International Journal of Business Analytics
Volume 10 • Issue 1

Sentic-Emotion Classifier on eWallet Reviews

Tong Ming Lim, Tunku Abdul Rahman University of Management and Technology, Malaysia*
Yuen Kei Khor, Tunku Abdul Rahman University of Management and Technology, Malaysia
Chi Wee Tan, Tunku Abdul Rahman University of Management and Technology, Malaysia
https://orcid.org/0000-0001-6828-4896

ABSTRACT

Emotion classification using hybrid framework using lexicon and machine learning algorithms have been proven to be more accurate. This research analyses emotions from reviews of a popular eWallet mobile application in Malaysia. The proposed Sentic-Emotion Classifier is evaluated on its performance as it analyses the code-switched reviews crawled that contain formal and informal or out-of-vocab words. The code-switched reviews are mainly made up of words and expressions in English and Malay language models. This research designs, implements, and investigates several novel techniques that have been shown to have reliable and consistent predictive outcomes, and these outcomes are validated with manually annotated reviews so that the proposed classifier can be evaluated objectively. The novel contributions of the Sentic-Emotion Classifier consist of 2-tier sentiment classification, extended emolex framework, and multi-layer discrete emotion hierarchical classes which is hypothesized to be able to yield better accuracy for emotion and intensity prediction for the proposed framework.

KEYWORDS

Code-Switched, Electronic Wallet, Emotion, Emotion Analysis, Lexicon-Based, Sentiment, Social Media Text, Supervised Machine Learning, TFIDF

INTRODUCTION

In recent years, the increase of cashless transactions in many countries such as China, Singapore, and Malaysia, are largely due to the rapid development of financial technology and higher consumer confidence on secured money-over-web activities. The adoption rate of fintech products such as eWallet by 21st century young consumers from cash-based to cashless has shifted rather quickly. These young consumers have always been regarded as tech-savvy users of the smartphone era. In Malaysia, 42 eWallet service providers have received official licenses from BNM (Bank Negara Malaysia) and six (6) of them are more popular and widely adopted. They are AEON Wallet, Boost, BigPay, GrabPay, WeChat pay, and Touch'n Go eWallet (Aji et al., 2020; Karim et al., 2020; Ray, 2017; Upadhayaya, 2012).

DOI: 10.4018/IJBAN.329928 *Corresponding Author

The number of digital payment providers increased by leaps and bounds during the COVID-19 pandemic period, as people tried to reduce physical contact with other people. Hence, understanding customers' needs are extremely important in order to drive business growth, provide better customer services, as well as deeply understanding the strengths and weaknesses of their products. The tech savvy generation, young and old, prefer to express their feelings and opinions on the software or services that they experience on the social media sites. In order to understand the perceptions of digital payment users, sentiment and emotion analysis can be analysed based on users' reviews, which can be collected from online app stores, such as Playstore or Appstore, social media sites, such as Facebook and Instagram, or product review platforms.

Medhat et al. (2014, p. 1094) defined sentiment analysis as a study of people's opinions, emotions, and attitudes toward an entity such as individuals, events, and topics. It evaluates the perception of humans towards entities and enables business organizations to employ effective decision-making. Sentiment analysis classifies the sentiment of a text document into three categories, which are: positive, negative, and neutral. For example, "The customer service is so poor! No one replies to me!" is a negative sentiment, and it is important to understand the customers' reaction towards the products and services they consumed. As a business grows, customer insight is vital, as it provides valuable information to the organization to improve the quality of services and products. In addition, emotion analysis is another dimension of affective analysis that can be conducted to further understand how customers feel based on the reviews collected. Emotion analysis is similar to sentiment analysis, but it is more specific because it classifies the reviews into one or more emotion categories, such as angry and/or happy. There are two emotion models, which are widely used in emotion analysis. Ekman's six basic emotions (Mohammad & Turney, 2013) contain anger, disgust, fear, happiness, sadness, and surprise emotions. On the other hand, Plutchik's wheel of emotion (Plutchik, 2003) defines a set of eight emotions, six of the emotion categories were adopted from Ekman, with two additional emotions added: trust and anticipation.

Malaysia is a multi-ethnic country (Vollmann & Soon 2018, p. 36) that is comprised of three keys ethnic groups: Malay, Chinese, and Indian. Malaysia is a multilingual, multicultural, and multi-region society, as most of the Malaysians can speak multiple languages, such as English, Chinese, and Tamil, other than the national language of Bahasa Melayu. Code mixing is not only common during Malaysian verbal conversations, but also in their social media texting. Social media texts are unstructured, which has posed some challenges in sentiment analysis and emotion analysis. The texts usually contain many formal abbreviations or acronyms, such as *etc* (et cetera) and *IoT* (Internet of Things), informal shorthand's, such as *yg* (yang) and *gmbr* (gembira), local dialect in Romanised alphabets, such as *chialat* (to describe a negative situation) and *kiasu* (afraid to lose out), internet slangs, such as *LoL* (lots of laugh) and *brb* (be right back), international slang (such as *to clap back*, which means to respond to another person's criticism or *yyds* (, y ng yu n de shén), which means "eternal god"), repeating characters or words such as *sooooo cooooool* or *pannndddaaaai tu,* and misspelled words (Kham, 2019). The study by binti Sabri et al. (2020) on internet slang used Malaysians between 15 and 30 years old in the social media world, pointed out that internet slangs can be further categorized into four more refined types: phonetic replacement, phrase abbreviations, word abbreviations, and inanity. For examples, internet slang includes phonetic replacement ("everyone" – "every1"), phrase abbreviations ("on the way" – "otw"), word abbreviations ("please" – "pls"), and inanity ("it is" – "itz"). Such slang terms may carry sentiment and emotion, but they complicate the process of sentiment and emotion analysis, making analysis more challenging, as they vary from the writing style of one person to another person. Even the youth generation may also have trouble understanding and interpreting internet slang.

This article investigates the emotion of eWallet users based on their opinionated reviews. The idea of this research was based on the work undertaken by Balakrishnan et al. (2020). According to the authors' emotion analysis on Malaysia, digital mobile payments are very much lacking. This research addressed the gaps found by carrying out reviews from Google Play Store to collect one of

the popular mobile eWallet service providers. For privacy reason, the provider will not be disclosed in this paper. The selected provider has about five million installations. It was chosen as the data source provider for this research. This paper reports the outcomes of the research based on code switching texts, considering informal or Out-of-Vocab terms (or words) from social media context where code-mixed texts are not being investigated.

REVIEW OF PAST RELATED WORKS

This section reviews past related works on electronic commerce and eWallet, sentiment and emotion analysis. The discussion also highlights the need to carry out this research due to a lack of works that analyse reviews or comments posted by Malaysians in a very unique way using a mix of formal and informal expressions in the social media communities.

Electronic Commerce and eWallet

Electronic commerce has been driving the development of eWallet as a convenient, easy-to-use, secure global payment system (Baker-Eveleth & Stone, 2015; Karim et al., 2020; Upadhayaya, 2012). Electronic payment systems or EPSs enable a customer to pay for goods and services online by using integrated hardware and software systems. One of the objectives of EPS is to provide the best quality of experience in addition to increasing efficiency, security, and enhancing services to customers with optimum ease of use (Upadhayaya, 2012). Upadhayaya(2012) stated some benefits of the functions provided by most eWallet, which include the ability to send and receive payments anywhere in the world, received email or SMS after transactions, easy recurring payments and transfers, the ability to manage one's account from their mobile phone, being able to withdraw money from any bank into eWallet, receiving wired funds/transfers directly into eWallet, transferring money from eWallet to eWallet without sharing personal account numbers, and many others. However, challenges of any eWallet implementation may include authentication processes, such as digital signatures, fingerprints, and passwords. Data integrity and confidentiality are the two other key challenges that providers of such services need to address.

Sentiment and Emotion Analysis

Sentiment analysis (Medhat et al., 2014) is a process of identifying the sentiment expressed from texts where sentiment can be classified into three main groups that include: positive, negative, and neutral. Though sentiment analysis enables the business community to understand whether the reaction from the public is positive or negative, this is not enough to understand how people actually feel. Emotion analysis, on the other hand, is a process that identifies the emotion expressed by texts (Hakak et. al, 2017) Emotions such as angry, happy, sad, and surprise are detected through emotion analysis on textual data.

Drus and Khalid's (2019) study undertaken between 2014 and 2019, found that Naïve Bayes and Support Vector Machine (SVM) learning models were frequently used to detect polarity from text documents on sentiment analysis using a machine learning approach. Schmidt & Burghardt (2018), pointed out that sentiment analysis and emotion analysis are often performed using supervised Machine Learning (ML) or lexicon-based approaches. Supervised machine-learning algorithms are used as classifiers with a set of labelled data in order to classify the incoming words or phrases to appropriate sentiment or emotion categories. Data is divided into two portions, training and testing data, and usually with a 70/30 or 80/20 split. After classifiers learn the patterns from labelled training data, testing data is used to perform classification or prediction based on the prior training given. Performance of the classifiers is measured in terms of accuracy, F1-score, precision, and recall. In order to conduct sentiment and emotion analysis, human power is required to annotate the sentiment and emotion of a set of unstructured text, as they could be reviews or comments. Human analysts tend to agree to 80% to 85% accuracy, which is usually known as the human baseline agreement

when evaluating the sentiment of a given text document (Arafat et al., 2014). However, the baseline agreement on emotion analysis will be 60% to 79% (Aman & Szpakowicz, 2007), slightly lower than the baseline agreement for manual sentiment annotation.

Aman and Szpakowicz (2007) proposed a knowledge-based approach to classify emotional and non-emotional sentences automatically. They compared the performance of Naïve Bayes and SVM to four feature groups and the results showed that SVM had better performance out of a total of three feature groups. The highest emotion classification accuracy was achieved by SVM, with 73.89%, which is higher than the baseline they defined. In the work conducted by Sharma and Dey (2012), they studied the performance of five feature selection techniques (Document Frequency, Information Gain, Gain Ratio, CHI statistics, and Relief-F) using seven machine learning-based classification techniques (Naïve Bayes, Support Vector Machine, Maximum Entropy, Decision Tree, K-Nearest Neighbor, Winnow, and Adaboost) for sentiment analysis on movie online reviews. Results showed that SVM outperformed other techniques, and the Naïve Bayes classifier provided better results with fewer features. Their experiment showed that Gain Ratio is the best feature, and proved that the high impact feature improved the performance of the analysis, but depended on the feature selection methods and the numbers of features selected. Muljono et al. (2016) claimed the model with better performances in four different classification methods: Naive Bayes (NB), J48, K-Nearest Neighbor (KNN), and Support Vector Machine-Sequential Minimal Optimization (SVM-SMO) for Indonesian text emotion detection was SVM-SMO. This experiment used 1000 sentences that consisted of six emotion classes: anger, disgust, fear, joy, sadness, and surprise in the Indonesian text corpus where TF-IDF is the feature extracted from the corpus. They compared the result that performed 10-fold cross validation and split validation in their experiments. SVM-SMO classifier delivered the best performance for both, where their results showed an accuracy of 85.5% for the 10-fold cross validation and 86% for split validation. However, the study by Moraes et al. (2013) obtained a completely different result. The comparison of classification accuracy between SVM and Artificial Neural Networks (ANN) on document-level sentiment analysis with several weighting schemes for different domains indicated that ANN either performed comparable or better than SVM for the movie review dataset.

