, pp. 330–338, 2018. https://doi.org/10.1016/j.jksues.2016.04.002.

[2] S. Alam, "Applying Natural Language Processing for detecting malicious patterns in Android applications," Forensic Science International: Digital Investigation, vol. 39, 2021. https://doi.org/10.1016/j.fsidi.2021.301270.

[3] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks", Knowledge-Based Systems, vol. 235, 2021. https://doi.org/10.1016/j.knosys.2021.107643.

[4] J. E. Copp, E. A. Mumford, and B. G. Taylor, "Online sexual harassment and cyberbullying in a nationally representative sample of teens : Prevalence , Predictors ,and Consequences," Journal of Adolescence, vol. 93, pp. 202–211, 2021. doi: 10.1016/j.adolescence.2021.10.003.

[5] S. Zalis. (2021). International Day Of The Girl: Helping Make The Internet A Safer Place. https://www.forbes.com/sites/shelleyzalis/2021/10/11/international-day-of-the-girl-helping-make-the-internet-a-safer place/?sh=6aed1f5964df.

[6] D. A. Griffith, P. van Esch, and M. Trittenbach, M, "Investigating the mediating effect of Uber's sexual harassment case on its brand: Does it matter?," Journal of Retailing and Consumer Services, vol. 43, pp. 111–118, 2018. https://doi.org/10.1016/j.jretconser.2018.03.007.

[7] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in Proceedings of the 8th International Conference on System Modeling and Advancement in Research Trends, 2020, pp. 266-270. https://ieeexplore.ieee.org/document/9117512.

[8] S. Ray. (2017). Understanding Support Vector Machine (SVM) algorithm from examples.

[9] S. Huang, N. Cai, P. P. Pacheco, S. Narrasndes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Cancer genomics & proteomics, vol. 15, pp. 41-51, 2018.

[10] W. Zheng and Q. Ye, "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm," in Third International Symoosium on Intelligent Information Technology Application, 2009.

[11] A. Yadav. (2018). Support Vector Machines (SVM). In Analytics Vidhya. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589.

[12] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, "Advantage and Drawback of Support Vector Machine Functionality," in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), 2014.

[13] A. Y. Nikravesh, S. A. Ajila, C. H. Lung, and W. Ding, "Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM," in 2016 IEEE International Congress on Big Data (BigData Congress), 2016.

[14] A. Garrett and N. Hassan, "Understanding the silence of sexual harassment victims through the #Whyididntreport movement," in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 649–652. https://doi.org/10.1145/3341161.3343700.

[15] N. Hassan, M. K. Mandal, M. Bhuiyan, A. Moitra, and S. I. Ahmed, S, "Can women break the glass ceiling?: An analysis of #metoo hashtagged posts on Twitter," in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 653–656. https://doi.org/10.1145/3341161.3343701.

[16] K. Budiman, N. Zaatsiyah, U. Niswah, F. Muhanna, and N. Faizi, "Analysis of Sexual Harassment Tweet Sentiment on Twitter in Indonesia using Naïve Bayes Method through National Institute of Standard and Technology Digital Forensic Acquisition Approach," Journal of Advances in Information System and Technology, vol. 2, 2020.

[17] E. Alawneh, M. Al-Fawa'Reh, M. T. Jafar, and M. A. Fayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," in Proceeding - 2021 International Symposium on Electronics and Smart Devices: Intelligent Systems for Present and Future Challenges, 2021. https://doi.org/10.1109/ISESD53023.2021.9501725.

[18] R. Ahirwar, M. Ajay, N. Sathyabalan and K. Lakshmi, "Online Harassment Detection using Machine Learning," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 1222-1224, doi: 10.1109/ICICT54344.2022.9850516.

[19] M. Islam, M. Rahman, M. T. Ahmed, A. Z. Muhammad Islam, D. Das and M. M. Hoque, "Sexual Harassment Detection using Machine Learning and Deep Learning Techniques for Bangla Text," 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ECCE57851.2023.10101522.

[20] E. U. Haq, J. Huang, H. Xu, K. Li, and F. Ahmad, "A hybrid approach based on deep learning and support vector machine for the detection of electricity theft in power grids," Energy Reports, vol. 7, pp. 349–356. https://doi.org/10.1016/j.egyr.2021.08.038.

[21] P. K. Illa, B. Parvathala, and A. Sharma, "Stock price prediction methodology using random forest algorithm and support vector machine," Materials Today: Proceedings, vol. 56, pp. 1776-1782, 2022. https://doi.org/10.1016/j.matpr.2021.10.460.

[22] Y. Jiang, X. Wang, Z. Zou, and Z. Yang, "Identification of coupled response models for ship steering and roll motion using support vector machines," Applied Ocean Research, vol. 110, pp. 102607. https://doi.org/10.1016/j.apor.2021.102607.

[23] S. Kannan and V. Gurusamy. (2014). Preprocessing Techniques for Text Mining. Available:https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining.

[24] G. Singhal. (2020). Importance of Text Pre-processing. Available: https://www.pluralsight.com/guides/importance-of-text-pre- processing (accessed Jan. 15, 2022).

[25] V. Kalra, "Importance of Text Data Preprocessing & Implementation in RapidMiner," vol. 14, pp. 71–75, 2018. doi: 10.15439/2018KM46.

[26] N. Hardeniya, J. Perkins, N. Joshi, and I. Mathur. (2022). Natural Language Processing: Python and NLTK. Available: https://books.google.com.my/books?hl=en&lr=&id=0J_cDgAAQBAJ&oi=fnd&pg=PP1&dq=nltk+word&ots=lfsruYmxYY&sig=DmnmnSqJqiYga7qnlSvIr0k1ZUY&redir_esc=y#v=onepage&q=nltk%20word&f=true.

[27] V. Balakrishnan and L.-Y. Ethel, "Stemming and Lemmatization: A Comparison of Retrieval Performances," Lecture Notes on Software Engineering, vol. 2, no. 3, pp. 262–267, 2014. doi: 10.7763/lnse 2014.v2.134.

[28] H. Jabeen. (2018). Stemming and Lemmatization in Python. Available: https://www.datacamp.com/community/tutorials/stemming- lemmatization-python.

[29] S. K. Sharma and X. Hoque, "Sentiment predictions using support vector machines for odd-even formula in Delhi," International Journal of Intelligent Systems and Applications, vol. 9, pp. 61– 69, 2017. doi: 10.5815/ijisa.2017.07.07.

[30] M. Pandey, R. Williams, N. Jindal, and A. Batra, "Sentiment analysis using lexicon based approach," in PDGC 2018 - 2018 5th International Conference on Parallel, Distributed and Grid Computing, 2018, pp. 13–18. doi: 10.1109/PDGC.2018.8745971.

[31] P. Huilgol. (2020). BoW Model and TF-IDF For Creating Feature From Text. Available: https://www.analyticsvidhya.com/blog/2020/02/quick-introduction- bag-of-words-bow-tf-idf/.

[32] D. Steen, (2020). Understanding the ROC Curve and AU. Available: https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb.