International Journal of Synergy in Engineering and Technology Vol. 3 No.2 (2022) 72-83

Combination of Lexicon Based and Machine Learning Techniques in the Development of Political Tweet Sentiment Analysis Model

Liyana Safra Zaabar[1*], Mohd Ridzwan Yaakub[2], and Muhammad Iqbal Abu Latiffi[3]

[1] Prime Minister's Department, Pusat Pentadbiran Kerajaan Persekutuan, Putrajaya
[2] Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Selangor
[3] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor

[*]Corresponding author: liyanasafrazaabar@gmail.com

KEYWORDS

Lexicon Based Approach
Sentiment Analysis
Opinion Mining
Twitter
Machine Learning
Feature Extraction
TF-IDF
Naive Bayes
Political Tweet

ABSTRACT

Twitter is a popular micro blogging social media platform and the largest data contributor in the analysis of political sentiments in the United States especially in Presidential elections. Lack of labeled data as well as requirements of testing data are major problems in political domain since due to their constant change according to current events. The contribution of this study is to compare two dictionary-based Lexicon approaches which are Bing Liu Opinion Lexicon and Textblob for tweets labelling. Some comparative models have been developed. Model based on Bing Liu Opinion Lexicon which used machine learning algorithm TF-IDF for feature extraction and also classified with Naïve Bayes gets the highest F1-Score with 93%, outperformed our baseline model with score of only 68%. Test results have shown the effectiveness of combining lexicon approaches and machine learning algorithms in the development of sentiment analysis model.

1.0 Introduction

The Internet was crowned the 'king of all political campaigns' after the 2008 U.S. presidential campaign and the election results were released. It has become a platform for online learning, exchanging opinions, and sharing opinions (Ahmad et al., 2020). The Internet plays an important role in the campaigns conducted by candidates running for the presidency. The use of social media is now getting a lot of attention from a variety of different personalities and entities such as candidates, political parties, the general public, and corporate news. Therefore, analysis of political scenarios on social networks is important for candidates, political parties, or the public. As noted by (Al-Saffar et al., 2018), results from social media sentiment analysis can help candidates and their parties know their position and general sentiment orientation.

Sentiment analysis is a text mining category in which the collected texts are classified based on the polarity of sentiment which can be positive, negative or neutral. There are various methods

Received October 2022; received in revised form November 2022; accepted December 2022

© 2020 University College TATI (UC TATI). All rights reserved.

of sentiment analysis including language Model, Part of Speech Tagging (POS) and semantic orientation (Awwalu, Bakar and Yaakub, 2019).

According to a report (Center, 2016), a research center that conducts public opinion polls, demographic research and social sciences based in the United States, in January 2016, 44 percent of adults in the United States stated that they learned about the presidential election through social media . Additionally, 24 percent of adults in the United States reported using both candidate social media as a source of news and information, of which more than 15 percent used both candidate websites or emails combined. 17.1 million tweets, the first presidential debate between Donald Trump and Hillary Clinton is the most talked about debate on the Twitter.

According to (Blake, 2016), an author in the Washington Post also mentioned the 2016 Presidential Election was “the most negative presidential election in our lives” and many people from around the world voiced their feelings for each candidate on different social networks including Twitter. The growth of Twitter users since the last election suggests that it may be a more accurate voting tool since the 2012 election (Joyce and Deng, 2018). According to (Statista.com, 2017), the number of monthly active Twitter users worldwide from the fourth quarter of 2012 to the fourth quarter of 2016 increased from 185 million to 328 million.

The large amount of text on websites has been proven to be a tool of communication between humans and machines as well as human-to-human interaction (Shaheen et al., 2014). The volume of this text has grown rapidly, creating large -dimensional specific content (Fersini, Messina and Pozzi, 2014; Choi and Lee, 2017). With the increasing availability of opinion -rich texts on Twitter, the challenges and opportunities in sentiment analysis are growing. Some of the challenges as noted by (Ranganathan and Bagavathi, 2020) include a lack of tweets, emoticons, and informal expressions. These challenges not only make the sentiment analysis process challenging but also affect the accuracy of the model for sentiment analysis (Awwalu, Bakar and Yaakub, 2019). Previous efforts to assess Twitter user sentiment have suggested that Twitter can be a valuable resource for studying political sentiment and it illustrates the real landscape of offline politics (Ebrahimi, Yazdavar and Sheth, 2017).

