eISSN: 2637-1138
© Penerbit UMT

SENTIMENT ANALYSIS OF PEOPLE'S ACCEPTANCE TOWARDS THE NEW MALAYSIAN GOVERNMENT USING NAÏVE BAYES METHOD

R.U GOBITHAASAN* AND NUR FARHANA SYAHIRA CHE HAMID

Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

*Corresponding author: *gr@umt.edu.my*

Abstract: Sentiment analysis is a field of research that has a significant impact on today's nations, politics and businesses. It is an algorithmic process to comprehend the opinions of a given subject based on the Natural Language Processing (NLP) methodologies. It has received much attention in recent years and is proven vital in various fields, e.g., online product reviews and social media analysis (Twitter, Facebook, etc.). This paper reports the outcome of sentiment analysis to investigate people's acceptance of Pakatan Harapan, as the new Malaysian government, spearheaded by Tun Dr. Mahathir Mohamad and Dr. Wan Azizah, with an influence of Dato Seri Anwar Ibrahim. The objective is to classify tweets into three types of sentiments; positive, neutral and negative using Naïve Bayes method which is readily available in Python. The first step is tweets extraction for a month (March to April 2019) using search queries: {Pakatan Harapan, Mahathir, Anwar Ibrahim, Wan Azizah}. It is followed by tweets wrangling using NLP library and lastly output visualization in the form of a word cloud. A word cloud is a visual representation of text data with various font sizes depending on its probabilities. Final results showed that the tweets related to new government consist of neutral sentiment (41%) followed by positive sentiment (30%) and negative sentiment (29%). Malaysians do prefer the new government. However careful mitigation steps must be crafted to overcome controversial issues such as the 'Rome Statute' to avoid negative digital footprint, hence winning the Malaysians' heart.

Keywords: Sentiment analysis, Naive Bayes method, natural language processing

Introduction

In early 2018, Cambridge Analytica Ltd (CA) was accused of harvesting personal data of millions of Facebook users' profile for political campaigning and advertisements (Hern, 2018). It is a simple scenario of identifying psychographics for behavioral targeted marketing. Psychographics is a refined term of lifestyle study to measure attitudes, beliefs, opinions, personality traits, etc. (Anderson & Golden, 1984). It is the latest strategy used to segment personality to influence human behavior without their awareness. Examples of strategies are notifications, microtargeted ads and auto-play plugins.

Data in the form of texts messages shared via social media expresses various sentiments. The rise of social media such as blogs and social networks has uplifted the interest in sentiment analysis. Sentiment analysis or opinion mining is the new way to analyze people's review and sentiments towards products, services, individuals, issues and events. It is a strategic algorithm to determine, extract, and classify a string of texts according to its polarity (Bird, 2009). This study employs sentiment analysis to forecast people's acceptance of Malaysia's new government by using the data acquired on Twitter.

The phrase of sentiment analysis first appeared in Nasukawa and Yi (2003), and the word of opinion mining first appeared in Dave *et al.* (2003). However, the research on sentiments and opinions appeared earlier in Pang *et al.* (2002). Sentiment analysis and opinion mining mainly focus on classifying opinions to neutral, positive or negative sentiments. Currently, this field has become a well-known and interesting research area to study due to the availability of big data and computation power.

There are various ways to execute text classification. Castilo (2006) proposed to solve text classification tasks by different graph-based representation. They used different methods such as Author Attribution, Authorship Verification, Author Profiling and Sentiment Analysis to represent text documents as a graph. Next, the two supervised learning approaches were used to classify text document using graph which is Feature-Vector Approach and similarity computation between graphs. (Hutto & Gilbert, 2014) proposed VADER (Valence Aware Dictionary for Sentiment Reasoning), a rule-based model for general sentiment analysis. They used a combination of qualitative and quantitative methods to produce and then validate a gold standard lexicon which is especially attuned to sentiment in microblog-like contexts. Violos *et al.* (2009) proposed a method called Word-Graph Sentiment Analysis Method (WSAM) to identify the sentiment that was conveyed in a microblog document using a sequence of the words.

