Big Data: Sentiment Analysis using Social Media Data

Nur Farhana Binti Ahmad
Faculty of Computer and Mathematical Sciences
UiTM Kampus Jasin
Melaka
nananurfarhana42@gmail.com

***Abstract*—An automated process that uses Artificial Intelligence (AI) to evaluate behaviours, feelings, sentiments, and opinions in natural written language is known as sentiment analysis. The emergence of the Internet of Things(IoT) and its widespread use in social media applications have opened new doors of chances to use data analytics to get valuable insights from unstructured data. As a result, the use of sentiment analysis in the big data era was useful in classifying opinion into various sentiment and assessing the mood of social media users. In this regard, this paper addressed an overview of sentiment analysis, its approaches, and its applications. Furthermore, this paper also highlighted the major challenges in sentiment analysis of big data.**

Keywords—Big data, sentiment analysis, social media.

I. INTRODUCTION

The development of digital sensors and modern communication tools has contributed to the advancement of the Internet of Things (IoT) that enables social networking sites and communication tools such as smartphones, laptops, and PCs communicating with each other to create vast quantities of big data. Social media posts are used for sentiment analysis, given a large number of participants, posts, and it has emerged as a source of free public data [1]. Almost 2.46 billion people are active in various social networks [2]. According to the 2011 report of the International Data Corporation, approximately 1 zettabyte (ZB) of data has been generated globally, and the rate of this volume is increasing. The amount of data produced is projected to be 44ZB by 2020, with not less than part of the text data generated by social media sites such as Facebook, Twitter, and Mobile Instant Messaging applications. Due to the growth of technologies, the decrease in both storage and computing power is one of the main circumstances that led to the development of big data that is characterized by the volume, velocity, veracity, volatility, and value of data. The analysis of big data can facilitate a rigorous and insightful understanding of business companies.

II. AN OVERVIEW OF SENTIMENT ANALYSIS

Sentiment analysis, commonly referred to as opinion mining, is a field of study within natural language processing (NLP) that extracts and classified opinions of text data. Aside from evaluating opinions, this approach also evaluates the user's emotions, such as polarity, in order to determine whether the opinion expressed is neutral, positive, or negative. Besides, the holder of the opinion shall be identified as the person or group giving the opinion.

The collected unstructured information will be automatically transformed into structured data. Some of the public sentiment regarding a product, service, or some other issue may be valuable to various commercial applications. The opinion can be classified into three main groups of opinion, which are comparative opinions, regular opinions, and suggestive opinions of a single entity or different entities. The regular opinion referred to a single entity and was mainly used to describe the positive or negative opinion for a particular item. On the other hand, comparative opinions compared and show the correlation among more than one entity and were primarily used for strategic intelligence as they explicate the relationship between multiple entities.

A. Sentiment Analysis Algorithm

The algorithms for the implementation of sentiment analysis systems are listed as follows:

- Rule-based systems that carry out a sentiment analysis on the basis of a collection of rules.
- Automated systems that rely on machine learning algorithms to learn from trained data.
- Hybrid systems which integrated both rule-based and automated solutions.

B. Sentiment Analysis scope

There are a variety of different levels of sentiment analysis that can be applied, such as:

- The document-level sentiment analysis interprets the sentiment of a whole text or paragraph.
- The sentence-level sentiment analysis evaluates the sentiment of one sentence.
- The sub-sentence level sentiment analysis that provides a subexpression sentiment between sentences.

C. Classes of sentiment analysis

The classes of sentiment analysis are as follows:

- Fine-grained sentiment analysis refers to the identification of sentiment on the sentence, sub-sentence, or aspect-based level. The level of polarity is more precise as the text is broken down into further categories, usually very positive to very negative.
- Emotion detection is classification strategies that involve a predefined set of emotion classes used to recognize the emotional state. Specific emotions are identified rather than positivity or negativity, including happiness, sorrow, frustration, and sadness.
- The intent-based analysis relies on grammar-parsing technology, which is a part of Natural Language Processing (NLP) that has the ability to understand conversation regardless of the content type. Besides, an intent analysis describes the

user's intention behind the text; it acknowledges the actions behind a text as well as the opinion

- The aspect-based analysis is when the text aspects, which are the attributes or components of a product or service, are broken down before allocating each one a sentiment level. It gathers the specific element that is positively or negatively mentioned. The rapid growth of opinions expressed on the web has made the mining aspect-level opinions a valuable source of online public opinion analysis.