A lexicon-based approach relied on the lexicons in order to aggregate the score of each word from the reviews, as they were found in the lexicons. Some lexicons contained sentiment polarity or/and emotion intensity of each word or words. The NRC Emotion Lexicon is widely used in sentiment and emotion classification for corpus of different subjects, such as political tweets (Bose et al., 2019), product reviews (Bose et al., 2020), and tweets with positive or negative hash-tagged words (Mohammad et al., 2013). The NRC Emotion Lexicon (Mohammad & Turney, 2013) is a word-level lexicon, which consists of 14,182 words that are associated with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The score ranges from minus-one to one. Drus and Khalid (2019)'s research found that SentiWordNet for sentiment classifications is widely used in the lexicon-based approach. SentiWordNet (Baccianella et al., 2010) is a lexical resource that consists of synsets from WordNet with three scores: positivity, negativity, and objectivity, which indicate positive, negative, and objective (i.e., neutral) of the terms where the score values range between zero and one. The higher the score, the higher the emotion or sentiment a word carries. However, these lexical resources may not be able to work accurately on reviews and comments posted by Malaysians due to their heavy use of mixed languages through either code switching or code-mixing mode with extensive informal expressions, such as abbreviation, dialects in Roman forms, and slangs terms.

There are three approaches to develop a lexicon: manual-, dictionary-, and corpus-based approach. Lexicon can be developed manually by hiring human annotators, but it is a high cost and time-consuming task. The dictionary-based approach expands the lexicon by searching for synonyms or antonyms from seed words in a corpus. Hu and Liu (2004) searched synonyms and antonyms of seeds in WordNet (Miller, 1995) and treated them as new opinion words. The process was repeated

until no new word could be added to the lexicon. The word coverage of corpus will affect the size of the lexicon to be created; however, this approach is not suitable to build an informal lexicon as only formal words can be found. The last approach is a corpus-based approach where it depends on the syntactic or co-occurrence of words or expressions in the text content of the input. It uses a seed list of opinion words to find other opinion words in a large corpus. This approach finds domain-specific opinionated words or phrases, as well as sentiment base on domain specific corpus (Amiri & Chua, 2012). For example, Hatzivassiloglou and McKeown (1997) and Kanayama and Nasukawa (2006) used a list of seed opinionated adjectives with a set of conjunctions (and, or, but, either-or, neither-nor) to identify and generate larger opinionated words. For example, "and" is a word that connects words with the same polarity (e.g., the delivery service is bad and slow). Therefore, the polarity of another word can be deduced if the polarity of one of the words is known. On the contrary, conjunctions like "but" and "however" can be used to indicate the changes of polarity.

A review of the past research had found that the difference between supervised machine learning and lexicon approaches is that the former requires labelled data, but does not rely on lexicon resources, while the latter relies heavily on lexicon resources without considering labelled data at all (Kamble & Itkikar, 2018). The advantage of a supervised machine-learning approach is that it outperforms other approaches, but a huge amount of labelled training data is needed to train the classifier (Khoo & Johnkhan, 2018). Sizeable and good-quality data will lead to high classification accuracy. If the training data is not large enough, a lexicon-based approach is suitable for sentiment and emotion categorization. Besides, the performance of classifiers can perform well in a domain, but drops precipitously when the same classifier is used in a different domain (Aue & Gamon, 2005). Chekima and Alfred (2018) stated that the drawback of supervised machine learning is more computationally expensive in terms of CPU processing, requirements, and training or classification time. Sometimes it takes a few hours to train the classifiers if complex models, such as SVM and deep learning models, are used. In contrast, the analysis speed of a lexicon-based approach is faster, as training is not required. Besides, a lexicon-based approach also provides the flexibility to improve classification accuracy by adding linguistic rules and syntactical or structural rules that handle features: negation, intensity, and question characteristics in the texts. Unlike supervised machine learning, a lexicon-based approach is difficult to modify because it derives features "behind the scenes" in a black box, which increases the difficulty for humans to interpret (Chekima & Alfred, 2018). Another drawback of the lexicon-based approach is that words can have multiple meanings and senses in a sentence depending on the context of the words, and they may be common in one domain, but not another (Khoo & Johnkhan, 2018). The classification accuracy also highly depends on the quality of the lexicons on many occasions.

As discussed earlier, supervised machine learning requires a large set of labelled data in order to boost performance of the model; but it is time-consuming and costly in the data annotation process. According to the emotion analysis survey study of Hakak et al. (2017), the common features used by most of the previous work are Part-of-Speech (POS) tag and bag of word. There is limited research work on exploring the effectiveness of other new features to generate a better emotion classification model. To address these gaps, this study introduces two mathematical formulas that have the ability to reduce human labour costs and time for annotating reviews. The formulas help in calculating the polarity score and emotion score of the reviews by leveraging the lexicon resources. Besides, additional numerical features, such as the emotion score for each emotion category, polarity score for positive and negative sentiment, review level emotion score, and polarity score, are also created to understand whether these features contribute a positive impact on model performance.

PROBLEMS, OBJECTIVES AND RESEARCH QUESTIONS

This research was motivated by several problems relating to the lexicons used as reported by Madhoushi et al. (2015). The issues on lexicon-based approaches included limited non-English resources and

limited word coverage, which failed to recognise emotion words (or terms), especially domain-specific words (Madhoushi et al., 2015). Tan et al. (2016) stated that local slangs and abbreviations highly occur in social media text, especially for multi-ethic countries like Malaysia, where the lack of such a lexicon that is capable of identifying these words may contribute to poor sentiment and emotion analysis. In order to tackle challenges, such as local slang and abbreviations, researchers (Chekima & Alfred, 2018; Kundi et al., 2014) attempted to build lexicons that consisted of informal words. Such a time-consuming task is costly and labour intensive. Therefore, a lot of research attempted to use deep-learning algorithms due to their scalability, learning capacity, and reliable accuracy (Canales et al., 2019) by investing in high computing costs and long model training.

This research aims to classify the emotions of mobile eWallet reviews and analyse the satisfaction of users about their services. In addition, four supervised machine-learning models are used to determine their accuracy and F1-score on the predicted sentiment and emotion, as well as the AUC. The comparative analysis of the models' performance of this research is also discussed with respect to outcomes from Balakrishnan et al. (2020) work. As mentioned in the literature review, feature selections will affect the performance of the classification results. The feature selection method used in Balakrishnan et al. (2020) was only TF-IDF. In this research, additional features are added to the six emotions and sentence intensity scores; negative, positive, and sentence sentiment polarity scores, and were included in the comparative analysis.

This article investigates three research questions:

Question One: Which machine-learning model can classify emotion with the highest accuracy rate?

Question Two: What is the performance improvement of the four different supervised machine-learning models when additional features, such as emotion scores, sentence emotion score, negative score, positive score, and sentence polarity score, are considered in addition to TF-IDF?

Question Three: Do the machine-learning models perform better than the baseline agreement on emotion analysis, which is 60% to 79%?

SENTIC-EMOTION CLASSIFIER ARCHITECTURE

The Sentic-Emotion Classifier (SEC) layered architecture consists of seven layers: eWallet data, text pre-processing, expand emotion lexicon, feature generation, manual annotation, data splitting, and model building. The detailed descriptions for each layer are discussed in this section. The processes of the lower layer support the higher layer (Richards, 2015) in a layered architecture (Figure 1).

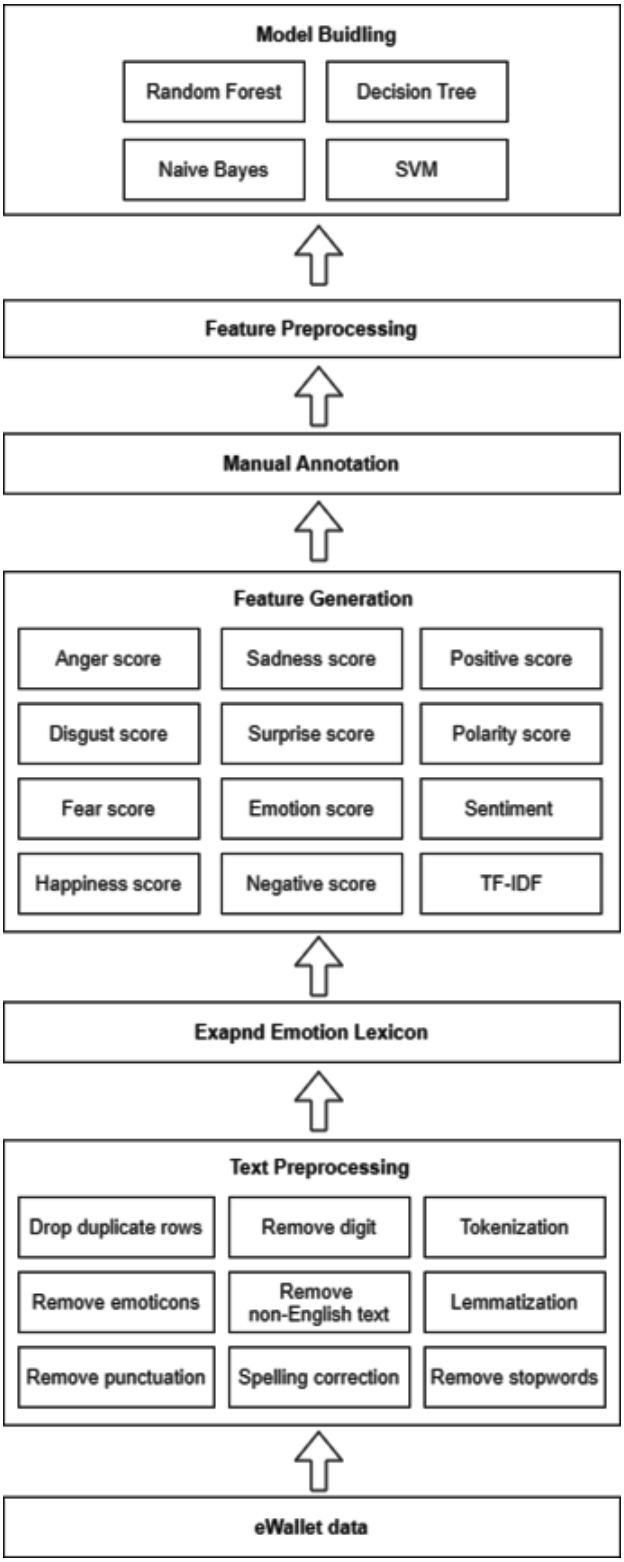
Dataset

EWallet is a popular Malaysia digital wallet app that has a total of one million installations available from Google Play Store. The dataset for this experiment was obtained from the Google Play Store using Python's Selenium package (Muthukadan, 2014). There were 2480 users' reviews collected as the sample for this research study.

