IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 11, 2019, accepted January 2, 2020, date of publication January 23, 2020, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968955

Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work

MUHAMMAD FAKHRUR RAZI ABU BAKAR[1], NORISMA IDRIS[1], LIYANA SHUIB[2], AND NORAZLINA KHAMIS[3]

[1]Department of Artificial Intelligence, Faculty of Computer Science and IT, University Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Information Systems, Faculty of Computer Science and IT, University Malaya, Kuala Lumpur 50603, Malaysia
[3]Language Engineering and Application Development Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu 88400, Malaysia

Corresponding author: Norisma Idris (norisma@um.edu.my)

This work was supported in part by the University Malaya under Grant IIRG003C-19SAH, and in part by the Cloud Connect Sdn. Bhd.

ABSTRACT Sentiment analysis (SA) is a study where people's opinions and emotions are automatically extracted in the form of sentiments from the natural language text. In social media monitoring, it is very useful because it allows user to gain an overall picture of the extensive public opinion behind many topics. Most works on SA are for the English text. Only a few works focus on the Malay language. Currently, a review on SA for the Malay language only focus on the SA approaches and the dataset. Some major issues such as the pre-processing techniques used to normalize the noisy text, the most employed performance measures for Malay SA, and the challenges for Malay SA has not been reviewed. Malaysians tend not to fully follow any abbreviations rules when writing on social media. Thus, a lot of noisy text can be found in social media sites like Facebook and Twitter which create some issues to SA process. Hence, the aim of this study is to investigate the state of the art, challenges and future works of SA for Malay social media text. This study provides a review on various approaches, datasets, performance measures, and pre-processing techniques used in the previous works on SA of the Malay text. More than 700 articles from journals and conference proceedings have been identified using the search keywords, however, only 17 relevant articles published from year 2013 to 2018 were reviewed. The findings from this review focus on three commonly used SA approaches which are lexicon-based, machine learning, and hybrid.

INDEX TERMS Hybrid, lexicon-based, machine learning, noisy Malay text, sentiment analysis.

I. INTRODUCTION

Social media has become a ubiquitous part of people's everyday life. It is one of the best mediums for them to stay informed. Every year, the growth in popularity of online social media has been outstanding. Social media sites such as Twitter, Facebook, forum and blog play an essential role in people's everyday life where they can communicate, collaborate and exchange information with each other. It is already become a trend where people express their thoughts, opinions, and emotions through social media. They will share, like or comment anything related to the current issues or anything that happened to them every day. Most people use the social media sites as a medium for them to communicate with real personality. Thus, the comments or opinions posted by the users are very valuable for the high-level administration in any organizations to study and understand on how to improve their services. For example, the higher educational institutions can know the students feedback on their new programmes [1] or the high level administration of any local government agencies can know how the public responds to whatever move that they made [2]. Hence, to identify whether the user's responds are neutral, negative or positive, the implementation of Sentiment Analysis (SA) is needed.

In the past few years, SA has received enormous attention where the interest in it has grown tremendously. According to [3] and [4], works on SA are mostly for the English language. Research works on SA for other languages such as the Malay language are still very limited although the usage of the Malay language on Twitter is very high. Malay language is consider as the fourth leading language used over Twitter [5]. SA is a computational study within the

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

domain of Natural Language Processing (NLP) where people's opinions, sentiments and emotions are automatically extracted in the form of sentiments from the natural language text. According to [6], SA can be done at aspect or feature level, sentence level or document level. SA has three fundamental approaches namely machine learning, lexicon-based, and hybrid [4]. Malaysians prefer to use noisy text when expressing their reactions on social media [7]. Noisy text is a text that is difficult to understand by any software or application to derive the actual meaning from the text. According to [8], noisy text has become one of the biggest obstacles in applying NLP applications such as sentiment analysis to User Generated Content (UGC). Pre-processing noisy and standard text are very different because noisy text does not have any standard rules or patterns whereas standard text followed the language standard format. To the best of author's knowledge, there is only one review regarding SA which focusing on Malay text which is a work by [9]. The work focused more on the proposed SA approaches and types of dataset for evaluation. However, it did not focus on how the previous works handled the noisy texts at the pre-processing phase. This is very crucial because most NLP applications are generally trained on formal and clean text [10]. Other than that, the performance measure used, and challenges also has not been reviewed and analyzed. Thus, to overcome this gap, this study presents the review on SA for noisy Malay text. The aim of this study is to investigate the state of the art, challenges and future works in SA of noisy Malay text. The objective of this study is addressed by the following research questions (RQs):

1) What are the most common approaches used in SA of noisy Malay social media text?
2) What are the most common pre-processing techniques used in SA of noisy Malay social media text?
3) What types of datasets used in SA of noisy Malay social media text?
4) Which performance measures are the most widely employed in SA of noisy Malay social media text?
5) What are the future research directions and challenges in the area of SA of noisy Malay social media text?

The structure of the article is as follows: First, details of the review process are discussed in Section II. The findings are discussed in Section III. The findings related to the research questions (RQs) are discussed in section IV and finally, the conclusion of the study is presented in Section V.

II. MATERIALS AND METHOD

The aim of this study is to investigate the state of the art, challenges and future works in SA for noisy Malay text.

A. TYPE OF STUDIES

Works on SA are mostly focusing on the English language and consider still lacking for the Malay language [3]. Besides than that, to the best of the author's knowledge, there is no review article in genuine academic databases on SA for noisy Malay text. Hence, to overcome this gap, this study presents the review on SA for noisy Malay text.

