Cyberbullying Detection in Twitter Using Sentiment Analysis

Chong Poh Theng[1†], Nur Fadzilah Othman[1†], Raihana Syahirah Abdullah[1†], Syarulnaziah Anawar[1†], Zakiah Ayop[1†], Sofia Najwa Ramli[2†]

[1]Information Security Forensics and Computer Networking (INSFORNET),
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka

[2]Faculty of Science and Information Technology
Universiti Tun Hussien Onn
Malaysia

Summary
Cyberbullying has become a severe issue and brought a powerful impact on the cyber world. Due to the low cost and fast spreading of news, social media has become a tool that helps spread insult, offensive, and hate messages or opinions in a community. Detecting cyberbullying from social media is an intriguing research topic because it is vital for law enforcement agencies to witness how social media broadcast hate messages. Twitter is one of the famous social media and a platform for users to tell stories, give views, express feelings, and even spread news, whether true or false. Hence, it becomes an excellent resource for sentiment analysis. This paper aims to detect cyberbully threats based on Naïve Bayes, support vector machine (SVM), and k-nearest neighbour (k-NN) classifier model. Sentiment analysis will be applied based on people's opinions on social media and distribute polarity to them as positive, neutral, or negative. The accuracy for each classifier will be evaluated.

Key words:
Sentiment analysis; Cyberbullying; Twitter; Machine Learning

1. Introduction

With the speedy evolution of the Internet, individuals' communication is no longer restricted by the need to be on-site [1]. With the latest advancement of social media, people have embraced new habits of broadcasting nasty comments or hate speech through social media such as Twitter, Instagram, and Facebook, which eventually lead to cyberbullying. The victims of cyberbullying will suffer from several mental issues, ranging from depression, loneliness, anxiety, and low self-esteem.

A study on cyberbullying prevention concentrates on detecting possibly dangerous information and developing intelligent systems to identify spoken aggressiveness shown with offences and threats. Text mining techniques are among the most encouraging tools applied in aggressive sentiment detection in short texts, such as comments and tweets. [2]. Sentiment analysis can be applied to relate to many different but related problems. Generally, it is used to describe the duty of automatically determining the valence or polarity of a piece of text, whether positive, negative, or neutral [3]. The progression of analysis in text analytics has enabled researchers to form algorithms and methods to find sentiments from the free text more efficiently.

As technology is evolving, the user of the technology also increases. Nowadays, people use technology and social media to build social networks, communicate with friends, share knowledge, update others on their activities and whereabouts, share photos, videos, archiving events, get updates on activities by friends, and send messages privately and posting public testimonials. An attractive way of online social interaction and communication offered by social media encourages users to use it. Twitter is one of the famous social to communicate by exchanging comments, thoughts, and messages. Moreover, Twitter has Retweet (RT) feature, enabling users to repost the comments or opinions shared by other users. Yet Twitter can be used by some to disseminate aggressive and bullying messages.

This study includes all highlights from mining texts from Twitter by applying sentiment analysis based on people's opinions expressed on Twitter to finally allot polarity to them as positive, neutral, or negative. Three types of machine learning classifiers, namely Naïve Bayes, support vector machine (SVM), and k-nearest neighbour (k-NN), will be used in this study. The accuracy, class precision and class recall of positive, neutral and negative tweets for each classifier will be evaluated.

The remainder of this paper is organized as follows. Section 2 describes the definition of cyberbullying, sentiment analysis, and classification techniques involved. In Section 3, we discussed the implementation of this study. Section 4 presents the findings and discussions, while section 5 concludes and suggestions for future works.

Manuscript received November 5, 2021
Manuscript revised November 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.11.1

2. Related Works

2.1 Cyberbullying

Cyberbullying is defined as a kind of offensive online behaviour involving a constant process such as a series of harsh words or messages sent from an attacker or bully to harm the victim [4]. The power of bullying among users has gone beyond the physical environment into the cyber world, a virtual space that is challenging to observe the actions and developments. It includes various technologies involving data such as e-mails, mobile phones, personal websites, or media like Tumblr, Twitter, Instagram, Facebook, etc. Twitter is recognized as a famous social media platform where most users experience cyberbullying. It allows users to post and read brief and informative messages defined as "tweets" on a website every day [5].

Cyberbullying has brought the world a severe impact, but most people have no idea how to handle it in the community. Cyberbullying is big trouble, and it needs to be stopped before it conquers the community because this situation may bring to the potential of the educational environment disruption and can happen in critical mental and physical outcomes for victims. Cyberbullying through Twitter has gained attraction in current years because of its connection with several terrible, severe suicides. According to [6], over half of teenagers and teens have been experienced cyberbullied or involved in cyberbullying, and $10\% - 20\%$ of them encountered it daily. The consequences of the victim who undergoes cyberbullying are severe such as anxiety, panic, lower self-esteem, depression, insanity, or even suicide.

2.2 Sentiment Analysis

Sentiment analysis, also defined as opinion or data mining, has been one of the numerous dynamic research fields in natural language processing since the early 2000. The views that are carried by any character of individuals are studied and analyzed using sentiment analysis. These reviews can be linked to an experience, label, symbol, or product.

Magazines and newspapers were used to express people's views during the old times. Nevertheless, with the improvement in technology, people have started to expose their feelings on social media and microblogging sites [7]. Sentiment Analysis technology has developed as the times require, drawing many learners at home and elsewhere to escort research due to the massive number of information data with a quick update rate [8]. The purpose of sentiment analysis is to assign automatic tools to obtain subjective information from texts in natural languages, such as comments, opinions, and sentiments [9]. To identify the overall sentiment of society, retrieval of data from sources or origin like Twitter, Facebook, Blogs are necessary. For the sentiment analysis, we concentrate on Twitter, a microblogging social networking website. Twitter produces massive data that cannot be managed manually to obtain some useful data or information. Therefore, the components of automatic classification are needed to address those data or information [10]. Researches on cyberbullying and Twitter usually summarized general cases of the aspect, with the potential for severe, harmful consequences for its victims [11].

