Analyzing Sentiment Level of Social Media Data Based on SVM and Naïve Bayes Algorithms

Hsu Wai Naing[1] and Phyu Thwe[2]

Aye Chan Mon[1] and Naw Naw[2]

[1]Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar
[2]Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar

hsuwainaing2054@gmail.com
pthwe19@gmail.com
polestar.mon20@gmail.com
nawnaw1986@gmail.com

Abstract. –Social media is a popular network through which users can share their reviews about various topics, news, products etc. People use internet to access or update reviews so it is necessary to express opinion. Twitter is a hugely valuable resource from which insights can be extracted by using text mining tools like sentiment analysis. Sentiment analysis is the task of identifying opinion from reviews. The system performs classification by combining Naïve Bayes (NB) and Support Vector Machine (SVM). The system is intended to measure the impact of ASEAN citizens' social media based on their usage behavior. The system is developed for analyzing National Educational Rate, Business Rate and Crime Rate occurred in Malaysia, Singapore, Vietnam and our country, Myanmar. The system compares the performance of these two classifiersin accuracy, precision and recall.

Keywords: Opinion Mining, Sentiment Analysis, Twitter, Support Vector Machine (SVM), Naïve Byes (NB), Text Classification.

1 Introduction

At the present time, millions of people are using social network sites like emotions, opinion and share Twitter, Facebook, etc, to express their emotions, opinion and share views about daily lives. People mostly depend upon user generated content over online to a great extent for decision making. The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used.

Sentiment analysis is a task to recognize writers' feelings as expressed in positive, negative and neutral comments. Machine learning sentiment analysis usually comes under supervised classification and under text classification techniques in specific.

Twitter is a social networking website which allows users to publish short messages that are visible to other users. These messages are known as tweets, and can only be 140 characters or less in length [1]. The system is developed to analyze Educational Rate, Business Rate and Crime Rate occurred in Malaysia, Singapore, Vietnam and our country, Myanmar through tweets. In daily life, many users share their opinions and experiences on social media. In this paper, at first, the system crawls the real time social media data from twitter. And then, the system uses Naïve Bayes and Support Vector Machine for the classifying tasks in Sentiment. After that, the system displays the sentiment scores by using visualization techniques. The performance of classification is also analyzed using precision, recall and accuracy.

This paper is organized as follow; second section gives related work; third section that describes preprocessing stage. Fifth section gives about the proposed feature extraction and classification processes followed by the sixth section that shows the experimental results of sentiment analysis about Education, Business and Crime.

2 Related Work

Mori Rimon [2] used the keyword based approach to classify sentiment. He worked on identifying keywords basically adjectives which indicates the sentiment. Such indicators can be prepared manually or derived from Wordnet.

Janice M. Weibe[3] performed document and sentence level classification. He fetched review data from different product destinations such as automobiles, banks, movies and travel. He classified words into positive and negative categories. He then calculated the overall positive or negative score for the text. If the number of positive words is more than negative then the document is considered positive otherwise negative.

3 System Design Overview

Firstly, the system fetches the real time social media data about education, business and crime from Twitter. And then, the system implements sentiment analysis on these collected data. The input of the system includes Education, Business and Crime tweets. The twitter data cannot classify directly because it consists of noisy information. So, this noisy information is removed by pre-processing. After that, the system uses Supervised Machine Learning Algorithm (Naïve Bayes and Support Vector Machine) that can achieve competitive accuracy when it is trained using feature. The main task of this system is to perform social media sentiment analysis by applying machine learning approach of Artificial Intelligence (AI). And then, this system can compare the rate of change of Crime Sector, Business and Education Sector occurred in Malaysia, Singapore, Vietnam and our country, Myanmar. In the system design, there are three main components. They are pre-processing, feature extraction and classification. Figure1 illustrates the overall system design.