B. SEARCH STRATEGY FOR THE IDENTIFICATION OF STUDIES

Scientific articles or conference proceedings related to SA for Malay language from six main academic databases were searched. These academic databases include Springer Link, Web of Science, Scopus, IEEE Explore, Google Scholar, and Science Direct. These databases are chosen because of the broad coverage of peer reviewed journals in various disciplines. Although most IEEE Explore papers were published in Scopus, there are a few papers like [11] and [12] exist in Scopus but cannot be found in IEEE Explore by using the selected search strategy. Other than that, Google Scholar is also included because there are a few papers which can be good references such as [3] that cannot be found in other selected academic databases using the selected search strategy. For efficient searching process, search keywords were identified. Then, the search strings were generated, and the search structure was formulated. Finally, the search process was conducted. To identify the hints or keywords relevant to this study, the keyword selection process was mainly performed using a snowballing process. The search was performed using seven keywords that are sentiment analysis, text classification, opinion mining, text mining, text categorization, sentiment classification and Malay language. Next, Boolean operator "OR" is incorporated to include synonyms and then Boolean "AND" operator is used to link the keywords and create the final search string. Almost all academic databases used required different search design to get relevant result. To reduce the number of search results and maintain the relevanceness of this study, only articles or conference proceedings published from year 2013 to 2018 were selected. Besides than that, the selected publications were only written in English or Malay language because works on SA are mostly for the English language [3], [4]. Works written in Malay language is included because there might be some relevant papers were written in Malay language. Finally, only publications that used noisy Malay text as one of the dataset's criteria were selected in this review for answering the research questions.

C. SCREENING AND SELECTION OF CRITERIA

A total of 773 publications were found after searching through all the academic databases based on the keywords stated above as shown in Fig 1. The software EndNote X8 was used to automatically find and remove duplicates. This process resulted 60 publications being excluded. Then, non-articles and conference proceedings were also removed resulted 97 publications being excluded. All the publications title and abstract were read manually for relevance checking. This process resulted 595 publications being excluded. Lastly, 14 eligible publications were selected and added 3 more from snowballing process. The analyzed publications

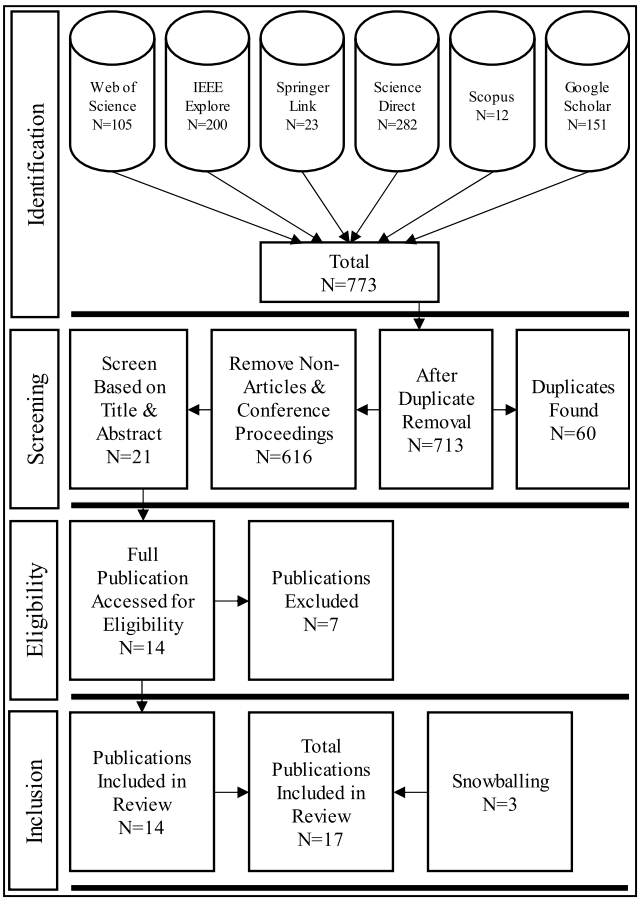

FIGURE 1. Review process.

were investigated based on the relevance to the research domain, availability, and publication years.

III. FINDINGS

In total, 17 articles and conference proceedings were selected for this review. All the selected publications were published from year 2013 to 2018. The selected language was English and Malay but no relevance publication for the review in Malay language was found. Other than that, there were six publications used a lexicon-based approach, four publications used a hybrid approach, and another seven publications used a machine learning approach. All the chosen publications handled noisy Malay text from several online sources such as Facebook, Twitter, blog, news, online forum, and other general websites or social media sites.

A. MALAY SENTIMENT ANALYSIS APPROACHES

SA has three fundamental approaches namely hybrid, machine learning and lexicon-based [4]. A lexicon-based approach is a practical, smooth, and feasible approach to SA where no prior training dataset is required. It detects any words or phrases that carry sentiment load and use them for determining the sentence sentiment value. A machine learning approach starts with the collection of data where it will be splitted into test and training datasets. A test dataset is used for evaluating the performance of the machine learning classifiers while the training dataset is used to train the classifier for making a distinction between attributes of text [13]. The third approach, hybrid, is a combination of two or more SA approaches or combination of two or more machine learning classifiers.

Based on the finding, there are six studies used the lexicon-based approach for Malay SA. The first study is by [3] where they proposed to quantify sentiments in Facebook by analyzing users' emotions based on their posted comments. Only comments written in English and Malay were chosen. The emotions used to quantify the sentiments were divided into three classes which are emotionless, unhappy, and happy. Other than that, there were two sentiment lexicons created based on manually identified unstructured Facebook 's contents namely happy and unhappy. For the sentiment classification, they tagged the words with the emotion categories. The percentage of every emotion's categories were then calculated. The sentence will be classified as unhappy if the percentage of unhappy emotion was higher than happy emotion and vice versa.

In another study by [2], they introduced a lexical based method to analyze sentiment of Facebook comments in Malay language. They created two types of score dictionary which are also known as sentiment lexicon namely Malay Adjective score dictionary which only contained the adjective of Malay words and Malay-English (BM-ENG) score dictionary where the English words in it were the translation of the Malay adjectives words. The commonly used Malay adjective words from their research data were extracted in order to construct the Malay Adjective score dictionary. After that, all the words inside the Malay Adjective score dictionary were manually categorized as positive (+1) and negative (-1). SentiWordNet was used as the resource for building the Malay-English (BM-ENG) score dictionary. The other Part-Of-Speech (POS) apart from adjectives were also included in this dictionary but they were considered as adjectives as they were derived from the Malay adjective word. There were three types of lexical based techniques implemented in their study namely Term Counting, Average on Comments, and Term Score Summation. Term Counting was solely depending on the Malay adjective score dictionary while Average on Comments and Term Score Summation used both Malay Adjective and Malay-English (BM-ENG) score dictionary.