The sentiment analysis tools can be integrated into any web application for the purpose of social media monitoring. Fig.1 depicts an example of a website dashboard that uses sentiment analysis tools

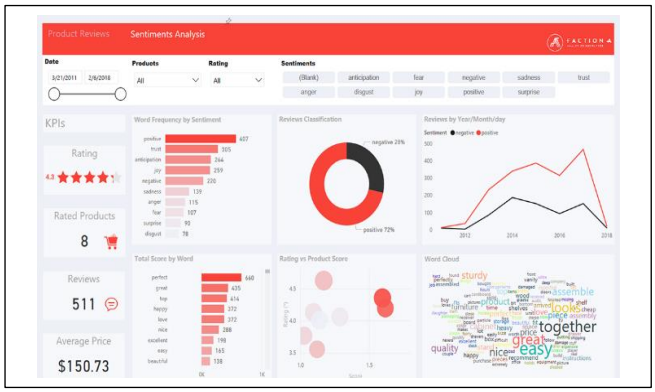

Fig.1. Example of sentiment analysis dashboard

III. Approaches for Sentiment Analysis

The general approaches of sentiment analysis will be reviewed in this section. First of all, the target set for employing the sentiment analysis is observed, and afterward, big data is collected from a single platform or multiple platforms of social media. The data collected should be pertinent to the function of the sentimental system. The next step is pre-processing the data to eliminate noise and irrelevant content, followed by the development and evaluation of the sentiment analysis model according to the approaches described below. After constructing the sentiment model and testing them with actual data sets, the model will be applied to the big data as an automatic classification. The most commonly used sentiment analysis approaches are described in the next subsections.

A. Machine Learning Approaches

Unsupervised learning and supervised learning is the classification technique for classifying text into classes used in machine learning [3].

- Unsupervised Learning

  Unsupervised learning has no right goals. Thus, it is dependent on cluster analysis. The model discovers patterns on its own and deals mainly with unlabeled data, including complex processing tasks.

- Supervised Learning

  Machine learning algorithms that are applicable to sentiment analysis are mainly a part of supervised learning, in which data sets that are required to train the classifier includes the training set and a test set. Once a supervised learning algorithm has been selected, the sentiment classification features will indicate how the documents are presented. The drawback of supervised learning is the need to provide a training instance that is sufficiently rigorous to make the algorithm efficient and highly credible in order to classify each of the data instances. The most frequently used sentiment classification features are the term and frequency of expressions, words of opinion, and part of speech information negations. The following subsections explained the steps involved in the development of the machine learning approach.

1) FEATURES EXTRACTION

In constructing an effective classifier, the extraction of the features must be conducted since the performance of the sentiment classification model depends greatly on the consistency of the features. In fact, if the extracted characteristics do not align with sentimental polarity and similar characteristics occur in positive or negative posts, the classification method will be less precise and reliable. During feature extraction, textual data (P1, P2, P3, P4,…Pn) are converted to word features (wf1,wf2,wf3,wf4,…wfn) by using various feature engineering approach. The most typically used features are Bag of Phrases (BoP), Bag of Words (BoW), Bag of Concepts (BoC), and n-gram. Lexical features are the second group of features that include the terms of opinions, negation words, and sentiment words. The third group of features differs according to the source of the data; for instance, social media data usually adds features such as the sum of hashtags and features related to social media such as social networking features such as acronyms and emojis [4].