2.0 Research Background

In the following sub-sections, we discuss the related research background on research problems and research objectives.

2.1 Research Problems

The manual annotation method is a commonly used method in sentiment analysis. The time-consuming and high-cost annotation process makes this method less relevant for political domains that require faster analysis, less cost and the availability of training data (Yaakub, Latiffi and Zaabar, 2019). Moreover, the annotation process is more challenging because the length of one tweet is no more than 140 characters which complicates the sentiment determination process according to this method (Ebrahimi, Yazdavar and Sheth, 2017).

Many good systems and models have been developed in the political domain but it is likely that these systems and models are less relevant or unusable after some events, propaganda or political scandals occur because political data is constantly changing and with the passage of time. Thus, to ensure that relevant systems are developed and always applicable, an important and key element is to provide training data sets that are always up to date in line with current events and agendas (Ebrahimi, Yazdavar and Sheth, 2017).

The second problem faced in this study is the amount of text in social media including twitter which is constantly growing rapidly and creating large-dimensional specific content (Fersini, Messina and Pozzi, 2014). This problem occurs because there are repetitive and irrelevant terms in the feature space (Ahmad, Bakar and Yaakub, 2019b, 2019a). Moreover, this condition can cause the classification algorithm to experience overfitting to train the sample and less effective on new samples. To overcome this problem, feature extraction methods can be used for text and sentiment classification (Ahmad, Bakar and Yaakub, 2019b).

The two main methods in sentiment analysis are the lexicon approach and machine learning (Alexander and Omar, 2017). The merging or integration between these two methods has resulted in high accuracy based on the improvement of the feature extraction process (Al-Saffar et al., 2018). (Hasan et al., 2018) in their study have used hybridization methods that combine lexicon approaches such as Textblob, SentiWordNet, W-WSD and machine learning to study political tweets in India. The experimental results showed that the Naïve Bayes algorithm achieved 76% accuracy for Textblob and 79% for W-WSD, while the Support Vector Machine (SVM) algorithm achieved 62.67% accuracy for Textblob and 62.33% for W-WSD. (Hutto and Gilbert, 2014) in their study using VADER model and conducted comparative studies with other lexicon approaches in the classification of social media text sentiment using machine learning methods. The results of the study found that this Vader model has surpassed the accuracy of manual annotation with an achievement of 96%.

2.2 Research Objectives

In particular, this study was implemented to achieve the following objectives. The first objective is to identify the best techniques for labeling tweets using the lexicon approach without using annotations manually. Next is to identify feature extraction techniques and machine learning algorithms that can produce sentiment classification accuracy. The final objective is to combine lexicon approach and machine learning methods to develop high accuracy sentiment analysis models.

3.0 Proposed Model

In order to carry out the study effectively and meet the objectives of the study, this study method includes seven important sub-phases starting with literature review, data analysis, data pre-processing, labeling, feature extraction using Document Term Matrix and Term Frequency Inverse Document Frequency (TF-IDF), sentiment classification, evaluation and output analysis as shown in Figure 1.