The Malay Adjective score created by [2] has been updated by [14] in their study which classified the sentiment of Facebook comments in Malay language based on the adjectives, verbs, adverbs, and negation. They used the term ''score dictionary'' for their version of sentiment lexicon. WordNet Bahasa has been used as the resource for creating the Verb, Adverb, and Negation score dictionary. In the Adjective and Verb score dictionary, all the words were manually categorized as positive (+1) and negative (-1). As for the Adverbs score dictionary, all the words were manually categorized to double (2) or half (0.5) the polarity score of the words paired

with them. There was no score being assign to the Negation score dictionary because they only act as the polarity inverter to the words paired with them. They implemented two types of lexical based techniques namely Term Counting and Term Counting Average in order to classify the sentiment of Facebook comments. A comment is classified as negative when the sentiment score value is negative and vice versa. The comment is classified as neutral if it has a sentiment score value of 0. They used several combinations of part of speech (POS) tags to perform the scoring method such as Adjective + Negation, Adjective + Adverb, Verb + Negation, and Verb + Adverb.

Reference [15] performed SA on a set of tweets related to local mobile telecommunication services in both English and Malay language. They produced their own sentiment lexicon, called SentiLexM. The content for the SentiLexM was derived from existing words and scores from AFINN-111, an English based sentiment lexicon [16]. The Malay translation of each English word from AFINN-111 was also being added to the SentiLexM and their polarity value strictly followed the polarity value of the English words. The polarity value of the words inside the SentiLexM was labelled with a score ranging from -5 to $+5$. For the method of Malay translation, they identified several prefixes and suffixes and added them to each of the existed Malay words in the SentiLexM. Besides than that, all the various forms of the Malay translations were added to the SentiLexM for the English words that had several Malay translations. Furthermore, only part of the Malay translations was used if the Malay translations comprised of two words because their SA program only analyzed each tweet in word by word. However, any Malay translations using repeated words connected with dash were considered as one word. Lastly, the SentiLexM was checked manually for spelling errors. There were no lexical based techniques used to calculate the total sentiment score for a tweet. They only calculated it by adding all the sentiment score for each word in a tweet. The tweets were classified into neutral, negative or positive by following the total sentiment score value.

Reference [4] introduced a SA on one of the derivatives of the Malay language, namely Sabah language on social media. They built their own Sabah language sentiment lexicon by retrieving the Malay language sentiment lexicon from Multilingual Sentiment in Data Science Labs2 which consists of positive and negative Malay words text files. The retrieved text files also consist of some English words. The Malay language sentiment lexicon was served as a foundation to their Sabah language sentiment lexicon. To expand their sentiment lexicon, they added any words in Sabah language that were found in the corpus which either acronyms or synonyms to the Malay language words. Besides than that, they omitted the duplicate entries where some of the Malay words existed in both positive and negative Malay words text files. Moreover, the English words found were translated to Malay before being removed from their sentiment lexicon. They assigned the polarity score of the words in their sentiment lexicon using three different methods namely Simple Polarity Score Assignation (Simple PSA), Strength-based PSA, and Simple PSA with Switch Negation (PSA-SN). Finally, bias aware classification was used in doing the SA classification and the performance were compared with simple classification.

In a recent work by [7], they constructed an effective sentiment lexicon namely RojakLex to handle noisy Malay text on social media. The RojakLex consists of Bahasa SMS lexicon, neologism lexicon, emoji lexicon, and English-Malay lexicon which also known as Bahasa Rojak lexicon. For English-Malay lexicon, they used MySentiDic which is a Malay lexicon with its translated version. Later, they formed a mix lexicon by combining both lexicons. For Bahasa SMS lexicon, they utilized the rules proposed by Dewan Bahasa dan Pustaka (Panduan Singkatan Khidmat Pesanan Ringkas (SMS) Bahasa Melayu, 2008). As Malaysian tend not to follow the rules when writing a Bahasa SMS, a neologism lexicon was constructed to handle endless emerging of neologism word in noisy Malay text. The construction process of the neologism lexicon was addressed using a semi-automatic approach where they adopted a human centric approach for translating the neologism terms and a frequent update was needed. As for the emoji lexicon construction, they combined all the identified resources to form a broad emoji lexicon where emotionless emoticons and duplicates were ignored. Valence shifter including diminisher, intensifier, contrast as well as negation were handled for helping in handling *Bahasa Rojak* effectively. They used a simple prediction method to classify a sentence by counting the amount of negative and positive words existed in the sentence. The sentence was classified as positive if it contains more positive words compare to negative words and vice versa.

For the machine learning approach, there are seven studies available in the research field. The first study is by [11] where they used a feature selection technique based on artificial immune system (AIS) namely Feature Selection based on Immune Network System (FS-INS) for SA. The result of the SA using FS-INS as a feature selection was compared with the accuracy when no feature selection was used. Besides than that, the results of the SA when other common feature selection techniques like Document Frequency (DF), CHI Square (CHI), Categorical Proportional Difference (CPD), and Information Gain (IG) were used also being compared with FS-INS. Reference [17] investigated on how feature selection methods lead to the enhancement of Malay sentiment classification performance. There were 7 feature selection methods used namely Support Vector Machines, CHI, uncertainty, Gini Index, Relief-F, Principal Components Analysis (PCA), and IG. Besides than that, the machine learning classifier used were k-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). Several experiments were conducted where each one of the three classifiers' overall performance on the Malay SA without feature selection were analyzed. After that, the performance of the classifiers using different feature selection methods were tested. Finally, all the results were compared and discussed.

Reference [18] proposed SentiRobo, a newly developed NB classifier algorithm. They examined the effectiveness of SentiRobo when doing the SA for contents in the social media broad datasets. The experiments were performed with the training and test sets were randomly separated into 70% and 30% of records respectively. Reference [19] used NB to classify the tweets. The classification process was based on trainers' perception which have been classified into three categories namely neutral, negative, or positive. There were 27 trainers participated in this process. They were assigned to classify half of the tweets which were obtained using 'Malaysia' and 'Maybank' as the keywords. After that, all the classified data were used as the training set for the NB classifier. The other half of the tweets were used as the test collection and the accuracy of the classifier were calculated.

