RESEARCH ARTICLE OPEN ACCESS

Hybrid Sentiment Analysis on Political Tweets

Dr. Yin Min Tun*, Dr. Myo Khaing**
*Faculty of Computer Science, UCS(Mdy), Myanmar
yinmintun.ymtkyaw@gmail.com
**Faculty of Computer Science
University of Computer Studies, Mandalay, Myanmar
Myokhaing.ucsy@gmail.com

--************************----------------------------------

Abstract:

On social media, the significantly increased number of online users generates a large amount of unstructured text in the form of posts, chats, blogs, and messages. Moreover, information exchange on social media is really convenient for describing opinions that gain popularity when admired by a large number of online users. This popularity may throw back the people's sentiments towards that organization, place, or person. Twitter generates a large number of texts with political insights that can be extracted for analysing people's opinions and predicting future election trends. The proposed system is implemented with the hybrid political sentiment analysis technique by combining Lexicon-based approach with a machine learning approach. Data from Presidents Obama, Donald Trump, Hillary Clinton, and Joe Biden are collected from Twitter for this system, and Apache Spark is used to analyse this large dataset. To get better performance on the multi-class political sentiment analysis, three supervised learning approaches—Decision Tree, Linear Support Vector Classifier, and Multinomial Naïve Bayes are used. According to the experimental result, the Support Vector Classifier is the best optimal sentiment classifier on a multi-class sentiment analysis system. The overall analysis of the evaluation results shows that the proposed system performs with an accuracy of 94.8% in the environment of Big Data Analytics.

Keywords — Big Data Analytics, Sentiment Analysis, Apache Spark, Machine Learning Approach, Lexicon based approach.

--************************----------------------------------

I. INTRODUCTION

Many researchers are interested in examining the popular political trends on social media. Social media can transform political communication, and society can develop the main platform on which all online users can post and share their comments and opinions based on the political cause. Among political events, elections generate a huge amount of data, which can then be analyzed using sentiment analysis methods. Sentiment analysis can be specified as the treatment of computational opinions and sentiments on the opinion text, and sentiment analysis can be implemented with the various approaches in political science [1].

Through social media, many academics can assess publicly published opinions, evaluate voters, gauge reactions, and analyze trends. Among the various social media networks, features of Twitter are most popular, and these features are a combination of web blogs and social networks [2]. The information can be collected from Twitter. It generates a vast amount of political text, which can be used to analyse the opinions of people and predict the sentiment trend for the elections. Because of the increasing volume of information, useful information can be retrieved from Twitter, and the modeling of complex attributes associated with sentiment words has commanded the use of machine learning approaches as one popular tool for sentiment analysis [3].

This proposed system is implemented to achieve great performance by combining Machine Learning approach with a lexicon-based approach for sentiment analysis. This framework was used to implement distributed and parallel system execution with batch data processing on various data notes. For the analysing step with a massive amount of tweet data, this system was built with Apache Spark from the Big Data Analytic frameworks. To achieve its great accuracy, the proposed system uses three machine learning techniques. In the experimental results, the accuracy of three classifiers—Decision Tree, Linear Support Vector Classifier, and Multinomial Naïve Bayes are evaluated for the sentiment classification.

II. RELATED WORK

In the previous researches on the sentiment analysis system, there has been developing the vast amount of related data through the online users. There are various previous related research papers of the sentiment analysis and then the reviews of these papers are discussed in this section.

M.Moh and et al. proposed the sentiment analysis system on the architecture of multi-tier classification: the first model for the classification for three classes and the last two models for the binary classifier. To get the greater performance, related features were selected by using the different techniques of feature selection. The posted movie reviews (150,000 reviews) on the social media were utilized for the training and testing of system's performance. In this system, four classifiers (Random Forest, Support Vector Machine, SGD and Naïve Bayes) are applied for the evaluation of experimental results. The results proved that the proposed architecture's performance was actually improved the accuracy of prediction on the single-tier model by over 10%. Their proposed sentiment analysis system was developed on the traditional analytics platform and classified the movies reviews into the specified multi-class by using supervised classifiers of Machine Learning approaches [4].

Alaouil and et al. implemented sentiment analysis system based on the three approaches, which were constructions of dynamic dictionary using the polarity of words on the given topic relied on the selected hashtags. In their system, tweets data from the 2016 US elections were defined to the negative class and positive class. According to their experimental results on traditional analytics approach, the great accuracy rate on the performance of prediction was reached but they couldn't support the accuracy improvement on the platform of Big Data Analytics [5].