Text Pre-Processing

Users' reviews are text that contain a lot of noise, such as emoticons and punctuations, and these cause computers a lot of difficulties in detecting the hidden patterns that can help to identify the emotion of the reviews. Therefore, removing noises from the text will be the most important steps, as these kinds of noises will affect the performance of the model development. First, emoticons, punctuations, digits (e.g., 1, 2, 3, etc.), and stop words had to be removed, followed by tokenization and lemmatization. This experiment focused on English reviews only, so all non-English text found in the reviews was removed. Additionally, misspelled words were corrected, as humans can sometimes create a lot of typo errors and abbreviations.

Figure 1. Sentic-emotion classifier layered architecture



Expand the Emotion Lexicon

The NRC Emotion Lexicon (NRC EmoLex) (Mohammad & Turney, 2013) is used to assign emotions to those emotion words from the reviews collected. It consisted of a list of words and their associated eight emotions: anger, disgust, fear, joy, sadness, surprise, trust, and anticipation. In this experiment, trust and anticipation emotions were removed as Ekman's six basic emotions are anger, disgust, fear, joy, sadness, and surprise. This was applied to the emotion annotation of reviews. The sentiment polarity (positive and negative) and score were added into the NRC EmoLex for each word. However, the words in emotion lexicon were limited, and this will lead to a wrong assignment of sentiment polarity and score for users' reviews. This is because some emotion words that could contribute significant emotions to the reviews were inadvertently left out. Hence, each original word in EmoLex was used to find the synsets from WordNet (Miller, 1998) in order to expand the lexicon. The richer and more complete the emotion words in the lexicon, the higher chances each emotion word can be recognised from the training and testing dataset.

Feature Generation

Ekman's six emotion types and their scores, sentence emotion score, negative score, positive score, and sentence polarity score are calculated by using the NRC EmoLex. The score of each emotion is calculated by totalling the intensity score of emotion words, which can be found in the NRC EmoLex. For example, if a review contains three surprise words, then the surprise score will be the total of the intensity score of these three surprise words. If two of the words are positive, then the positive score will be the total of the sentiment score for these two words, and negative scores are calculated the same way. The formula for calculating the sentence emotion score and the sentence polarity score can be referred to as Implementation of Sentic-Emotion Classifier and Analysis of Result. The positive or negative value of the polarity score will determine the polarity of the sentence in order to classify it as positive, negative, or neutral. In addition, the top 1000 unigram features were selected based on the Term Frequency Inverse Document Frequency (TF-IDF) score for model training.

Manual Annotation

This experiment built four different supervised machine-learning models to predict sentiments and emotions from the test dataset. Supervised machine learning requires a labelled dataset for training purposes, so manual emotion annotation was completed before training. Reviews with the correct sign of emotion score and polarity score were separated into different files according to the polarity and perform manual annotation. To be more specific, if the review had a positive emotion score and a positive polarity score, it was assigned with a positive polarity. This review will be saved in a CSV file, which only contains positive reviews. In contrast, if the emotion score and polarity score had different signs, such as a negative emotion score and a positive sentiment score, this review was not saved into any files, as the polarity of the review was considered to be confusing. Overall, 2286 reviews were annotated with Ekman's six basic emotions. For example, "Very poor customer service!" would be labelled as an "anger" emotion.

Feature Pre-Processing

Features were selected after pre-processing steps were completed before they were used for model training. For example, numerical features were standardized to prevent biased result. Categorical features were converted to numerical features, as most of the algorithms worked better with numerical features.

Model Building

Random Forest, Naïve Bayes, Decision Tree, and Support Vector Machine (SVM) algorithms were the supervised machine-learning techniques implemented for emotion analysis in this research. Each of these models is discussed to provide an overview of the model architecture and risks in selecting them.

Decision Tree is a model with a tree-like structure that can be used for both classification problems and regression problems in predicting classes. The main components of Decision Tree are root node, branches, internal nodes, and leaf nodes. Root node is the starting node of Decision Tree and does not have incoming branches. The outgoing branches from the root node are then fed into few internal nodes. Each branch represents the outcomes of the root node and internal nodes. Each path passes through from root node to internal nodes, and reach leaf node as a decision rule. The path can be interpreted using "if-then" rules (Song & Ying, 2015). For example, "if condition 1 and condition 2… condition n, then outcome k occurs." Leaf nodes are called end nodes, which represent the result of prediction. Internal nodes are called decision nodes, which are the nodes in between the root node and leaf node. Decision Tree is a top-down approach, which branches out from the root node and continues to the sub-nodes until the leaf node is reached. All the nodes are selected based on evaluation metrics, such as Information Gain, Gini impurity, and Gain Ratio (Patel & Upadhyay, 2012). For example, the highest gain ratio element is selected as the root node, and then it continues calculating for the sub-node for splitting until the prediction is made. It is possible that the accuracy of Decision Tree drops and overfitting could happen when the trees grow deeper. A deeper tree will result in long computational time. Therefore, pruning is important in Decision Tree to maximize the accuracy and optimize the computational efficiency. Pruning is a process to reduce the depth of the tree to its optimal size. To avoid overfitting or underfitting, parameters such as the depth of the tree and minimal samples in the leaf nodes could be tuned to obtain a better result in the prediction.

Random Forest is a machine-learning model that evolved from Decision Tree. It is an ensemble learning method that is made up of a sequence of weak correlated decision trees and aggregating the outcomes of the trees to identify the most popular prediction. In other words, the classification result of Random Forest is determined by choosing the most voted class from a multitude of decision trees. The decision trees are trained using a bootstrapping method, which creates each tree by selecting different subsets of features from available features. Random Forest will just randomly select a subset of features to grow the tree at each node, which helps to reduce overfitting, overall variance, and results in a more accurate prediction. Prediction of new data varies depending on the type of problems (Liaw & Wiener, 2002). For regression problems, the prediction is made by averaging the individual decision trees. For classification problems, Random Forest chooses the majority votes from a collection of decision tress as a result. Since the prediction of Random Forest obtains the highest number of votes from multiple decision trees, the prediction performance would be better than Decision Tree. However, the use of a collection of decision trees will slow down the processing speed and requires more memory consumption. Multiple trees generated from resampling the same dataset increase the difficulty in understanding the rules used to generate the classification results. Random Forest can handle big data efficiently with thousands of input variables as it is relatively robust to outliers and noise, estimates what variables are important to the classification, etc. (Rodriguez-Galiano et al., 2012).

Naïve Bayes is a probabilistic machine-learning model based on the Bayes Theorem by assuming that features are independent given class (Rish, 2001). In other words, a Naïve Bayes assumes that the presence of a particular feature in a class is independent with the presence of any other feature. For example, a review emotion is classified as happy if the review polarity score is a positive value and the happiness score is greater than zero. Despite any potential correlations between the variables of polarity score and happiness score, Naïve Bayes classifier considers each of these features to contribute independently to the likelihood that the emotion of review is happy. The statement of

Bayes theorem is as follows: Let $X = \left(X_1, X_2, \ldots, X_n\right)$, $P\left(X|C\right) = \prod_{i=1}^{n} P\left(X_i|C\right)$ where X is the event and C is the class. This Naïve Bayes equation is used to calculate probability of occurrence of an event for each class, and the class with the highest probability would be the prediction outcome. According to the study by Stern et al. (1999), Naïve Bayes is easy to use, as fewer parameters need to be set. It requires short computational time for training and makes prediction fast. Besides, Naïve Bayes performs well in multi-class predictions compared to other algorithms, so it is widely used in text classification, sentiment analysis, and spam filtering (Dey et al., 2016; Karimovich & Salimbayevich, 2020). One of the problems of Naïve Bayes is the probability will return zero if the event has a class that never appeared in the training dataset and will lead to inaccurate predictions being made (Boyko & Boksho, 2020). This is called zero probability issues. In simple terms, it is impossible to ensure the training dataset includes every word as well as their sentiment and emotion class. This problem can be overcome through a smoothing process, such as Laplace smoothing, by adding a count when calculating the probability. However, the assumption made in this model is not realistic in a real world application as it is almost impossible to get a set of features that are totally independent of one another. It is believed that the features used in this study are related to some other features.

Support Vector Machine (SVM) is a supervised machine-learning model that learns by example to assign labels to objects for classification and regression problems. It can solve linear and non-linear problems. SVM maps data into a high-dimensional feature space and a line or separating hyperplane is created to classify the data into their respective classes. SVM will try to find the best line that can maximize the separation of data to the potential class they belong to in an n-dimensional space. It could be infinite lines to separate data and the accuracy of SVM depends on how the hyperplane is selected, which can make the model become more generalized. The optimal hyperplane is determined by selecting the hyperplane with maximum margins (Boser et al., 1992). Margin is the distance between the hyperplane and support vectors that are the closest points from two classes to the hyperplane. The purpose of selecting a hyperplane with the maximum margins is to maximize correction prediction on unseen data. In real word problems, it has the possibility that data are not linearly separable, so drawing a straight line could not help to classify data. To convert the data to linearly separable, a kernel function allows for the adding of additional dimension to the data. Simply put, the kernel function enables SVM to find the optimal hyperplane in a higher dimensional space (e.g. 3-dimensional space) where originally it is in a 2-dimensional space. Examples of kernel function include linear, sigmoid, polynomial, and radial basis function. The selection of a kernel greatly affects the classification result and needs to keep on trial and error to find the optimal kernel, which is a time-consuming task (Noble, 2006). SVM is primarily designed for performing binary classification, but it is also capable of solving multiclass problems (Mayoraz & Alpaydin, 1999). The multiclass approach used by the SVM model in this study is one-to-one approach. It will break down the multiclass problems into multiple binary problems. This study employs the C-Support Vector Classification model in Python's scikit-learn package, where a linear kernel is chosen and the rest of the parameters remain in default to align with the work of Balakrishnan et al. (2020). The risk of selecting this model includes the selection of parameters such as kernel function, which affects the model accuracy, and the probability estimates are not provided directly because it is calculated using expensive five-fold cross-validation (Support Vector Machines, n.d.).