Reference [20] presented the effects of the common-used feature selection methods with three machine learning classifiers namely NB, k-NN, and SVM for Malay sentiment classification. The feature selection methods used were CHI, IG, and Gini Index. First, all the machine learning classifiers were experimented without using any feature selection method. After that, the feature selection methods were applied, and their performance were compared. Reference [21] conducted several experiments using three different types of machine learning classifiers namely NB, SVM, and k-NN on Malay short text. All the machine learning classifiers were investigated on four types of features namely ITC, TF-IDF, Bag of Words (BOW), and smoothed TF-IDF. Lastly, they used 20% and 50% of their test data to conduct the experiments. Reference [22] proposed an enhanced ensemble of machine learning classification methods for Malay SA. There were three machine learning classifiers namely NB, SVM, and k-NN and five ensemble classification algorithms namely Bagging, Voting, Stacking, MetaCost, and AdaBoost were used in this study.

For the hybrid approach, there are four studies available in the research field. The first study is by [12] where they used a combination of machine learning and lexicon based approach for Malay SA. They used the English WordNet as their sentiment lexicon resource. The English words were translated to Malay language and used as the sentiment words in the sentiment lexicon. A Malay native speaker assigned manually the polarity value and a synonym for each phrase and word was stored. Every word in the sentiment lexicon were assigned a score ranging from -5 which is strongly negative to 5 which is strongly positive. The features were created by using the Malay sentiment lexicon. Besides than that, there were eleven features used which classified under four categories namely features based on the conditional probability of subjective words, features based on the polarity level of sentiment words, features based on the level of the sentence, and features based on the presence and frequency of sentiment words. Lastly, k-NN was used as the machine learning classifier in this study.

Reference [23] proposed a hybrid approach which combined the knowledge base and machine learning approach for improving accuracy in SA for Malay language. They have identified several common sentiment challenges in order to conduct meaningful SA of text namely false negatives, metaphore, named entity sentiments, hyperbole, slang, context dependency, and negation. In this approach, the machine learning was used to classify models representing negative and positive classes while the knowledge based was used to get matching properties and concepts. Lastly, for evaluation purpose, the performance of the proposed approach was compared against commonly used machine learning classifier namely k-NN, SVM, and NB.

Reference [24] used both lexicon-based and machine learning approach for determining the customer intention on purchasing the services or products by analyzing the communications particularly using Malay language on social media. The initial dictionary was used to initialize the sentiment lexicon construction process automatically. Besides than that, every word in the sentiment lexicon were given a polarity score of negative 1, positive 1, and 0 for neutral. They assumed any words that did not exist in the sentiment lexicon had a polarity score of 0. The feature selection methods used were influenced by the data size, data consistency, and the need to investigate the most efficient feature selection method. Lastly, two machine learning classifiers were used in their study namely SVM and NB.

Reference [25] conducted a Malay SA using a combination of several machine learning classifiers namely Deep Belief Network (DBN), NB, and SVM which operated on document level. Besides than that, Malay sentiment lexicon was used for tagging the words inside their dataset along with its polarity value for the feature extraction phase. The English WordNet was used as the Malay sentiment lexicon resource where the English words were translated to its corresponding Malay words. Moreover, there were more than one Malay native speaker assigned manually the polarity value and synonym for each translated Malay phrase and word. The polarity value assigned was ranging from -5 which is strongly negative to 5 which is strongly positive. For the overall sentiment classification, the simple majority voting was used where two out of three classifiers should agree on the document class. Lastly, the performance of the combination approach was compared with its own individual classifier with combination of various feature sets. Table 1 shows the summary of approaches used in Malay SA.

TABLE 1. Summary of malay sentiment analysis approaches.

No	SA Approach	Article
1.	Lexicon-based	[3], [14], [2], [15], [4], [7]
2.	Machine Learning	[21], [20], [19], [17], [11], [18], [22]
3.	Hybrid	[12], [24], [23], [25]

B. PRE-PROCESSING TECHNIQUES USED IN MALAY SENTIMENT ANALYSIS

The informal writing style used by many users of social media sites like Facebook and Twitter caused problem to many NLP

applications like SA because they were generally trained on clean text [10]. Therefore, the pre-processing phase played a very crucial role in SA process because it cleaned the noisy text into a text that can be computationally process. In this review, there were 17 pre-processing techniques used by all the selected studies namely stop words removal, case folding, tokenization, spelling correction, lemmatization, stemming, spam removal, diacritics removal, intrinsic words removal, repeated characters removal, emoticons removal, punctuation marks removal, symbols removal, social media tags removal, characters removal, non-words removal, and others as shown in Table 2.

Stop words removal technique has been used by [4], [7], [11], [12], [17], [20], [23]–[25] in their study. By using this technique, [7], [11], [12], and [25] removed both Malay and English stop words from their dataset where [7] excluded some of the words from the English stop words list like "*Although*", "*and*", "*cannot*", "*can*", "*but*", "*could*", "*couldn't*", "*cry*", "*has*", "*should*", "*not*", and " *hasn't*" which they believed will reduce the SA accuracy if they were used. Other than that, [23] only removed the Malay stop words while [4] removed all the common Sabah Malay words.

Next, tokenization technique which helps to simplify the SA process has been used by [4], [12], [17], [22]–[25] in their study. Based on the review, [12], [17], and [24] tokenized their dataset based on the location of the punctuation marks or whitespace between the words while [4] tokenized their dataset based on the location of whitespace only. Other than that, case folding technique has been used by [2], [4], [11], [15], and [19] in their study. Based on the review, [2], [11], [15], and [19] converted their dataset into the lower case format while [4] did not stated clearly in what format their dataset has been converted.

Another technique is spelling correction which corrects spelling errors such as abbreviations. This technique has been used by [4], [14], [17], and [24] in their study. In the review, [14] created manually a dictionary named "full form dictionary" for converting the noisy word into its meaningful word. Other than that, [17] solved their abbreviations problem by implemented the spelling correction algorithm. For [4], they only corrected the misspelled Malay words because Sabah language has no fix spelling. Lemmatization and stemming techniques both have been used by [4] in their study.

