# Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers

**Naw Naw**

DOI: 10.29322/IJSRP.8.10.2018.p8252
http://dx.doi.org/10.29322/IJSRP.8.10.2018.p8252

***Abstract*****—** At the present time, social media has become an important popular communication medium among all online suffers. Twitter is one of the most popular social networking services where thoughts and opinions about various aspects and activities can be shared by the millions of users. Social media websites are rich sources of data for opinion mining. Such data can be applied for sentiment analysis. Sentiment analysis is the study of human behavior by extracting user opinion and emotion form plain text. Among machine learning techniques, Support Vector Machine (S.V.M) classifier and K-Nearest Neighbour (K-N.N) classifier is used in this system. The system provides the analytical results of education, business, crime and health for Educational Authorities, Economists, Government Organization's needs and Health. And then, the system predicts the conditions of selected ASEAN countries (Malaysia, Singapore, Vietnam and Myanmar) according to the tweets. In this system, accuracy, precision, recall and f1-score is also compared by using these two classifiers.

***Index Terms*****—**K-NN, Opinion Mining, Sentiment Analysis, SVM, Twitter.

## 1) Introduction

Nowadays, social media gives the very large effect to the digital improvement in terms of global communications. The emergence of social media has provided a place for web users to share their thoughts and express their opinions on different topics in an event. Micro blogging has become a very popular communication tool among Internet users. Millions of messages are appearing daily in popular web sites that provide services for micro blogging such as Facebook, twitter, Tumblr and so on. Twitter is an online social networking medium, popular since 2006, where registered users share or post messages under 140 characters known as tweets.

Sentiment Analysis or opinion mining is the computational study of the opinions, attitudes and emotions of the entity. The entity may describe and individual, event or topic. The topic is likely to be a review. The system is developed to analyze Educational Rate, Business Rate, Health Rate and Crime Rate occurred in Malaysia, Singapore, Vietnam and our country, Myanmar according to the tweets. As a first step, the system crawls the real time data as the input from Twitter. As a second step, sentiment analysis is implemented by using the crawled data. The system needs to build classifier model (Support Vector Machine and K-Nearest Neighbour) to implement sentiment analysis. Finally, the system outputs the percentage score of these sectors and displays according to their scores by using visualization techniques. The performance of classification is also analyzed using accuracy, precision, recall and f1-score.

## 2) Related word

AaratiPatil and SrinivasaNarasimhaKini [1] proposed Evolution of Social Media from the era of Information Retrieval. Some insightful information that data scientists and social users are showing interest in sentiment analysis of social media data is got. Different approaches have been implemented to automatically detect sentiment on texts [2]. An active research on Sentiment analysis on selective micro blogging sites has explored in [3].

Barbosa and Feng [4] presented robust sentiment detection on Twitter from biased and noisy data. The subjectivity of social media messages based on traditional features with the inclusion of some social media site specific clues such as retweets, hash tags, links, uppercase words, emoticons, and exclamation and question marks has been classified. Further, Agarwal, Xie, Vovsha, Rambow and Passoeau introduced a Part-Of-Speech (P.O.S) specific prior polarity features and a tree kernel to obviate the need for tedious feature engineering.

Agarwal et al. [5] approached the task of mining sentiment form twitter, as a 3-way task of classifying sentiment into positive, negative and neutral classes. They experimented with three types of models: unigram model, a feature based model and a tree kernel based model. For the tree kernel based model they designed a new tree representation for tweets. The feature based model that uses 100 features and the unigram model uses over 10,000 polarity of words with their parts-of-speech tags are most important for the classification task. The tree kernel based model outperformed the other two.

## 3) System Design Overview

The system design is mainly composed of two parts such as Training and Testing. In training phase, the system trains the input raw dataset of tweets with classifier model. In testing phase, the system crawls the real tweets of Education, Business, Crime and Health data and then classifies the positive, negative or neutral class based on Classifier Model. In the system design, there are four main components. They are preprocessing, feature selection, feature extraction and classification. The overall system design is illustrated in Fig. 1.

Firstly, tweets about Education, Business, Crime and Health are extracted from Twitter. The language is English using Twitter Streaming API. The system needs to perform preprocessing step. The preprocessing step performs transformation, negation handling, tokenization, filtering and normalization. And then, features are selected by comparing with Knowledge Base. After that, meaning features are extracted from Term Frequency-Inverse Document Frequency (T.F-I.D.F). And then, features are selected as the input features of classification. At last, the system builds Classifier Model

(Support Vector Machine and K-Nearest Neighbour) in order to perform the training process.