The Synthetic Minority Oversampling Technique (SMOTE) described in the paper of Chawla et al. (2002) was used to re-sample the imbalanced data to prevent the models from producing biased outcomes. In this experiment, reviews with anger and happiness emotions were found to have more observations than reviews with fear emotions. For example, there were only 18 reviews labelled with fear emotions out of 2286 reviews. There were two approaches that solved the sample imbalance problem: under-sampling and oversampling. In this experiment, oversampling was used to re-sample the data by increasing the minority class proportion to be the same as the

majority class proportion. Next, users can determine the sample size of the training and testing sets. Random Forest, Naïve Bayes, Decision Tree, and Support Vector Machine (SVM) were the supervised machine-learning techniques implemented for emotion analysis. Accuracy and F1-scores of each model were compared to evaluate the models' performance. Other than that, stratified k-fold validation (Brownlee, 2020), one of the cross-validation techniques, was used to ensure that the training set and testing set had the same proportion of labels in each fold. In other words, it solved multi-class problems, as all labels were randomly selected as samples in every fold. The dataset was then split into k parts of equal size, whereby k is the number of the group that the data was to be split to, and one of the parts was randomly selected to be the testing set. If the k parameter was set to 10, it meant the model would be trained for 10 times with different training and testing sets. The final performance was calculated by averaging the evaluation metrics, which included accuracy, F1-score, and Area Under the ROC Curve (AUC) score of all iterations. Stratified k-fold validation is better than just one split because predictions on all the data could be made, as it was not just one part of the data being tested; hence, each sample would not miss each type of label. In the experiment, the number of k was set to 10. At the end, each model was evaluated by accuracy and F1-score to compare performance. The model with higher accuracy in stratified 10-fold validation was chosen as the best model in this experiment.

CONCEPTUAL AND FUNCTIONAL DESIGN FOR SENTIC-EMOTION CLASSIFIER

This section discusses the conceptual system and detailed functional class diagrams for the Sentic-Emotion Classifier in Figure 2 and 3, respectively. The architecture of the Sentic-Emotion Classifier consists of several novel designs, and they have contributed new knowledge in the sentiment and emotion computational computing space.

Figure 2 provides the high-level system design for the Sentic-Emotion Classifier. The overall conceptual design proposes EmoLex Token Matric (ETM), 2-tier Sentiment Classification (2-tier SC), Extended EmoLex Framework (EEF), and Multi-layer Discrete Emotion Hierarchical Classes (MDEHC) as its key engines.

The 2-tier Sentiment Classification (2-tier SC) takes lexicon resources through the fusion of NRC EmoLex and SenticNet 4 to improve the sentiment classification and produces a very reliable set of polarity scores with a much improved recall rate. The 2-tier SC injects these scores as part of the features for the model training.

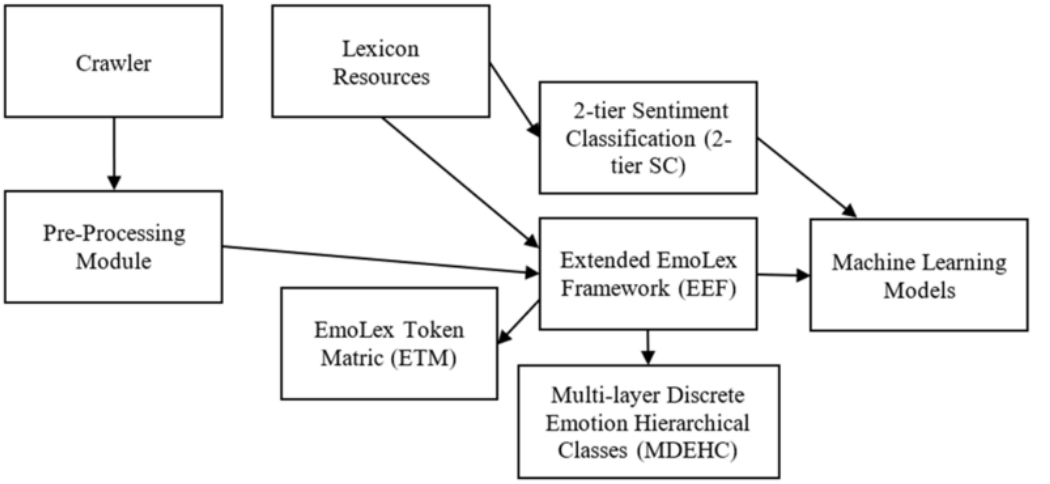
The Extended EmoLex Framework (EEF) takes input from post-processed data and lexicon resources in conjunction with synsets extracted from WordNet (Fellbaum, 2010; Miller, 1995) to enlarge the EmoLex Token Matric (ETM) so that the expanded tokens can generate more comprehensive emotions and sentiment-related data for model training purposes. EEF works closely with the novel Multi-Layer Discrete Emotion Hierarchical Classes (MDEHC) to capture and improve the prediction of reviews' sentiments and emotions more accurately. MDEHC is a multi-hierarchical emotion classifications framework designed to classify emotions and scores at different layers based on Ekman's six basic emotions.

Figure 3 illustrates the classifier design to explain the architecture using a class diagram. The class diagram illustrates core classes of the engine that hypothesizes its ability to predict sentiments and emotions of the reviews with a much higher accuracy.

The class diagram for Sentic-Emotion Classifier illustrates classes' functional characteristics and the relationships between a main class and sub classes. Main_Class is the main program of this experiment, and other classes' functions were called from the main program. Main_Class class had a one-to-one relationship with Text_Preprocessing, Expand_EmoLex, process_features, Validation, one-to-many relationship with createFeature, and one-to-many relationship with Training.

The main purpose of Text_Preprocessing class is removing noises such as stop words, emoticons, and digits from reviews collected. Expand_EmoLex is functioned to expand the number of emotion

Figure 2. Conceptual system design for sentic-emotion classifier



words in NRC Emotion Lexicon using the synsets from WordNet. Since the emotion words coverage of NRC Emotion Lexicon is limited, this class is used to enlarge NRC Emotion Lexicon. Synsets of the emotion lexicon seed words are added with respective sentiment scores retrieved from SenticNet4 and Vader. Create_feature class and will generate features such as Ekman's six emotion scores, sentence emotion score, negative score, positive score, sentence polarity score, and TF-IDF for training purpose. After calculating the different scores needed, manual annotation is performed by an annotator to label the emotion for each review. Once annotation is completed, Validation class will validate the emotion and sentiment polarity to ensure they tally. For example, polarity must be negative for reviews with emotions such as anger, fear, disgust, sadness, and surprise. Before training machine-learning models, process_features class is called to perform data transformation, such as standardizing the numerical features and converting categorical features to numerical values. In training class, four machine-learning models, such as Random Forest, Decision Tree, Naïve Bayes, and SVM, are trained with labelled data to predict the emotions of the reviews. The performance of each model is evaluated in order to select the best predictive model.

IMPLEMENTATION AND ANALYSIS OF RESULTS

The performance of the Sentic-Emotion Classifier is benchmarked against manually annotated reviews with the predicted outcomes of the implemented models. This section discusses the manual annotation and the implemented classifier.

Manual Annotation

Reviews that have been pre-processed contain 2286 observations, and based on manual annotation, the annotated reviews are compared with the machine predicted outcomes to examine the number of correct sentiments and emotions achieved. For example, "EWallet is a stupid app, their customer support is not well trained as well." was predicted to have anger emotion with a positive sentiment. It was not correct, because emotions such as happiness and surprise were expected to have a positive sentiment, where anger is a negative sentiment. From the annotation outcomes, 981,865 and 440 are positive, negative, and neutral reviews, respectively. Each of the reviews is also manually annotate with the correct emotion.

Figure 3. Class diagram of sentic-emotion classifier

Implementation of Sentic-Emotion Classifier and Analysis of Result

The design of the Sentic-Emotion Classifier has been implemented and results are discussed in this section. The research uses four supervised machine-learning models. They are Random Forest (RF), Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM). The models developed have been trained and tested with 1072 observations after pre-processing the original population of 2480 reviews. Manual emotion annotation was completed for each review by categorizing each into one of these emotion categories such as anger, disgust, fear, happiness, sadness, surprise, and neutral (Mohammad & Turney, 2013). One of the objectives of this experimentation was to study the performance of machine-learning models developed on their -ability to predict emotion in the reviews.

In the pre-processing stage, duplicated reviews were dropped and the total dataset was reduced to 2465 from its population. Reviews were carefully examined and it was found that emoticons, punctuation, and digits were noises that contributed to the poor predictive capability of the models. Hence, they were removed from the reviews. In addition, non-English, such as Chinese and Malay words and expressions, were also removed from the reviews. At the same time, misspelled English words were corrected. Next, each review was tokenized and lemmatized into its root form. Stopwords, such as "a," "an" and "the," as well as digits, were removed using a Natural Language Toolkit (NLTK) package from the lists of tokens.

This research devises a novel sentiment and emotion classification technique by applying Plutchik's (Plutchik, 1991; Plutchik, 2003) 8-class wheel of emotions into Multi-layer Discrete Emotion Hierarchical Classes (MDEHC) in order to capture and improve the prediction of reviews' sentiments and emotions more accurately. The NRC EmoLex (Mohammad & Turney, 2013) was utilized as the sentiment and emotion lexicon in the process of affective classification. A total of 9921 words were labelled with their associated emotion and intensity score obtained from the NRC EmoLex. Eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) were used to classify the tokens harvested.

Some tokens may carry multiple emotions based on the EmoLex lexicon. For example, "treat" carries eight emotions while "disappointment" only carries disgust and sadness emotions. There were 1765 tokens that carried fear emotions: trust, anger, sadness, happiness, disgust, anticipation, and surprise have 1564, 1483, 1298, 1268, 1094, 864, and 585 tokens, respectively.

In this research, the number of emotion classes was reduced from eight to six by adopting Ekman's six basic emotions (Ekman, 1999). Due to the need to align to other lexicon resources in this research, "anticipation" and "trust" were not included. A total of 9921 tokens were reduced to 7493, which were labelled based on Ekman's Six Basic Theory of Emotion with their emotion intensity score.

For sentiment classification, this research employed VADER sentiment lexicon in the NTLK package (Hutto & Gilbert, 2014) to rate tokens either as positive, negative, or neutral. In order to obtain much higher sentiment accuracy, SenticNet 4 (Cambria et al., 2014; Cambria et al., 2016; Cambria et al., 2018) was used to strengthen the sentiment classification of each token. There were a total of 23,682 tokens being labelled with positive, negative, or neutral in SenticNet, and the range of polarity score fell between minus-one to one.

This research used a 2-tier Sentiment Classification (2-tier SC) approach, where it improved the sentiment classification and produced a much more reliable set of polarity scores. With the fusion of the NRC EmoLex and SenticNet 4 in the 2-tier sentiment classification framework, recall rate was improved. Tier 1 was validated by VADER, whereas Tier 2 was enhanced by SenticNet 4. The outcomes of the novel contributions as shown in Figure 4, show samples of the emotion, intensity score, sentiment, and polarity score.