Next is basic removal technique which includes symbols removal, punctuation marks removal, social media tags removal, spam removal, diacritics removal, intrinsic words removal, repeated characters removal, emoticons removal, characters removal, and non-words removal. The symbols removal technique has been used by [2], [4], [7], and [21] in their study. By using this technique, [2] removed excessively used symbols from their dataset while [7] used regular expression to remove unwanted or specific symbols without affecting the emoji present in their tokenized dataset.

The next technique is to remove punctuation marks which has been used by [17], [24], and [25]. Reference [25] has stated clearly that they cleaned the punctuation marks such as periods and commas in their study. Next, social media tags removal technique has been used by [20] and [24] in their study. For [24], the social media tags removed were user id and hashtags only. Another technique is to remove

TABLE 2. Summary of pre-processing techniques used in malay sentiment analysis.

No	Pre-processing Techniques		Reference
1.	Stop Words Removal		[25], [23], [24], [12], [11], [17], [20], [7], [4]
2.	Case Folding		[11], [19], [4], [15], [2]
3.	Tokenization		[25], [23], [24], [12], [17], [4], [22]
4.	Spelling Correction		[24], [17], [4], [14]
5.	Spam Removal		[2]
6.	Lemmatization		[4]
7.	Stemming		[4]
8.	Diacritics Removal		[20]
9.	Intrinsic Words Removal		[17]
10.	Repeated Characters Removal		[20]
11.	Emoticons Removal		[2]
12.	Punctuation Marks Removal		[25], [24], [17]
13.	Symbols Removal		[21], [7], [4], [2]
14.	Social Media Tags Removal		[24], [20]
15.	Characters Removal		[21]
16.	Non-words Removal		[24], [17]
17.	Others	Cleaned unwanted tags to obtain the words including abbreviations	[3]
		Choose only tweets that were written in English, Malay or Indonesian language and contained only one of the subjects being monitored in their project	[15]
		Filtered out any tweets that did not include any sentiment words which included in their sentiment lexicon	[15]
		Choose only tweets that can be encoded and decoded in UTF-8	[15]
		Merged the word "tidak" with the next word to cater for negative words in the Malay language	[11]
		The duplicates that may deviate or modify the SA's result were removed	[25]
		Any word in a tweet that contained "www" or "https://", "#hashtag", and "@username" were converted to "URL", "hashtag" and "AT_USER" respectively	[19]
		Trim	[19]

non-words which has been used by [17] and [24]. Spam which mostly consist of irrelevant words and emoticons were removed by [2] in their study. Besides than that, all repeated characters and diacritics were removed by [20] in their study. Lastly, both characters removal and intrinsic words removal technique were done by [21] and [17] respectively in their study.

Beside the above techniques, there are other pre-processing techniques used by previous studies. Based on the review, [3] cleaned their dataset which consisted of unwanted tags to obtain the words including abbreviations. For [15], their dataset was filtered by choosing tweets that were written in English, Malay or Indonesian language only. Other than that, they only chose the tweets which contained only one of the subjects being monitored in their project. To avoid huge numbers of neutral tweets, they filtered out any tweets that did not include any sentiment words which included in their sentiment lexicon. To avoid unreadable symbols and characters, the selected tweets must able to be encoded and decoded in UTF-8. In [11], they merged the word ''tidak'' with the next word to cater for negative words in the Malay language. Next, for [25], the duplicates that may modify the sentiment analysis's result were removed. Lastly, for [19], any words in a tweet that contained ''www'' or ''https://'', ''#hashtag'', and ''@username'' were converted to ''URL'', ''hashtag'', and ''AT_USER'' respectively. They also used another technique called trim.

C. DATASET USED IN MALAY SENTIMENT ANALYSIS

All reviewed datasets were collected from several online resources such as Facebook, Twitter, blog, news, online forum, and other general websites or social media sites as shown in Table 3. In this review, [2]–[4], [7], [11], [14], and [23] used Twitter as their source or one of their source of dataset. Next, [2]–[4], [7], [11], [14], and [23] used Facebook as their source or one of their source of dataset.

Moreover, [12], [17], [22], [23], and [25] used blog as one of their source of dataset. Furthermore, movie feedbacks or any reviews from online forums were used by [11], [12], and [25] as one of their source of dataset. Next, there were three studies used general websites as their source or one of their sources of dataset. References [20] and [22] gathered movie reviews from several web pages in Malay while the third study is by [23]. Other than that, [23] also used news as one of their source of dataset. Finally, [17] used reviews collected from several online social media as one of their source of dataset.

TABLE 3. Summary of dataset used in malay sentiment analysis.

No	Dataset's Source	Reference
1.	Facebook	[3], [14], [2], [4], [7], [11], [23]
2.	Twitter	[15], [4], [7], [21], [19], [11], [18], [24], [23]
3.	Blog	[17], [12], [23], [25], [22]
4.	News	[23]
5.	General Websites	[20], [23], [22]
6.	Online Forum	[11], [12], [25]
7.	General Online Social Media	[17]

TABLE 4. Summary of performance measure used in Malay sentiment analysis.

No	Performance Measure	Reference
1.	Accuracy	[14], [2], [15], [4], [7], [19], [11], [18], [24], [23]
2.	Precision	[2], [21], [12], [24], [23]
3.	Recall	[2], [21], [12], [24], [23]
4.	F-measure	[21], [20], [17], [12], [24], [25], [22]
5.	Macro-F1 measures	[20], [17], [22]

D. PERFORMANCE MEASURE USED IN MALAY SENTIMENT ANALYSIS

Performance measure is very important for determining how efficient a SA system is functioning. In this review, there were 10 studies used a metric known as accuracy as their or one of their performance measures. All of the studies were [2], [4], [7], [11], [14], [15], [18], [19], [23], and [24] where [15] combined the accuracy with confidence level and [19] combined the accuracy with standard deviation. Next, there were 5 studies used a metric known as precision as one of their performance measures. All of the studies were [2], [12], [21], [23], and [24]. Other than that, there were also 5 studies used a metric known as recall as one of their performance measures. All of the studies were [2], [12], [21], [23], and [24]. The next performance measure is F-measure. There were 7 studies used F-measure as their or one of their performance measures. All of the studies were [12], [17], [20]–[22], [24], and [25]. Lastly, there were 3 studies used Macro-F1 measures as one of their performance measures. All of the studies were [17], [20], and [22]. This review found that a publication by [3] did not state clearly on how the performance measure is being used in their study. Table 4 shows the summary of performance measure used in Malay SA