Hence, cyberbullying detection on social media using sentiment analysis will benefit and efficiently assists that social media to detect any cyberbully threats.

2.3 Classification Algorithm

Classification algorithms have been created in machine learning, which employs various methodologies to classify unlabeled data. Classifiers could require training data. Naive Bayes, Support Vector Machine, and k-Nearest Neighbour are examples of machine learning classifiers. It's worth noting that effectively training a classifier will make future predictions easier.

a) Naïve Bayes

According to [3], Naive Bayes classifiers can be defined as Bayesian network classifiers, a well-known supervised classification model based on Bayes' Theorem and significant (naive) feature independence assumptions. Naïve Bayes was founded below another name into the text retrieval population and remains a recommended method for text classifying, the difficulty of assessing documents as relating to one class or the other with word repetitions as the feature. [12] highlighted that Naïve Bayes is utilized to foretell the probability for a given word to fit into a specific class. It is easy to apply both through training and classifying procedures. The Bayes hypothesis is a method of computation that allows you to distinguish likelihood P(A|B) from P(A), P(B), and P(B|A).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (1)$$

Where P(A|B) is the posterior probability of class A given predictor B and $P(B|A)$ is the likelihood; the probability of predictor B given class B. The prior probability of class A is denoted as P(A), and the prior probability of predictor P is denoted as $P(B)$. The Naïve Bayes algorithm is frequently used for classifying texts into several categories, and it was recently applied for sentiment analysis classification.

b) Support Vector Machine

Support Vector Machine (SVM) is one of the supervised machine learning algorithms applied for classification and

regression. It intends to classify or categorize data by searching for proper hyperplanes that divide data by the highest margin [13]. SVM builds a tool to classify data into various classes by an N-dimensional hyperplane that estimates based on a given training dataset [14]. The purpose of the SVM algorithm is to determine a hyperplane in N-dimensional space which N is the number of features that distinctly classify the data points.

c) k-Nearest Neighbour

K-Nearest Neighbor (kNN) is one of the most famous supervised classifier models as it is straightforward to implement and easy to apply. [15] stated that the function of kNN is to classify unlabeled data by assigning them to different classes with labeled samples. The class label is specified for an unknown data sample, depending on most of its k-nearest neighbours regarding an interpreted data collection. Before the classification process begins, there are two crucial options to be made: the value of k will be applied on and to find an optimal value by using cross-validation or distance metric. Euclidean distance is recommended to be used to compute distance as it is an adequately enigmatic concept, and the precise metric to handle is always going to be resolved by the dataset and the classification duty. Euclidean distance is the most common way to calculate distance. It is approximately the magnitude of the vector taken by subtracting the training data point from the point that needs to be classified.

3. Implementation

Seven stages are required in detecting cyberbullying tweets on Twitter, such as data collection, pre-processing, folding, automated training set classifier, extraction, tweets classification, analysis, and evaluation are shown in Figure 1.

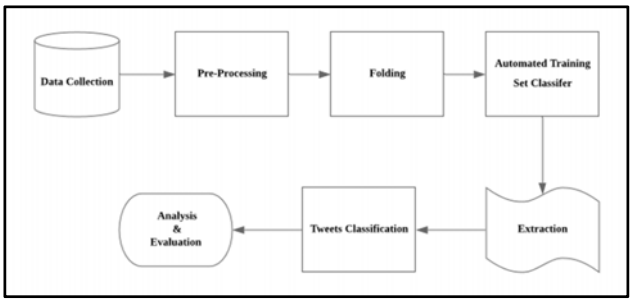


Figure 1: Cyberbullying detection process

Data collection is the process of accumulating and covering information on variables of the case within a precise authorized procedure that allows one to acknowledge stated study topics, test hypotheses, and appraise outcomes. This research collects relevant data from Twitter using RapidMiner Studio. The access token can be obtained by connecting to Twitter. Once the access token has been verified, users can collect the relevant data from Twitter smoothly. The collection process involves the data using various keywords such as 'gemuk', 'anjing', 'bodoh', 'babi', and 'sial', which indicate cyberbully in Malay. The selection of the five keywords is based on a study conducted by [16], who found that most users often use those keywords to attack or bully people through Twitter in Malay language. The retweets and any website or URL link are avoided during the collection process as the commands such as rt, HTTP, and HTTPS are applied in the query. In addition, 2000 recent tweets are collected from Twitter using the specific keyword, as shown in Figure 2.

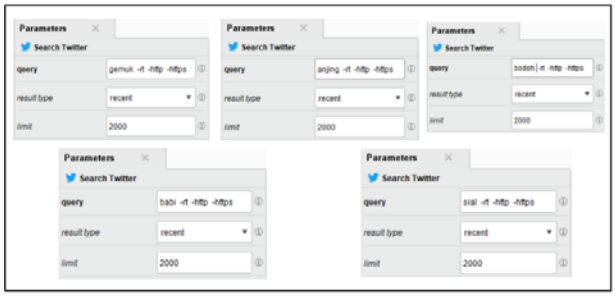


Figure 2: Data retrieve using specific keyword

b) Pre-Processing

Pre-processing is the process of detecting, modifying, or eliminating misleading and inaccurate data from a record set, table, or database referring to identifying inadequate, inaccurate, mistaken, or unnecessary parts of the data and then replacing, altering, or removing the messy or poor data. Before generating a dataset, some attributes are selected to retrieve from Twitter, such as 'Text' and 'Id'. Then, a generate attributes operator is chosen to create specific attributes, which are 'Text,' 'Id', and 'sentiment' for further process. Next, the duplicates operator is removed to avoid repeated tweets occurring. The steps are illustrated in Figure 3.