In order to provide better coverage for tokens extracted from the reviews, this research designed and developed an Extended EmoLex Framework (EEF) where synsets for each token extracted from WordNet (Fellbaum, 2010; Miller, 1995; Miller, 1998) were used to enlarge the EmoLex Token Matric (ETM) where it held the source token, expanded tokens generated from synsets, as well as emotions and sentiment-related information. For example, the emotions of "hateful" in the

Figure 4. Emotion and sentiment classification with scores

	word	emotion	emotion-intensity-score	Sentiment	Sentiment_score
0	outraged	anger	0.964	Negative	-0.5423
1	brutality	anger	0.959	Negative	-0.6124
2	hatred	anger	0.953	Negative	-0.6369
3	hateful	anger	0.940	Negative	-0.4939
4	terrorize	anger	0.939	Negative	-0.6486
...	...	...	...	...	...
7488	spa	surprise	0.086	Neutral	0.0000
7489	leisure	surprise	0.086	Positive	0.5110
7490	picnic	surprise	0.078	Positive	0.0250
7491	tree	surprise	0.078	Positive	0.0740
7492	worm	surprise	0.055	Negative	-0.0200

7493 rows × 5 columns

NRC EmoLex are anger, disgust, sadness, and fear, while the synsets for "hateful" are ["terrorize," "terrify," "terrorise"]. Each of these synonyms were added into the Multi-layer Discrete Emotion Hierarchical Classes (MDEHC) and 2-tier Sentiment Classification (2-tier SC), which were based on the base lexicon (NRC EmoLex, SenticNet, and VADER) repository, if it was not yet included in the lexicon. It could also be tokens from synsets that exist in the NRC EmoLex, but different emotions with "hateful." As such, the token was added again into the NRC EmoLex with "hateful" emotions. The Extended EmoLex Framework demonstrated that a much comprehensive and rich sentiment and emotion lexicon resource was made available after the expansion process.

Features generation is another major step for the Sentic-Emotion Classifier. For each review, every token extracted was POS tagged prior to the sentiment and emotion computational routine taking place. The Sentic-Emotion Classifier takes emotion and sentiment-carried tokens, with their emotion intensity and sentiment polarity scores, and applies the following formula to obtain the review level sentiment and emotion scores:

Emotion score = total emotion score with positive polarity – total emotion score with negative polarity
Polarity score = total negative polarity score – total positive polarity score

This research assumed that if the emotion score was > 0, it was classified as a positive sentiment; negative if < 0, and neutral if the emotion score was equal to zero.

A review level emotion score was obtained by obtaining the difference between the sum of the emotion intensity score of all emotion words that have a positive sentiment score and the sum of the emotion intensity score of all emotion words with a negative sentiment score. This research assumed that if the review level emotion score was greater than zero, the review tended to be positive. This assumption was made because when the review level emotion score was a positive value, it was

deducted that more positive words were in the reviews, or the reviews had a strong positive emotion score, thus the review would naturally carry positive emotions and sentiments. So, the review was negative if it the scores were smaller than zero, and neutral if they were equal to zero. Deducting the sum of the polarity scores from the words that elicited positive sentiment by the sum of polarity scores from the words that elicited negative emotions yielded the review level polarity scores. If the final polarity score was smaller than zero, it could be deduced that it was a negative review, and positive if the score was greater than zero; otherwise, the review was neutral. The formulas were designed to reduce the effort of human annotation on the future opinion mining tasks. The effectiveness of the formulas was quite acceptable, as 92.7% of the reviews had the same sign of emotion score and polarity score, either a positive emotion score with positive polarity score, a negative emotion score with negative polarity score, or a neutral emotion score with neutral polarity score.

To obtain good predictive capability, the Sentic-Emotion Classifier scores offered the top N high-impact features based on TF-IDF; the N chosen in this research was 1000. The TF-IDF score stands for "Term Frequency - Inverse Document Frequency," where this technique quantifies a token in the corpus of reviews by computing a weight to each token, signifying the importance of the token in the corpus. The selected features were combined with eleven sentiment and emotion features: Anger_score, Disgust_score, Fear_score, Happiness_score, Sadness_score, Surprise_score, Emotion_score, positive_score, negative_score, polarity_score, and polarity for model training. The following describes the computational mechanics of TF-IDF:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

where token t for a review d in a corpus of D reviews.

$$tf\left(t,d\right) = \frac{f_{t,d}}{\sum_{t' \epsilon d} f_{t',d}}$$

and:

$$idf(t,D) = \log \frac{N}{\left|\left\{d \in D : t \in d\right\}\right|}$$

Prior to model training, the eleven sentiment and emotion features were scaled through a standardization technique to prevent bias during the model training process for Random Forest, Naïve Bayes, Decision Tree, and the SVM model training.

DISCUSSION AND EVALUATION

In this section, the research questions are discussed and answered based on the outcomes obtained.

Question One: Which machine-learning model can classify emotion with the highest accuracy rate?
Question Two: What is the performance improvement of the four different supervised machine-learning models when additional features, such as emotion scores, sentence emotion score, negative score, positive score, and sentence polarity score, are considered in addition to TF-IDF?
Question Three: Do machine-learning models perform better than the baseline agreement on emotion analysis, which is 60% to 79%?

From the 1,072 reviews predicted by the Sentic-Emotion Classifier, the majority were classified as negative with N = 632 or 59%, whereas positive and neutral were N = 322 or 30%, and N = 118 or 11%, respectively. Results showed that the frequency for each emotion type was as follows: neutral, anger, disgust, fear, sadness, surprise, and happiness emotions were N = 118 or 11%, N = 385 or 36%, N = 77 or 6%, N = 18 or 2%, N = 237 or 22%, N = 35 or 3%, and N = 319 or 30%, respectively. As stated previously, this experiment assumed happiness was considered a positive emotion, emotion types such as anger, disgust, fear, sadness were considered negative emotions, and surprise emotion might have either positive or negative sentiment. This is because customers were shocked to find that this EWallet service had no customer service phone number to seek for assistance, or they were surprised to receive a high amount of cash back. Figure 5 illustrates the frequency distribution for each emotion type. Obviously, most of the EWallet users were not satisfied with the services provided, as the total number of negative emotions was less than the positive emotions.

This research aimed to study the algorithm that could classify emotion categories with optimum performance by comparing the accuracy rates and F1 scores of the four machine learning algorithms: Random Forest, Naïve Bayes, Decision Tree, and the Support Vector Machine trained. Accuracy and F1-scores were evaluated for each algorithm and the average accuracy from stratified 10-fold validation with the supported metric was used to determine the best algorithm. Table 1 shows the confusion matrix of the four machine-learning algorithms with accuracy and F1-scores.

There are four basic terms of confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positive and true negative indicates the outcome of a model correctly predicting positive and negative classes. In contrast, false positive and false negative are the results of a model incorrectly predicting positive and negative classes. According to the confusion matrices in Table 1, the horizontal axis represents the number of reviews of each emotion class being predicted, while the vertical axis represents the total number of reviews being predicted correctly for each emotion class. The diagonal values represent the number of reviews being correctly predicted for the respective emotion classes, while off-diagonal values are those misclassified by the four models. The colour of values is darker if the value is higher. To further illustrate the diagonal