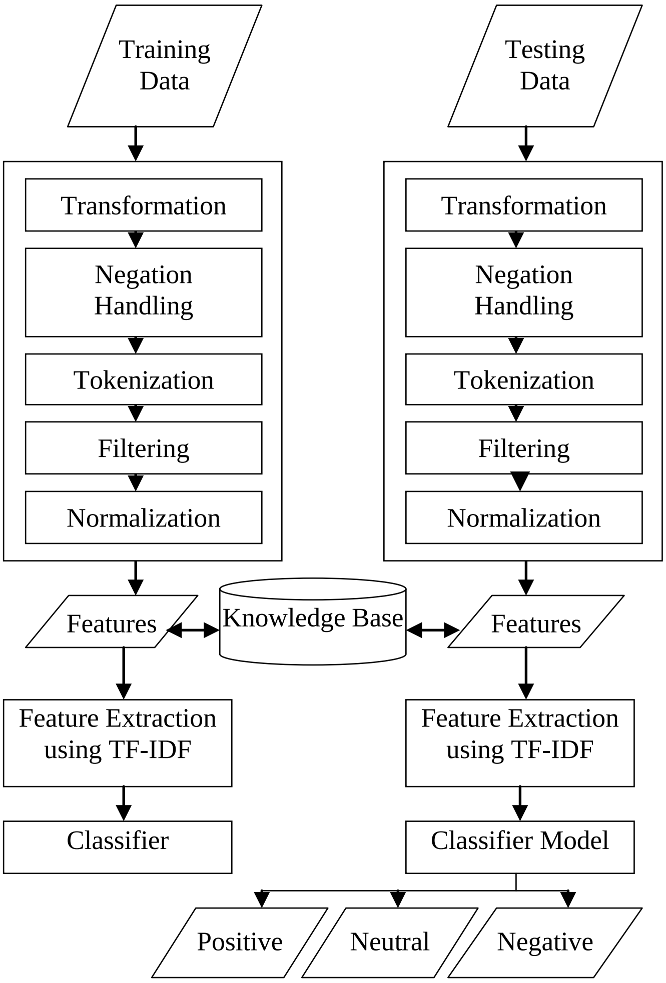
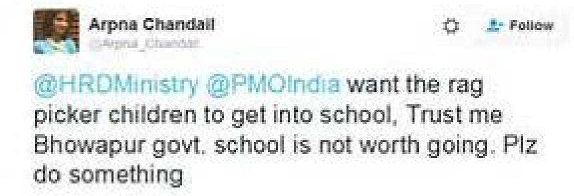


Fig. 1. The System Design

In Testing phase, the real data about Education, Business, Crime and Health is crawled from Twitter as the input of the system. And then, Preprocessing, Feature Selection and Feature Extraction are also performed like Training phase. The output features are executed in order to classify the positive or negative or neutral of class based on Classifier Model. And then, the system displays according to their scores by using visualization techniques. The system also compared the performance of these two classifiers in accuracy, precision, recall and f1-score.

### 1) *Preprocessing*

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. If there is much irrelevant and redundant information or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preprocessing is the most important phase of a machine learning project, especially in computational biology [6]. Sample tweet about education is shown in Fig. 1.



Arpna Chandail
@Arpna_Chandail

@HRDMinistry @PMOIndia want the rag picker children to get into school, Trust me Bhowapur govt. school is not worth going. Plz do something

Fig. 2. Sample Tweets about Education

#### a) *Transformation*

The following steps are performed in Transformation step. 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, replaced URLs with generic word URL, replaced @username with generic word AT_USER, replaced #hashtag with the exact same word without the hash, removed punctuations at the start and end of the tweets and replaced multiple whitespaces with a single whitespace. The resultant tweets from Transformation step are shown in Table 1.

Table 1. Tweets after Performing Transformation Step

| Step | Tweet |
|---|---|
| Transformation | AT_USER AT_USER want the rag picker children to get into school trust me bhowapur govt school is not worth going plz do something |

#### b) *Negation Handling*

Negations are those words which affect the sentiment orientation of other words in a sentence. Examples of negation words include not, no, never, cannot, should not, would not, etc.

Negation handling is an automatic way of determining the scope of negation and inverting the polarities of opinionated words that are actually affected by a negation [7]. The resultant tweets after performing negation handling step are shown in Table 2.