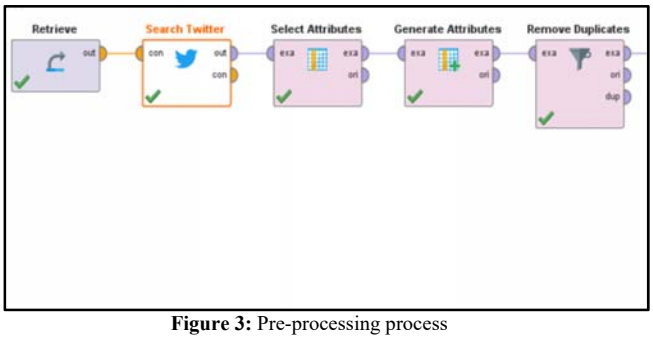


Figure 3: Pre-processing process

c) Folding

Folding is when the data will be split into training and validation or test sets while building a machine learning

model using some data. The machine learning model is trained using the training set while the validation or test set is applied to validate data that it has never seen before. In this research, the dataset is split into two types which are training data and testing data. The occurring ratio used is 80:20, which means 80% for training data while 20% for testing data. The sum of tweets collected from Twitter is 10 000 tweets; therefore, 8 000 tweets will be used to train the classifier models, while 2 000 tweets will be used to test the polarity of the tweets upon the classifier models. The split operator separates the dataset into two categories and is stored in an excel format in Figure 4.

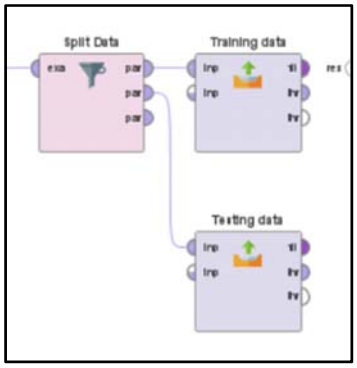


Figure 4: Dataset split into two categories

d) Automated Training Set Classifier

These stages are essential to training a machine learning model to classify data into few classes manually or automatically. After the data was separated into a training and testing dataset, the training dataset was labeled with various classes such as positive, neutral, and negative, respectively manually. This step consumes a long period of duration to complete but will generate a very stable glossary. The labeled dataset is depicted in Figure 5, while Figure 6 shows the unlabeled dataset for testing purposes. Both labeled and unlabeled datasets were stored in the RapidMiner Studio database. A non-relational database is more suitable for a massive amount of datasets and is simple to store than a relational database.

Text	Id	sentiment
Kai memiliki seekor anjing cowok ras Terrier Airedale, namanya MeongGu	12707402	Positive
Dua tiga tangkap malingBacot banget gua anjing	12707402	Positive
setelah harimau,anjing,kucing, skrg alpaca nanti gw iri sama hewan apalagi	12707402	Positive
@bukankocenkk gmana tu anjing yg ketawa	12707402	Negative
@affriskii @sacaaex @ananotyou @memukulmuu @ahhdut @sashuttt @nakanjingg @aripudin_ @heyzaduum @Astrou	12707402	Neutral
anjing kenala ngegas unggie astaga ??	12707401	Positive
Good job untuk barista, kepada owner, kalian semua anjing babi.	12707401	Negative
anjing lah	12707401	Neutral

Figure 5: Labelled training data

Text	Id	sentiment
tadi live suara nya ngapa gema sih anjing udah kayak hajatan aja	127074215	
@mycgrttsnsweet nopal anjing wkwkwkkw	12707421	
@kdrama_menfess VOICE 1 !!!!!!!!!!jin hyuk anjing gila??????	12707419	
kok goblok sih anjing	12707418	
@Indomiegorenq @jefrinichol anjing wkwk????	12707418	
@anakgembalaa anjing, untung g ikut rep	12707418	
@DrRozanizam @currentlyzhaff Haritu time tgh iv sekali jiran belakang punya anjing menyalak p	12707414	
Liat di base ttg anjing yang diracun dan baca repnya, ternyata banyak kasus kucing diracun juga,	12707414	

Figure 6: Unlabeled training data

e) Extraction

The extraction process is vital to decrease the number of resources required for processing without dropping necessary information. Using various operators such as tokenize operator, transform cases operator, and filter stop words operator, the unnecessary information in the dataset was eliminated. The function of tokenize operator is to break the sentence in the document into a series of words so that the word lists can be applied in the next sub-process. The transform cases operator converts all the sentences into lower case, and the filter stop words operator filters away the words that did not symbolize any sentiment. The process of feature extraction is shown in Figure 7.

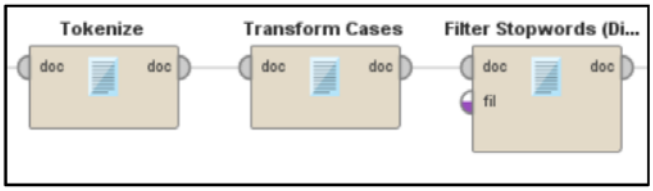


Figure 7: Feature extraction

f) Tweets Classification

Before running through the machine learning classifying models, the training dataset must be retrieved from the database and classified. The process of tweets classification is illustrated in Figure 8.

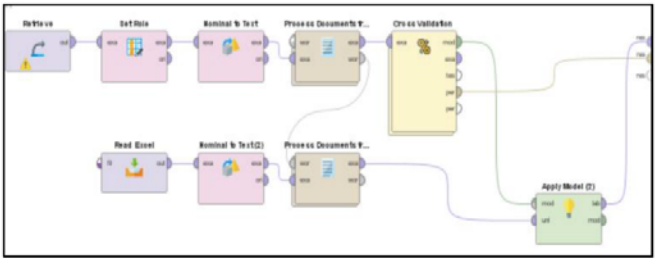


Figure 8: Tweets classification process

The set role operator is used to modify the role of one or more attributes. Additionally, the nominal to text operator is applied to convert all nominal attributes to string attributes as every value is used as a string value of the new attribute. The process document operator is to extract the features of the dataset. At the same time, the function of the cross-validation operator is to measure the model's predictive performance on unlabelled data. It has two sub-processes which are the Training sub-process and Testing sub-process. Besides, the apply model operator is to apply a training model on the testing dataset. The performance for each model is estimated during the Testing stage shown in Figure 9, Figure 10, and Figure 11.

