Journal of Pharmacognosy and Phytochemistry
Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2018; SP1: 609-613

Naw Naw
Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar

Aye Chan Mon
Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar

Correspondence
Naw Naw
Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar

Social media data analysis in sentiment level by using support vector machine

Naw Naw and Aye Chan Mon

Abstract
A social media is an intermediary for communication among people. Twitter is one of the most popular social networking services. All types of users can share their thoughts and opinions on various aspects of day to day activities. So social media websites are regarded as rich sources of data for opinion mining. Such data can be well used for sentiment analysis. Sentiment analysis or opinion mining is the computational study of the opinions, attitudes and emotions of the entity. The entity may describe an individual, event or topic. Support Vector Machine (SVM) is able to identify the separated hyperplane which maximize margin the different classes. 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 and Crime Rate occurred in Malaysia, Singapore, our country, Myanmar. The system is also aimed to perform social media sentiment analysis by applying machine learning approach of Artificial Intelligence (AI).

Keywords: opinion mining, sentiment analysis, twitter, Support Vector Machine (SVM), text classification

Introduction
Nowadays, social media has become information exchange center. In everyday life communication happen over social media sites, people initiate conversation and others open up with their opinions. Hence, social media web-sites are rich sources of data for opinion mining. Since widespread of World Wide Web, internet and extensive growth of social media, organizations feel need to study public opinions for decision making. However to analyze polarity of opinions the exact intelligent information needs to be filtered. Hence automated opinion mining and sentiment analysis systems are needed [1].

Sentiment analysis is treated as a classification task which classifies the orientation of a text into positive, negative or neutral. Sentiment Analysis is conducted at any of the three levels: the document level, phrase level or the phrase level.. In document level, summary of the entire document is taken first and then it is analyze whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment.

Twitter is a social networking service and it can be easily accessible by all people. A particular user has to create his/her own account and that user can read and post messages in twitter. It allows the user to post messages of 140 characters or less [2]. The system is developed to analyze Educational Rate and Crime Rate occurred in Malaysia, Singapore and our country, Myanmar through tweets.

In everyday life, all types of users share their opinions and experiences via social media. Some people have positive aspects about a particular topic or some have negative aspects or some are neutral at their opinions. At first, the system crawls the real time social media data from twitter. The system performs the preprocessing processes to filter the noises from the extracted tweets as a second step. And then, the system classifies these data based on their features as positive, negative or neutral. And the system displays the sentiment scores by using visualization techniques. The performance of classification is also analyzed using precision, recall and accuracy.

Classification techniques can be applied at document, sentence or phrase level as per requirements, to classify opinion as positive, negative or neutral. Machine learning based classification provides an accurate prediction and also improves the performance of the results. This system uses Support Vector Machines (SVM), a supervised machine learning method, to classify tweets. SVMs are able to identify the separated hyperplane which maximize margin the different classes [3].

In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new testing data [4]. SVM is effective, accurate, and can work well with small amount of training data.

This paper is organized as follow; third section gives overview of system design followed by the fourth 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 and Crime.

II. Related Work

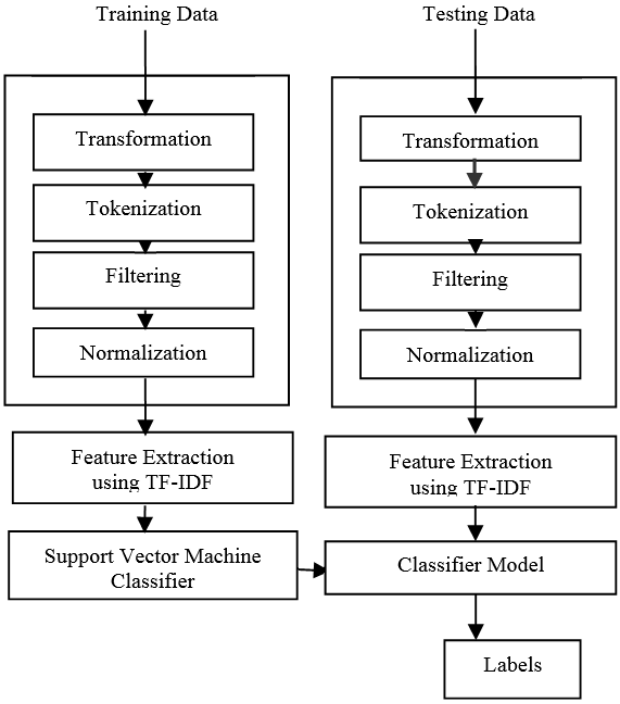
Several techniques were used for opinion mining tasks in history. Pang *et al*. [5], Mukras R.J [6] use the data of movie review, customer feedback review and product review. They use the several statistical feature selection methods and directly apply the machine learning techniques. These experience show that machine learning algorithm only is not well perform on sentiment classification. They show that the present or absent of a word seems to be more indicative of the content rather than the frequency of a word.