E. CHALLENGES

There were three main challenges existed for the SA on noisy Malay text. Noisy Malay text have a lot of informal patterns or rules which are not easy to handle because they are often ambiguous and messy. Some of the noisy Malay text can be grouped as one pattern but some are not because their pattern is too individualistic. Other than that, noisy Malay text from any social media sites like Facebook and Twitter will evolve from time to time and some of the rules or patterns cannot be used anymore because the trend is always changing. Hence, it is not easy to handle the noisy Malay text where it keeps on evolving and becoming more complex. The next challenge is lack of relevant resource for Malay SA. Most of the SA has been done for the English language but still lacking for the Malay language domain. There is lack of

relevant Malay sentiment lexicon exist for the lexicon-based approach. Other than that, the machine learning approach has lack of relevant training data for the Malay language domain. Therefore, any research on the Malay SA will need a longer time to complete compare to research in English language domain because more time is needed to prepare a relevant sentiment lexicon or training data. The last challenge is no available techniques for handling Malay sarcasm. Sometimes people tend to use sarcasm text in the social media sites like Facebook and Twitter. Thus, it is not easy to detect whether a text is considered as sarcasm or not because the true meaning really depends on the individual desire.

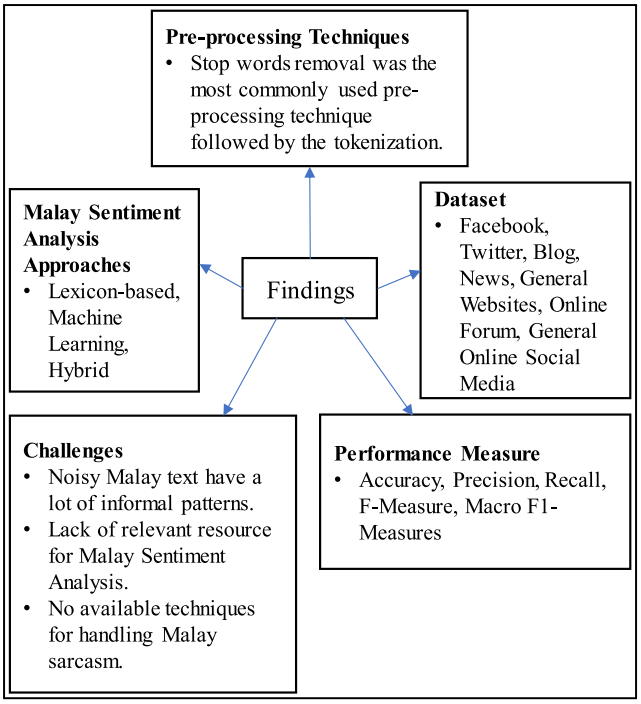

FIGURE 2. Summary of findings.

Figure 2 presents the summary of findings based on the Malay SA approaches, pre-processing techniques, types of dataset, performance measure, and challenges.

IV. DISCUSSION

Works on SA is mostly done for English language and still consider lacking for Malay language. For this review, we focused on research study that done the SA for noisy Malay text. Hence, based on the review, four main modules were needed in order to do a complete process and evaluation for noisy Malay text SA namely dataset, pre-processing, sentiment classification, and evaluation. For the dataset module, it focused more on the type of dataset used and determined how big is the workload needed to be done by the pre-processing module in order to produce a clean text. A clean text is very crucial to many NLP tools because they were generally trained on clean text in order to produce a good result. For the sentiment classification module, it focused more on the type of SA approach used. There were three main approach for SA namely machine learning, hybrid, and lexicon based. The last module, which is the evaluation module focused more on the performance measure used in order to determine how well the SA system is functioning.

Selecting the most suitable SA approach to detect sentiment value for noisy Malay text is quite challenging for researchers because every SA approach has its own advantages and disadvantages over each other. Based on the review, machine learning approach is the most common approach used to detect sentiment value for the noisy Malay text. Next is lexicon-based approach followed by hybrid approach. There were 41.2% of the publications reviewed used the machine learning approach, another 35.3% used the lexicon-based approach, and 23.5% used the hybrid approach. A lexicon-based approach is more flexible compare to the machine learning approach because it does not focus on a specific domain. Although we can train the machine learning classifiers using a generic dataset, it will require a very high-quality training corpus which is not easy to find for the Malay language. Same goes to the hybrid approach which used the machine learning classifier as one of the approach combinations. Other than that, it is easier to handle or improve a wrong sentiment value classification when using a lexicon-based approach by adding in certain rules such as negation handling, intensifier handing, or even contrast handling. Hence, for a specific domain, a machine learning approach will perform better compare to the lexicon-based approach and vice versa. For the hybrid approach, it still depends on the combination of SA approaches or machine learning classifiers used. Although a lexicon-based approach is more suitable for noisy Malay text SA, the availability of the sentiment lexicon is still lacking. Therefore, it will be a very tough challenge for researchers in doing the SA for language other than English, especially Malay language.

Most of the NLP tools were generally trained on a clean data. Therefore, pre-processing phase plays a very crucial role in SA because it helps to transform the noisy text into a cleaner text which will help in improving the SA result. In the publications reviewed, almost all researchers had their own compilation of pre-processing techniques used in a single pre-processing phase. Regarding RQ2, this review can conclude that stop words removal was the most commonly used pre-processing technique followed by the tokenization. There were 52.9% of the publications reviewed removed the stop words and another 41.2% tokenized their dataset during the pre-processing phase. Other than that, there were other pre-processing techniques used in the publication reviewed such as lemmatization, case folding, stemming, and many more as explained in section III(B). The stop words removal technique is helpful in the pre-processing phase because it helps to remove words that carry no important meaning with respect to SA, then speed up the classification process. The tokenization process will help to simplify the SA process. However, focusing only on those two techniques are not

enough to fully handle the noisy Malay text. Noisy Malay text in social media sites like Facebook and Twitter have many informal patterns which cannot be fully handled using those methods specified above only. There are a lot of ambiguities where using a dictionary to map between the noisy text to its normalized form is still not enough because some noisy words may have more than one meaning. Hence, all these challenges need to be handled in order to improve the SA result.