Table 2. Tweets after Performing Negation Handling Step

| Step | Tweet |
|---|---|
| Negation Handling | AT_USER AT_USER want the rag picker children to get into school trust me bhowapur govt school is not not_worth not_going not_plz not_do not_something |

#### c) *Tokenization*

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens [8].

The system tokenizes the uniformed sentence from Negation Handling Step which got into smaller components (unigram) as shown in Table 3. These resultants words become the input for the next preprocessing.

Table 3. Features after Performing Tokenization Step

| Step | Features |
|---|---|
| Tokenizati on | want, the, rag, picker, children, to, get, into, school, trust, me, bhowapur, govt, school, is, not, not_worth, not_going, not_plz, not_do, not_something |

#### d) *Filtering*

Stop words are words which are filtered out before or after processing of natural language data (text). A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query [9]. The resultant features from Filtering step are shown in Table 4.

Table 4. Features about Performing Filtering Step

| Step | Features |
|---|---|
| Filtering | want, rag, picker, children, get, school, trust, bhowapur, govt, school, not, not_worth, not_going, not_plz, not_do, not_something |

*e) Normalization*

In the Normalization step, lemmatization is performed. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form [10].

After normalization step is performed, the root words are got as shown Table 5 and they are used for feature extraction step.

Table 5. Features about Performing Normalization Step

| Step | Features |
|---|---|
| Normalization | hrdministry, pmoindia, want, rag, picker, children, get, school, trust, bhowapur, govt, school, not, not_worth, not_go, not_plz, not_do, not_something |

*2) Feature Selection*

After preprocessing step, features are selected by comparing with Knowledge Base to improve accuracy. Several words related to hundreds education, business, crime features are collected and added to the Knowledge Base. Finally, the system selects features from Knowledge Base. In this way, the system can get the essential features for the system as shown in Table 6 and performs the best accuracy.

Table 6. Features about Performing Feature Selection Step

| Step | Features |
|---|---|
| Feature Selection | Want, school, trust, not, not_worth, not_go |

*3) Feature Extraction*

The system trains Support Vector Machine and K-NN classifier based on T.F-I.D.F (Term Frequency-Inverse Document Frequency) weighted word frequency. TF is how frequently a word occurs in a document. IDF decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.

This can be combined with term frequency to calculate a term's tf-idf, the frequency of a term adjusted for how rarely it is used. When the feature extraction is performed using T.F-I.D.F, the system selects features as the input features of classification.

*4) Classification*

Text classification models are used to categorize text into organized groups. Text is analyzed by a model and then the appropriate tags are applied based on the content. Machine learning models that can automatically apply tags for classification are known as classifiers.

Classifiers can't just work automatically; they need to be trained to be able to make specific predictions for texts. Once enough texts have been trained, the classifier can learn from those associations and begin to make predictions with new texts.

There are two main approaches to sentiment classification: lexicon-based and machine-learning. A lexicon-based approach tokenizes data into individual words which are checked with a sentiment lexicon containing a polarity value for individual words. The sum of the polarities is passed to an algorithm that determines the overall polarity of the sentence. A machine-learning approach utilizes a labeled training set to adapt a classifier to the data domain of the training set. The trained classifier can then predict the result of the problem and the success rate of the prediction depends on how well the problem is contained within the same domain.

There are three types of machine learning algorithms: supervised learning, unsupervised learning and reinforcement learning. Among them, this system uses supervised learning approach. This algorithm consists of the target or outcome variable which is to be predicted from a given set of predictors. Using these set of variables, a function that map inputs to desired outputs is generated. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning are decision tree, K-Nearest Neighbour (K-N.N), Support Vector Machine (S.V.M), Naïve Bayes (N.B) , Maximum Entropy(MaxEnt) and so on. In this system, S.V.M and K-N.N classifier are used.

*a) Support Vector Machine (One-Versus-One)*

SVMs are often employed for binary sentiment detection because they are binary classifiers. In order to perform multi-class classification, the problem needs to be transformed into a set of binary classification problems. There are two approaches to do this: One vs. Rest Approach (O.V.R) and One vs. One Approach (O.V.O).

In the system, OVO strategy is used. In the O.V.O strategy, one trains K(K-1)/2 binary classifiers for a K-way multi-class problem; each receives the samples of a pair of classes form the original training set, and must learn to distinguish these two classes. At prediction time, a voting scheme is applied: all K(K-1)/2 classifiers are applied to an unseen sample and the class that got the highest number of "+1" predictions gets predicted by the combined classifier. Like O.V.R, O.V.O suffers from ambiguities in that some regions of its input space may receive the same number of votes [11].