Machine learning is a notable approach which is gaining momentum recently in text classification environment. It is an algorithmic practice for computers to learn from data. Therefore, there is an increasing number of case studies carried out to understand the effectiveness of machine learning algorithm, such as Twitter sentiment classification using distant supervision as introduced by Go *et al.* (2009). Their approach is by using different machine learning classifiers (Naïve Bayes, Support Vector Machine and feature extractors (unigrams, bigrams, etc). Boutet *et al.* (2012) proposed a practical algorithm to identify political inclination of users using the amount of Twitter messages which seem related to political parties. Violos *et al*. (2016) concluded that the Gaussian Bayes classifier can be trained easier and produces more accurate predictions than other classifiers.

Reverend Thomas Bayes (1702–61) developed the Bayes' theorem and proposed a method to compute distribution for the probability parameter of a binomial distribution. In 1763, Richard Price modified the method of solving a Problem in the Doctrine of Chances. Recently, Bayesian classifiers have been quite popular and were reported to perform well (Langley *et al*. (1992), Sahami (1996), Friedman (1997)). This probabilistic model uses a collection of labelled training examples to estimate the parameters of the generative model. Classification of new examples is performed with Bayes' rule by selecting the most likely class. The naive Bayes classifier assumes that all features of the examples are independent of each other given the context of the class.

On 10th May 2018, the 14th Malaysian general election results were announced where Tun Dr. Mahathir's coalition had officially won 121 seats of 222 seats to establish a new government of Malaysia. For the first time, the *Barisan National* political party that has been ruling since 1957 only won 79 seats and their counterpart *Parti Islam Malaysia* won 18 seats, hence were voted of out power. The coalition spearheaded by Mahathir for *Pakatan Harapan* won 113 seats along with other candidates won 12 seats (https://election.thestar.com.my/). Similar to other citizens in the world, Malaysians too utilize social media to show their interest or opinion towards the general election result. The statistics of social media usage in Malaysia from the year 2017-2018 that we obtained from StatCounter GlobalStats shows that Malaysians use Facebook (89.32%), followed by Twitter (3.23%), Pinterest (2.82%), YouTube (2.27%), Instagram (1.03%), Tumblr (0.78%), Reddit (0.23%), and others (0.32%). Twitter is one of the media which can be used without any fee to extract people's acceptance of the new Malaysian government. Since there were charges imposed to access the Facebook data, we used the free text messages from Twitter to do the sentiment analysis.

Materials and Methods

Naïve Bayes Method

Naive Bayes classifiers mostly used in text classification have a higher success rate as compared to other algorithms. The advantage of this method is, each feature independently contributes to the decision of the finalized label, hence the features has interaction with each other (Bird *et. al*, 2009). Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features that exist. Naive Bayes classifiers can be stated as follows (Manning, 2008):

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

where,

- d denotes the document (strings of Tweets)
- c denotes the classes (positive, negative or neutral sentiments)
- is the probability of class given in the document
- is the prior probability of class
- is the likelihood which is the probability of
- document of the given class
- is the prior probability of the document

In our case, a tweet d is represented by a vector of k attributes such as $d = (w_1, w_2, ...w_k)$. In Naive Bayes, we will assume that all of the feature values w_j are independent given the category label c; for $i \neq j$, w_i and w_j are conditionally independent given the category label c. So the Bayes rule can be rewritten as,

$$P(c|d) = P(c) \times \frac{\prod_{j=1}^{K} P(w_j|c)}{P(d)} \quad (2)$$

Based on this equation, the maximum a posterior (MAP) classifier can be constructed by seeking the optimal category which maximizes the posterior *P* (*c*|*d*) such as:

$$c *= arg \max_{c} P(c|d) \quad (3)$$

$$c *= arg \max{}_c \left\{ P(c) \times \frac{\prod_{j=1}^{K} P(w_j|c)}{P(d)} \right\} \quad (4)$$

$$c *= arg \max_{c} \left\{ P(c) \times \prod_{j=1}^{K} P(w_j|c) \right\} \quad (5)$$

However, *P*(*d*) is removed since it is constant for every category c. The prior distribution *P*(*c*) can be used to incorporate additional assumptions about the relative frequencies of classes. It is computed by:

$$P(c) = \frac{N_i}{N} \quad (6)$$

where *N* is the total number of training tweets and N_i is the number of training tweets in class *c*. The likelihood $P(w_j|c)$ is calculated using the formula:

$$P(w_j|c) = \frac{1 + count(w_j, c)}{|V| + N_i} \quad (7)$$

where, (w_j, c) is the number of times that word w_j occurs within the training tweets of class, $|V| = \Sigma\, w_j$ the size of the vocabulary. The simplest smoothing method can be applied to solve the zero-probability problem that occurs when our model does not have this word in the training data, which is Laplace or add-one since we use 1 as constant. Below is a standard framework of the steps involved in the sentiment analysis.