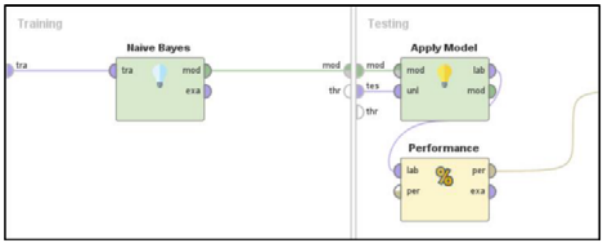


Figure 9: Cross-validation process for Naïve Bayes

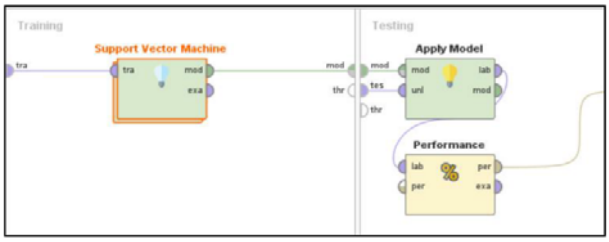


Figure 10: Cross-validation process for SVM

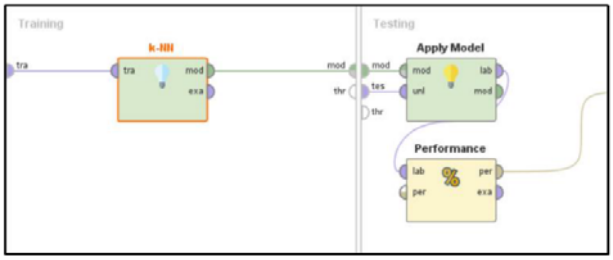


Figure 11: Cross-validation process for Naïve Bayes

g) Analysis and Evaluation

The training data is retrieved from the database and run by three trained machine learning classifier models, and the polarity is detected by running the classifier models as mentioned. Next, the performance of three machine learning classifier models is compared and recorded. Based on the result given, the accuracy of the classifier, class precision, and class recall of positive, neutral and negative tweets is recorded. The class precision is defined as the fraction of relevant instances among the retrieved instances. In contrast, the class recall is the fraction of the total amount of retrieved relevant instances. Meanwhile, a system with high-class precision but low-class recall indicates that most predicted labels are true compared to the training labels in contrast to a system with low-class precision. Still, high-class recall means most predicted labels are incorrect compared to the training labels.

4. Results and Discussions

The following section will present the testing results and analysis for the keywords 'anjing', 'babi', 'bodoh, 'gemuk' and 'sial'. The class precision represents the percentage of the relevant results, while the class recall indicates the relevant classified results.

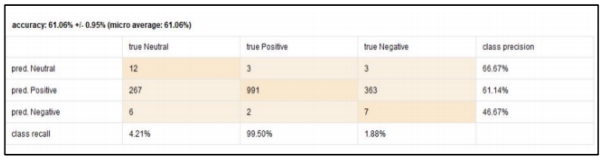
The result in Figure 12 highlights that Naïve Bayes has achieved a positive class precision (69.28%) and a positive class recall (53.21%) for the keyword 'anjing'. This explains that the classifier obtained 69.28% of relevant positive tweets while only 53.21% of the related positive were obtained. At the same time, Naïve Bayes presents a negative class precision (27.33%) and a negative class recall (23.59%) for the keyword 'anjing'. This shows that 27.33% of relevant negative tweets, but only 23.59% were related tweets were obtained. The negative class recall is low might be because there are many wrong classified negative tweets. Meanwhile, Naïve Bayes analyzed a neutral class precision (21.69%) and a neutral class recall (43.16%) explains that the neutral class recall is higher than the neutral class precision for keyword 'anjing'. The accuracy in the Naïve Bayes classifier model is 44.8% for the keyword 'anjing'.

accuracy: 44.79% +/- 4.50% (micro average: 44.80%)

	true Neutral	true Positive	true Negative	class precision
pred. Neutral	123	297	147	21.69%
pred. Positive	97	530	138	69.28%
pred. Negative	65	169	88	27.33%
class recall	43.16%	53.21%	23.59%	

Figure 12: Naïve Bayes classifier results for Keyword' anjing'

The result in Figure 13 highlights that Support Vector Machine (SVM) depicts 61.14%, 46.67%, and 66.67%, respectively for positive, negative, and neutral class precision. In comparison, the percentage for the class recall is 99.50%, 1.88%, and 4.21%, respectively, for the keyword 'anjing'. The positive class recall is higher because the classifier model obtained positive tweets are more significant than related positive tweets. The positive class precision is slightly higher than the negative and neutral class precision because the training data is more solid to the positive polarity. Finally, the accuracy in the SVM classifier model is 61.06% for the keyword 'anjing'.

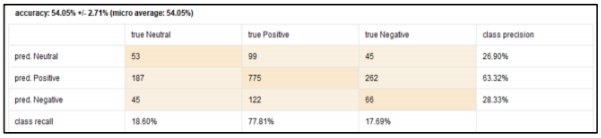


accuracy: 61.06% +/- 0.95% (micro average: 61.06%)

	true Neutral	true Positive	true Negative	class precision
pred. Neutral	12	3	3	66.67%
pred. Positive	267	991	363	61.14%
pred. Negative	6	2	7	46.67%
class recall	4.21%	99.50%	1.88%	

Figure 13: SVM classifier results for keyword 'anjing'

Based on Figure 14, the class precision value measured by k-Nearest Neighbor (k-NN) is 63.32%, 28.33%, and 26.90%, respectively. In comparison, the value for the class recall is 77.81%, 17.69%, and 18.60%, respectively, for positive, negative, and neutral in keyword 'anjing'. Additionally, Figure 9 presents that the false positive is 22.20% while the false neutral is 81.40% because the class recall is low. Moreover, the negative class recall has 82.31% false negative, which means many positive sentiment tweets. The accuracy in the k-NN classifier model is 54.05% for the keyword 'anjing'.