The datasets used in the publications reviewed were collected from several online sources for different studies such as Facebook, Twitter, blog, news, online forum, and other general websites or social media sites as explained in section III(C). All the datasets especially from the Twitter and Facebook are informal and unstructed. Therefore, most of the datasets are not correctly follow the grammatically correct syntactic format and often ambiguous and messy. These types of dataset make a SA process become more challenging because unstructed or informal dataset will reduce the SA accuracy compare to when using more formal or standard dataset. Hence, these types of dataset need to be analyzed properly in order to get a new combination of informal patterns or rules which will be used to handle them.

Regarding RQ4, this review can conclude that 58.8% of the publications studied have used accuracy as one of the metrics or a single metric for their performance measure. It is considered the most widely employed performance measure in this review. Accuracy refers to a measure of how close the measured value to the known or standard value. It is often confused with precision. Precision refers to how close the two, three or more measurements to each other. If you have three measurements which are the value close to each other but far from the known value, then you have precision without accuracy. Another metric, known as recall, refers to a measure of how many correctly predicted positive observations over all observations in real class. Finally, another metric, knows as the F-Measure refers to a weighted average of precision and recall. The F-Measure is a harmonic mean of the recall and precision. From all these matrices, accuracy will work best if the values of the false negative and false positive are almost the same. If the values of the false negative and false positive are different to each other, F-Measure is needed to be used because both recall, and precision are needed to be looked after.

Several main challenges existed for the noisy Malay text SA as explained in section III(E). All the challenges need to be handled properly in order to get a relevant Malay SA result. According to [26], non-English text like Indonesian also faces the same issue, lack of resources. Both machine learning and lexicon-based approaches have difficulties in terms of having a labelled data for training and dictionary which contains list of positive and negative sentiment words respectively [26]. Other than that, there is no fully reliable techniques for pre-processed the noisy text before SA can be applied [26]. Same issue also faces by Urdu language where they were lacking of lexical resources with a reliable scoring mechanism [27]. For the future work, a normalizer to normalize noisy Malay text is needed before proceeding to the SA phase. Other than that, there are several dictionaries are needed in order to help a normalizer to solve the informal pattern which are consider as individualistic and not consistent. The dictionaries will focus more on the Malay dialect, Malay trend word, and informal Malay texts which cannot be grouped in general pattern because there are not consistent and have an individualistic criterion. Furthermore, a specific technique is needed in order to handle a Malay sarcasm. Moreover, a relevant Malay sentiment lexicon is needed in order to handle a more generic dataset. The sentiment lexicon needs to have a diverse collection of Malay words with reliable sentiment value. Lastly, huge reliable Malay training data is needed for many important domains. The huge reliable training data is very important for the machine learning classifier to learnt and classify the dataset effectively.

V. CONCLUSION

This paper presented a review on various approaches, datasets, pre-processing techniques and performance measures used including challenges and future works in the previous works on SA of noisy Malay text. There were three fundamental SA approaches have been focused in this review namely lexicon-based, machine learning, and hybrid. Hence, other SA approaches and works using standard Malay dataset only are not included. As a conclusion, there are still a lot of works needed to be done in order to improve the noisy Malay SA overall methods. As for future research directions, a normalizer for Malay text is needed in order to clean the noisy Malay texts before proceeding to the SA phase. In addition, a specific technique is needed in order to handle a Malay sarcasm. Other than that, a reliable Malay sentiment lexicon and training data for many important domains are also needed in order to help speed up and facilitate the progress of improvement for Malay SA. Researchers in this domain can use this review as a reference for their current or forthcoming research since most of the datasets used contained huge amount of noisy text because Malaysians prefer to use noisy text when writing on social media [7]. In addition, this review also very useful for them because works on SA for malay language is still limited [3] , [4]. Finally, they also can use this review for increase their understanding on the current methods used to handle noisy Malay text because the existence of the noisy Malay texts have develop into one of the main obstacles in put into use SA applications to UGC [8] where they were largely trained on formal text [10].

VI. DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the authorship, research, and/or publication of this article.

REFERENCES

[1] M. Roblyer, M. Mcdaniel, M. Webb, J. Herman, and J. V. Witty, "Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social networking sites," *Internet Higher Edu.*, vol. 13, no. 3, pp. 134–140, Jun. 2010.