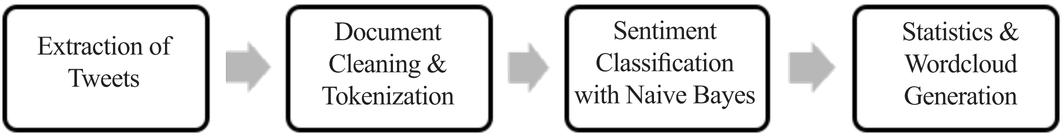
Extraction of Tweets
Document Cleaning & Tokenization
Sentiment Classification with Naive Bayes
Statistics & Wordcloud Generation

Figure 1: A standard framework of sentiment analysis

Extraction of Tweets

We extracted tweets for a month (March to April 2019). First, we registered a developer account with Twitter to extract the data. We proceeded the extraction process with a given key using Python. Tweepy is an open-source package available online to handle all the interaction with the Twitter APIs (application programming interfaces). A string of tweets is called a document. Tweets can be extracted based on keywords and stored as a comma-separated values (CSV) format. However, these tweets are in a raw format, filled with icons, unwanted characters and Uniform Resource Locator (URL),

Document Cleaning & Tokenization

We used Beautifulsoup and RegEx packages to clean the tweets by removing links and special characters respectively. First, the Hypertext Markup Language (HTML) decoding must be executed where raw tweets containing HTML encoding that has not been converted to text, seen as text field '&amp','&quot', etc. The second part of the preparation is dealing with @mention. Even though @mention carries certain information about a user that the tweet mentioned, this information does not give any value for sentiment analysis. There are also some unwanted patterns of character such as "\xef\xbf\xbd" that were detected in some of the entries. This is known as UTF-8 BOM which is a sequence of bytes (EF BB BF) that allows the reader to identify it as being encoded in UTF-8. By decoding text with 'utf-8-sig', this Byte Order Mark will be replaced with Unicode unrecognisable special character, then we can process this as "?". Next, we remove the hashtag without removing the text that was used with the hashtag as it may contain information about the tweets. Once these steps are done, we would be able to carry out tweet statistics. A simple boxplot would be able to give the summary of the tweet's length and word counts of each tweet which were cleaned as shown in the next section. The outcome of this step is a CSV file in the form of a string of tweets with simple words, which are called documents.

Natural Language Toolkit (NLTK) is a popular package for building Python programs to work with tweets. It has interfaces to over 50 corpora and lexical resources such as WordNet and a complete package of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, etc (Bird *et. al*, 2009). The next step will be on tokenization, stemming/lemmatization, stop words identification. Tokens are individual terms or words, and tokenization is the process of splitting a string of text into tokens. The stemming process is a rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word. In this step, we use WordPunctTokenizer which is available in the NLTK package to splits all punctuations into separate tokens. For example, "cook", "cooks", "cooked", "cooker" and "cooking" are the different variations of the word – "cook". Next, we remove the short words from the cleaned data that have length 3 or less such as "hmm" and "ohh.". Finally, we have tweets which have been cleaned and striped to words for sentiment classification.

Sentiment Classification with Naive Bayes

TextBlob is a python package available free for natural language processing (NLP) routines, namely for part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, etc. It has a class called NaiveBayesClassifier to train a set of tweets which can be customized either as positive, neutral or negative sentiments. However, in this paper, we used the standard classifier which is readily available with the library without carrying out any training where each word has been labelled with a sentiment. As shown in the previous section, the calculations are simple yet robust, hence the percentage of positive, negative and neutral classification can be computed in a short time.

Statistics and Wordcloud Generation

To get a better insight into the sentiment, simple statistics can be applied to show the common words that have been used in the cleaned tweet, thus identifying its sentiment. A wordcloud is a visualization method wherein the most frequent words appear large and the less frequent words appear in smaller sizes based on the sentiments. By using wordcloud, it is easier to detect the sentiment by judging the size of a word.

Results and Discussion

The total tweets accumulated during the 32 days from 4 March until 6 April are 2212 tweets. The search queries are Mahathir, #mahathir, Pakatan Harapan, Anwar Ibrahim and Wan Azizah. Based on the output, there are 641 data classified as negative sentiments, 664 data belongs to positive sentiments and 907 data belongs to neutral sentiments. Tables below show some output of the pre and post cleaned tweets by HTML decoding, removing @mention, removing URL links, decoding UTF-8 BOM, and removing non-letters characters.