*b) K-Nearest Neighbour (K-N.N)*

K-NN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms. K-NN is a non-parametric, lazy learning algorithm. Its goal is to use a database in which the data points are separated into several classes to predict the classification of a new sample point [12].

In the case of classification, the output is class membership (the most prevalent cluster may be returned), the object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. This rule simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbours in the training sets.

The Nearest Neighbour rule (N.N) is the simplest form of K-NN when K=1. Given an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample.

Therefore, the unknown sample may be classified based on the classification of the nearest neighbor. The K-N.N is an easy algorithm to understand and implement, and a powerful tool that there is at our disposal for sentiment analysis.

### 4) PERFORMANCE ANALYSIS

The system calculates the performance of the system. It performs accuracy, precision, recall and f1-score of the two classifiers.

Accuracy is not only the metric for evaluating the effectiveness of a classifier. The system calculates the accuracy of S.V.M and K-N.N Classifiers. It is calculated by number of correctly selected positive, negative and neutral words divided by total number of words resent in the corpus.

Precision measures the exactness of a classifier. Recall measures the completeness, or sensitivity of a classifier.

The system calculates them by using sklearn libraries based on Support Vector Machine and K-Nearest Neighbor Classifier. F1-score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

### 5) EXPERIMENTAL RESULTS

The crawled training dataset consists of 3484 tweets about education, 5455 tweets about business, 2460 tweets about crime and 8078 tweets about health. These crawled data are preprocessed in transformation, negation handling, tokenization, filtering and normalization.

After that they system selects meaning features by comparing with knowledge base. And then, the system performs feature extraction step. The output features are the input features of Support Vector Machine and K-Nearest Neighbour classifiers.

For testing data, tweets about education, business, crime and health are extracted from a particular Twitter account after getting prior permission.

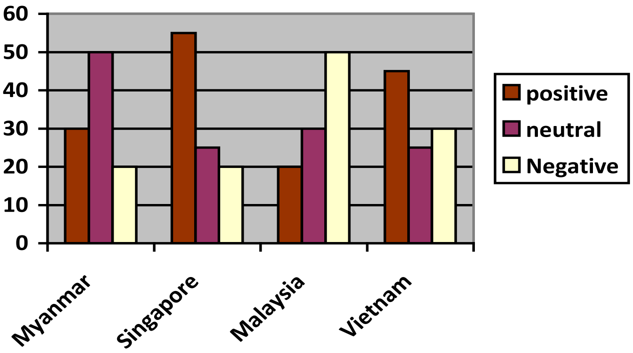


Fig. 2. Graphical Analysis about Education

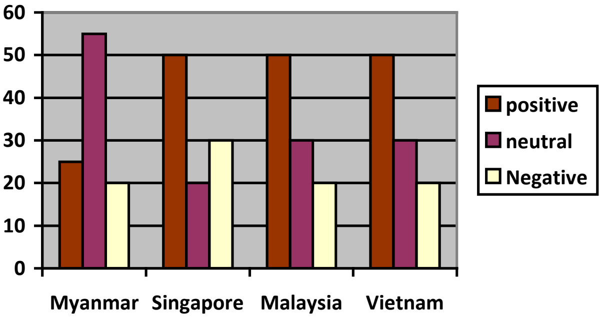


Fig. 4. Graphical Analysis about Business

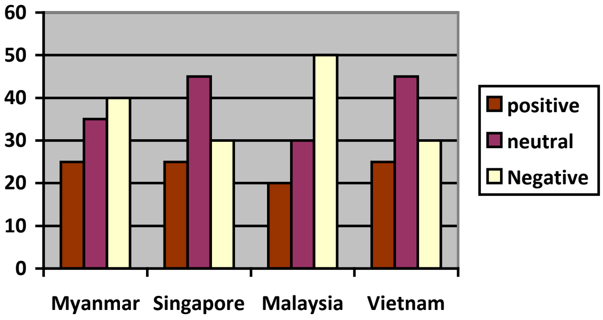


Fig. 5. Graphical Analysis about Crime

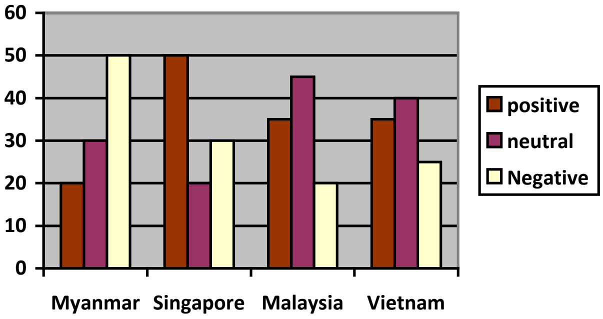


Fig. 6. Graphical Analysis about Health

### 6) PERFORMANCE COMPARISON

The system is aimed to perform the accuracy, precision, recall and f1-score of Support Vector Machine and K-Nearest Neighbour Classifier on the same training dataset.