Yaquba and et al. presented the political sentiment analysis system to discourse from Twitter that discovered the presented sentiment words and topics to achieve the better opinions for the election events. However, there were some limitations as they applied tagging tool the sentiment words in their system [6]. In addition, Singh and et al. proposed the political sentiment analysis system for the outcomes prediction of US presidential election event. In this system, the raw data from Twitter are collected through the social media within the limited duration. They created the classification model in the WEKA 3.8 for the sentiment words classification (negative words or positive words) with the help of supervised learning approach (Support Vector Machine classifier) for the prediction of the political sentiment analysis system [7].

Ansaria and et al. created the supervised Machine Learning approach-based sentiment analysis system for the India General Election event using the political sentiment dataset from Twitter. The author used Long Short-Term Memory (LSTM) classifier for the sentiment classification model and then evaluated the analysis experiments on the other Machine Learning classifiers. They also did not implement on the Big Data Analytic architecture for the classification of multi-class [8]. Hasan and et al. proposed the hybrid sentiment analysis technique that was the combination of Machine learning approaches and sentiment analyzer on political data with the two Machine Learning classifiers (Naïve Bayes and Support Vector Machine). In their proposed system, the researchers analyzed the performance on their applied two classifiers but they also didn't develop for classification on multi-class [9]. Moreover, Baltas and et al. implemented the sentiment analysis tool for analysis of microblogging messages using their specified sentiment words. In their sentiment analysis system, three classifiers (Logistic, Decision Tree and Naïve Bayes) were applied for the output classification using MLib. According to the experiment, the accuracy of Naïve Bayes classifiers reached the higher performance result than the other classifiers but they didn't consider the sentiment analysis system on the multi-class [10].

After a review of previous research on the sentiment analysis system, a hybrid approach for the sentiment analysis system for multi-class using political data from Twitter for the US election event is proposed. The multi-class political sentiment analysis system is built on Apache Spark using Big Data Analytics frameworks. The proposed system achieved a better performance with the effective support of the Apache Spark framework.