Table 1: Examples of the tweet cleaning process

	Before	After
HTML decoding	intention to capture Sabah &amp; Sarawak.......no more autonomy https://t.co/ zxeqntMWY4'	intention to capture Sabah & Sarawak....... no more autonomy https://t.co/ zxeqntMWY4
Removing @mention	@syahredzan It was indeed one of the PH Minister who failed to observe the procedure when agreeing on the Rome Statute. https://t.co/ Rnsxt09p8X	It was indeed one of the PH Minister who failed to observe the procedure when agreeing on the Rome Statute. https://t.co/ Rnsxt09p8X'
Removing URL links	"Exclusive: EU risks \'trade war\' with Malaysia over palm oil - Mahathir" - https://t.co/ Fo1ncRr67V'	"Exclusive: EU risks \'trade war\' with Malaysia over palm oil - Mahathir" - '
Removing UTF-8 BOM	Equanimity sold to Genting for US$126m, AGC confirms ï¿½ https://t.co/ XM6ds5fdy	Equanimity sold to Genting for US$126m, AGC confirms? https://t.co/ XM6ds5fdy
Removing non-letters characters	#NSTnation: After selling off superyacht Equanimity, the government is considering selling land to pare down its de https://t.co/ pDUveEeLVT'	NSTnation After selling off superyacht Equanimity the government is considering selling land to pare down its de https t co pDUveEeLVT

Boxplot is a visual output that we used to show the overall length of pre-clean and post-cleaned tweets. Figure 2 illustrates boxplot with five minimum parts. (3 times less the size of the body), first quartile (Q1), median, third quartile (Q3) and maximum (3 times more the size of the body). There are many outliers marked by small circles which are located beneath the minimum values as shown in the pre-cleaned tweets on the left. Outliers can cause errors because they differ significantly from the lengths of tweets. The boxplot on the right represents tweets upon cleaning, with a median of 70 words per tweet and there are no outliers (the length of tweets are within 3 times the size of the body) hence ready for sentiment analysis. Figure 3 shows a pie chart which represents the sentiments of the leaders of the new Malaysian government.

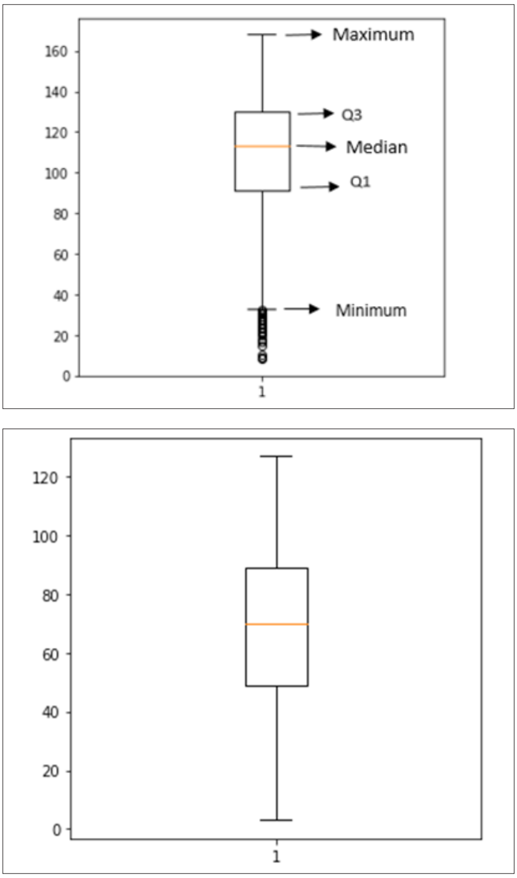

Figure 2: Boxplot of the length of pre-cleaned (left) and post-cleaned tweets

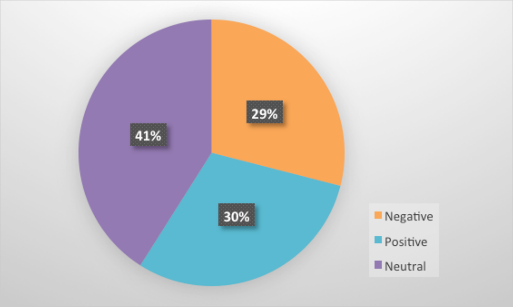

Figure 3: Summary of the leaders of the new Malaysian government

Figure 4 illustrates the wordcloud of all the tweets. The font size and the colour of each word are different based on the frequency in the tweets. The most frequent words appear large and the less frequent words appear in smaller sizes. Different colours have been used to make it easier for us to observe the wordcloud.