In [7] they have classified the subjectivity of social media messages based on traditional features with the inclusion of some social site specific clues such as retweets, hash tags, links, uppercase words, emotions, and exclamation and question marks. Further a Part-Of-Speech (POS) specific prior polarity features and a tree kernel to obviate the need for tedious features engineering is introduced in [8]

Researchers have paid attention to this problem to some extent. In this paper [9], authors looking at popular microblogging Twitter, here the authors build models for two classifying tasks. These are a binary task of classifying sentiment into positive, negative classes and three-way task means to classify sentiment into positive, negative and neutral classes. They also performed an experiment with unigram model, a feature based model, and a tree kernel-based model. They were combining unigrams with their features and features with the tree kernel. In this paper, they presented extensive feature analysis of the 100 features they propose.

III. System design overview

The input of the system is the tweets about Education and Crime. These tweets are crawled from various twitter users that post on twitter about Education and Crime and then the system implements sentiment analysis on these collected data. The twitter data cannot classify directly because it has noisy information. So, this noisy information is removed by pre-processing. After that, the system uses Supervised Machine learning algorithm (SVM) that can achieve competitive accuracy when it 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 and Education Sector occurred in Malaysia, Singapore and our country, Myanmar. There are three maincomponents in the design of the system. They are pre-processing stage, feature extraction and classification. Figure 1 illustrates the overall system design.



First of all, the system requires twitter data about Education and Crime. The language is as English using Twitter Streaming API. These tweets are crawled from various twitter users that post on twitter about Education and Crime. The extracted social media data is needed to preprocess. In the Pre-processing stage, Transformation, Tokenization, Filtering and normalization processes are performed. After that, meaningful features are extracted with 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. And then, the system selects features as the input features of classification. In this system, Support Vector Machine classifier is used. Support Vector Machine (SVM) belongs to the class of Supervised Learning algorithm in which the learning machine is given a set of examples (input) with the associated labels (output values). Finally, Support Vector Machine labels sentiment scores as positive or negative or neutral in these Education and Crime sectors. The system displays according to their scores by using visualization techniques. The performance of classification is also analyzed using precision, recall and accuracy.

IV. Pre-Processing Stage

Preprocessing the data is the process of cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, and advertisements. Pre-processing stage consists of four main processes:

Fig 2: Sample Tweets about Education

A. Transformation

In general, a clean tweet should not contain URLs, hashtags (i.e, #studying) or mentions (i.e. @Irene). Firstly, the tweets extracted from twitter are converted from upper case to lower case. And then, URLs are replaced with generic word URL.URLs does not contribute to analyze the sentiment in the informal text. Then, @username is replaced with generic word AT_USER. After that, #hashtag is replaced with the exact same word without the hash. Punctuations are removed at the start and end of the tweets because they are part of grammar constructs such as the genitive. Finally, multiple whitespaces are replaced with a single whitespace. In this step, the system executes all the operations to make the text uniform. It is important because features can be correctly chosen during the classification process. Tweets about Education and Crime from Myanmar, Singapore and Malaysia are preprocessed in the Transformation step as shown in figure 3.

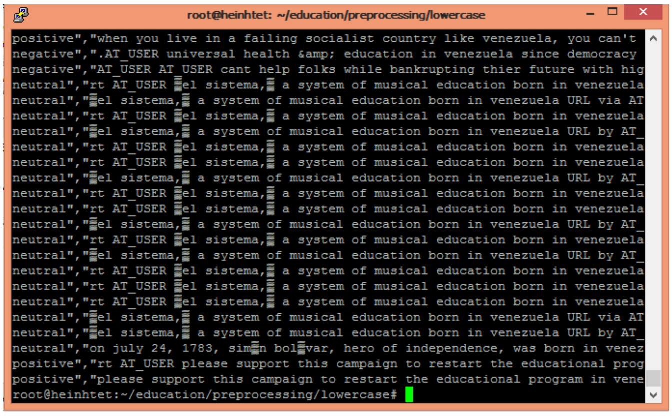


Fig 3: LowerCase Conversion

B. 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. Tokenization helps to divide the textual information into individual words. The list of tokens becomes input for further processing such as parsing or text mining [10]. Tokenization is one of pre-processing steps. The system performs Tokenization to break the text into smaller components (unigram).