2) MACHINE LEARNING ALGORITHMS

This section outlines the most commonly used machine learning algorithms for sentiment analysis.

a) ARTIFICIAL NEURAL NETWORK(ANN)

The artificial neural network (ANN) is an algorithm based on a mathematical modeling approach developed to replicate how the human brain processes information and analyzes it. A typical ANN layer is made up of input, output, and often a hidden layer in which the implicit feature information is extracted. The inputs are what teach the ANN layer to produce the desired output. It is an outstanding algorithm for finding patterns that are too complex because the hidden layers are capable of performing non-linear transformations of the inputs entered into the network. However, the main problem is that neural networks are "black boxes," where the user inputs data and receives responses. The responses can be fine-tuned, but the actual decision-making process is not accessible for the user. Many researchers are actively working on this issue, but it has become more challenging as artificial neural networks play an increasingly large role in our lives.

b) RANDOM FOREST

The random forest (RF) is a regression and classification technique that applies the ensemble of the rapid growth of decision trees. Ensemble learning, also known as bagging, is a process that can construct multiple classifiers and then produce cumulative results. Define RF is a classifier composed of structured tree classifiers {h(x, Fy), y=1,...} with (Fy) is independently identical distributed random vectors, and each tree casts a voting unit for the class at input x.

The algorithm constructs trees on the subset of data and later integrates multiple outputs of all trees. As a result, the accuracy of the classifier model will be improved, and the problems of overfitting and the ambiguity in decision trees will be minimized. Nevertheless, the random forest algorithm is far more complex than the decision tree as it demands more computing power and resources. The training process also takes time, as many trees are produced.

Some of the RF classifiers carry out classification by ensembles from random partitions (CERP) that are explicitly designed to handle high-dimensional data sets that can estimate a random portion of the entire collection of predictions. Multiple classifiers will then combine to boost the accuracy of the model as compared to the original classifier. On the other hand, the ensemble-based classifier is built with the support of logistic regression trees (LR-T CERP) and classification trees (C-T CERP). C-T CERP was less reliant on the threshold, unlike LR-T CERP, but the results of both these classifiers showed high accuracy.

c) SUPPORT VECTOR MACHINE (SVM)

The support vector machine (SVM) is a supervised machine learning technique that uses classification algorithms to solve classification problems. The algorithm produces the most accurate result in traditional text categorization by selecting the best possible difference between positive and negative training samples. There are various extensions available for the algorithm to enhance its accuracy and to be more adaptable to real-world problems. A soft margin classifier is one of the SVM extensions that classifies the majority of the data. It eliminates any outliers and noisy data as most data are linearly segregated and occasionally linearly observable for multi-dimensional problems.

The non-linear classifier is another extension of the SVM in which kernel is used to maximize the hyperplane margin. SVM and its extensions are used for the functions of binary classification. However, for multi-class problems, multi-class SVM extension is used with labels designed for objects that are drawn from a finite collection of multiple elements. By using a suitable kernel such as various kernel functions, SVM can handle linear separation at high dimensional non-linear input. Nevertheless, the SVM algorithm does not perform well when there are large data sets.

d) GENETIC ALGORITHM

Genetic algorithm (GA) is an algorithm to solve both constrained and unconstrained problems of optimization that have been created as an attempt to emulate the process that drives biological evolution concept, including selection, crossover, and mutation. It operates on string structures similar to a biological structure that evolves over time by using a randomized yet structured information exchange in accordance with the rule of survival of the fittest. This rule implies that the species that can adapt to environmental changes are able to survive and reproduce. The GA randomly selects individuals from the existing population to be the parents at each step and uses them to produce the next generation of children. The population is evolving towards an optimal solution over the successive generation. Thus, a new set of strings is created in each generation, using parts of the old set's fittest members. GA is commonly used to produce the most reliable solutions for optimization and search problems.

GA has many advantages in contrast to traditional techniques of optimization. Among these, the most prominent is the ability to handle complex problems and parallelism. GA can solve various types of optimization problems, whether the fitness function is linear or non-linear. However, GA also has some drawbacks. Choosing important parameters such as mutation rate and crossover and the new population selection criteria should be cautiously done. Any inappropriate choice will make it hard to converge the algorithm or produce pointless results.

e) NAÏVE BAYES (NB)

NAÍVE BAYES (NB) is a classification technique based on the Bayes' theorem that assumes any feature is unrelated to the occurrence of a particular feature in a class. It is a theorem that operates using the conditional probability that predicts the possibility that something will happen, given that something else has already happened using its prior knowledge. The classifier model is suitable for large amounts of data [5], where the probabilities P of two events c and x represented as $P(c)$, and $P(x)$ and the conditional probability of event c by event x are represented as $P(c|x)$. Thus, the Bayes' formula is as in (1).