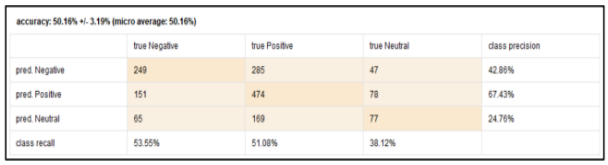


accuracy: 54.05% +/- 2.71% (micro average: 54.05%)

	true Neutral	true Positive	true Negative	class precision
pred. Neutral	53	99	45	26.90%
pred. Positive	187	775	262	63.32%
pred. Negative	45	122	66	28.33%
class recall	18.60%	77.81%	17.69%	

Figure 14: k-NN classifier results for keyword 'anjing'

From Figure 15, Naïve Bayes read a positive class precision (67.43%) and a positive class recall (51.08%) for the keyword 'babi'. This explains that the classifier obtained 67.43% of relevant positive tweets while only 51.08% of the related positive were obtained. Next, Naïve Bayes presents a negative class precision (42.86%) and a negative class recall (53.55%) for keyword 'babi', which means that 42.86% of related negative tweets obtained by the classifier and 53.55% related negative tweets were obtained. Thus, this explains that there are many false positives were detected in the negative class. Meanwhile, Naïve Bayes analyzed a neutral class precision (24.76%), and a neutral class recall (38.12%) shows that the neutral class precision is higher than the neutral class recall for keyword 'babi'. The accuracy in the Naïve Bayes classifier model is 50.16% for keyword 'babi'.

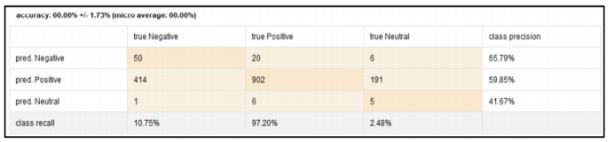


accuracy: 50.16% +/- 3.19% (micro average: 50.16%)

	true Negative	true Positive	true Neutral	class precision
pred. Negative	249	285	47	42.86%
pred. Positive	151	474	78	67.43%
pred. Neutral	65	169	77	24.76%
class recall	53.55%	51.08%	38.12%	

Figure 15: Naïve Bayes classifier results for Keyword 'babi'

Based on Figure 16, Support Vector Machine (SVM) presents 59.85%, 65.79%, and 41.67%, respectively, for positive, negative, and neutral class precision. In comparison, the percentage for the class recall is 97.20%, 10.75%, and 2.48%, respectively, for the keyword 'babi'. The value of positive class recall is higher because the classifier model obtained positive tweets are more significant than relevant positive tweets. The negative class recall value is small, giving rise to false negatives (89.25%) for negative tweets. The accuracy in SVM classifier model is 60.00% for the keyword 'babi'.

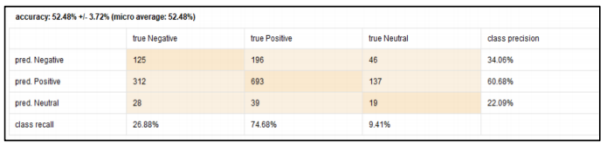


accuracy: 60.00% +/- 1.73% (micro average: 60.00%)

	true Negative	true Positive	true Neutral	class precision
pred. Negative	50	20	6	65.79%
pred. Positive	414	902	191	59.85%
pred. Neutral	1	6	5	41.67%
class recall	10.75%	97.20%	2.48%	

Figure 16: SVM classifier results for keyword 'babi'

The class precision value calculated by k-Nearest Neighbor (k-NN) in Figure 17 is 60.68%, 34.06%, and 22.09%, respectively. In comparison, the value for the class recall is 74.68%, 26.88%, and 9.41%, respectively, for positive, negative, and neutral in keyword 'babi'. Then, Figure 17 highlights that the false neutral is 90.59% and the 73.12% false negative, indicating that both neutral and negative class recall is low. The accuracy in the k-NN classifier model is 52.48% for the keyword 'babi'.

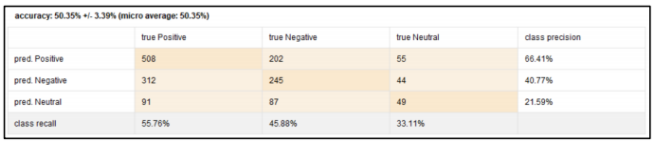


accuracy: 52.48% +/- 3.72% (micro average: 52.48%)

	true Negative	true Positive	true Neutral	class precision
pred. Negative	125	196	46	34.06%
pred. Positive	312	693	137	60.68%
pred. Neutral	28	39	19	22.09%
class recall	26.88%	74.68%	9.41%	

Figure 17: k-NN classifier results for keyword 'babi'

The result in Figure 18 indicates that the Naïve Bayes classifier model reads a positive (66.41%), negative (40.77%), and neutral (21.59%) class precision, respectively, while provides a positive (55.76%), negative (45.88%) and neutral (33.11%) class recall for keyword 'bodoh'. The value of positive class recall is more than the positive class precision as the classifier model obtained positive tweets are more significant than the related positive tweets. The root of the negative class precision is low because most of the contents in the training data are positive tweets compared to hate speech tweets. The value of negative class recall is low, leading to 54.12% of false-negative because some negative tweets are classified inaccurately. In addition, the accuracy for keyword 'bodoh' in the Naïve Bayes classifier model is 50.33%.



accuracy: 50.35% +/- 3.39% (micro average: 50.35%)

	true Positive	true Negative	true Neutral	class precision
pred. Positive	508	202	55	66.41%
pred. Negative	312	245	44	40.77%
pred. Neutral	91	87	49	21.59%
class recall	55.76%	45.88%	33.11%	

Figure 18: Naïve Bayes classifier results for Keyword' bodoh'

Based on Figure 19, Support Vector Machine (SVM) presents 59.85%, 65.79%, and 41.67%, respectively for class precision positive, negative, and neutral. In comparison, the percentage for the class recall is 97.20%, 10.75%, and 2.48%, respectively, for the keyword 'bodoh'. The value of positive class recall is higher because the classifier model obtained positive tweets are more

significant than relevant positive tweets. The negative class recall value is small, giving rise to false negatives (89.25%) for negative tweets. The accuracy in SVM classifier model is 58.88% for the keyword 'bodoh'.