C. Filtering

Stopword filtering is a common step in preprocessing text for various purposes. This step is to reduce the noise of textual data by removing stopwords. The system removes stopwords from a keyword phrase to provide the most relevant result. Stopwords are words that are generally considered useless. 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. Words like ", We, are, am, was" etc are not important for processing the text.The system tokenizes the output tweets from Transformation step and then removesstopwords as shown in figure 4.

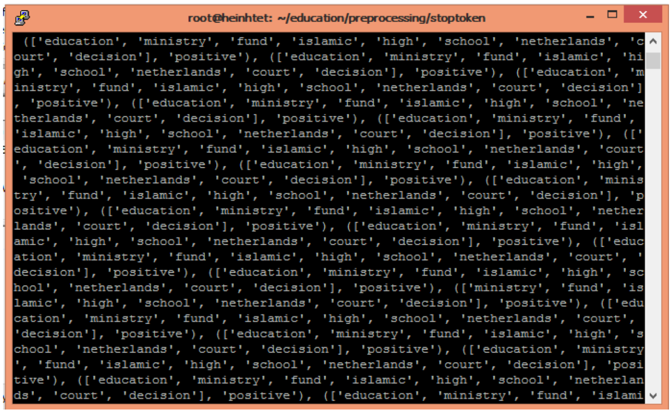


Fig 4: Tokenization and Stop Word Filtering

D. Normalization

In the normalization step, lemmatization is performed. Lemmatization is closely related to stemming. In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence [11]. Lemmatization returns depending on whether the use of the token was as a verb, an adjective or a noun. After the lemmatization the root words are got and they are used for feature extraction step.Figure 5 shows that the output words from the Filtering step are transformed to the base of that word.

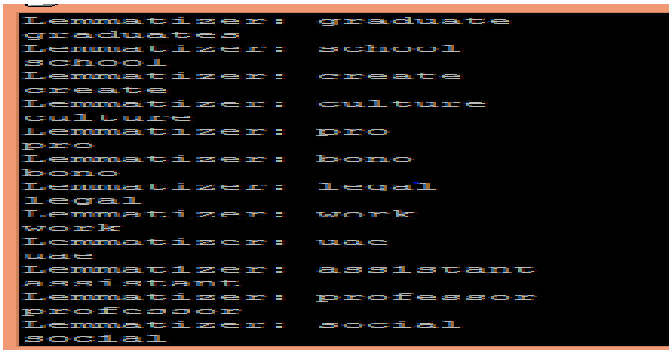


Fig 5: Normalization

V. Feature Extraction

Transforming the input data into the set of features is called Feature Extraction. There are several ways to assess the importance of each feature by attaching a certain weight in the text. The most popular ones are: feature frequency (FF), Term Frequency Inverse Document Frequency (TF-IDF), and feature presence (FP). In this system, Support Vector Machine Classifier is trained on tf-idf weighted word frequency features. What Term Frequency-Inverse Document Frequency (tf-idf) is how important is a word to a document in a

collection, and that's why tf-idf incorporates local and global parameters, because it takes in consideration not only the isolated term but also the term within the document collection [12]. 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 most useful features for the system and perform the best accuracy. In figure6, the most useful features for the system are displayed.

Fig 6: Extracting Features

VI. Classification

There are two text classification methods using ML approach such as supervised and unsupervised learning methods. In supervised technique, the main task is to build a classifier. The classifier requires training datasets which can be labeled manually or obtained from a user-generated user-labeled online source. 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 machine learning classification, "training sets" and "testing sets" are required.So, the classifier is trained by using "training set" and then the system tests the performance of the classifier on unseen "testing set" [13]. There are many supervised machine learning approaches such as Support Vector Machine (SVM), Naïve Bayes Classifier and Maximum Entropy Classifier. In this system, Support Vector Machine (SVM) Classifier is used.