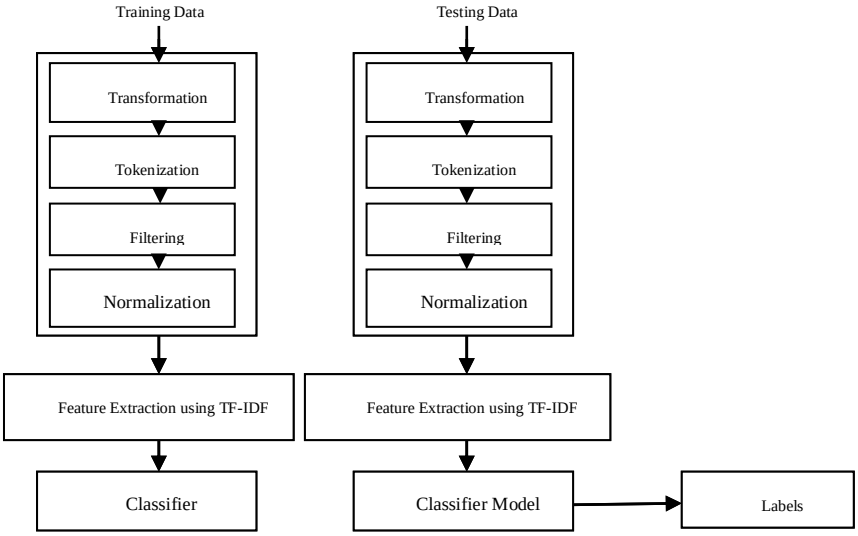

Fig. 1. The System Design

At first, the system crawls tweets about Education, Business and Crime from Twitter. The language is as English using Twitter Streaming API. The extracted twitter data is needed to preprocess. In the pre-processing stage, transformation, tokenization, filtering and normalization are performed. And then, the system extracts meaningful features by using Term Frequency-Inverse Document Frequency (TF-IDF). Feature extraction can make the classifier more effective by reducing the amount of data to be analyzed to identify the relevant features for further processing. After that, the system selects features as the input features of classifier (Naïve Bayes and Support Vector Machine). Finally, the system provides the percentage score of Education, Business and Crime sectors and displays according to their scores by using visualization techniques. The system also compared the performance of these two classifiers in accuracy, precision and recall.

4 Pre-Preprocessing Stage

In preprocessing, the extracted data is cleaned and made ready for feeding it into the classifier. In this stage, there are four main processes:

4.1 Transformation

In Transformation stage, a clean tweet should not contain URLs, hashtags (i.e, #studying) or mentions (i.e. @Irene). The input tweets are transformed to lowercase. URLs are replaced with generic word URL. Then, @username is replaced with generic word URL. Then, @username is replaced with generic word AT_USER. Then, #hashtag is replaced with the exact same word without the hash. After that, all the punctuations are removed at the start and ending of the tweets. Additional whitespaces are replaced with a single whitespace. The next operation is to remove the vowels repeated in sequence at least three times [4]. Tweets about Education, Business and Crime from Myanmar, Singapore, Vietnam and Malaysia are preprocessed in the Transformation step.

4.2 Tokenization

Tokenization may be defined as the process of splitting the text into smaller parts called tokens, and is considered a crucial step in NLP [5]. The system tokenizes the uniformed sentence which got into smaller components (unigram).

4.3 Filtering

One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words word (such as “the”, “a”, “an”, “in”). Most search engines ignore these words because they are so common that including them would greatly increase the size of the index without improving precision or recall [6]. The system tokenizes the output tweets from Transformation step and then removes stopwords.

4.4 Normalization

In preprocessing step, Normalization process performs the important step. Normalization is a process that transforms a list of words to a more uniform sequence. By transforming the words to a standard format, the system leads to a more accurate classification. In the normalization step, lemmatization is applied. Lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence

[7]. After the lemmatization, the root words are got and they are used for feature extraction step.

5 Feature Extraction

Transforming the input into the set of features is called Feature Extraction. If the extracted features are correctly chosen, it is expected that the features set will perform the desired task using the reduced representation instead of the full size input. In this system, Naïve Bayes Classifier and Support Vector Machine Classifier are trained on tf-idf weighted word frequency features. Term Frequency-Inverse Document Frequency (tf-idf) is a popular feature extraction method which reflects the relevance of a word in a particular document among the corpus. After Feature Extraction step with TF-IDF, the system selects features as the input features of classification. In this way, the system can get the essential features for the system and perform the best accuracy.

6 Classification

Sentiment Analysis is a current research area in text mining. It is the stem of natural language processing or machine learning methods. There are two sentiment analysis techniques such as unsupervised and supervised techniques. In unsupervised technique, classification is done by a function which compares the features of a given text against discriminatory-word lexicons whose polarity are determined prior to their use. In supervised technique, the main task is to build a classifier. The classifier needs training examples which can be labeled manually or obtained from a user-generated user-labeled online source. The system performs Sentiment Analysis by using supervised technique, Naïve Bayes Classifier and Support Vector Machine Classifier.