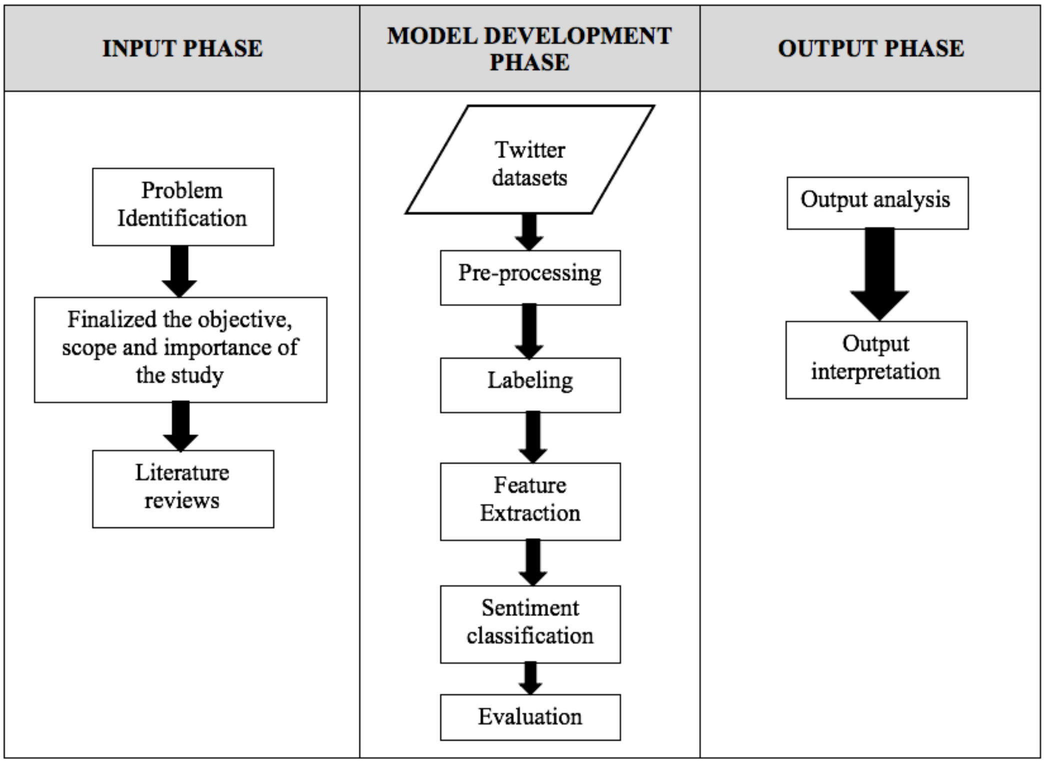

Figure 1: Research Proposed Model

3.1 Input Phase

The input phase involves several work processes such as identifying problems involving the analysis of sentiments in the political domain as well as setting the objectives, scope and importance of the study. The most important subphase in this phase is the literature review which will help identify problems and help set the scope, objectives and importance of the study.

3.2 Model Development Phase

The model development phase includes two main subphases namely labeling and Tweets classification. Figure 2 shows the model development phase framework used in the study.

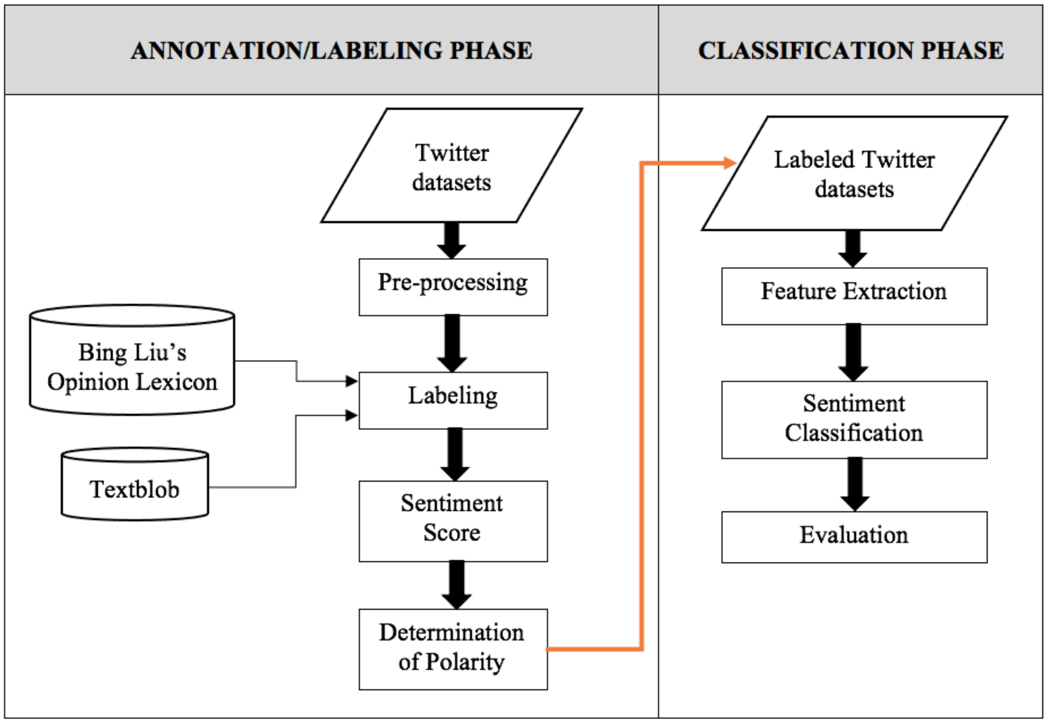

Figure 2: Framework for Model Development Phase

The model development phase begins with identifying data that fits the focus of the study, namely the twit data of the two main candidates for President of the United States in 2016, namely Donald Trump and Hillary Clinton. This data was obtained from Ben Hamner's Kaggle repository where the data was originally collected from Twitter's application programming interface (API) in the political domain. This data is unstructured and unlabeled data. This Tweets data contains 6,444 twits from the two candidates mentioned above as well as contains 28 attributes that need to be processed. Once the raw data is provided, the data will go through a pre-processing process that involves data cleaning, data filtering, data transformation, attribute generation and attribute selection. There are 9 data cleaning methods used in this study as in Table 1. All 9 steps are used because they meet the objectives of the study to produce a clean data set for the next phase.

