Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception

Mohd Naim Mohd Ibrahim
College of Information Technology
Universiti Tenaga Nasional
Putrajaya, Malaysia
naim.ibrahim@gmail.com

Mohd Zaliman Mohd Yusoff
College of Information Technology
Universiti Tenaga Nasional
Putrajaya, Malaysia
zaliman@uniten.edu.my

***Abstract*— This paper presents strategy to classify tweets sentiment using Naïve Bayes techniques based on trainers' perception into three categories; positive, negative or neutral. 50 tweets of 'Malaysia' and 'Maybank' keywords were selected from Twitter for perception training. In this study, there were 27 trainers participated. Each trainer was asked to classify the sentiment of 25 tweets of each keyword. Results from the classification training was then be used as the input for Naïve Bayes training for the remaining 25 tweets. The trainers were then asked to validate the results of sentiment classification by the Naïve Bayes technique. The accuracy of this study is 90% ± 14% measured by total number of correct per total classified tweets.**

Keywords—Naïve Bayes, Twitter Sentiment Classification, Supervised Learning.

I. INTRODUCTION

Social media is new way of communication among people all over the world. People express their opinion on social media such as Facebook, Twitter, Instagram and WhatsApp. The advancement of computing power, internet speed and mobile devices is main contributor for this phenomenon. Issues that going viral will stir limitless public expression of opinion, this will be good or bad to the affected side. In the last decade, there is increase the interest to extract sentiment from the text.

Since explosion of the social media, the interest in sentiment analysis is spur into analyzing the sentiment that being express by the netizens in order to gauge the public perception in the issues. Many government, corporation, NGO, politician, celebrities and individual would like to know what public opinion towards things that could affect their interest.

As other aspect of Natural Language Processing (NLP) research, there is quite a significant challenge to train computer to handle data that are easily differentiated by human. The step that must undergo to make the opinion can be analyzed by automated system is it need to be quantified, turning the fuzzy of emotion into something that are measurable, then it can be calculated and classified.

II. RELATED WORKS

A. Twitter Sentiment Classification

Numerous study has been done in determine and classify sentiment of tweets in Twitter. Both supervised and unsupervised technique are used. Supervised technique such as Nearest Neighbor, Naïve Bayes, TF-IDF and Support Vector Machine (SVM) as proven can give good results. Study by Horn [3] shown that SVM can give 80% accuracy, Pablo and Marcos [4] shown that Naïve Bayes could also give 80% accuracy, Abdul Wahab et al. [5] demonstrate how TF-IDF archive 85.71% accuracy. While Marco and Ana-Maria [6] is using Gradient Boosted Decision Tress (GBDT) to classify political orientation, ethnicity of tweet user and business fan detection.

B. Naïve Bayes for Classification

The Naïve Bayes has been developed by C.T Yu and G. Salton [1] and also S. Roberson and K. Spark [2] in the 1970 respectively. The first model in probabilistic information retrieval is Binary Independence Model.

In Naïve Bayes, each document can represented by a vector $\underset{y}{\rightarrow} = (y_1, \ldots y_m)$ where $y_t = 1$ if a certain term *t* is occur in the document and $y_t = 0$ if not. Not dependent means that the model assumes that the terms are not related with each other. Using the Bayes rule, we can calculate the probability of relevance of a certain document. It achieves good results in practice although seems quite simple. The Binary Independent Model and the Naive Bayes classifier are closely related. The Naive-Bayes classifier assumes that the features are not associated with each other in general. To help to identify an object a feature can be defined as an attribute. For example, the features of a motorcycle are, amongst others, 'vehicle', '2 tires', 'engine', 'moves on ground'. But, the Naive Bayes classifier will assume that none of those features related to each other. This is a rather simplifying assumption (that come the name Naive), we can say that this algorithm performs remarkably well in practice. It is worth also to mention that the more state-of-the-art classification algorithm such as Okapi BM25 is also based on the Binary Independence Model, but use more advance weighting scheme [8].