$$P(c|x) = \frac{P(x|c) \,.\, P(c)}{P(x)} \quad (1)$$

The prediction of test data set class is easier and faster in NB than other algorithms. Besides, it also works well in the multi-class estimation, although fewer training data is needed. The weakness of the NB classifier is the supposition of independent predictors. In real life, having a set of wholly independent predictors is almost impossible

B. Lexicon-Based Approaches

The lexicon-based approaches for classifying sentiments usually are based on the idea that a piece of text's polarity can be acquired on the basis of the polarity of the words it composes. It uses a sentiment lexicon to classify the word polarity for textual content as positive, negative, or neutral. This approach is more comprehensible and can be easily implemented as compared to machine-based learning algorithms. Nevertheless, the major downside is that it needs human participation in the process of analyzing the text. The more prominent the volume of information, the more critical the test will be for identifying sentiment, sifting noise, and distinguishing useful data from different sources of content. Unfortunately, a simple approach is

likely to fail due to the complexity of natural languages, as most facets of the language, such as sarcasm and thwarted expressions, are not considered. The typical library for the lexicon-based approaches for processing textual data is TextBlob. It is a very convenient NLP library that comes prepackaged with its sentiment analysis functionality based on NLTK.

The following are two sub-categories for this method:

- Dictionary-based
  The dictionary-based approach employed the terms compiled and manually annotated. The synonyms and antonyms from the dictionary will be scanned after the terms of opinion from a review text are found. However, due to the absence of the sentiment words in the lexicon, many terms cannot be analyzed by traditional dictionary-based sentiment classifiers. For this reason, expanding the lexicon is necessary. Nowadays, the thesaurus lexicon has been built using a dictionary-based approach for classifying sentiments. It uses online dictionaries to collect thesauruses dependent on seed words, and stores only concurrence words in the thesaurus lexicon to improve the thesaurus lexicon 's reliability.
- Corpus-based
  The corpus-based approach uses data from the corpus, which can be further divided into statistical and semantic approaches. It suggests a data-driven approach in which users will have access to the labels of sentiment.

C. Hybrid Approach

A hybrid approach is the amalgamation of machine learning and lexicon-based approaches for optimum results. This approach generally yields better results because it combines a lexicon-based approach that performs better whenever there is a clear distinction between positive and negative sentiments of input dataset and robust machine learning algorithms.

IV. Applications of Sentiment Analysis

A. Applications that Use Reviews from Websites

Sentiment analysis can extract sentiments from an extensive collection of reviews and feedbacks from the internet, including product reviews, feedbacks, and comments about services. The extracted information can help both the users and the vendors to streamline the collection of feedback or rating for the given products, items, or services.

B. Applications for Subcomponent Applications

A sentiment predictor application could be useful in a recommendation system. Items or services that received a lot of poor reviews or lower ratings will not be recommended. Violence language and other harmful elements can also be identified easily by recognizing a somewhat negative sentiment and taking action against it accordingly.

C. Applications for Business Intelligence

As for most businesses, online reviews decide the successes or failures of the product. Sentiment analysis creates opportunities for owners of a business to determine their popularity among the customers and what they think about their products or services. This information can help business owners to assess the efficacy and capability of business brand communication in addition to enabling them to analyze the business stock price flow via social media [6]. Thus, Sentiment Analysis plays a remarkable role in businesses. The extracted sentiment from online reviews can help improve the business' reputation and maximize customer satisfaction.

D. Applications In Smart Homes

Sentiment Analysis is finding its way on the Internet of Things (IoT). In a smart home system, many different home automation devices might need to communicate with the external network commands to perform various actions. The network sentiment framework, as in Fig. 2, is introduced not long ago to enhance the security and privacy for a smart home. In this case, if the level of network sentiment of the smart home network raised due to the possibility of an attack, this information shall be disseminated to adjacent smart home networks. For example, different locations of the same building, which could be counteracted automatically by creating an intrusion prevention system (IPS) in real-time.