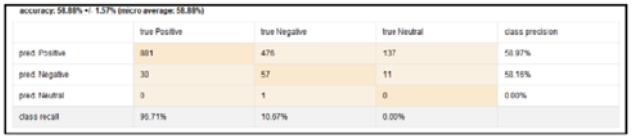


Figure 19: SVM classifier results for keyword 'bodoh'

The positive class precision value measured by k-Nearest Neighbor (k-NN) in Figure 20 is 58.22%, while the value of positive class recall is 74.68% for the keyword 'bodoh'. The value of positive class precision is lower than the positive class recall because the positive tweets obtained by the classifier model are massive than the entire relevant positive tweets. Meanwhile, the value of negative class precision (37.31%) is higher than the negative class recall (31.65%). Thus, 37.31% are related negative tweets, while 31.65% of the related tweets were detected. The neutral class precision and the neutral class recall analyzed by the k-NN classifier model are null as they did not detect any neutral tweets for the keyword 'bodoh'. The accuracy in the k-NN classifier model is 50.85% for the keyword 'bodoh'.

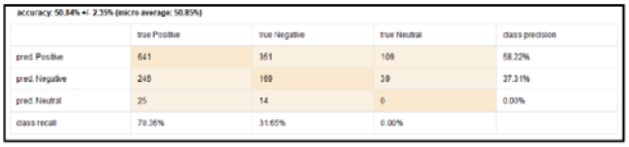


Figure 20: k-NN classifier results for keyword 'bodoh'

Figure 21 illustrates that the Naïve Bayes classifier model generates a positive (91.93%) and negative (10.23%) class precision while providing a positive (89.93%) and negative (9%) class recall for keyword 'gemuk'. The Naïve Bayes classifier model obtained a higher value of positive class precision because most of the predicted sentiment was true compared to the training dataset. The value of negative class precision and class recall is low stated that the value of false negative is 91%, indicating many tweets are classified inaccurately. Additionally, both values of neutral class precision (39.20%) and the neutral class recall (59.04%) analyzed by the Naïve Bayes classifier model are higher than the negative class precision and the negative class recall. The accuracy in the Naïve Bayes classifier model is 83.12% for the keyword 'gemuk'.

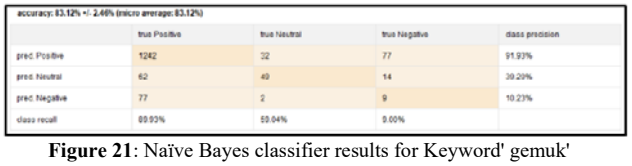


Figure 21: Naïve Bayes classifier results for Keyword' gemuk'

The class precision value calculated by Support Vector Machine (SVM) in Figure 22 is 90.79% and 97.73% for positive and neutral. The value for the class recall is 99.93% and 51.81% for positive and neutral for the keyword 'gemuk'. The higher class recall value for positive tweets shows that the classifier's obtained positive tweets are related more significantly than the whole related positive tweets. In addition, Support Vector Machine (SVM) detected 0% for both negative class precision. Negative class recall indicates that the sentiment predicted is mostly positive tweets without hate speech or unable to detect due to the slang or short forms in the tweets. The accuracy in the SVM classifier model is 90.98% for the keyword 'gemuk'.

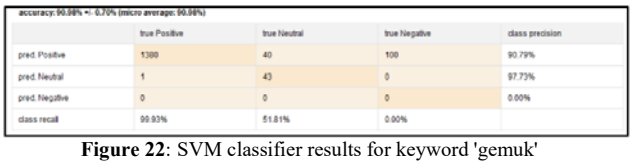


Figure 22: SVM classifier results for keyword 'gemuk'

Based on Figure 23, the class precision value calculated by k-NN Classifier Model is 89.42% and 45.45% positive and neutral. The value for class recall is 97.97% and 24.10% for positive and neutral for keyword 'gemuk'. The class recall value for positive tweets is more than the class precision means the obtained related positive tweets were larger than all related positive tweets. In addition, the k-NN classifier model detected 0% for both negative class precision and negative class recall, indicating that most of the tweets were classified as positive or neutral. The accuracy in k-NN classifier model is 87.79% for the keyword 'gemuk'.

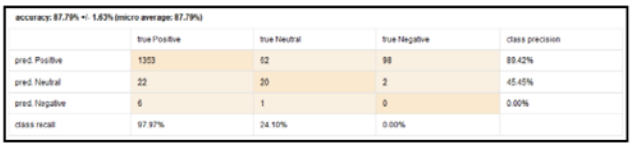


Figure 23: k-NN classifier results for keyword 'gemuk'

Figure 24 represents that Naïve Bayes classifier model produces a positive (77.49%), negative (38.81%), and neutral (33.33%) class precision while provides a positive (89.93%), negative (47.35%), and neutral (51.88%) class recall for keyword 'sial'. At the same time, the positive class precision is higher because most of the predicted sentiment was true compared to the training dataset. The classifier model can better predict the positive sentiments correctly while maintaining the value of false positives low. The cyberbullying detection accuracy in Naïve Bayes classifier model is 61.19% for keyword 'sial'.

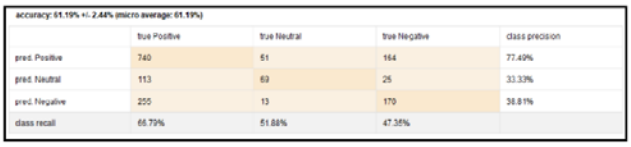


Figure 24: Naïve Bayes classifier results for Keyword 'sial'

The class precision value calculated by Support Vector Machine (SVM) in Figure 25 is 70.01%, 70.83%, and 100% respectively for a positive, negative and neutral while produces a class recall of 99.46%, 4.74%, and 1.50%, respectively for positive, negative and neutral for keyword 'sial'. The positive tweets have the least precision among the other two classifiers as the training data was more impenetrable against the negative polarity. The detection accuracy in the SVM classifier model is 70.06% for the keyword 'sial'.