The labeling phase is the second subphase in the model development phase. The labeling process is done on the Tweets data sets of Donald Trump and Hillary Clinton after going through the data pre-processing process. Two lexicon approaches are used in the labeling phase, namely the dictionary-based method, namely the lexical database known as Bing Liu Opinion Lexicon and Textblob which is a library developed in Python programming language to label Tweets data.

Table 1: Techniques for Pre-Processing

No.	Technique
1	Remove #
2	Remove @
3	Convert to small letter
4	Remove hyperlink
5	Remove punctuation
6	Remove special char
7	Remove numbers
8	Remove underscore _
9	Remove stopwords

Sentiment score is a value of polarity strength that will determine the polarity of a tweet whether positive, negative or neutral. The labeling phase ends with the distribution of sentiments for the entire data set which will produce a set of 6,444 labeled Tweets data that will be used for the next phase which is the feature extraction that operated by Term Frequency Inverse Document Frequency (TF-IDF).

TF-IDF is one of the most popular termweight schemes today. The TF-IDF weighting scheme has two components. Term Frequency measures the number of occurrences of a term in a document. Inverse Document Frequency calculates the logarithm of the number of documents in a corpus to the number of documents in which the term appears (Qazi and Goudar, 2018). The following is the TF-IDF formula:

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{total\ words\ in\ document\ j}$$

$$IDF(i) = \log_e \frac{Total\ documents}{documents\ with\ term\ i}$$

The output produced in the feature extraction phase will go through a classification phase using three popular algorithms in text and sentiment classification namely Random Forest (RF), Support Vector Machine (SVM) and Naïve Bayes (NB).

3.3 Output Phase

The output phase will analyze and interpret the labeling and classification results obtained from the model development phase and subsequently determine the best model for the classification of sentiment in political Tweets.

4.0 Result and Discussion

4.1 Labelling Using Bing Liu's Opinion Lexicon

4.1.1 Sentiment Score Distribution

In the labeling method using Bing Liu's Lexicon of Opinion, the sentiment score is calculated based on the total frequency of a sentiment word appearing in a tweet. Figure 3 shows the distribution of sentiment scores for the entire data set where positive sentiment scores are highest and dominate the data.

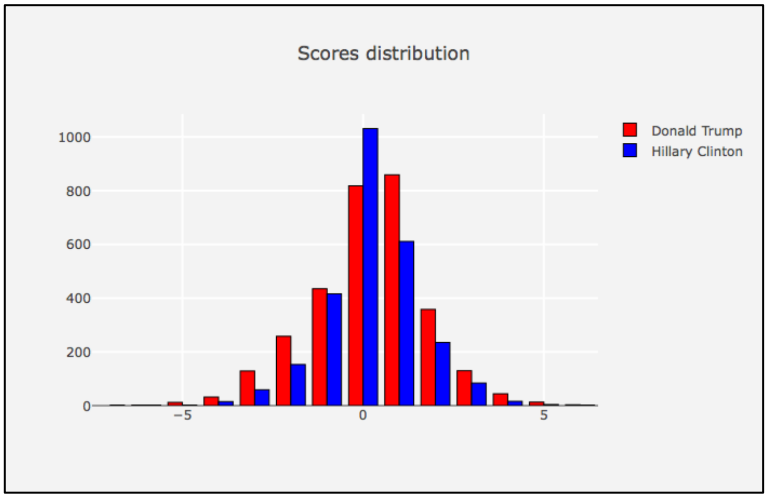

Figure 3: Distribution of Sentiment Scores

4.1.2 Sentiment Polarity Distribution

The polarity distribution visualized in Figure 4 shows positive sentiment is highest in this data set followed by negative and neutral. The findings of the study also show that the sentiment played out in Donald Trump's tweets is more positive than Hillary Clinton's.