III. METHODOLOGY OF THE PROPOSED SYSTEM

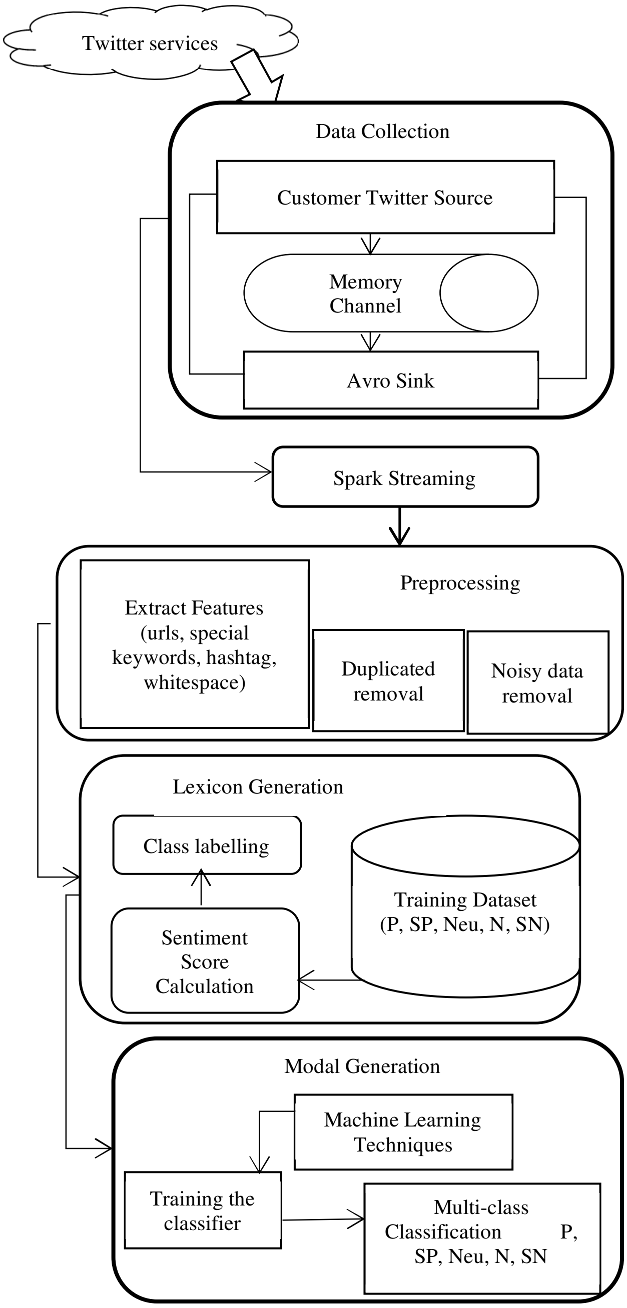

Figure. 1 Process Flow Diagram for Multi-Class Classification

The political tweets data is the association activities to make decisions for the groups. To develop effective extraction of useful related political information, the proposed system implementation has four steps: related data collection, tweet data preprocessing, class labelling, and sentiment word classification. The flow diagram of the proposed system is presented in Figure 1.

A. *Data Collection*

In this proposed political sentiment analysis system, political stream data from Twitter is collected by Apache Flume with the batch stream configuration, and then these collected data are assigned to an HDFS sink. English tweets are filtered and collected using keywords. Twitter data and Apache Flume are the main sources for data collection.

B. *Data Preprocessing*

These collected raw data contain useless information and duplicates; they must be cleaned and preprocessed before being used in an analysis system. The purpose of data preprocessing is for tweet text feature selection, duplicate and noise data removal, stop words, and repeated characters. In this step, the task of classification is simplified, and the cost of processing is decreased for the training stage.

C. *Class Labelling*

A generated political lexicon is used for labelling the training data of the learning-based classifier. It uses additional (non-lexical) linguistic information and rules to detect sentiment strength in short, informal English texts. For each political text, lexicon outputs a positive sentiment score from 1 to 5 and a negative score from -1 to -5. For multiclass classification, based on the values of the positive and negative sentiment scores, the tweets are classified as Positive, Negative, Strong Positive, Strong Negative and Neutral.

Procedure: Class Labeling

```
    Input: raw tweets
    Output: tweets, class label
    1.  Begin
    2.  Calculate total polarity strength on each sentence by
        applying generated political lexicon with
        SentiStrength_ Data
    3.  If (score > 1) then print “strongly positive”
    4.  Else if (score > 0 && score <= 1) then print
        “positive”
    5.  Else if (score == 0) then print “neutral”
    6.  Else if (score>=-1 && score< 0) then print “negative”
    7.  Else if (score < -1) then print “strongly negative”
    8.  emit (tweets, class label)
    9.  End
```

- If (sem-score>1), the output label => “Strong Positive”
- If (sem-score > 0 && sem-score <= 1), the output is “Positive”
- If (sem-score<-1), the output label => “Strong Negative”
- If (sem-score > 0 && sem-score <= -1), the output is “Negative”
- If (sem-score==0), the output label => “Neutral”

D. *Sentiment Classification Model with Machine Learning Techniques*

At the final step, the classification model is used for the class labels and feature vectors. The best model with the highest accuracy is chosen from among the other developed classification models in this system. In the training and testing phrase, three different classification methods (Multinomial Naïve Bayes, Decision Tree and Linear SVC) are implemented.

1) Multinomial Naïve Bayes (MNB): The Multinomial Naive Bayes classification is a Bayesian learning approach from Machine Learning and an algorithm for probabilistic learning method that is mostly applied in Natural Language Processing (NLP). MNB evaluates the probability of each tag for a given sample and then returns the tag with the highest probability as the output. Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. It's straightforward and can be used to forecast real-time applications. It's very scalable and can handle enormous datasets with ease.

Maximum likelihood

$$\theta_{yi} = \frac{N_{yi}+\alpha}{N_y+\alpha n} \qquad (1)$$

Posterior Probability

$$P(y|x_i) = P(y)\prod P(x_i|y) \qquad (2)$$

where $\theta_{yi}$is the probability $P(x_i|y)$ of feature i appearing in a sample belonging to class y.