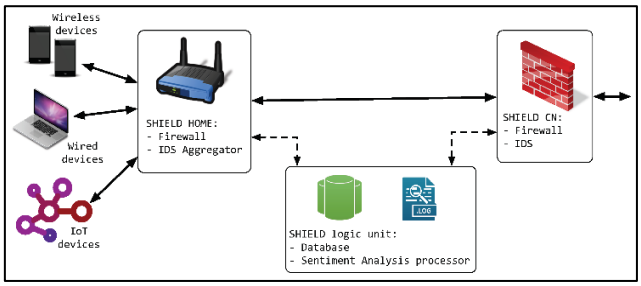

Fig. 2. Network sentiment framework in a smart home

V. Major Challenges

Sentiment analysis is a challenging task. In this section, challenges related to sentiment analysis of twitter data were addressed.

A. Analysis Uncertain, Incomplete and Sparse Data

Data sparsity is one of the characteristics of big data in which the data contain much noise due to the extensive use of acronyms and misspellings. This occurrence may have an impact on the accuracy of the sentiment classification. Moreover, privacy restriction also affects the clarity of the big data since information such as location is a useful feature that can enhance the functionality of sentiment classifiers. For instance, creating an overview of the sentiment of political posts within a particular country typically requires the location where the post is created. However, the geolocation information may only appear in specific posts. Therefore, the sentiment classifiers must be able to predict incomplete information in such a way, so that more precise details can be provided, along with a classifier that can categorize the sentiment polarity of a post.

B. *Identifying the Context of a Homograph Sentence*

Homographs are words with the same spelling but different meanings and sometimes origin and pronunciation. For example:

Sentence 1- When you get to the end of the road, turn left.

Sentence 2- She left the office at 5 pm.

The word "left" refers to the direction in the first sentence and past tense of leave in the second sentence. Yarowsky's semi-supervised machine learning algorithm may be applied to resolve ambiguities of Homographs in Sentiment Analysis in order to analyze the sentiments of online tweets. The features of this algorithm can be exploited to improve the performance of sentiment classification as compared to the traditional algorithms.

C. *Dependency of Domain*

Some sentences or phrases may have multiple meanings in different contexts. For instance, the term "unexpected" is positive in the movie's domain. However, if the same term is used in the domain of the steering of a vehicle, then it will leave a negative impression.

D. *Detection of Sarcasm*

Sarcastic sentences possess unique behavior. In comparison with a simple negation, a sarcastic sentence expresses a negative sentiment by using a positive connotation of words. A sentence such as "great perfume, you have to be bathed in it." The sentence contains only positive phrases, but it clearly expresses a negative sentiment. Naive bayes classifier and AdaBoost algorithms are two of the suggested techniques that can be used to detect sarcasm on twitter. By using Naive bayes classification, the tweets are classified into sarcastic and non-sarcastic. The AdaBoost algorithm is used to render a weak statement to strong statements.

E. *Thwarted expressions*

The overall polarity of some sentences can be specified by only a part of the text in a paragraph. For example, "This series should be great. It sounds like a good plot, the famous actors, and the supporting cast is talented as well." In this case, a simple bag-of-words approach will term it as positive sentiment, but the whole sentiment is negative. The machine learning-based approach can be used to detect thwarting in expressions by using the domain ontology to define domain-related key features. The approach involves two main steps, namely the learning the weights and creation of a model that classifies the opinions using the weights learned.

F. *Dependence on the order of the sentence*

Discourse structure is a term used to denote how an entire text is organized, and the analysis of it is necessary for sentiment analysis. For instance, an entity A is higher than B, conveys the exact opposite opinion from, B is higher than A.

VI. DATA AND METHODS

The association between the severity of the coronavirus disease (Covid-19) and its impact on public sentiment, particularly on social media, was explored in order to develop an understanding of how the public responds, especially the outrage that most people felt as the pandemic approached their countries.

The dataset used in this case study is Covid-19 tweets that include tweets regarding Covid-19 provided by Panacea Lab, and it is available publicly on Zenodo. It consists of timestamped tweets related to the Covid-19 pandemic. Besides, the dataset contained the features of the tweet's date and the tweet's text. Data cleaning of the text is applied to every tweet in the data frame, by making all the text in lowercase and by removing hyperlinks, hashtags, mentions, and retweets. The experimental results will be discussed in the next section.