6.1 Naïve Bayes Classifier

Naïve Bays Classifier is one of Supervised Machine Learning Algorithm. Naïve Bayes Classifier can predict whether a new text message can be categorized as positive or negative or neutral. It is used to predict the probability for a given word to belong to a particular class. Pre-processed and Feature Extraction data is given as input to train input set using Naïve Bayes Classifier. That trained model is applied on test to generate positive or negative or neutral of Education, Business and Crime. First, Naïve Bayes Classifier computes the prior probability. Second, Naïve Bayes Classifier computes the conditional probability/Likelihood of each word attribute. Third, Naïve Bayes Classifier computes the posterior probability. Finally, Naïve Bayes Classifier determines the class.

6.2 Support Vector Machine Classifier

In machine learning, Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training data, each marked as belonging to one or the other, an SVM training algorithm builds a model that assigns new data to one category or the other. An SVM model is a representation of the data as points in space, mapped so that the features of the separate categories are divided by a clear gap that is as wide as possible. New testing data are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The goal of a Support Vector Machine (SVM) classifier is to find a linear hyper plane (decision boundary) that separates the data in such a way that the margin is maximized**.**

6.3 Accuracy, Precision and Recall

Accuracy is not the only metric for evaluating the effectiveness of a classifier. There are two other useful metrics (precision and recall). They can provide much greater insight into the performance characteristics of a binary classifier. Precision measures the exactness of a classifier. Recall measures the completeness of a classifier [8].

The system computes the accuracy of Support Vector Machine Classifier and Naïve Bayes Classifier. It is calculated by number of correctly selected positive, negative

and neutral words divided by total number of words present in the corpus. The system measures precision and recall of Support Vector Machine Classifier and Naïve Bayes Classifier by using NLTK metrics module.

7 Experimental Results

Sample tweet messages about education, business and crime are extracted from a particular Twitter account after getting prior permission. The extracted training dataset has 7540 tweets about education, 5414 tweets about crime and 8000 tweets about business. The extracted tweets are preprocessed such as transformation, tokenization, filtering, lemmatization. And then, the output words need to be meaningful features for the system.

The output feature words are input features of Naïve Bayes Classifier and Support Vector Machine Classifier. The system uses real time testing data on twitter. The testing datasets used in the application were retrieved from twitter using twitter4j API's.

Table 1. Experimental Results Using Support Vector Machine and Naïve Bayes

Location	Positive Percentage	Negative Percentage	Neutral Percentage	Total Percentage	SearchKey
Myanmar	15	15	69	99	Education
Malaysia	27	18	53	98	Education
Singapore	52	17	29	98	Education
Vietnam	34	35	27	96	Education
Myanmar	33	22	44	99	Crime
Malaysia	31	24	43	98	Crime
Singapore	35	20	44	99	Crime
Vietnam	24	35	33	92	Crime
Myanmar	45	35	10	90	Business
Malaysia	45	25	15	85	Business
Singapore	46	25	26	97	Business
Vietnam	45	15	30	90	Business

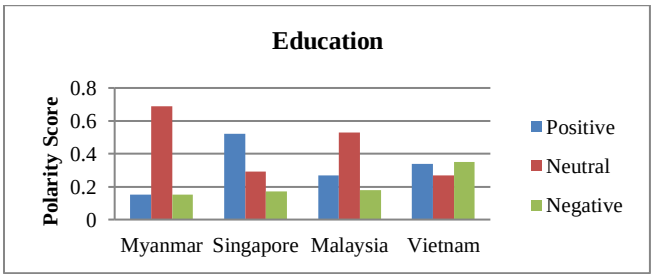

Fig. 2. Graphical Analysis of Education

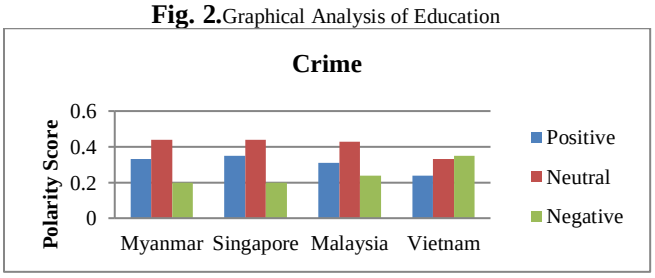

Fig. 3. Graphical Analysis of Crime

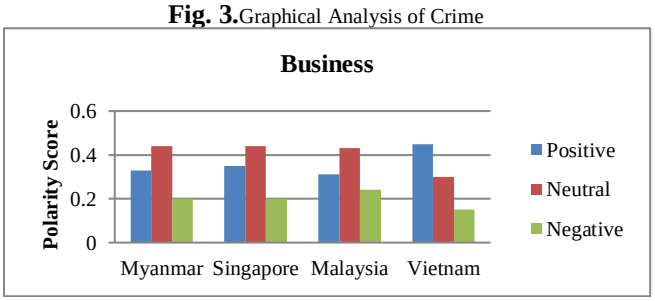

Fig. 4. Graphical Analysis of Business

7.1 Performance Comparison

The system is intended to perform the accuracy comparison of Support Vector Machine Classifier and Naïve Bayes Classifier on the same training dataset. The system compares the precision and recall of these two classifiers on the same dataset.