A. Support Vector Machine

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. SVM is a method for classification of linear data. It uses a non linear mapping to transform the training data into a higher dimension. Within the higher dimension it searches for the linear optimal separating hyperplane. Data from two classes can always be separated by a hyperplane. The SVM finds hyperplane using support vectors. This approach was developed by Vladimir vapnik, Bernhard Boser and Isabelle Guyan in 1992 [14]. A support vector machine is considered thehighly effective at traditional text classification method.

Support Vector Machines attempt to find the best possible surface to separate positive, negative and neutral training samples. Support Vector Machine perform sentiment classification task on twitter data about Education and Crime Support Vector Machine Classifiers are inherently two class classifiers. The traditional way to do multiclass classification with SVM s is to use one of the methods (One-Versus-Rest Classification and One-Versus-One Classifiers. In this system, One-Versus-Rest Classifier is used. OVR Classifier is to train N different classifiers and it separates one class label from all of the rest. When it is desired a new testing data, the N classifiers are run, and the classifier which outputs the largest value is chosen. OVA Classifier is extremely powerful, producing results that are often at least as accurate as other methods. The rationale behind SVM's is that if we choose the one that maximizes the margin we are less likely to misclassify unknown items in the future.Figure 7 shows the positive, negative and neutral percentage result of testing data about Education.

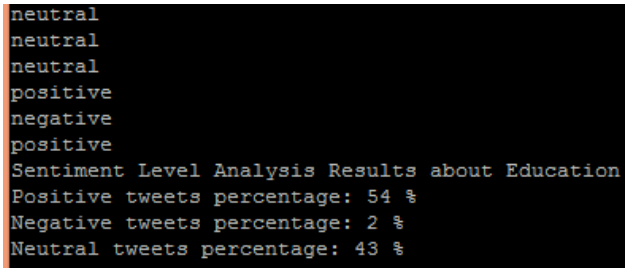


Fig 7: Support Vector Machine Testing Result

B. Accuracy, precision and recall

Accuracy is the performance evaluation parameter for the system. 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. A higher precision means less false positives, while a lower precision means more false positive. Recall measures the completeness of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can decrease precision [15].

The system computes the accuracy of Support Vector Machine 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 by using NLTK metrics module. The NLTK module provides functions for calculating these metrics. In figure 8, the system displays accuracy, precision and recall result of Education Data based on training data 7540 and testing data 1500. In figure 9, the system displays accuracy, precision and recall result of Crime Data based on training data 5414 and testing data 1500.

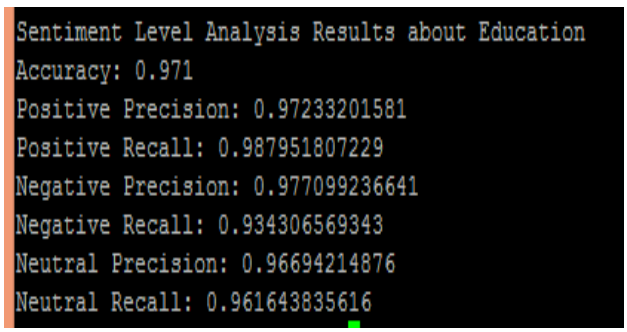


Fig 8: Performance Analysis Result about Education

```
Sentiment Level Analysis Results about Crime
Accuracy: 0.978571428571
Positive Precision: 0.990384615385
Positive Recall: 0.919642857143
Negative Precision: 0.996
Negative Recall: 0.98031496063
Neutral Precision: 0.962427745665
Neutral Recall: 0.997005988024
```

Fig 9: Performance Analysis Result about Crime

VII. Experimental Result

The experimental results are presented based on extracting the particular positive, neutral and negative keywords of Education and Crime in Myanmar, Malaysia and Singapore. The language is as English using Twitter Streaming API. The experimental results are displayed as Bar Graph.

Tweets about Education and Crime are extracted from a particular Twitter account. The extracted training dataset consists of 7540 tweets of Education and 5414 tweets of Crime. The extracted tweets are preprocessed in these steps such as Transformation, Tokenization and Lemmatization. After that, the system extracts meaningful features using TF-IDF. The feature words are reduced to 6024 about Education and 4924 about Crime. It is manually labeled for classification. The output feature words are input features of Support Vector Machine Classifier.