[2] N. F. Shamsudin, H. Basiron, Z. Saaya, A. F. N. Abdul Rahman, M. H. Zakaria, and N. Hassim, "Sentiment classification of unstructured data using lexical based techniques," *J. Teknologi*, vol. 77, no. 18, pp. 113–120, 2015.
[3] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment analysis: Determining people's emotions in Facebook," in *Proc. 13th Int. Conf. Appl. Comput. Appl. Comput. Sci.*, 2014, pp. 111–116.
[4] M. H. A. Hijazi, L. Libin, R. Alfred, and F. Coenen, "Bias aware lexicon-based sentiment analysis of Malay dialect on social media data: A study on the Sabah language," in *Proc. 2nd Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2016, pp. 356–361.
[5] L. Hong, G. Convertino, and E. H. Chi, "Language matters in Twitter: A large scale study," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, no. 1, pp. 518–521.
[6] M. S. M. Vohra and P. J. B. Teraiya, "A comparative study of sentiment analysis techniques," *J. Inf., Knowl. Res. Comput. Eng.*, vol. 12, no. 2, pp. 313–317, 2013.
[7] K. Chekima and R. Alfred, "Sentiment analysis of Malay social media text," *Comput. Sci. Technol.*, vol. 488, pp. 205–219, Feb. 2018.
[8] M. A. Saloot, N. Idris, and R. Mahmud, "An architecture for Malay Tweet normalization," *Inf. Process. Manage.*, vol. 50, no. 5, pp. 621–633, Sep. 2014.
[9] D. Handayani, N. S. Awang Abu Bakar, H. Yaacob, and M. A. Abuzaraida, "Sentiment analysis for Malay language: systematic literature review," in *Proc. Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M)*, Jul. 2018, pp. 305–310.
[10] T. Baldwin and Y. Li, "An in-depth analysis of the effect of text normalization in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 420–429.
[11] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in *Proc. World Congr. Eng.*, vol. 3, 2013, pp. 3–5.
[12] A. Alsaffar and N. Omar, "Integrating a lexicon based approach and K nearest neighbour for Malay sentiment analysis," *J. Comput. Sci.*, vol. 11, no. 4, pp. 639–644, Apr. 2015.
[13] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 628–632.
[14] N. F. Shamsudin, H. Basiron, and Z. Sa'aya, "Lexical based sentiment analysis-verb, adverb & negation," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 2, pp. 161–166, 2016.
[15] Y. F. Tan, H. S. Lam, A. Azlan, and W. K. Soo, "Sentiment analysis for telco popularity on twitter big data using a novel Malaysian dictionary," in *Proc. Frontiers Artif. Intell. Appl.*, vol. 282, 2016, p. 112.
[16] F. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proc. CEUR Workshop*, vol. 718, 2011, pp. 93–98.
[17] A. Alsaffar and N. Omar, "Study on feature selection and machine learning algorithms for Malay sentiment classification," in *Proc. 6th Int. Conf. Inf. Technol. Multimedia*, Nov. 2014, pp. 270–275.
[18] V. A. Rohani and S. Shayaa, "Utilizing machine learning in sentiment analysis: SentiRobo approach," in *Proc. Int. Symp. Technol. Manage. Emerg. Technol. (ISTMET)*, Aug. 2015, pp. 263–267.
[19] M. N. M. Ibrahim and M. Z. M. Yusoff, "Twitter sentiment classification using Naive Bayes based on trainer perception," in *Proc. IEEE Conf. e-Learn., e-Manage. e-Services (IC3e)*, Malacca, Malaysia, Aug. 2015, pp. 187–189.
[20] T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar, "Feature selection methods effects on machine learning approaches in Malay sentiment analysis," in *Proc. 1st ICRIL-Int. Conf. Innov. Sci. Technol. (IICIST)*, 2015, pp. 1–2.
[21] S. Tiun, "Experiments on Malay short text classification," in *Proc. 6th Int. Conf. Electr. Eng. Informat. Sustain. Soc. Digit. Innov. (ICEEI)*, vol. 2017, Nov. 2017, pp. 1–4.
[22] T. Al-Moslmi, N. Omar, M. Albared, and A. Alshabi, "Enhanced Malay sentiment analysis with an ensemble classification machine learning approach," *J. Eng. Appl. Sci.*, vol. 12, no. 20, pp. 5226–5232, 2017.
[23] A. A. Sadanandan, N. A. Osman, H. Saifuddin, M. K. Ahamad, D. N. Pham, and H. Hoe, "Improving accuracy in sentiment analysis for Malay language," in *Proc. 4th Int. Conf. Artif. Intell. Comput. Sci.*, Nov. 2016, pp. 28–29.
[24] M. I. Eshak, R. Ahmad, and A. Sarlan, "A preliminary study on hybrid sentiment model for customer purchase intention analysis in social-commerce," in *Proc. IEEE Conf. Big Data Anal. (ICBDA)*, Kuching, Malaysia, 2017, pp. 61–66.
[25] A. Al-Saffar, S. Awang, H. Tao, N. Omar, W. Al-Saiagh, and M. Al-bared, "Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0194852.
[26] F. Djatmiko, R. Ferdiana, and M. Faris, "A review of sentiment analysis for non-English language," in *Proc. Int. Conf. Artif. Intell. Inf. Technol. (ICAIIT)*, Mar. 2019, pp. 448–451,
[27] M. Z. Asghar, A. Sattar, A. Khan, F. M. Kundi, A. Ali, and S. Ahmad, "Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language," *Expert Systems*, vol. 36, no. 3, 2019, Art. no. e12397.

MUHAMMAD FAKHRUR RAZI ABU BAKAR received the bachelor's degree in computer science (artificial intelligence) from the University of Malaya, Kuala Lumpur, Malaysia, where he is currently pursuing the master's degree with the Faculty of Computer Science and Information Technology. His research interests include text normalization and sentiment analysis.

NORISMA IDRIS received the Ph.D. degree in computer science from the University of Malaya, in 2011. She joined the Faculty of Computer Science and Information Technology, University of Malaya, in 2001. She is currently the Associate Professor with the Artificial Intelligence (AI) Department. Her research interest is on natural language processing (NLP), where the main focus is on developing efficient algorithms to process texts and to make their information accessible to computer applications, mainly on text normalization, and sentiment analysis. She is working on a few projects, such as Malay Text Normalizer for Sentiment Analysis with an Industry, and Implicit and Explicit Aspect Extraction for Sentiment Analysis under the Research University Faculty Grant. For the past five years, she has published more than 15 articles on NLP and AI in various WoS-indexed journals.

LIYANA SHUIB received the master's degree in information system (data mining) from Universiti Kebangsaan Malaysia, in 2005, and the Ph.D. degree from the University of Malaya, Malaysia, in 2013. She is currently a Senior Lecturer with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya. She has published a number of journal articles and proceedings locally and internationally. Her research interests include personalization, e-learning, recommender system, data science, data mining, artificial intelligence application, and educational technology. She is also a member of IEEE computing society, an active blogger and presently, and the principal investigator of multiple research grant in the Faculty. She has received more than 20 awards from reputable innovation competition internationally.

NORAZLINA KHAMIS received the Bachelor of Information Technology (BIT) degree (Hons.) from the University of Malaya, in 1999, the M.Sc. degree in realtime software engineering from Universiti Teknologi Malaysia, in 2001, and the Ph.D. degree from Universiti Kebangsaan Malaysia, in 2012. She is currently attached with Universiti Malaysia Sabah. Her research interests include software engineering, intelligent software engineering, the Internet of Things, and software quality. She also involved in research related to ICT in disaster management system. She is also a Fellow in Natural Disaster Research Centre, UMS. She is also working with several project related with the Internet of Things.