2) Decision Tree: Decision Tree algorithm is one of the most popular classifier in Machine Learning approaches that recursively divide the recorded dataset with depth first greedy and breadth first approaches until all the data items classified to a specific particular class. Its structure is created with root, internal nodes and leaf nodes as a tree. This structure is applied for the recorded unknown data classification. At each internal node, best split decision is performed with impurity measure. And the leaves of tree are defined the class labels for grouping the data items. In this proposed system, Decision Tree is used to get the great opinion in the classification of sentiment tweets data. Decision Tree based classification model, also known as statistical classifier is an approach for classification of data.

$$Entropy(S) = \sum_{i=1}^{n} -P_i \times \log_2 P_i \qquad (3)$$

$$Gain\ (S, A) = Entropy(S) \sum_{i=1}^{n} \frac{|s_i|}{|s|} \times Entropy(S)(4)$$

Where, P is class proportion for output, S is a set of case, A is a case attribute, $|s_i|$ is a count of cases to i, and $|s|$ is count of cases in the set.

3) Linear SVC: The Linear Support Vector Classifier (SVC) uses a linear kernel function for classification and it develops well using the vast amount of samples. Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function. The kernel method cannot be transformed in linear SVC, because it is based on the kernel linear method. Linear SVC creates the binary classification model using the Linear SVM classification model. The hyperplane of SVC generates the good separation that achieves the highest distance for the nearest points of training data on any class. Hinge Loss is optimized with the help of OWLQL optimizer.

$$L(W; x, y) = \max(0, 1 - yW^T x) \qquad (5)$$

$$f(w) = \tau R(w) + \frac{1}{n} \sum_{i=1}^{n} l(W; x_i, y_i) \qquad (6)$$

Where, $L(W; x, y)$ is loss function for Linear SVC, $R(w)$ is raw tweets, W^T is word vector, y is sentence label and x is emotional words for each sentence.

IV. EXPERIMENTAL RESULTS

The evaluation performance of the three applied classification techniques is compared based on the experimental results to determine the best classifier for the political sentiment analysis system on multi-class. The unstructured Trump and Clinton datasets (containing 30000 tweets for each person) and the unstructured Obama and Joe Biden datasets (containing 30000 tweets for each person) for the American presidential election are used in this system. From each dataset, 80% of the total dataset is used for training and 20% of the total dataset is used for testing. The popular performance measures (accuracy, recall, F-measure, percentage, and precision of tweets) are evaluated for this proposed system analysis.

As the result of analysis on each dataset, Strong Positive, Positive, Neutral, Negative, and Strong Negative are calculated percentages of lexicon-based classification and manual classification in the Clinton, Trump, Joe Biden, and Obama datasets, which are not much different. The overall error rate is less than 1%. As a result, the proposed system performance is achieved with the expected high accuracy.

Table 1. Performance Comparisons of Three Classifiers on Clinton Dataset

	Clinton		
Machine Learning Algorithm	Precision	Recall	F-measure
Multinomial Naive Bayes	0.92	1	0.87
Linear SVC	0.92	1	0.94
Decision tree	0.89	1	0.94

The performance comparisons of three classifiers on Clinton, Trump, Joe Biden, and Obama dataset are illustrated in Table 1, 2, 3 and 4.

Table 2. Performance Comparisons of Three Classifiers on Trump Dataset

	Trump		
Machine Learning Algorithm	Precision	Recall	F-measure
Multinomial Naive Bayes	0.83	0.92	0.89
Linear SVC	0.91	1	0.83
Decision tree	0.87	0.89	0.85

Table 3. Performance Comparisons of Three Classifiers on Joe Biden Dataset

	Joe Biden		
Machine Learning Algorithm	Precision	Recall	F-measure
Multinomial Naive Bayes	0.92	1	0.87
Linear SVC	0.92	1	0.94
Decision tree	0.89	1	0.94

Table 4. Performance Comparisons of Three Classifiers on Obama Dataset

	Obama		
Machine Learning Algorithm	Precision	Recall	F-measure
Multinomial Naive Bayes	0.83	0.92	0.89
Linear SVC	0.91	1	0.83
Decision tree	0.87	0.89	0.85

Table 5. Overall Accuracy with Proposed Political Lexicon for Clinton, Trump, Joe Biden, and Obama Dataset

Dataset	Multinomial Naïve Bayes	Linear SVC	Decision Tree
Trump	84.2	89.6	78.3
Clinton	89.6	90.5	80.3
Obama	75.7	85.8	73.5
Joe Biden	89.5	94.8	71.8