Figure 5. Frequency of all the emotion types from EWallet reviews

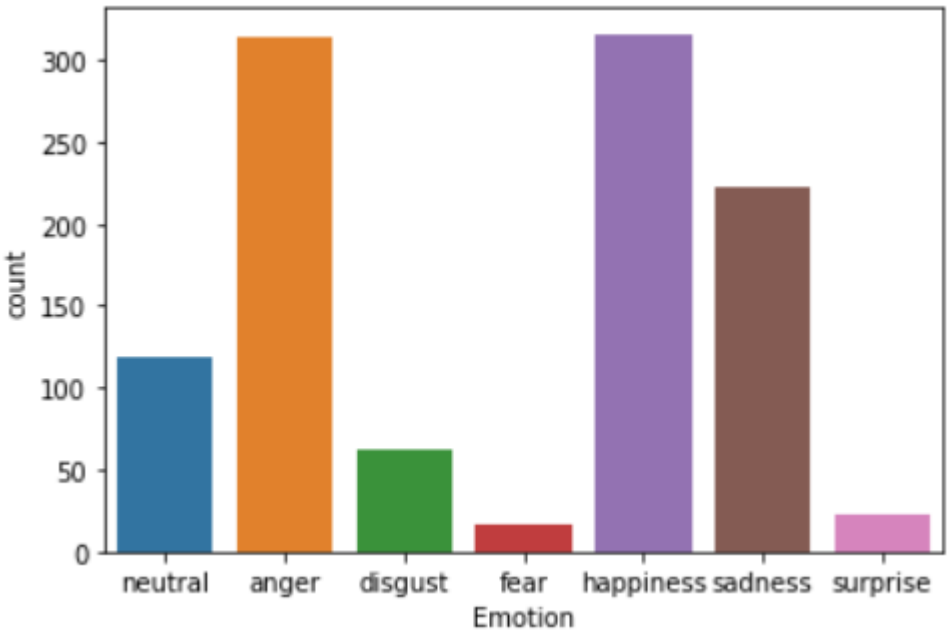
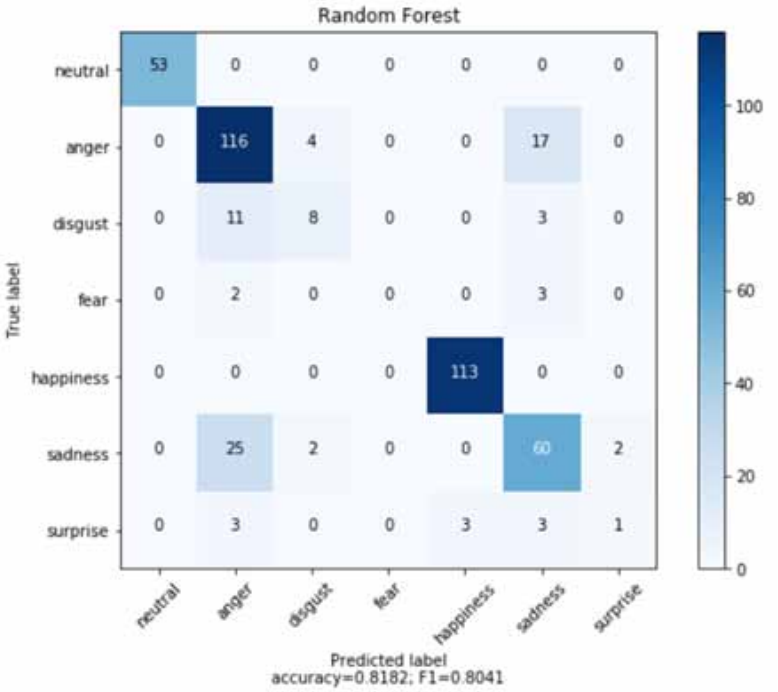
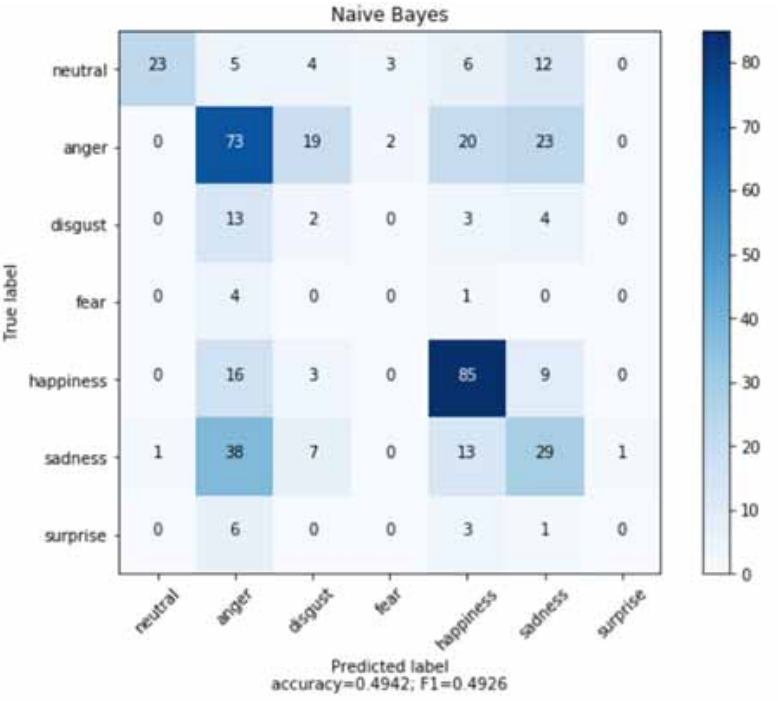
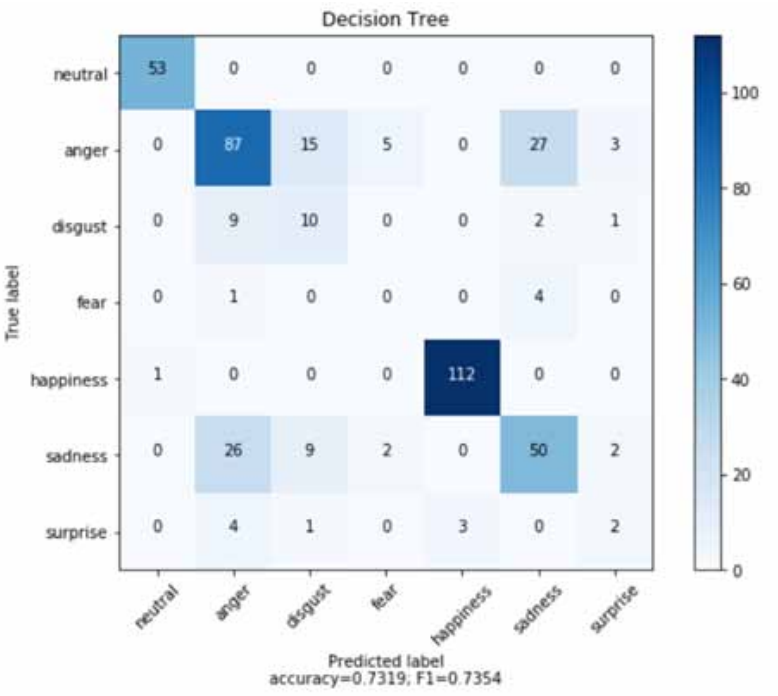
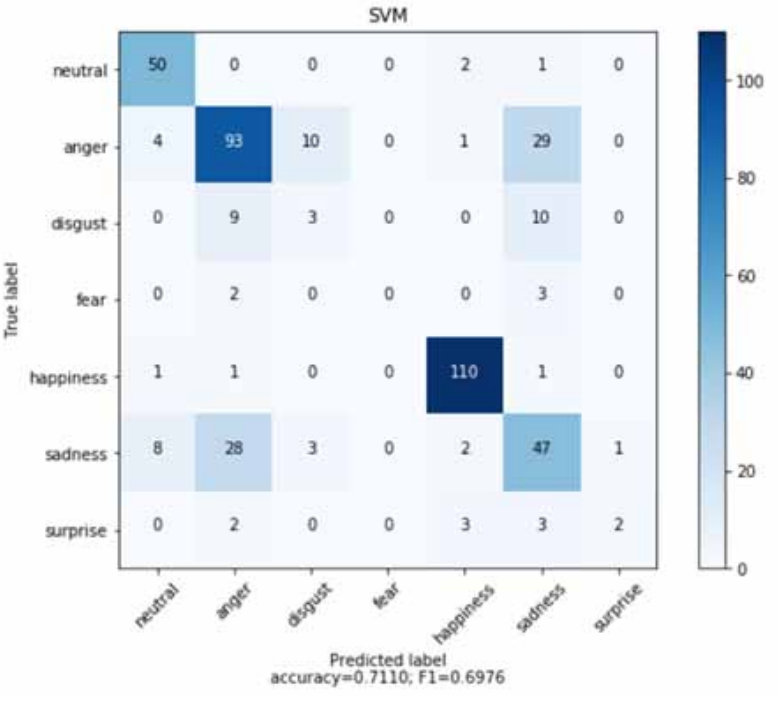


Table 1. Confusion matrix of four machine learning algorithms

Model	Confusion Matrix
Random Forest	
Naïve Bayes	

continued on following page

Table 1. Continued

Model	Confusion Matrix
Decision Tree	
SVM	

elements, the number of diagonal values of an emotion class is directly proportional to the correct prediction of that emotion. For example, the correct prediction of anger reviews in Random Forest confusion matrix is 116. By looking horizontally, 4 anger reviews are misclassified as disgust and 17 as sadness. While looking at the graph vertically, 11 disgust reviews, 2 fear reviews, and 25 reviews are wrongly predicted as anger. In the overall view, all models are good at predicting anger and happiness reviews due to sufficiently labelled data being provided. Reviews with anger and happiness emotions are found to be the most frequent emotions towards the EWallet app service, as shown in Figure 5. Since there are fewer fear reviews (2%) in the dataset, the models could not learn well due to less training data for the fear emotion. Poor prediction results in all models are shown in Table 1. None of the fear reviews are predicted correctly, because the diagonal values for the fear emotions are zero in all of the models.

Table 2 compares the models' average performance between the Sentic-Emotion Classifier and Balakrishnan et al. (2020) after 10-fold cross validation in terms of accuracy and F1-scores. Each model in the Sentic-Emotion Classifier was also evaluated by AUC score. From Table 2, the accuracy and F1-scores for Random Forest was the highest compared to the other three algorithms. Naïve Bayes had the lowest accuracy, F1-score, and cross validation score. Random Forest randomly creates decision trees to carry out emotion classification where the outcomes are voted by various decision trees by the algorithm. Since Random Forest is an ensemble method that predicts by taking the average or the mean of the output from various trees, increasing the number of decisions trees increases the precision of the outcome. This always reduces the overfit problem. As for Naïve Bayes, the reason it had the worst accuracy in this experiment could be that zero probability occurred when calculating the probability of the testing data as being any of the emotion classes. The testing data contained attributes that were not observed in the training data, so zero probability occurred and led to the misclassification of emotions. However, the results are contrary to Balakrishnan et al.'s (2020) study, wherein they reported a higher accuracy for Naïve Bayes compared to Decision Tree. This is because the Laplace smoothing technique does not apply for this experiment, unlike their study. A smoothing technique is an approach to overcome zero probability issues. Therefore, this research also proved that a smoothing technique was required to apply the Naïve Bayes algorithm in order to prevent zero probability issues that occurred. On the other hand, the findings in this experiment were in line with their study, as this experiment's result shows that SVM outperforms Naïve Bayes. This experiment was a success, as the accuracy performance of three models, such as Random Forest, Decision Tree, and SVM were greatly increased compared to their studies with an increment of 20.7%, 25.3%, and 10.1%, due to additional features being used for model training. Other than using the same features, such as TF-IDF and sentiment from the reference study, new features, such as emotion scores, sentence emotion score, negative score, positive score, and sentence polarity score, were also used to train models in order to evaluate the performance of each model with a new set of features.

Using accuracy as an evaluation metric to rank how often a model could make correct predictions, Random Forest (83%) was the best model, followed by Decision Tree (79.2%), SVM (70.8%), and

Table 2. Summarizes the performance of ten-fold cross validation for four algorithms based on their accuracies, F1 scores, and AUC scores

Models	Sentic-Emotion Classifier			Balakrishnan, Selvanayagam and Lok (2020)	
	Accuracy (%)	F1-score (%)	AUC	Accuracy (%)	F1-score (%)
Random Forest	83.0	79.6	0.91	62.3	58.9
Naïve Bayes	48.5	49.9	0.61	54.4	53.1
Decision Tree	79.2	78.6	0.77	53.9	53.7
SVM	70.8	69.3	0.87	60.7	55.5

Naïve Bayes (48.5%), but the dataset is imbalanced, as shown in Figure 5. Reviews with anger emotions and happiness emotions have accounted for 36% and 30%, but reviews with emotion types such as fear and surprise were less than 5%. Models would be biased to the anger and happiness class and result in better accuracy because there were fewer samples for other emotion types. Therefore, the AUC score was evaluated to measure how well a model was capable of distinguishing between emotion classes. The range value of the AUC fell between zero to one. The higher the AUC, the better the model was at classifying emotions of the reviews. According to the interpretation of the AUC score applied in Mandrekar (2010), 0.7 to 0.8 was considered acceptable, 0.8 to 0.9 was considered excellent, and it was considered outstanding if greater than 0.9. From Table 2, Random Forest still ranked first with an AUC of 0.91, and the score showed no overfitting on the dataset. Naïve Bayes was the worst model with an AUC of 0.61. Even though Decision Tree (79.2%) had better accuracy compared to SVM (70.8%); the AUC score of Decision Tree (0.77) was lower than SVM (0.87). This showed that SVM was more convinced in its prediction compared to Decision Tree, and it is believed that SVM is more likely to have better accuracy than Decision Tree when predicting future samples. In conclusion, there is no doubt that Random Forest is the best predictive model for emotion analysis and reflected in the reference study of this experiment with accuracy of 62.3%.

From the human experts' analysis, the emotion analysis baseline had been agreed to be 60% to 79%. The average of 69.5% falls in the range of 60–79%, which was the baseline for emotion analysis in this study for comparative analysis. Based on the results shown in Table 2, none of the models from the Balakrishnan et al.'s (2020) study was able to reach 69.5%. Additionally, Random Forest (62.3%) and SVM (60.7%) barely passed the acceptance rate of the human baseline of 60%. However, it was a surprise that this research showed the accuracy of Random Forest (83%), Decision Tree (79.2%), and SVM (70.8%) as being above average, except for Naïve Bayes, due to zero probability issues. Performance of Random Forest and Decision Tree even exceeded the maximum percentage (79%) in the range of the baseline set. Therefore, features, such as the six emotion scores, review level emotion score, negative score, positive score, and review level polarity score are important to contribute high prediction accuracy for emotion analysis when a machine-learning model is applied compared to only using TF-IDF.