Table 7. Performance Comparison about Education

| Training Data 2788 Testing Data 696 | Support Vector Machine | K-NN |
|---|---|---|
| Accuracy | 0.7446197991391679 | 0.6958393113342898 |
| Precision | 0.7130609476214108 | 0.7780602580671679 |
| Recall | 0.7080479012071043 | 0.6339499441631286 |
| F1-Score | 0.7080479012071043 | 0.6339499441631286 |

Table 8. Performance Comparison about Business

| Training Data 4364 Testing Data 1091 | Support Vector Machine | K-NN |
|---|---|---|
| Accuracy | 0.7241063244729606 | 0.7057745187901008 |
| Precision | 0.6727147115713413 | 0.7585090117040704 |
| Recall | 0.6740008345575266 | 0.5960035137416786 |
| F1-Score | 0.7239724741913184 | 0.6819787503355964 |

Table 9. Performance Comparison about Crime

| Training Data 1968 Testing Data 492 | Support Vector Machine | K-NN |
|---|---|---|
| Accuracy | 0.7215447154471545 | 0.5873983739837398 |
| Precision | 0.694264733332539 | 0.6301859690017584 |
| Recall | 0.6764180216193556 | 0.5892023932506887 |
| F1-Score | 0.7175716931918386 | 0.5992940183025018 |

Table 10. Performance Comparison about Health

| Training Data 6463 Testing Data 1615 | Support Vector Machine | K-NN |
|---|---|---|
| Accuracy | 0.709389492039024 | 0.5920848023027024 |
| Precision | 0.5842024027422439 | 0.4703402840230202 |
| Recall | 0.5700240203207324 | 0.4028508204830833 |
| F1-Score | 0.6409402930294028 | 0.5084348028302832 |

## 7) CONCLUSION

Support Vector Machine and K-Nearest Neighbour classifiers are performed on twitter to classify about Education, Business, Crime and Health. The system is intended to measure the impact of ASEAN citizens' social media usage behavior. The main purpose of the system is to understand how to perform social media sentiment analytics on big data environment by applying machine learning approach of Artificial Intelligence (A.I). The system is developed for analyzing Business Rate, Crime Rate, Educational Rate and Health Rate occurred in Malaysia, Singapore, Vietnam and our country, Myanmar. The rate of change of these sectors can be compared by looking at these conditions. The system can be expected to contribute a lot of advantages for the Ministry of Education, Commerce, Home Affairs and Health in each country's government.

## ACKNOWLEDGEMENT

Firstly, I would like to appreciate Dr. Aung Win, 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. Last but not least, many thanks are extended to all persons who directly and indirectly contributed towards the success of this paper.

## REFERENCES

[1] AaratilPatil, SrinivasaNarasimhaKini: Evolution of Social Media from the era of Information Retrieval, International Journal Science and Research (*IJSR*), 4 (14) (2015) 2326-2331.

[2] Bo P., Lillian L., and Shivakumar V.: Sentiment Classification Using Machine Learning Techniques, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2002).

[3] Bhayani A., Huang R., L: Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford (2009).

[4] Barbosa, L., Feng, J: Robust sentiment detection on Twitter from biased on noisy data, In: Proceedings of COLING, (2010) 3644.

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media, 2011, pp. 30-38.

[6] https://en.wikipedia.org/wiki/Data_pre-processing.

[7] https://www.researchgate.net/publication/314424838_Negation_Handling_in_Sentiment_Analysis_at_Sentenence_Level

[8] https://www.techopedia.com/definition/13698/tokenization

[9] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

[10] htts://en.m.wikipedia.org/wiki/Lemmatisation

[11] https://medium.com@adi.bronshtein/a-quick-indroduction-to-k-nearest-neighbors-algorithm-62214cea29c7

[12] https://sadanand-singh.github.io/posts/svmpython/