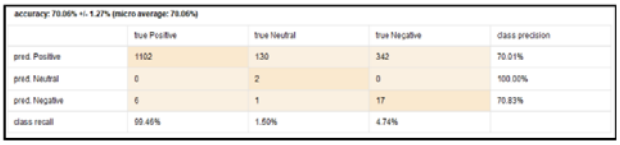


Figure 25: SVM classifier results for keyword 'sial'

Figure 26 mentions that the class precision value calculated by k-Nearest Neighbor (k-NN) is positive (68.78%), negative (17.41%), and neutral (16%) while generates a class recall of positive (85.29%), negative (9.75%) and neutral (9.75%) for keyword 'sial'. The figure shows the slightest class recall in negative and neutral causes 96.99% false neutrals and 90.25% false negatives, while the value in the positive class recall is higher with only 14.71%. The accuracy in the k-NN classifier model is 61.50% for keyword 'sial'.

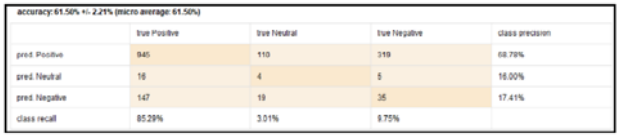


Figure 26: k-NN classifier results for keyword 'sial'

Table 1 represents the accuracy comparison of each keyword for different types of machine learning classifier models.

Table 1: Accuracy Comparison of Machine Learning Classifier for different keyword

Keyword	***Accuracy (%)***		
	Naïve Bayes	***Support Vector Machine (SVM)***	***k-Nearest Neighbor (k-NN)***
anjing	44.88	61.06	54.05
babi	50.16	60.00	52.48
bodoh	50.35	58.88	50.85
gemuk	83.12	90.98	87.79
sial	61.19	70.06	61.50
Average	**57.92**	**68.20**	**61.33**

In conclusion, the Support Vector Machine (SVM) model has a more reliable performance than Naive Bayes and k-Nearest Neighbor (k-NN). That might be because k-Nearest Neighbor (k-NN) is a dull algorithm that depends on statistics and comparison as it must trace massive features. Support Vector Machine (SVM) works offline learning to gain the optimal hyperplane that indicates that the Support Vector Machine (SVM) relies on the training set to find an equation that divides between the two categories or hyperplanes. Then, the Support Vector Machine (SVM) begins to apply this equation and stops based on the training. The performance of the Naive Bayes model is lower than the Support Vector Machine (SVM) model might be because of the variety of collected data or tweets. Users have insufficient range to write syntactically and proper tweets due to the maximum tweet length up to 280 words only. Meanwhile, some users apply short forms and unintentionally put whitespace to separate words.

Table 2 shows sample prediction analysis of three machine learning classifier models. Support Vector Machine (SVM) has outstanding performance among the other two machine learning classifier models, which are Naïve Bayes and k-Nearest Neighbor (k-NN). Fifteen samples were taken from the 10,000 tweets to depict the dissimilarities in the classification results in three machine learning classifier models. Generally, the Support Vector Machine (SVM) predicted eleven corrects tweets out of fifteen tweets, and the k-Nearest Neighbor (k-NN) predicted nine correct tweets out of fifteen tweets. In comparison, the Naïve Bayes predicted only eight correct tweets out of fifteen tweets.

Table 2: Sample Prediction Analysis of Machine Learning Classifier

Tweets	***Labeled Sentiment***	***Naïve Bayes***	***Support Vector Machine (SVM)***	***k-Nearest Neighbor (k-NN)***
Sume perangai cm anjing	Negative	Negative	Negative	Positive
@undertheskyie Punyaaa. Kucing sama anjing, yg kucing namanya manis yg anjing namanya snowie hahahah	Positive	Positive	Positive	Positive
@rizaleko_ Termasuk babi dan anjing ??	Positive	Negative	Positive	Positive
@junholeyyyyyyy BABI BAH KAUU DIAM BAHHH ???	Negative	Negative	Positive	Negative
@jiminparks07 Aku babi juga ingin dimanja dan dicintai -babi	Positive	Negative	Positive	Negative
Kadang rasa bodoh sebab senang sangat percaya orang.	Positive	Negative	Positive	Positive
bodoh punya pondan tahu nak report aje	Negative	Negative	Negative	Positive
Bodoh kau tu simpan la sikit macam gampang sial perangai	Negative	Positive	Positive	Negative

Bodoh jiran jenis rembat kucing ni. Mampos kau kena maki dengan bapak aku.	Negative	Negative	Negative	Positive
Siapa yang quarantine ini tambah gemuk???	Positive	Positive	Positive	Positive
maksudnya terima kita seadanya tak kisah la gemuk atau kurus.	Positive	Positive	Positive	Positive
@faliiqq Mcm biasa la setan gemuk tu ??	Negative	Positive	Positive	Positive
@_jkwithluv pergi mampus la sial	Negative	Negative	Negative	Positive
@fareastzs Puas hati sial tengok	Positive	Negative	Positive	Positive
masih ada orang yang sayang sama orang yang gemuk?	Positive	Neutral	Positive	Positive
Prediction (Correct)		8	11	9
Prediction (Incorrect)		7	4	6

5. Conclusion

The performance of machine learning approaches in cyberbullying detection from social media using sentiments has been presented in this research. Therefore, the system is capable of detecting cyberbullying using three classifier models. Support Vector Machine (SVM) has the highest accuracy, the k-Nearest Neighbor (k-NN) is the second, while the Naïve Bayes has the lowest accuracy. The system also formed a dataset of tweets containing cyberbullying and assessed a methodology for the proper data classification. The test will be run repeatedly to find out the most reliable results after the simulation is enhanced.

This research's aim of cyberbullying detection on Twitter is to decrease and weaken potential cyberbully threats to overcome standard patrolling works on social media. Furthermore, this research also focuses on utilizing optimizing class precision and class recall. The system can also classify the tweets based on cyberbullying categories such as harassment, insult, blackmail or curse. This will allow user penetration into which variety of bullying is more notable on social media. Additionally, other models or algorithms, except for Naïve Bayes, Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN), will be considered to implement in this research to achieve more accurate and efficient results.