Another finding was that some reviews were assigned positive by system, but the annotator labelled them as negative emotions. Such contradictions will affect the quality of the training result; therefore these kinds of reviews were filtered and not considered as a part of the training data. It could be said that manual emotion annotation is a second layer filtering to validate the alignment between sentiment and emotion, and produces a high quality of training data. Table 3 shows the total number of reviews with the right emotions and sentiments after manual annotation. As previously mentioned, the sentiment is assigned to each review by calculating the polarity score using the formula introduced in this research. Results show 981, 865, 440 reviews are potential positive, negative, and neutral, respectively, where the sentiment was assigned according to the result of the review level polarity score formula used. These emotions obtained from the reviews were then annotated. However, Table 3 shows that only 32.82% (322) reviews were classified correctly out of 981 potential positive reviews, 748 (86.47%) out of 865 reviews were classified correctly as negative reviews, and only 118 (26.8%) reviews were correctly predicted as neutral reviews. The reasons for the high wrong sentiment classification rate is shown in Table 7 and discussed in the last section.

In summary, Random Forest was the best model, which had the highest accuracy to classify the emotions obtained from the reviews among the four models reviewed. To answer the second question, the performance improvement of Random Forest, Decision Tree, and SVM were 20.7%, 25.3% and 10.1%, except for Naïve Bayes. Since the human baseline had agreed that emotion analysis was 60% to 79%, the average value of 69.5% was used as the baseline of this study. Results show Random Forest (83%), Decision Tree (79.2%), and SVM (70.8%) met the baseline requirement, and all three were an above average set; thus this research question has been answered.

Table 3. Total number of reviews with the right emotion and sentiment

Emotion	Sentiment	Total
Happiness	Positive	319
Surprise	Positive	3
Anger	Negative	385
Disgust	Negative	77
Fear	Negative	18
Sadness	Negative	237
Surprise	Negative	31
Neutral	Neutral	118

CONTRIBUTIONS, FUTURE RESEARCH, AND CONCLUSION

This research concludes that Random Forest performs better in terms of accuracy and F1-score when predicting emotion. The performance of SVM is comparable, with outcomes shown by Balakrishnan et al. (2020), while Random Forest and Decision Tree performed much better than the original authors in the terms of accuracy and F1-scores using the same input source. This is because the training features used in their study were TF-IDF only and sentiment, but this research included six emotion scores, emotion_score, positive_score, negative_score, polarity_score, and polarity as the features for model training. These 11 features contained relevant information that helped to improve the emotion prediction. The models discussed in this paper had been trained to learn more patterns from reviews labelled on sentiments and their scores. Therefore, the accuracy and F1-score of the algorithms reviewed had greatly been improved in this research. In this study, emotion labeling, based on Ekman's six basic emotions, was applied for the reviews' annotation. However, this research found that some emotions were very uncommon, such as fear and surprise, in the reviews collected. As a result, true negative rate and true positive rate for these emotions in the testing data are very low. In other words, reviews with fear and surprise emotions are too few in the sample, so the algorithms found it difficult to correctly predict fear and surprise emotions from the test data. It was found that there were only 3, 9, and 10, out of 22 reviews predicted correctly for disgust, sadness, and anger, respectively. Hence, this will form part of future research work.

To improve the accuracy of models, hyper parameter tuning can be used for future experiments in order to tune using more appropriate parameters for each model to increase its performance. As expected, the larger a quality sample size is, the more precisely the models can be trained, as long as the models are not overfitted. It is necessary to collect more EWallet reviews from multiple sources for a longer period of time so that there are more samples with infrequent emotions. Reviews written in Bahasa Melayu and rojak words were not included in this research. Future work would consider EWallet reviews that are code-mixed with Bahasa Melayu and rojak words to better understand users' satisfaction.

This research also highlights a contradictory and debateable scenario. The Extended EmoLex Framework (EEF) module is designed to generate a more comprehensive emotion lexicon with sentiment through adding the WordNet synsets of NRC emotion words for lexicon expansion. However, this module is not working well and problems have been identified and discussed in the following paragraphs. Table 4 provides an example wherein errors were found because some words had negative emotion scores, but with positive sentiment. In Table 5, there were 414 out of 2466 rows of the reviews paired with the wrong polarity when the expanded version of EmoLex was being applied. This is due to the incomparable samples that were used during model training between NRC

EmoLex, VADER, and SenticNet 4. For example, "fierce" was an "anger" word, and the intensity score was 0.812 (Table 4); however, VADER produced neutral sentiment. When the sentiment yielded by VADER does not align with the anger emotion predicted by NRC EmoLex, which should carry a negative polarity, SenticNet 4 is the next lexicon to be examined in order to verify the consistency. The design of the improved model has been engineered in such a way that when the sentiment score returned from SenticNet4 is also a positive value, then this score is the final sentiment score of "fierce" as the sentiment. As a result, in some circumstances, the final outcome is inaccurate for "fierce."

Therefore, this study compared the total number of wrong polarities before and after expanding the NRC EmoLex lexicon. Table 5 clearly shows that the NRC EmoLex without expansion had better results, as only 180 reviews or 7.3% were assigned the wrong polarity. This experiment selected the non-expanded lexicon to generate features instead of the expanded version of the NRC EmoLex. This research found that the expanded version of the NRC EmoLex consisted of more words where the sentiment was not compatible with the emotion, which is a serious challenge that needs to be addressed in future work. For example, words with anger emotions should have a negative polarity while a positive polarity for words with happiness emotions. The sentiment become inaccurate because different sentiment lexicons produced different sentiment types and scores based on the different sets of training datasets used. Another problem found was that there is always only one sentiment for a word in the sentiment lexicon, but in actual application with different contexts, a word can have more than one emotion. Table 6 shows the emotion and sentiment of "honest" from the expanded version of the NRC EmoLex. It has six different emotions, but the sentiment score shows a positive value for all emotions. This creates serious contradiction that will cause wrong calculations when generating features such as sentiment and sentiment score. In general, "honest" is not a word that carries a negative emotion, such as anger, disgust, fear, and sadness. There are two possible conditions on how the lexicon is going to be expanded: (1) Synonym of the emotion seed word does not exist in NRC EmoLex. (2) The synonym of emotion seed word does exist, but consists of different emotions.

This result in the emotion categories of words in the NRC EmoLex are excessively broad, where most of the emotions are irrelevant to them. 'Honest' is one of the examples that have many irrelevant emotions when the expanding lexicon process is completed in the Extended EmoLex Framework (EEF).

This study examines the reasons for wrong polarity assignment of the 180 reviews and draws the following conclusions. Table 7 concludes four factors that impact the emotion and sentiment analysis. The first reason is multiword expressions are not being recognised properly. Some reviews consist of multiword expressions such as *getting on my nerves* (become extremely annoying to someone), which have unpredictable meaning from the individual words. The computation of emotion score and polarity score would not be accurate if the emotional multiword expressions are not treated as a single unit. Beside this, the NRC EmoLex is made up of single emotion words only. The second

Table 4. 'Fierce' emotion, intensity score, sentiment, and sentiment score

Word	Emotion	Emotion Intensity Score	Sentiment	Sentiment Score
fierce	anger	0.812	positive	0.0250

Table 5. Total number of reviews where emotions are paired with wrong sentiment polarity after using expanded emotion lexicon compared to lexicon without expansion

NRC EmoLex	Wrong Polarity
Without expansion	180
With expansion	414

Table 6. "Honest" emotion, intensity score, sentiment, and sentiment score

Word	Emotion	Emotion Intensity Score	Sentiment	Sentiment Score
honest	anger	0.087	positive	0.5106
honest	disgust	0.055	positive	0.5106
honest	fear	0.047	positive	0.5106
honest	happiness	0.303	positive	0.5106
honest	sadness	0.062	positive	0.5106
honest	surprise	0.0547	positive	0.5106

review show an example that multiword expressions are not being treated as a single unit. The word, *reward* is recognised instead of *reward system,* which contribute a small number of positive values to the emotion score and polarity score, because *reward* carries positive emotions and polarity. Insufficient word coverage in the emotion lexicon also results in inaccurate emotion score and polarity score computations. Common emotion words, such as *poor*, *lousy,* and *slow,* do not exist in the NRC EmoLex. Other than that, intensifier and negation handling have a significant impact on emotion and sentiment analysis, as they could change the intensity of emotion and polarity, and even invert the results. For example, the emotion and polarity of the fifth review in Table 7 should be *sadness* and *negative*. If *user friendly* is detected, but *not* is being ignored, the emotion and polarity will become *happiness* and *positive*. Lastly, the quality of a lexicon is also one of the factors. The inclusion of neutral words, such as *family* and *service,* in a lexicon affect the performance in emotion and sentiment detection. Additionally, some words have an extremely high polarity score, which is not supposed to be as high. For example, the polarity score for *service* and *money* are 0.84 and 0.57, respectively.

This study discovered four factors that have a significant impact on emotion analysis and sentiment analysis. The four factors include failure to recognise multiword expressions, insufficient word coverage in the emotion lexicon, application of intensifier and negation handling, and wrong emotion assignment to words. The factors described above negatively affect model performance due to the score computation for the 11 numerical features; for example, emotion_score, anger_score, and polarity_score are not accurate. The root cause can be attributed to the quality of a lexicon is barely satisfactory. It is essential that the identification of multiword expressions and frequently used emotion words from online reviews could help to construct high quality lexicons. Also, the ambiguity in human text poses a challenge in designing a model for emotion detection. Apart from selecting features using TF-IDF, future research would also conduct experiments using BERT (Bidirectional Encoder Representations from Transformers), which can address ambiguity issues effectively.

Briefly, this study extends Balakrishnan et al.'s (2020) work on emotion analysis and proposes a new framework called the Sentic-Emotion Classifier to improve model accuracy in emotion detection. The first contribution is eleven new features, which includes: Anger_score, Disgust_score, Fear_score, Happiness_score, Sadness_score, Surprise_score, emotion_score, positive_score, negative_score, polarity_score, and polarity, have yielded better model accuracy for Random Forest, Decision Tree, and SVM, with an increment of 20.7%, 25.3%, and 10.1% compared to the reference study. Results show Random Forest (83%), Decision Tree (79.2%), and SVM (70.8%) have exceeded 69.5%, which is the average of the human baseline agreement on emotion analysis with the range of 60%–79%, Random Forest and Decision Tree even outperformed the baseline agreement. The evaluation of the AUC metric showed Random Forest and SVM in the Sentic-Emotion Classifier with an AUC score of 0.91 and 0.87, which are excellent in predicting emotions.

Initially, this research study was intended to expand the NRC EmoLex by assigning sentiment scores retrieved from VADER and SenticNet 4 to each emotion word. In this stage, the researchers discovered that the sentiment was not compatible with the emotion for some words due to different

Table 7. Four major factors lead to wrong polarity assignment to reviews

Reasons	Reviews
Unable to recognise multiword expressions	For a starter it's good. Now it's *getting on my nerves*.
	Keep depreciating coins value and the *reward system.*
Insufficient word coverage in emotion lexicon	Very *poor* customer service! No reply at fb messenger and email.
	lousy support... *slow* response.
Intensifier and negation handling are not considered	*Not* as user friendly as before.
	It *sooo* shameful.. please improve this issue.
Wrong emotion and sentiment assignment to words	It's very useful for me and my *family* members to pay any bills and do shopping.
	Service for bill payment is unavailable since last night!

sentiment resources having different interpretations on sentiment for some words. In addition, each word from the sentiment resources that were applied for expanding the NRC EmoLex had only one sentiment, but the word could have multiple meanings in different contexts, which led to a word possibly having more than one sentiment. This study concludes that four factors need to be considered for emotion analysis: failure to recognise multiword expressions, insufficient word coverage in emotion lexicons, application of intensifier and negation handling, and wrong emotion assignment to words.