VII. EXPERIMENTAL RESULTS

The subjectivity and polarity of each tweet are classified using the TextBlob model. Polarity in sentiment analysis refers to the identification of sentimental orientation, either positive, neutral, or negative in a natural text language. In this case study, the polarity of a text is indicated by -1 as extreme negative, 0 as neutral, and 1 as extreme positive. Subjective expressions are opinions that characterize people's thoughts regarding a particular subject or topic. The subjectivity of the tweets was defined as 0 indicating fact, and 1 indicating opinion. The relationship between subjectivity and polarity is illustrated in the graphs, as in Fig. 3, Fig. 4 and Fig. 5.

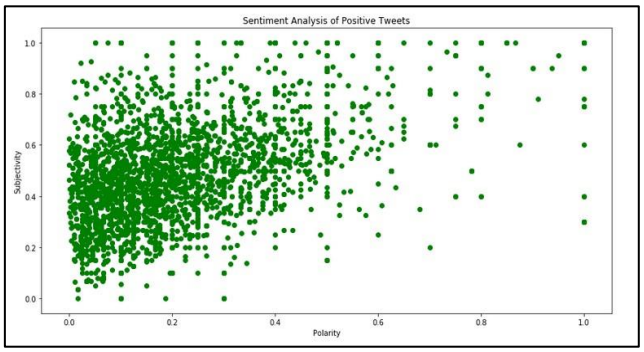

Fig. 3. Subjectivity and polarity of positive tweets

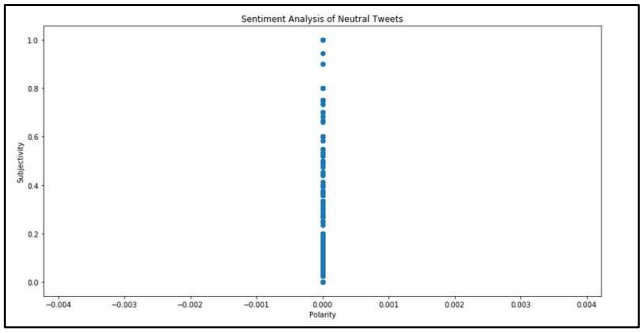

Fig. 4. Subjectivity and polarity of neutral tweets

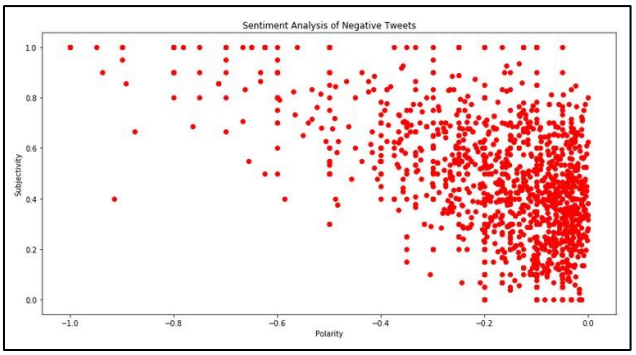

Fig. 5. Subjectivity and polarity of negative tweets

After classifying the tweets data, a total of 3417 tweets were classified as positive, 2265 classified as neutral, and 1818 of tweets as negative. The results are plotted into a bar chart, as in Fig. 6.

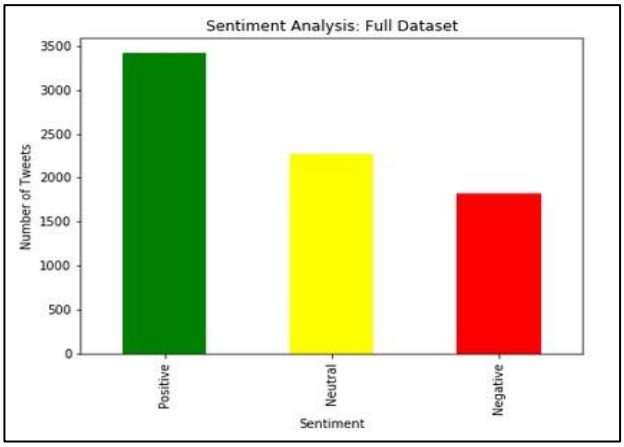

Fig. 6. The overall sentiment of the dataset

The tweets are classified according to February, March, and April sentiment, as illustrated in Fig. 7, the green line refers to tweets classified as positive, the red line refers to tweets classified as negative, and the yellow line refers to tweets classified as neutral.