978-1-4673-9437-6/15/$31.00 ©2015 IEEE

III. Proposed Sentiment Classification Based on Trainer Perception

The proposed sentiment classification based on trainer perception as shown in Figure 1 and describe as follow

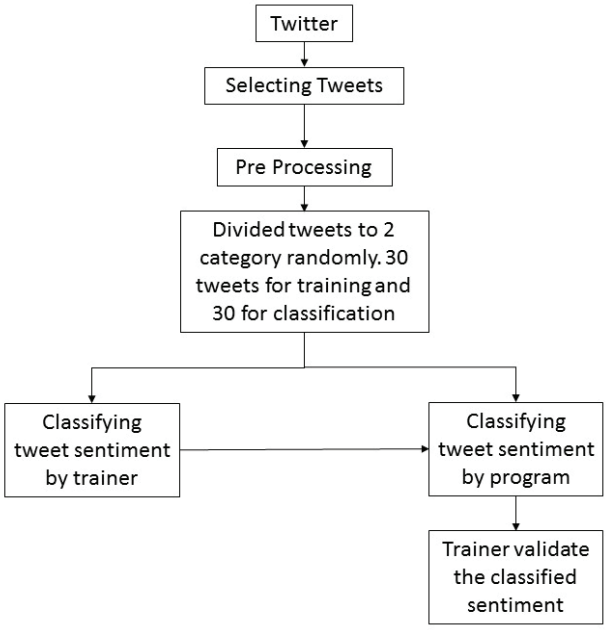

Figure 1. Overall process

A. Collecting Tweets From Twitter

The tweets which consist keyword 'Maybank' and 'Malaysia' is being grab from twitter stream using Python Twitter library named Twython. All the tweets are stored in database using MongoDB.

B. Selecting Tweets

50 of most relevant tweets for each 'Maybank' and 'Malaysia' keyword is being selected for this study. These two keyword is chosen for demonstrating ability of Naïve Bayes for sentiment classification in political and business environment. Most of the tweets are mixed language between Malay and English sentences / word. This is nature of current social media trend among netizens to mixed up languages.

C. Preprocessing

The preprocessing stages is where all the tweets get clean up. The process as shown below:

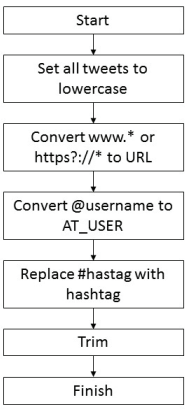

Figure 2. Preprocessing

D. Divided tweets

The tweets are being divided to two parts, one parts which consist 25 tweets for training purpose and another part which consist 25 tweets for classification using Naïve Bayes. The tweets that are selected for each part is randomly picked.

E. Classifying By Trainer

The 25 tweets are shown to the trainer for them to train. The training is done by trainer to select sentiment whether Positive, Negative or Neutral that the trainer think/feel for each tweet. The selection interface is shown in screen shot below. Then the trainer need to click 'Train' button to save the training. These training data will be used in next step for classification of the tweets sentiment.

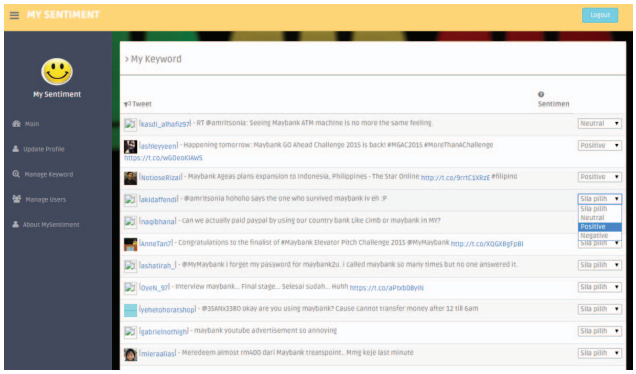

Figure 3. Screen shoot for training session

F. Classifying By Naïve Bayes

The remaining 25 tweets will be classified their sentiment using Naïve Bayes algorithm. The training data is being used as features to the Naïve Bayes to calculate the probability of remaining tweets for classification purpose. The automatic classification is being done by utilizing Python NLTK Library.

G. Sentiment Classification Validation

The classified tweets then being shown to the trainer to validate. The trainer than validates whether the

classification of the tweets is right or wrong based on their perception on the tweets sentiment. The feedback is being used to calculate the accuracy of Naïve Bayes in our study.