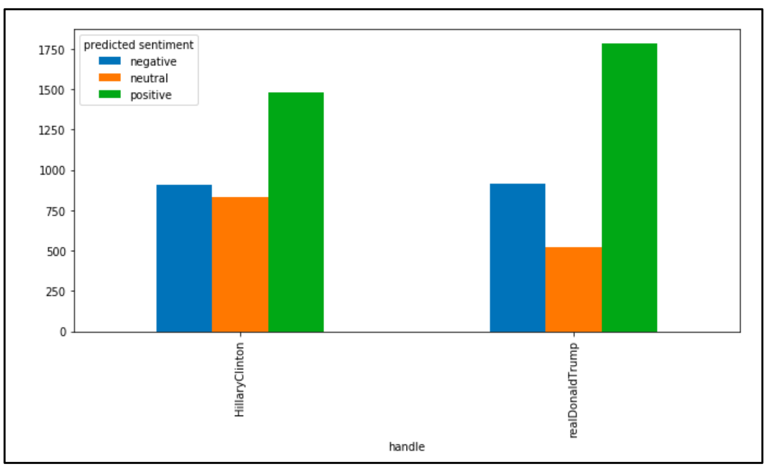

Figure 4: Distribution of Sentiment Polarity

4.2 Labelling Using the Textblob Library

4.2.1 Sentiment Score Distribution

The textblob library method differs slightly from the previous approach where sentiment scores were calculated through polarity strength in tweets in the range [-1,1] as shown in Figure 5. The results showed positive sentiment was highest in this data set followed by neutral and negative.

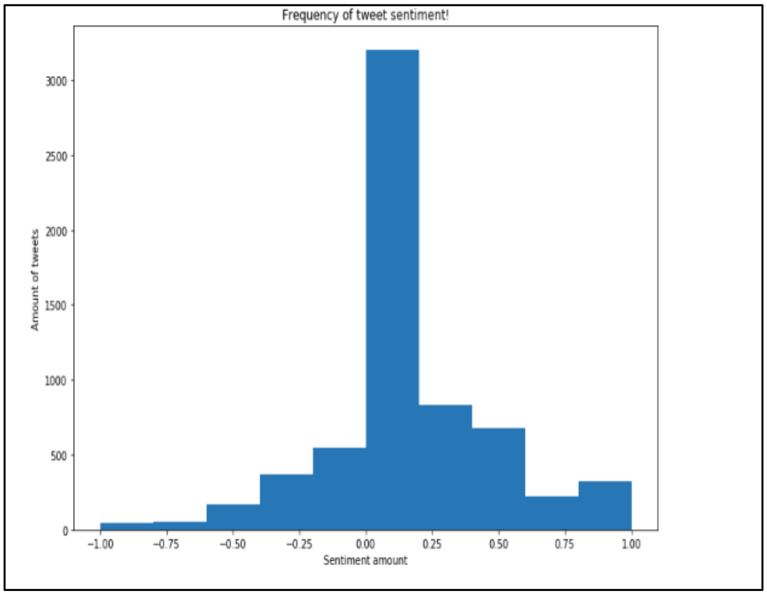

Figure 5: Distribution of Sentiment Score

4.2.2 Sentiment Polarity Distribution

Figure 6 shows the polarity distribution of sentiment in tweets where the results show positive sentiment is the highest sentiment for both candidates followed by neutral and negative sentiment. Studies also show Trump uses more positive sentiment than Hillary.