AUTHOR NOTE

The authors of this publication declare there is no conflict of interest.

Authors thank the Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology for financial support and resources to carry out this research.Correspondence concerning this article should be addressed to Corresponding Author's Name, Address of University, Malaysia. Email: Corresponding Author's Email.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING INFORMATION

Authors thank the Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology for financial support and resources to carry out this research.

REFERENCES

Aji, H. M., Berakon, I., & Md Husin, M. (2020). COVID-19 and eWallet usage intention: A multigroup analysis between Indonesia and Malaysia. *Cogent Business & Management*, *7*(1), 1804181. doi:10.1080/23311975.2020.1804181

Aman, S., & Szpakowicz, S. (2007, September). Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue* (pp. 196-205). Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-74628-7_27

Amiri, H., & Chua, T. (2012). Mining slang and urban opinion words and phrases from cQA services: An optimization approach. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 193-202). ACM. doi:10.1145/2124295.2124319

Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different feature selection for sentiment classification. *International Journal of Information Science and Intelligent System*, *1*(3), 137–150.

Aue, A., & Gamon, M. (2005, September). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (Vol. 1, no. 3.1, pp. 2-1). ACM.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources* (Vol. 10, pp. 2200-2204). ACM.

Baker-Eveleth, L., & Stone, R. W. (2015). Usability, expectation, confirmation, and continuance intentions to use electronic textbooks. *Behaviour & Information Technology*, *34*(10), 992–1004. doi:10.1080/0144929X.2015.1039061

Balakrishnan, V., Selvanayagam, P. K., & Yin, L. P. (2020). Sentiment and emotion analyses for Malaysian mobile digital payment applications. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis* (pp. 67-71). ACM. doi:10.1145/3388142.3388144

Binti Sabri, N. A., bin Hamdan, S., Nadarajan, N. T. M., & Shing, S. R. (2020). The usage of English internet slang among Malaysians in social media. *Selangor Humaniora Review, 4*(1), 15-29.

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2019). Analyzing political sentiment using Twitter data. In *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018,* Volume 2 (pp. 427-436). Singapore: Springer.

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2020). Sentiment analysis on online product reviews. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018* (pp. 559–569). Springer. doi:10.1007/978-981-13-7166-0_56

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152). ACM. doi:10.1145/130385.130401

Boyko, N., & Boksho, K. (2020, November). Application of the Naive Bayesian Classifier in work on sentimental analysis of medical data. In IDDM (pp. 230-239).

Brownlee, J. (2020, May). *A gentle introduction to k-fold cross-validation*. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

Cambria, E., Olsher, D., & Rajagopal, D. (2014, June). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (Vol. 28, No. 1). ACM. doi:10.1609/aaai.v28i1.8928

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016, December). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (pp. 2666-2677). https://aclanthology.org/C16-1251

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018, April). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (Vol. 32, no. 1, pp. 1795-1802). ACM. doi:10.1609/aaai.v32i1.11559

Canales, L., Daelemans, W., Boldrini, E., & Martínez-Barco, P. (2019). EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text. *IEEE Transactions on Affective Computing*, *14*(8), 1–14.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Chekima, K., & Alfred, R. (2018). Sentiment analysis of Malay social media text. In *Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November* (pp. 205–219). Springer. doi:10.1007/978-981-10-8276-4_20

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982.*

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707–714. doi:10.1016/j.procs.2019.11.174

Ekman, P. (1999). Basic emotions. Handbook of cognition and emotion, 98(45-60), 16.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT press. doi:10.7551/mitpress/7287.001.0001

Fellbaum, C. (2010). WordNet. In *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer Netherlands. doi:10.1007/978-90-481-8847-5_10

Hakak, N. M., Mohd, M., Kirmani, M., & Mohd, M. (2017, July). Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (COMPTELIX)* (pp. 397-402). IEEE. doi:10.1109/COMPTELIX.2017.8004002

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *35th annual meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 174-181). IEEE.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177). ACM.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, no. 1, pp. 216-225). ACM. doi:10.1609/icwsm.v8i1.14550

Kamble, S. S., & Itkikar, A. R. (2018). Study of supervised machine learning approaches for sentiment analysis. [IRJET]. *International Research Journal of Engineering and Technology*, *5*(4), 3045–3047.

Kanayama, H., & Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 355-363).Karim, M. W., Haque, A., Ulfy, M. A., Hossain, M. A., & Anis, M. Z. (2020). Factors influencing the use of E-wallet as a payment method among Malaysian young adults. *Journal of International Business and Management*, *3*(2), 1–12.

Karimovich, G. S., & Salimbayevich, O. I. (2020, November). Analysis of machine learning methods for filtering spam messages in email services. In *2020 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-4). IEEE. doi:10.1109/ICISCT50599.2020.9351442

Kham, N. N. (2019). Lexicon based emotion analysis on Twitter Data. *International Journal of Trend in Scientific Research and Development*, *3*(5), 1008–1012.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44*(4), 491–511. doi:10.1177/0165551517703514

Kundi, F. M., Ahmad, S., Khan, A., & Asghar, M. Z. (2014). Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. *Life Science Journal*, *11*(9), 66–72.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In 2015 Science and Information Conference (SAI) (pp. 288-291). IEEE. doi:10.1109/SAI.2015.7237157

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. doi:10.1097/JTO.0b013e3181ec173d PMID:20736804

Mayoraz, E., & Alpaydin, E. (1999, June). Support vector machines for multi-class classification. In *International Work-Conference on Artificial Neural Networks* (pp. 833–842). Springer.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. doi:10.1016/j.asej.2014.04.011

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Mohammad, S. M., & Turney, P. D. (2013). NRC emotion lexicon. National Research Council, Canada, 2, 234.

Muthukadan, B. (2014). *Selenium with Python*. goodreads. https://www.goodreads.com/en/book/show/25332181

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. doi:10.1038/nbt1206-1565 PMID:17160063

Patel, N., & Upadhyay, S. (2012). Study of various decision tree pruning methods with their empirical comparison in WEKA. *International Journal of Computer Applications*, *60*(12), 20–25. doi:10.5120/9744-4304

Plutchik, R. (1991). *The emotions*. University Press of America.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association.

Ray, A. (2017). *What are the different types of e-commerce payment systems?* Amazon. https://services.amazon.in/resources/seller-blog/different-types-of-e-commerce-paymentsystems.html

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. doi:10.1016/j.isprsjprs.2011.11.002

Schmidt, T., & Burghardt, M. (2018). An evaluation of lexicon-based sentiment analysis techniques for the plays of gotthold ephraim lessing. *In Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp.139-149). ACM.

Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (pp. 1-7). ACM. doi:10.1145/2401603.2401605

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Jingshen Yixue*, *27*(2), 130. PMID:26120265

Stern, M., Beck, J., & Woolf, B. P. (1999). *Naive Bayes classifiers for user modeling*. Center for Knowledge Communication, Computer Science Department, University of Massachusetts.

Support Vector Machines. (n.d.). *1.4. Support Vector Machines*. scikit learn 1.3.0. https://scikit-learn.org/stable/modules/svm.html

Tan, Y. F., Lam, H. S., Azlan, A., & Soo, W. K. (2016, April). Sentiment analysis for Telco Popularity on Twitter Big Data using a Novel Malaysian Dictionary. In *International Conference on the Applications of Digital Information and Web (ICADIWT),* (pp. 112-125). ACM.

Upadhayaya, A. (2012). Electronic Commerce and E-wallet. *International Journal of Recent Research and Review*, *1*(1), 37–41.

Vollmann, R., & Soon, T. W. (2018). Chinese identities in multilingual Malaysia. *Grazer Linguistische Studien*, *89*, 35–61.

Winarsih, N. A. S., & Supriyanto, C. (2016, August). Evaluation of classification methods for Indonesian text emotion detection. In *2016 International Seminar on Application of Technology for Information and Communication (ISemantic)* (pp. 130-133). IEEE.

Tong Ming Lim has 10 years of industry experiences in the design, development, implementation and maintenance of commercial software from 1989 to 1999 after returning from Mississippi State University USA. In 1999, Professor Lim left the software industry and he joined as an academics with Monash University, UTAR and Sunway University. Having many years of exposure in the software industry and academics world, Professor Lim understands very well the needs and requirements of accounting, distribution, manufacturing and point-of-sales software segments. In the past 20+ years, Professor Lim had always focused in object oriented software and databases technologies while he was with University of Malaya and Monash. He also had spent more than 10 years in peer to peer technologies related research as he was with UTAR and Sunway University. In the last 12 years, his work has been focusing on organizational knowledge sharing and technology acceptance, social media analytics and social influence maximization while he was with Sunway University and Tunku Abdul Rahman University College (TAR UC) (since 2008). Professor Lim has graduated more than 20 master and 2 PhD students while he was with Monash, UTAR and Sunway University. He is currently the Director for CBIEV at TAR UC, Professor at FOCS at TAR UC and Head for Big Data Analytics Centre supervising 5 PhD candidates and leading big data analytics projects with Webqlo, MoT, CMG, and DynaFront. Professor Lim's vast experience from the software industry and academics allows him to provide leadership in teaching and learning, research and industry projects development, and academics administration services in the University he attaches to.

Yuen Kei Khor is a data scientist who work on Natural Language Processing tasks. Her research and task focuses on analyzing Malaysia's text reviews such as identify code-mixed multiword expressions, sentiment and emotion analysis. She is pursuing a master's degree in Computer Science at Tunku Abdul Rahman University College.

Chi Wee Tan received BCompSc(Hons) and PhD degrees in year 2013 and 2019 respectively in Universiti Teknologi Malaysia. Currently, he is a Senior Lecturer cum Programme Leader at Tunku Abdul Rahman University College and AI and Computer Vision Consultant for Imagine AI Sdn Bhd. Also, he is the research group leader for Audio, Image and Video Analytics Group under Centre for Data Science and Analytics (CDSA). Dr Tan's main research areas are Computer Vision (CV) and Natural Language Processing (NLP) and Artificial Intelligence (AI). Being a meticulous and analytical researcher with many years of educational and hands-on experience, he was invited to Université d'Artois (France) under Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) programme for collaborative research between European countries with Southeast Asian countries on motion detection and computer vision. Being a certified Train-The-Trainer (TTT), Dr Tan also involved in multiple research and industrial projects such as AI based Document Content Extraction, AI based Social Distancing Monitoring System, Malaysian Rojak Lexicon using Machine Learning, Product Defects Detection and Object counting.