IV. Results

Total of 116 trainers have participate but only 27 managed to complete all the required task in this study. The rest either not completed training or validating the sentiment for both 'Maybank' and 'Malaysia' tweets. The result as shown in Figure 4.

TRAINER NAME	Total trained for Maybank	Total correct for Maybank	Maybank accuracy	Total trained for Malaysia	Total correct for Malaysia	Malaysia Accuracy
HAZIM ROZAMAN	25	25	1.0000	25	25	1.0000
TAJUL AZHAR	25	25	1.0000	25	25	1.0000
MOHD SHAFARUDIN	25	25	1.0000	25	23	0.9200
MOHAMAD ARQAM	25	25	1.0000	25	25	1.0000
SYAHMAN MOHAMAD	25	25	1.0000	25	25	1.0000
ISMAIL CHE ANI	25	25	1.0000	25	15	0.6000
ZAZA	25	25	1.0000	25	25	1.0000
AKMAL AHMAT	25	25	1.0000	25	25	1.0000
AZRUL	25	18	0.7200	25	19	0.7600
AMINAH	25	21	0.8400	25	25	1.0000
SUHAILAN	25	15	0.6000	25	22	0.8800
LINDA	25	19	0.7600	25	19	0.7600
RAY	25	19	0.7600	25	23	0.9200
HABIBURRAHMAN	25	16	0.6400	25	20	0.8000
SHUKOR SANIM	25	25	1.0000	25	25	1.0000
SING	25	25	1.0000	25	22	0.8800
NORITA	25	25	1.0000	25	25	1.0000
SANDRA	25	19	0.7600	25	15	0.6000
JUSTIN R. D.	25	11	0.4400	25	19	0.7600
HAKIM	25	25	1.0000	25	25	1.0000
EDMUND	25	21	0.8400	25	25	1.0000
NORSAN BINTI SAMSUL	25	25	1.0000	48	2	1.0000
ZURAIN	25	25	1.0000	49	1	1.0000
HENNI JUMITA	25	25	1.0000	25	25	1.0000
FAZREENA IDAYU	25	25	1.0000	25	25	1.0000
ARNOLD	25	15	0.6000	25	22	0.8800
	Average accuracy for Maybank		0.8831	Average accuracy for Malaysia		0.9138
	Std Deviation		0.1653	Std Deviation		0.1234

Average Accuracy	0.8985
Std Deviation	0.1443

Figure 4. Result the accuracy of sentiment classification

From the result, we discover that our classification of tweets using Naïve Bayes could give us 90% of accuracy with standard deviation of 14%. The total accuracy is calculated by averaging the average accuracy of 'Maybank' and 'Malaysia' sentiment.

V. Conclusion

From our study, we learned that by training and verifying the sentiment classification by the same person, we could archive a high degree of accuracy using Naïve Bayes technique. This method is suitable to train and classify sentiment from twitter and other social network data. This method also is a good candidate to assist human / operator to classify a large number of tweets. We also learn that this technique is suitable to political or business sentiment classification. From our results also could negate some perception that Naïve Bayes is weak compare to SVM [4].

VI. References

[1] C. Christopher, R. Prabhakar and S. Hinrich, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.

[2] C. Horn, "Analysis and Classification," Graz University of Technology, Graz, Austria, 2010.

[3] G. Pablo and G. Marcos, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014.

[4] R. d. Groot, "Data Mining for Tweet Sentiment Classification," Utrecht University, Utrecht, Netherlands, 2012.

[5] A. W. Muhammad Faheem Khan, A. K. and A. K. , "IMPORTANT FEATURES SELECTION DURING SENTIMENT," *Sci.Int(Lahore),* vol. 26, no. 2, pp. 959-964, 2014.

[6] P. Marco and A.-M. Popescu, "A Machine Learning Approach to Twitter User Classification," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

[7] C. Yu and G. Salton, "Precision Weighting—An Effective Automatic Indexing Method," *Journal of the ACM (JACM),* vol. 23, no. 1, pp. 76-88, 1976.

[8] S. J. K. and R. S. E., "Relevance weighting of search," *Journal of the American Society for Information Science,* vol. 27, no. 3, pp. 129-146, 1976.