Acknowledgments

This publication has been supported by Center of Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM). The authors would like to thank UTeM and INSFORNET research group members for their supports.

References

[1] V. Balakrishnan and T. Fernandez, "Self-esteem, empathy and their impacts on cyberbullying among young adults," Telemat. Informatics, vol. 35, no. 7, pp. 2028–2037, 2018, doi: 10.1016/j.tele.2018.07.006.

[2] F. K. Ventirozos, I. Varlamis, and G. Tsatsaronis, "Detecting aggressive behavior in discussion threads using text mining," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 10762 LNCS, pp. 420–431, doi: 10.1007/978-3-319-77116-8_31S. Encrypt, "What is Privacy Protection? [Updated for 2020]," Search Encrypt Blog, Dec. 20, 2019.

[3] S. M. Mohammad, Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. Elsevier Ltd, 2016.

[4] M. Yao, C. Chelmis and D. S. Zois, (2019) 'Cyberbullying ends here: Towards robust detection of cyberbullying in social media', in The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019. New York, New York, USA: Association for Computing Machinery, Inc, pp. 3427–3433. doi: 10.1145/3308558.3313462.

[5] V. Balakrishnan , S. Khan and H. R. Arabnia (2020) 'Improving cyberbullying detection using Twitter users' psychological features and machine learning', Computers and Security. Elsevier Ltd, 90, p. 101710. doi: 10.1016/j.cose.2019.101710.

[6] C. Chelmis, D. S. Zois, and M. Yao (2017) 'Mining Patterns of Cyberbullying on Twitter', IEEE International Conference on Data Mining Workshops, ICDMW, 2017-Novem, pp. 126–133. doi: 10.1109/ICDMW.2017.22.

[7] P. Tyagi, et al. (2019) 'Literature Review of Sentiment Analysis Techniques for Microblogging Site', SSRN Electronic Journal. doi: 10.2139/ssrn.3403968.

[8] P. Yang and Y. Chen (2018) 'A survey on sentiment analysis by using machine learning methods', in Proceedings of the 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2017, pp. 117–121. doi: 10.1109/ITNEC.2017.8284920.

[9] M. V. Mäntylä , D. Graziotin and M. Kuutila (2018) 'The evolution of sentiment analysis—A review of research topics, venues, and top cited papers', Computer Science Review, pp. 16–32. doi: 10.1016/j.cosrev.2017.10.002.

[10] H. Parveen and S. Pandey (2017) 'Sentiment analysis on Twitter Data-set using Naive Bayes algorithm', Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016. IEEE, pp. 416–419. doi: 10.1109/ICATCCT.2016.7912034.

[11] D, Chatzakou et al. (2017) 'Mean Birds: Detecting Aggression and Bullying on Twitter'. doi: 10.1145/3091478.3091487.

[12] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm," World Appl. Sci. J., vol. 35, no. 1, pp. 54–59, 2017, doi: 10.5829/idosi.wasj.2017.54.59.

[13] A. D. M. Africa, A. R. V. Tabalan, and M. A. A. Tan, "Speech emotion recognition using support vector machines," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 4, pp. 1212–1216, 2020, doi: 10.30534/ijeter/2020/43842020.

[14] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Syst. Appl., vol. 67, pp. 296–303, 2017, doi: 10.1016/j.eswa.2016.09.041

[15] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," Ann. Transl. Med., vol. 4, no. 11, pp. 1–7, 2016, doi: 10.21037/atm.2016.03.37.

[16] Z. Zainol, et al. (2018)' Association Analysis of Cyberbullying on Social Media using Apriori Algorithm', International Journal of Engineering and Technology, 7(December), pp. 72–75. doi: 10.14419/ijet.v7i4.29.218

Chong Poh Theng is a degree student in Bachelor of Computer Science (Computer Security) Universiti Teknikal Malaysia Melaka (UTeM). Her research interest includes data analytics and security application development.

Nur Fadzilah Othman received a degree in Computer Engineering in 2008 and master's in educational technology in 2011 at Universiti Teknologi Malaysia (UTM). In 2017, she obtained her PhD in Information Security at Universiti Teknikal Malaysia Melaka (UTeM). She started her career as a senior lecturer at the Faculty of Information Technology and Communication, UTeM from March 2018. Her research interests include information security, usable security and privacy and Internet of Things (IoT).

Raihana Syahirah Abdullah is currently a senior lecturer at the Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. She received her PhD in Network Security from Universiti Teknikal Malaysia Melaka (UTeM). Her Research interests include intrusion detection, network security malware analysis and design.

Zakiah Ayop holds BSc. in Computer Science (2000) from UTM and MSc in Computer Science (2006) from UPM. Currently, she is a senior lecturer in Faculty of Information and Communication Technology (FTMK), Universiti Teknikal Malaysia Melaka (UTeM). She is a member of the Information Security, Digital Forensic, and Computer Networking research group. Her research interest is Information System, Internet of Things (IoT) and Network and Security.

Syarulnaziah Anawar holds her Bachelor of Information Technology (UUM), Msc in Computer Science (UPM), and PhD in Computer Science (UiTM). She is currently a Senior Lecturer at the Faculty of Information and Communication Technology, UTeM. She is a member of the Information Security, Digital Forensic, and Computer Networking (INSFORNET) research group. Her research interests include human-centered computing, participatory sensing, mobile health, usable security and privacy, and societal impact of IoT.

Sofia Najwa Ramli received her PhD degree in Information Security from Universiti Teknikal Malaysia Melaka, Malaysia, in 2016. She received her Master's degree (M. Eng) in Electrical – Electronics & Tele-communications Engineering and Bachelor's degree (B. Eng) in Biomedical Engineering from Universiti Teknologi Malaysia in 2011 and 2009. She is currently a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. Her current research interests include authentication systems, biometrics, biomedical signal processing, cryptography, and information security. She has been actively involved as a program committee of an international conference and a technical editorial committee of the OIC-CERT Journal of Cyber Security. She has delivered articles in various international conferences and journals.