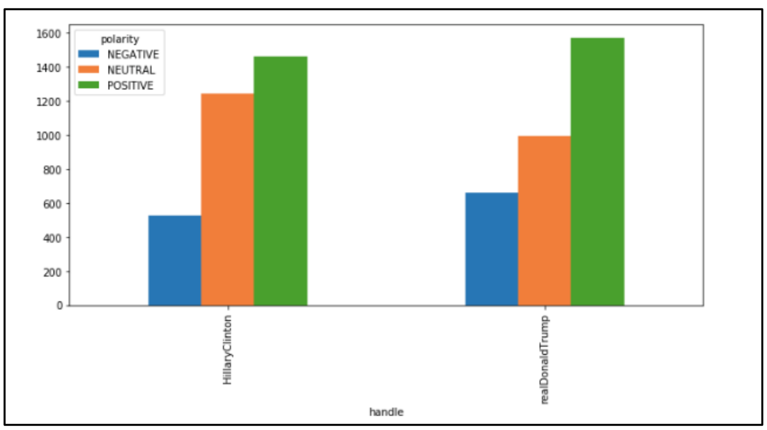

Figure 6: Distribution of Sentiment Polarity

4.3 Sentiment Classification Based on Bing Liu's Opinion Lexicon Model

Table 2 shows the performance of the classification model using Bing Liu's Lexicon of Opinion compared to the policy model that has been developed by Pavan Raj using the Random Forest classifier. Prior to undergoing the classification phase, the labeled data sets underwent a feature extraction process using TF-IDF. The highest accuracy, retrieval and F1-Score were on the Bing Liu model alongside the Naïve Bayes classifier with a score of 0.93% followed by Bing Liu model with SVM classifier with a score 0f 0.78% and the last were Bing Liu model with RF classifier and baseline model which obtained a score of 0.68% for both. What can be seen from these results, the NB classifier is more likely to show better performance that the SVM classifier for this data set. The overall performance of the model is visualized in Figure 7.

Table 2: Performance for Baseline Model and Bing Liu Opinion Lexicon Model

Model	Accuracy	Recall	F1-score
Baseline Model +RF	0.68	0.71	0.68
Bing Liu +TF-IDF+RF	0.68	0.70	0.68
Bing Liu +TF-IDF+SVM	0.78	0.71	0.74
Bing Liu +TF-IDF+NB	0.93	0.93	0.93

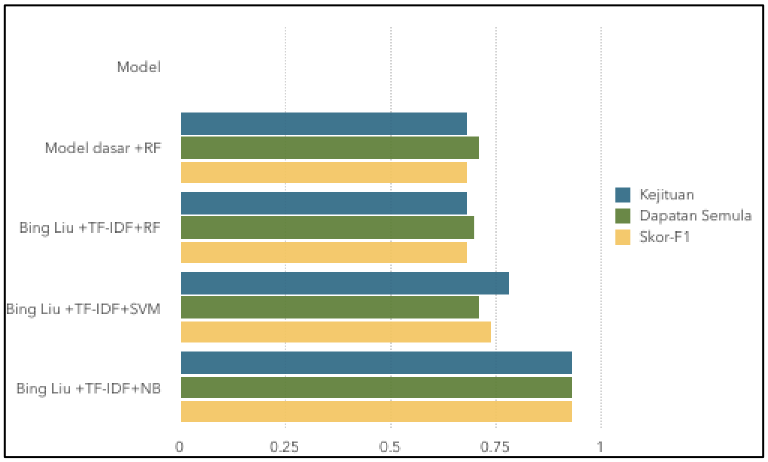

Figure 7: Performance for Baseline Model and Bing Liu Opinion Lexicon Model

4.4 Sentiment Classification Based on Textblob Library

Table 3 shows the performance of the classification model using the textblob library compared to the policy model that has been developed by Pavan Raj using the Random Forest classifier. Prior to undergoing the classification phase, the labeled data sets underwent a feature extraction process using TF-IDF. The highest accuracy, retrieval and F1-Score were on the textblob model with the Naïve Bayes classifier with a score of 0.93% followed by Bing Liu model with SVM classifier with a score 0f 0.78% and the last were Bing Liu model with RF classifier and baseline model which obtained a score of 0.68% for both. What can be seen from these results, the NB classifier is more likely to show better performance that the SVM classifier for this data set. The overall performance of the model is visualized in Figure 8.

Table 3: Performance for Baseline Model and Textblob Model

Model	Accuracy	Recall	F1-score
Baseline Model +RF	0.68	0.71	0.68
Bing Liu +TF-IDF+RF	0.68	0.70	0.68
Bing Liu +TF-IDF+SVM	0.78	0.71	0.74
Bing Liu +TF-IDF+NB	0.93	0.93	0.93