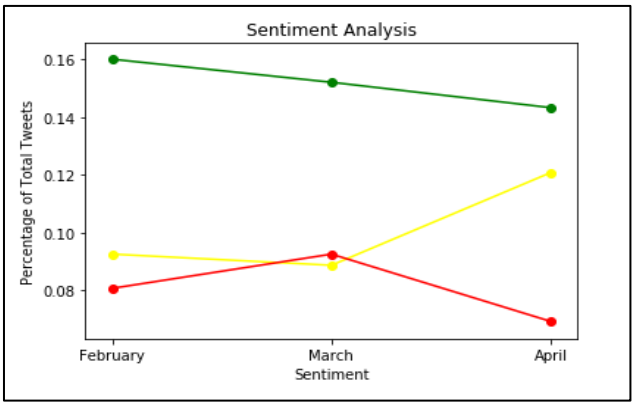

Fig. 7. Percentage of total tweets according to months

The TextBlob model shows that the percentage of positive tweets decreased substantially from February to April. As the pandemic spread, which started to take place rapidly in March and April, the public sentiment reflected a negative sentiment. A noticeable inverse relationship can be observed between the percentage of negative tweets and the percentage of neutral tweets. The percentage of negative tweets dropped significantly from March to April, while the percentage of neutral tweets increased significantly. One possible reason for this trend is that many people began self-quarantining during March. Another possible implication of the model trends was that, as people began to live with the virus present throughout the countries, they were used to living adjustments, making the perspectives less negative and more neutral. Future work may be able to obtain some insight into the spread of the virus by monitoring the sentiment on the twitter data.

VIII. Summary and Future Work

Due to the advancement in data storage, analytics, and access enabled through the big data frameworks, the sentiment analysis approaches using social media data have been exploited by enterprises as an essential tool for marketing planning. This paper discusses sentiment analysis approaches and their applications. The issues in the existing approaches should be improved in future work. Upon observation, the currently existing system is fixed and restricted to the quantity of data collected over time with a lot of missing information. In reality, text extracted from social media includes short-handed abbreviations, emoticons, and symbols that contribute significantly to the sentiment and emotion of the text. A dynamic system should offer additional recommendations based on any social media topic, including user's emotions and sentiment using full posts rather than just selecting exact words according to natural language from the posts, including possible future research to recommend a range of preferred topics to users.

The future of sentiment analysis should give deeper and broader insight as social media is increasingly becoming more emotive and expressive through the usage of emoticon. Rather than segmenting markets based on age, gender, and income, the organizations can further segment according to the audience expressed opinions.

Acknowledgment

This research paper was written as a requirement for the subject Special Topics In Computer Science (CSC649). I owe a massive amount of gratitude to Madam Nurazian Binti Mior Dahlan that ensures I got the help when needed and for her extraordinary compassion shown throughout the entire semester.

References

[1] K.Ali, H.dong, A.Bouguettaya, A.Erradi, and R.Hadjidj, "Sentiment Analysis as a Service: A social media based sentiment analysis framework",2016.

[2] K.Sailunaz and R.Alhajj, "Emotion and sentiment analysis from twitter text", 2019.

[3] J.Tang, "The Theory and Method of Sentiment Analysis Approaches for Application in the Big Data Frameworks",2017.

[4] S.Shahid, N.I Jaafar, S.Bahri, A.Sulaiman, P.S.Wai, Y.W.Chung, A.Zahid.Piprani, and M.A.Al-Garadi. "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges", IEEE Access, 2018.

[5] M.Ahmad, S.Aftab, S.Shah Muhammad, and S.Ahmad, "Machine Learning Techniques for Sentiment Analysis: A Review",2017.

[6] Z.Drus, H.Khalid, "Sentiment Analysis in Social Media and Its Application : Systematic Literature Review", 2019.