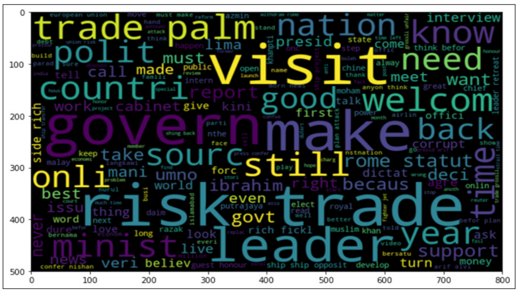

Figure 4: Wordcloud of all words

Figure 5 illustrates a wordcloud that consists of positive words in the tweets. The words such as ‘leader’, ‘support’, ’government’ and ‘good’ are in large font, so these words have been mentioned several times in the positive tweets. The less mentioned words such as ‘well’, ’believe’ and ‘respect’ have only been repeated once or twice in the tweets.

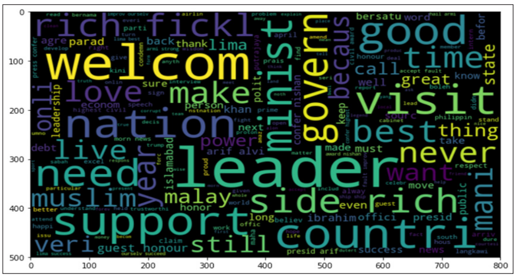

Figure 5: Wordcloud of positive words

Figure 6 consists of negative words wordcloud. The words such as 'Rome statute', 'risk', 'UMNO' and 'trade palm' are in large font, so these words have been disclosed commonly in the negative sentiment tweets. The less mentioned words such as 'fail' and 'unfair' are not repeated many times by the users on Twitter.

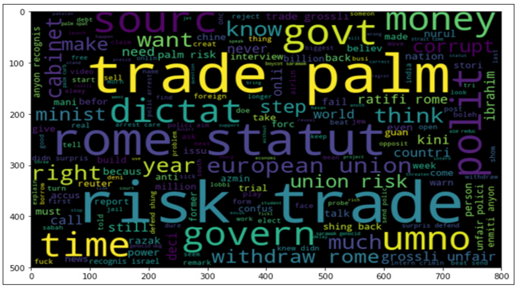

Figure 6: Wordcloud of negative words

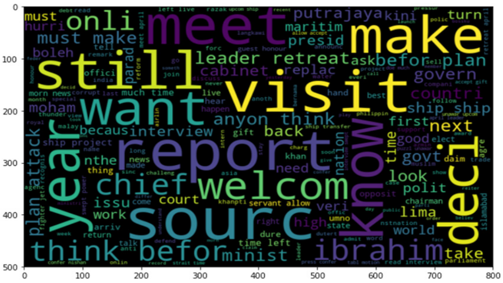

Figure 7: Wordcloud of neutral words

Figure 7 is a wordcloud that consists of neutral words in the tweets where 41% of the tweets related to the new government consists of neutral sentiment. Hence, the majority of words are tweeted in this category. The words such as' know', 'want', 'year' and 'visit' in large font, so these words have been stated quite frequently in the neutral tweets. The less mentioned words in this wordcloud are 'tell', 'back' and 'fighter'

Discussion

This study is carried out almost a year after the election to understand whether Malaysians are satisfied with the new government. Out of 2212 tweets collected in a month, keywords related to the new government were; 'Mahathir', 'Pakatan Harapan', 'Anwar' and 'Wan Azizah', there are 641 tweets belong to the negative sentiment (29%), 664 data belongs to positive sentiment (41%) and 907 data belongs to the neutral sentiment (30%). In the wordcloud of positive words, the words such as 'leader' and 'support' have been mentioned frequently while in the wordcloud of negative words, we can see that controversial issues such 'palm oil trade' and 'Rome statute' have been repeated in the tweets.