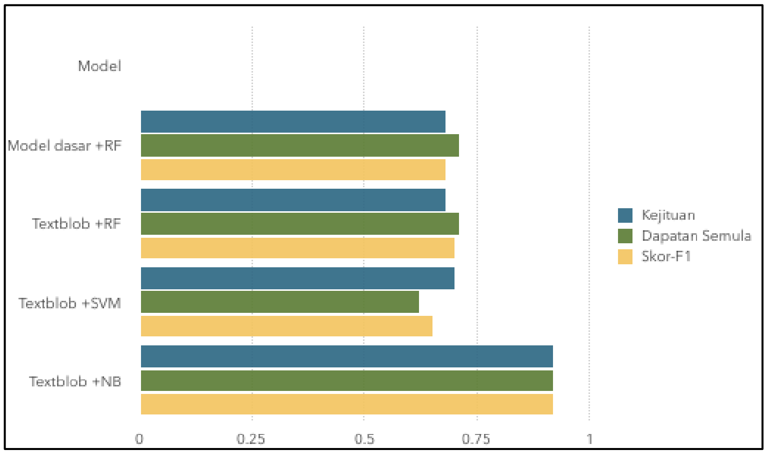

Figure 8: Performance for Baseline Model and Textblob Model

The results showed that the sentiment analysis model developed together with the extraction of the TF-IDF feature and the Naïve Bayes classifier achieved the highest accuracy in both the lexicon approach methods used. The advantage of Naïve Bayes in text classification is that it is easy to use and does not require a lot of training data. Since the data used in this study is only 6,444, it does not require complex algorithms such as Vector Support Machine (SVM) for classification. The nature of naivety allows these algorithms to be easily developed without having to think of complicated iterative classification estimates (Ibrahim and Yusoff, 2018; Ahmad et al., 2020). SVM is an algorithm of high complexity that requires large amounts of testing data and training to produce good classification accuracy. The Random Forest (RF) algorithm is more in favor of attributes that have more levels for variable data and are suitable for use in predictive models.

4.5 Average 'retweets' and 'favorites' by Sentiment

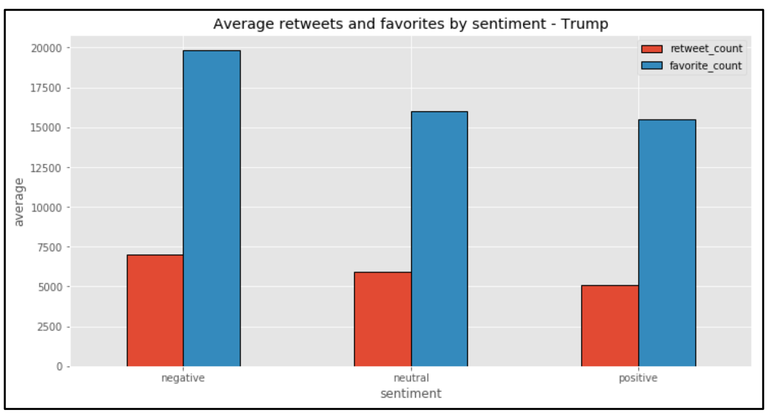

Figure 9: Average Retweets and Favourites by Sentiment

Figure 9 shows the average of 'retweets' and 'favorites' by sentiment where negative sentiments are most favored. As for the number of retweets, negative sentiment was the highest indicating that twitter users prefer to retweet negative sentiment.

5.0 Conclusion

The main problems identified in this study are the lack of labeled data and the need for testing data in the political domain for the purpose of sentiment analysis. In addition, the development of social media has been a major contributor to the amount of large-dimensional unstructured data that complicates the sentiment classification phase.

Three objectives of the study have been stated and this study has met all the objectives of the study and answered the problem statements that arose. Tweets data can be labeled without manual annotations using Bing Liu & Textblob's lexicon of opinions. Feature extraction using TF-IDF helps provide an optimal feature set that results in sentiment classification accuracy using the Naïve Bayes algorithm. The combination of lexicon and machine learning approaches in a supervised manner has successfully developed a high accuracy sentiment analysis model.

Acknowledgement

The authors gratefully acknowledge Universiti Kebangsaan Malaysia, as well as the Industrial Grant Scheme with Perunding Tamadun Teras Pte Ltd for supporting this research project through grant no. TT-2020-008.

References

[1] Ahmad, I. S. et al. (2020) 'Beyond Sentiment Classification : A Novel Approach for Utilizing Social Media Data for Business Intelligence', International Journal of Advanced Computer Science and Applications (IJACSA), 11(3), pp. 437–441.

[2] Ahmad, S. R., Bakar, A. A. and Yaakub, M. R. (2019a) 'A review of feature selection techniques in sentiment analysis', Intelligent Data Analysis, 23(1), pp. 159–189. doi: 10.3233/IDA-173763.

[3] Ahmad, S. R., Bakar, A. A. and Yaakub, M. R. (2019b) 'Ant colony optimization for text feature selection in sentiment analysis', Intelligent Data Analysis, 23(1), pp. 133–158. doi: 10.3233/IDA-173740.

[4] Al-Saffar, A. et al. (2018) 'Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm', PLoS ONE, 13(4), pp. 1–18. doi:10.1371/journal.pone.0194852

[5] Alexander, N. S. and Omar, N. (2017) 'Generating a Malay sentiment lexicon based on wordnet - UKM Journal Article Repository', Journal of Information Technology and Multimedia, 6(1), pp. 126–140. Available at: http://journalarticle.ukm.my/11848/ (Accessed: 16 September 2021).

[6] Awwalu, J., Bakar, A. A. and Yaakub, M. R. (2019) 'Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter', Neural Computing and Applications. Springer London. doi: 10.1007/s00521-019-04248-z.

[7] Blake, A. (2016) Welcome to the next, most negative presidential election of our lives, The Washington Post. Available at: https://www.washingtonpost.com/news/the-fix/wp/2016/07/29/clinton-and-trump-accept-their-nominations-by-telling-you-what-you-should-vote-against/ (Accessed: 15 September 2021).

[8] Center, P. R. (2016) 2016 presidential candidates differ in their use of social media to connect with the public | Pew Research Center, Pew Research Center. Available at: https://www.pewresearch.org/journalism/2016/07/18/candidates-differ-in-their-use-of-social-media-to-connect-with-the-public/ (Accessed: 15 September 2021).

[9] Choi, Y. and Lee, H. (2017) 'Data properties and the performance of sentiment classification for electronic commerce applications', Information Systems Frontiers. Springer New York LLC, 19(5), pp. 993–1012. doi: 10.1007/S10796-017-9741-7.

[10] Ebrahimi, M., Yazdavar, A. H. and Sheth, A. (2017) 'Challenges of Sentiment Analysis for Dynamic Events', IEEE Intelligent Systems. Institute of Electrical and Electronics Engineers Inc., 32(5), pp. 70–75. doi: 10.1109/MIS.2017.3711649.

[11] Fersini, E., Messina, E. and Pozzi, F. A. (2014) 'Sentiment analysis: Bayesian Ensemble Learning', Decision Support Systems. Elsevier, 68, pp. 26–38. doi: 10.1016/J.DSS.2014.10.004.

[12] Hasan, A. et al. (2018) 'Machine Learning-Based Sentiment Analysis for Twitter Accounts', Mathematical and Computational Applications. MDPI AG, 23(1), p. 11. doi: 10.3390/MCA23010011.

[13] Hutto, C. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp. 216–225. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (Accessed: 16 September 2021).

[14] Ibrahim, M. N. M. and Yusoff, M. Z. M. (2018) 'The impact of different training data set on the accuracy of sentiment classification of Naïve Bayes technique', 2017 IEEE Conference on Open Systems, ICOS 2017, 2018-Janua, pp. 17–20. doi: 10.1109/ICOS.2017.8280267.

[15] Joyce, B. and Deng, J. (2018) 'Sentiment analysis of tweets for the 2016 US presidential election', 2017 IEEE MIT Undergraduate Research Technology Conference, URTC 2017. Institute of Electrical and Electronics Engineers Inc., 2018-January, pp. 1–4. doi: 10.1109/URTC.2017.8284176.

[16] Qazi, A. and Goudar, R. H. (2018) 'An Ontology-based Term Weighting Technique for Web Document Categorization', Procedia Computer Science. Elsevier, 133, pp. 75–81. doi: 10.1016/J.PROCS.2018.07.010.

[17] Ranganathan, J. and Bagavathi, A. (2020) 'Action rules for sentiment analysis using Twitter', Int. J. Social Network Mining, 3(1), pp. 35–51.

[18] Shaheen, S. et al. (2014) 'Emotion Recognition from Text Based on Automatically Generated Rules', IEEE International Conference on Data Mining Workshop. doi: 10.1109/ICDMW.2014.80.

[19] Statista.com (2017) Twitter: monthly active users worldwide | Statista, Statista.com. Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (Accessed: 15 September 2021).

[20] Yaakub, M. R., Latiffi, M. I. A. and Zaabar, L. S. (2019) 'A Review on Sentiment Analysis Techniques and Applications', IOP Conference Series: Materials Science and Engineering, 551(1). doi: 10.1088/1757-899X/551/1/012070.