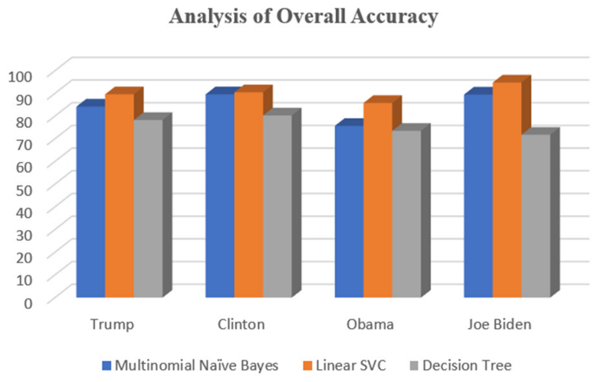

Figure 2. Performance Analysis of multi-class Classifiers with Different Dataset

The comparison of accuracy between each method and the proposed lexicon-based classification on the Clinton, Trump, Joe Biden, and Obama datasets is illustrated in Table 5. The experimental results show that the precision, recall, F-measure, and accuracy of Linear SVC achieve better accuracy than Naïve Bayes and Decision Tree on Clinton, Trump, Joe Biden, and Obama tweets in multi-class sentiment analysis system. According to the experimental results, Linear SVC with one VsRest approach classifier and the proposed political lexicon-based classification model achieved greater accuracy than the other two developed classifiers. The Linear SVC classification model is the best classifier in the evaluation of this system.

V. CONCLUSION

In this proposed political sentiment analysis system, the hybrid approach is applied to achieve the greater performance system for the multi-class classification on the architecture of Machine Learning Model that is presented for useful information extraction from the vast volume of social political data. This approach is created on the Apache Spark framework from the Big Data Analytic platform for the high analysing velocity and vast number of tweets using effective manner. To get the greatest performance sentiment analysis system, the hybrid approach is proposed by using the combination of lexicon-based and Machine Learning based classification model (Decision Tree, Linear SVC, and Multinomial Naïve Bayes) based on the architecture of Machine Learning Model. The evaluation results of these three methods are compared and analysed. According to the experiment on the classifiers, the accuracy of the Linear SVC classifier achieved the higher performance than the other two applied methods. For the future work of this proposed system, the political based sentiment analysis system implementation will be developed for real time streaming data.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, ‘‘Thumbs up?: Sentiment classification using machine learning techniques,’’ in Proc. ACL Conf. Empirical Methods Natural Lang. Process. (EMNLP), vol. 10, 2002, pp. 79–86.

[2] “Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 [online] Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [Accessed 4-April-2019]

[3] A.S. Khattak, R. Batool, F.A.Satti and et al. “Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation”, Hindawi, Complexity, Volute 2020, A rticle ID 8892552, 11 pages, 2020.

[4] M.Moh, A.Gajjala, S.C.R.Gangireddy, T.S.Moh, “On Multi-Tier Sentiment Analysis using Supervised Machine Learning”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2015.

[5] I. E. Alaoui1, Y. Gahi, R. Messoussi, Y. Chaabi1, A. Todoskoff, and A. Kobi, “A novel adaptable approach for sentiment analysis on big social data”, Journal of Big Data, 2018.

[6] U. Yaquba, S. A. Chunb, V. Atluria, and J. Vaidyaa, “Analysis of political discourse on twitter in the context of the 2016 US presidential elections”, Government Information Quarterly (2017), 2017.

[7] P. Singh, R. S. Sawhney, and K. S. Kahlon, “Forecasting the 2016 US Presidential Elections using Sentiment Analysis”, 16th Conference on e-Business, e-Services and e-Society(I3E), pp.412-423, 2017.

[8] M. Z. Ansaria, M. B. Aziza, M. O. Siddiquib, H. Mehraa, and K. P. Singha, “Analysis of Political Sentiment Orientations on Twitter”, International Conference on Computational Intelligence and Data Science (ICCIDS 2019), pp.1821-1828, 2019.

[9] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, “Machine Learning-Based Sentiment Analysis for Twitter Accounts”, Mathematical and Computational Applications, Volume 23, Issue 1 (March 2018), 2018. https://doi.org/10.3390/mca23010011.

[10] A. Baltas, A. Kanavos, A. K. Tsakalidis, “An Apache Spark Implementation for SA on Twitter Data”, Algorithmic Aspects of Cloud Computing, pp.15-25, April 2017.