Conclusion and Future Work

As a conclusion, Malaysians do prefer the new government with a difference of 1% between positive and negative sentiments. Negative sentiments are due to controversial issues occurring during the study such as 'Rome statute' and chemical pollution of Kim river outbreak. However, the majority of sentiments,41% were under neutral classification indicating doubts over the new Malaysian government's functionalities. Hence, careful mitigation steps must be crafted to avoid negative digital footprints to win the Malaysians' heart who have neutral sentiments. Future works include customizing the tweets using Naïve Bayes Classifier classification to obtain a better result and analyse tweets in the Malay language.

Acknowledgments

Authors thank the anonymous reviewers for their insightful comments and suggestions.

References

Alec Go, R. B. (2009). Twitter Sentiment Classification using Distant Supervision. *Natural Language Processing*, 50-62.

Anderson T. W., & Golden L.L. (1984) ,"Lifestyle and Psychographics: a Critical Review and Recommendation", in NA - Advances in Consumer Research Volume 11, eds. Thomas C. Kinnear, Provo, UT : Association for Consumer Research, Pages: 405-411.

Andrew McCallum, K. N. (1998). *A Comparison of Event Models for Naive Bayes Text Classification.* Washinghton: AAAI Technical Report.

Antoine Boutet, H. K. (2012). What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. *ASONAM '12 Proceedings of 2012 International Conference on Advances in Social Networks Analysis and Mining* (pp. 132-139). Washington,DC: IEEE Computer Society.

Bird, Steven, Edward Loper & Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Bo Pang, L. L. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 79-86).

Castillo, O. C. (2006). Text Analysis Using Different Graph-Based Representations. *SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 485-492). Seattle: ACM New York.

Christopher D. Manning, P. R. (2008). Text classification and Naive Bayes. In *Introduction to Information Retrieval* (pp. 253-287). Cambridge University Press.

Friedman, J. H. (1997). On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 55-77.

Hern, Alex (April 10, 2018). "How to check whether Facebook shared your data with Cambridge Analytica". The Guardian. https://www.theguardian.com/technology/2018/apr/10/facebook-notify-users-data-harvested-cambridge-analytica

Hays, J. (June, 2015). *POLITICS AND POLITICAL PARTIES IN MALAYSIA*. Facts And Details: http://factsanddetails.com/southeast-asia/Malaysia/sub5_4d/entry-3668.html

Hutto, C. a. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)* (pp. 123-134). Ann Arbor, MI: Association for the Advancement of Artificial Intelligence.

John Violos, K. T. (2009). Sentiment Analysis using Word-Graphs. *WIMS '16 Proceedings of the 6th International Conference on Web Intelligence,Mining and Semantics.* Nimes,France: ACM New York.

Kushal Dave, S. L. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *WWW'03 Proceedings of the 12th International conferenc on World Wide Web* (pp. 519-528). Budapest,Hungary: ACM New York.

Nir Friedman, D. G. (1997). Bayesian network classifiers. *Machine Learning*, 131-163.

Pat Langley, W. I. (1992). An analysis of Bayesian classifiers. *AAAI '92 Proceedings of the tenth national conference on Artificial Intellegence* (pp. 223-228). Son Jose,California: AAAI Press.

Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook.* Birmingham,UK: Packt Publishing Ltd.

Russell, R. (2018). *Machine Learning:Step-by-Step Guide To Implement Machine Learning Algorithms with Python.* California: CreateSpace Independent Publishing Platform.

Sahami, M. (1996). Learning limited dependence Bayesian classifiers. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 335-338). AAAI Press.

Sue, A. (6 May, 2013). *Analisa Keputusan Pilihanraya Umum ke 13 Malaysia : BN menang 133 kerusi, PR 89 kerusi*. Wanista. https://www.wanista.com/2013/12524/analisa-keputusan-pilihanraya-umum-ke-13-malaysia-bn-menang-133-kerusi-pr-89-kerusi/

Teacher, L. (Friday February, 2018). *The Political Parties in Malaysia*. Free Law Essay. https://www.lawteacher.net/free-law-essays/administrative-law/the-political-parties-in-malaysia-administrative-law-essay.php

Tetsuya Nasukawa, J. Y. (2003). Sentiment Analysis :Capturing favorability using Natural Language Processing. *K-CAP '03 Proceedings of the 2nd International Conference on Knowledge capture* (pp. 70-77). Sanibel Island,FL: ACM New York.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 417-424). Philadelphia: Association for Computational Linguistic.