Training Data800 Testing Data 100	Support Vector Machine(SVM)	Naïve Bayes(NB)
Accuracy	0.91625	0.88125
Positive precision	0.967302452316	0.917098445596
Negative precision	0.97619047619	0.602564102564
Neutral precision	0.86189258312	0.904761904762

Positive recall	0.878712871287	0.876237623762
Negative recall	0.694915254237	0.796610169492
Neutral recall	1.0	0.902077151335

Training Data700 Testing Data 100	Support Vector Machine(SVM)	Naïve Bayes(NB)
Accuracy	0.94	0.888571428571
Positive precision	0.993243243243	0.772093023256
Negative precision	0.947368421053	0.710144927536
Neutral precision	0.923232323232	0.978365384615
Positive recall	0.821229050279	0.927374301676
Negative recall	0.870967741935	0.790322580645
Neutral recall	0.995642701525	0.886710239651

Training Data600 Testing Data 100	Support Vector Machine(SVM)	Naïve Bayes(NB)
Accuracy	0.951666666667	0.913333333333
Positive precision	0.983146067416	0.957575757576
Negativeprecision	0.991666666667	0.902985074627
Neutral precision	0.917218543946	0.893687707641
Positive recall	0.916230366492	0.82722513089
Negative recall	0.908396946565	0.923664122127
Neutral recall	0.996402877698	0.967625899281

The system performs the accuracy comparison of Support Vector Machine and Naïve Bayes Classifier on the same training dataset. Support Vector Machine classifier performs well in large dataset and gets the best accuracy. Naïve Bayes classifier performs well in small dataset and gets the best accuracy. Naïve Bayes classifiers is fast when decision making. But, Support Vector Machine classifier is slow when decision making.

8 Conclusion

This paper has presented Naïve Bayes and Support Vector Machine Classification on twitter to classify about Education, Business and Crime. The system is aimed to study machine learning model in the case of mining social media data for sentiment analysis. The system is developed for analyzing Educational Rate, Business Rate and Crime Rate occurred in Malaysia, Singapore and our country, Myanmar. The rate of change of these three sectors can be clearly compared by analyzing these conditions. The system is intended to contribute a lot of advantages for the Ministry of Education, Economists and Home Affairs in each country's government. The system can also analyze the accuracy, precision and recall of Support Vector Machine and Naïve Bayes Classifier.For Further Extension, the system can perform for Facebook and other social media application. The system can extend that input data is not only for text data but also for image data.

Acknowledgement

Firstly, I would like to appreciate Dr. SoeSoeKhaing, Pro-Rector, University of Technology (Yatanarpon Cyber City), for her vision, chosen, giving valuable advices and guidance for preparation of this article. And then, I wish to express my deepest gratitude to my teacher Dr. Hninn Aye Thant, Professor, Department of Information Science and Technology, University of Technology (Yatanarpon Cyber City), for her advice. I am also grateful to Dr. Yi Yi Myint, Assistant Lecturer, co-leader of our Research Development Team, University of Technology (Yatanarpon Cyber City), for giving me valuable advices. Last but not least, many thanks are extended to all persons who directly and indirectly contributed towards the success of this paper.

References

1. TWITTER, http://www.businessdictionary.com/definition/Twitter.html.
2. Rambocas M., Gama J., S.: Marketing Research: The Role of Sentiment Analysis. April2013, ISSN-0870-8541.
3. MOVIE REVIEW DATASET, http://www.cs.cornell.edu/people/pabo/movie-review-data, accessed Octorber 2013.
4. Angiani G., Ferrari L., Paolo Fornacciari T., Eleonoralotto, Magiliani F. and Manicard S., A Comparision between Preprocessing Techniques for Sentiment Analysis in Twitter.
5. Tokenization,https://www.packpub.com/mapt/book/big_data_business_intelligence/ 978172585101/20/ch01/vl1sec008/tokenization.
6. STOP-WORDS-COLLOCATION,https://streamhacker.com/2010/05/24/text-classification -sentiment-analysis-stopwords-collocation.
7. LEMMATISATION, https://en.m.wikipedia.org/wiki/Lemmatisation.
8. PRECISION-RECALL,https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall/.