The system also requires real time testing data from Twitter. The testing datasets used in the application were retrieved from Twitter using Twitter4j API's.The total number of tweets extracted is 119 tweets for testing about education. It consists of 33, 69 and 17 tweets from Myanmar, Malaysia and Singapore. The total number of tweets extracted is 84 tweets for testing about education. It consists of 9, 41 and 34 tweets from Myanmar, Malaysia and Singapore. For Education, the system outputs 15, 15 and 69 as positive, negative and neutral percentages in Myanmar. The system outputs 27, 18, 53 as positive, negative and neutral percentages in Malaysia and then 52, 17, 29 in Singapore. For Crime, the system outputs 33, 22 and 44 as positive, negative and neutral percentages in Myanmar. The system outputs 31, 24, 43 as positive, negative and neutral percentages in Malaysia and then 35, 20, 44 in Singapore.

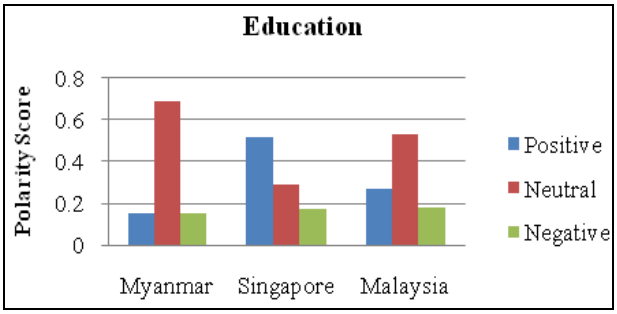


Fig 10: Graphical Analysis of Education

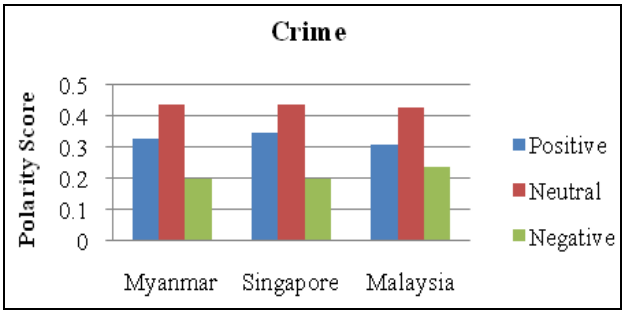


Fig 11: Graphical Analysis of Crime

VIII. Conclusion

In this paper, Support Vector Machine Classification on twitter to classify about Education and Crime has been presented. The proposed system would be useful for the selected ASEAN countries to know their citizens' social media usage behavior. The system is to study machine learning model in the case of mining social media data for sentiment analysis. The system is developed for analyzing Educational Rate and Crime Rate occurred in Malaysia, Singapore and our country, Myanmar. By analyzing these conditions, the system can easily check the selected ASEAN countries' conditions in these education and crime. The system is intended to contribute a lot of advantages for the Ministry of Education and Home Affairs in each country's government. As a future extension, the system can also analyze the accuracy of Support Vector Machine with other machine learning techniques.

References

1. Bo Pang, Lillian Lee. Opinion Mining and Sentiment Analysis, 2008.
2. A machine learning based classification for social media messages R. Nivedha and N. Sariam.
3. Sentiment analysis of smartphone product review using SVM algorithm-based particle swarm optimization
4. docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm /introduction_to_svm.html
5. Pang B *et al*. Thumbs up? Sentiment Classification Using Machine Learning Techniques, Procs. Of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL Press. 2002, pp 79-86.
6. Mukras R, Carroll J. A comparison of machine learning techniques applied to sentiment classification. 2004, pp 200-204.
7. Barbosa L, Feng J. Robust sentiment detection on Twitter from biased and noisy dat. In: Proceedings of COLING. 2010, pp. 3644.
8. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of Twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. 2011; Pp. 3038.
9. Apoorv Agarwal, Boyi Xie, llia Vovsha, Owen Rambow. Rebecca Passonneau: Sentiment Analysis of Twitter Data. In: Proceedings of the workshop on Language in Social Media (LSM). 2011, pages 30-38.
10. https://streamhacker.com/2010/05/24/text-classification-sentiment -analysis-stopwords-collocation/
11. https://en.m.wikipedia.org/wiki/Lemmatisation
12. http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/
13. https://www.quora.com/What-is-a-training-data-set-test-data-set-in-machine-learning-What-are-the-rules-for-selecting-them
14. Simeon M, Hilderman R. Categorical proportional difference: A feature selection method for text categorization In Aus DM. 2008, pages 201-298.
15. https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall/