Twitter Sentiment Analysis For Emotional Behavior Monitoring System Based on Naïve Bayes
(Social Media Data: Business, Technology)

Hsu Wai Naing[#1], Dr. PhyuThwe[*2]
[#]*Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar*
[*]*Department of Information Science, University of Technology (Yatanarpon Cyber City), Myanmar*
[1]hsuwainaing2054@gmail.com
[2]pthwe19@gmail.com

Abstract **—Most of the people are using social media applications such as Twitter, Facebook as the daily activity of their life. Sentiment Analysis area becomes the innovative research area for the analysis of public opinions behind certain topics, feelings, emotions and attitudes of people. Sentiment Analysis is the task of identifying whether the opinion expressed in a document is positive or negative or neutral about a given topic. The system needs to build Supervised Classifier Model to perform the sentiment analysis. The proposed system uses Naïve Bayes Classifier model. The system is intended to measure the impact of ASEAN citizens' social media usage behavior. The system is aimed for analyzing National Business Rate, and Technology Rate occurred in Korea, Vietnam and our country, Myanmar. The main purpose of the system is to understand how to perform social media sentiment analytics by applying machine learning approach of Artificial Intelligence (AI).**

Keywords **– Twitter, Text Classification, Sentiment Analysis, Opinion Mining, Naïve Bayes Classifier**

I. INTRODUCTION

Nowadays, the micro blogging has become a very popular messaging tool between internet users. Millions of users can share their opinions in different aspects of life every day in popular websites like Twitter and Facebook. Twitter supports brief explanation of ideas via short messages of tweets that are no longer than 140 characters [1]. It allows for valuable and well-timed statement of information. Sentiment analysis is the automated extraction of expressions of positive or negative or neutral attitudes from text. Sentiment Analysis involves extraction, classifying and presenting the opinions expressed by the users [2].

This paper is related to the subject of sentiment analysis for the emotional behavior monitoring system of Business and Technology sectors. The proposed system analyzes the rate of change of Business Sector and developed and undeveloped percentage in Technology Sector occurred in Korea, Vietnam and our country, Myanmar. Sentiment Analysis performs these tasks such as preprocessing, features extraction and classification based on twitter data. The system uses Naïve Bayes Classifier model for the classifying tasks in Sentiment. The system provides the analytical results of business and technology data for Economists and Technology Department' needs.

A supervised classification technique is a systematic approach for classification models from an input data set. The classification techniques include: Decision Tree Induction, Rule-Based Classifiers, Neural Networks, Support Vector Machines, Naive Bayesian Classifiers, Nearest-Neighbor Classifiers. Among them, the proposed system uses the Naive Bayes Classifier. Naive Bayes Classifier is a simple model which works well on text categorization. Naive Bayes Classifier that is based on the popular Bayes' probability Theorem. Naive Bayes Classifier is easy to understand and implement. It is no complicated optimization required and easily updateable if new training data is received. Although the independence assumption may seem sometimes unreasonable, its performance is usually good [3].

This paper is organized as follow; second section gives overview of system design followed by the third section that describes Transformation, Tokenization, and Filtering and Normalization process. Fourth section gives about the proposed Feature Extraction and Fifth section describes Sentiment Classification processes. Sixth section that shows the experimental results of the system.

II. SYSTEM DESIGN OVERVIEW

The input data of the system consists of Business and Technology tweets. These tweets are captured from various twitter users that post on twitter about business and technology. The main task of this system is to classify the rate of change of Business and Technology sectors occurred in Korea, Vietnam and our country, Myanmar. There are three main components in the design of the system. They are Pre-processing stage, Features Extraction and Classification. Figure 1 illustrates the overall system design.

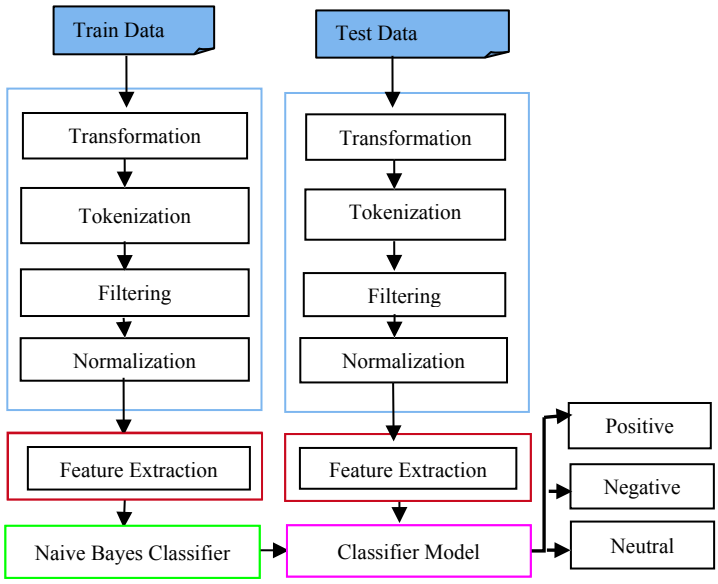

Fig.1 The System Design

In the Pre-processing stage, Transformation process, Tokenization process, Filtering process and Normalization process are performed. In the Feature Extraction stage, the system extracts meaningful features by using Term Frequency-Inverse Document Frequency (TF-IDF). After this step, the system selects features as the input features for Naive Bayes Classifier Model. Finally, Naive Bayes Classifier labels sentiment scores as positive or negative or neutral in these Business and Technology sectors. The system displays these sentiment scores by using visualization techniques. The performance of classification is analyzed using precision, recall and accuracy.

III. PRE-PROCESSING STAGE

Pre-processing stage consists of four main processes:

A. *Transformation*

Transformation process manages basic cleaning operations, which consists in removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words. To provide only significant information, a clean tweet should not contain URLs, hashtags (i.e. #happy) or mentions (i.e. @BarackObama). The input tweets are converted to lower case. URLs are 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.

Furthermore, tabs and line breaks should be replaced with a blank and quotation marks and apexes. After this step, all the punctuations are removed punctuation at the start and ending of the tweets. Additional white spaces that are replaced multiple white-spaces with a single white-space. The next operation is to remove the vowels repeated in sequence at least three times. For example, two words are written in a different way (i.e. cooooool and cool) will become equals [4].

The system executes all the operations in this step to make the text uniform. This is important because during the classification process, features can be chosen correctly.

Fig.2 Transformation Image

B. *Tokenization*

Tokenization is the process of segmenting text into words and sentences. Text is a linear sequence of symbols (characters or words or phrases).Text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numerics, etc. This process is called tokenization. Tokenization is a kind of Pre-processing. The system performs Tokenization to break the text into smaller components (unigram) [5].

C. *Filtering*

The Filtering step is to reduce the noise of textual data by removing stop words. Stop words are common words that carry less important meaning than keywords. The system removes stop words from a keyword phrase to provide the most relevant result. The system removes all stop words right from Determiners that tend to mark nouns where a determiner usually will be followed by a noun (e.g. the, a, an, another) to Coordinating conjunctions that connect words, phrases, and clauses (e.g. for, an, nor, but, or, yet, so) to Prepositions that express temporal or spatial relations (e.g. in, under, towards, before). This step is important within the classifier model because these stop words lead to a less accurate classification [6].

D. *Normalization*

In Pre-processing step, Normalization process performs the critical step. Normalization is a process that converts 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. At least two tasks are commonly applied as part of any normalization process such as Lemmatization and Stemming.

Lemmatization is the task of determining that two words have the same root, despite their surface differences. The words *am, are*, and *is* have the shared lemma *be* and the words *dinner* and *dinners* both have the lemma *dinner [7]*.

Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. The system performs Lemmatization task in Normalization process step [7].

IV. FEATURE EXTRACTION

Transforming the input data into the set of features is called Feature Extraction. If the features extracted 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.

Binary representation that is commonly used and only count the presence or absence of a word in the document. The process that is commonly used is the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a popular Feature Extraction method which reflects the relevance of a word in a particular document among the corpus. It is a numeric statistical approach which is often considered as as weighing factor in Information Retrieval and Text Mining and its vale is directly proportional to the number of times a word appears in particular document. Denote a term by ‘t’, a document by ‘d’ and a corpus by ‘D’, the Term Frequency TF (t, d) is defined as the number of times the term ‘t’ appears in document ‘d’ while Document Frequency DF(t, D) is defined as the number of documents that contains the term ‘t’.

However, some frequent terms may not provide any

relevance for the task of feature extraction and the weight of such terms should be diminished. For this, the 'Inverse Document Frequency ' approach is used to distinguish relevant and non-relevant keywords which results in minimization of weight of frequently occurring non-relevant terms and maximization of weight for terms that occur rarely. The idf gives the measure of specificity of a term which can be expressed as the inverse function of the number of documents in which the term occurs [8].

V. SENTIMENT CLASSIFICATION

Sentiment Analysis is a current research area in text mining. It is the stem of natural language processing or machine learning methods. The main goal is to connect on Twitter and search for the tweets that contain a particular keyword and then evaluate the polarity of the tweets as positive or negative or neutral [9]. In this paper, the keywords of Business and Technology are collected from Twitter using Twitter API and the extracted raw data are preprocessed using Natural Language Toolkit techniques. The sentiments of the online tweets are evaluated based on features' weights. The system uses Term Frequency-Inverse Document Frequency (TF-IDF) to select the features and Naïve Bayes Classifier is used for training and testing the features and also evaluating the sentiment polarity. The proposed system is implemented using Python.

A. *Different Classes of Sentiment Analysis*

There are three classes of sentiments, i.e. positive, negative and neutral sentiments. The positive sentiments that refer to positive attitude of the speaker about the text. Emotions with positive sentiments reflect good, raise, grow etc. In case of Business and Technology sector, if the positive sentiments about the Business and Technology are more, it means these Business and Technology sectors are developing.

The negative sentiments refer to negative attitude of the speaker about the text. In case of Business and Technology sector, if the negative sentiments are more, it means these sectors are undeveloped.

The neutral sentiments that no emotions are reflected about the text. It is neither preferred nor neglected. Although this class doesn't imply anything, it is very important for better distinction between positive and negative classes.

B. *Sentiment Analysis Techniques*

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. Most used supervised algorithms as Support Vector Machines (SVM), Naïve Bayes Classifier and Maximum Entropy Classifier. The system performs Sentiment Analysis by using supervised technique, Naïve Bayes Classifier [9].

C. *Naïve Bayes Classifier*

Naïve Bayes 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. Naïve Bayes Classifier that is based on the popular Bayes' probability theorem. It is used to predict the probability for a given words 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 Business and Technology [9].

First, Naïve Bayes Classifier computes the prior probability.

$$P(C) = N_c/N \quad (1)$$

where, P(C) is the number of documents classified into the class category divided by the total number of documents. For positive class, the number of positive documents are divided by the total number of documents.

Second, Naïve Bayes Classifier computes the conditional probability / Likelihood of each word attribute.

$$P(w|c) = \text{count}(w,c) + 1 \,/\, \text{count}(c) + |V| \quad (2)$$

where, P(w|c) is the Likelihood where w is the word attribute and c is the class, count (w,c) is the total count of word attribute occurs in c class and +1 is Laplace Smoothing, count (c) is the total count of word attribute in a particular class occurs in training set and |V| is the vocabulary.

Third, Naïve Bayes Classifier computes the posterior probability.

$$C_{MAP} = \text{argmax}\ P(x_1, x_2, \ldots, x_n)\ P(c) \quad (3)$$

Finally, Naïve Bayes Classifier determines the class.

D. *Accuracy, Precision, Recall*

Accuracy is the performance evaluation parameter for the system. Two other useful metrics are precision and recall that can provide much greater insight into the performance characteristics of classifier.

Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives.

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases [10].

The system computes the accuracy of Naïve Bayes Classifier. The system measures precision and recall of Naïve Bayes Classifier by using NLTK metrics module that provides functions for calculating these metrics.

```
*****************************************************
Sentiment Analysis Results for Business
Positive tweets percentage: 29 %
Negative tweets percentage: 10 %
Neutral tweets percentage: 60 %
Accuracy: 0.978
Positive Precision: 0.964028776978
Positive Recall: 0.978102189781
Negative Precision: 0.981132075472
Negative Recall: 0.928571428571
Neutral Precision: 0.983766233766
Neutral Recall: 0.986970684039
```

Fig.3 System Performance for Business Data

```
*****************************************************
Sentiment Analysis Results for Technology Data
Positive tweets percentage: 19 %
Negative tweets percentage: 6 %
Neutral tweets percentage: 73 %
Accuracy: 0.983333333333
Positive Precision: 0.972222222222
Positive Recall: 0.985915492958
Negative Precision: 0.972222222222
Negative Recall: 0.921052631579
Neutral Precision: 0.989583333333
Neutral Recall: 0.994764397906
```

Fig.4 System Performance for Technology Data

In Fig.3, the system calculates accuracy, precision and recall of business data on the amount of input business data 8000 instances that are trained and testing data 2000 instances. In Fig.4, the system calculates accuracy, precision and recall of Technology data on the amount of input technology data 7864 instances and testing data 2000 instances.

VI. EXPERIMENTAL RESULTS

The experimental results are presented based on extracting the particular positive, negative and neutral keywords of Business and Technology in Korea, Vietnam and Myanmar. The language is as English using Twitter Streaming API. The experimental results are shown as bar graph.

A number of Business and Technology tweets from Korea, Vietnam and Myanmar are collected and stored them in database. These tweets are shown in Fig5.

```
"RT @business: Pentagon spokesman says Afghanistan bombing was the first time the GE
"RT @Jamal_Mousavi: 2 local entrepreneurs in #Helmand turned their homemade \"cooler
"RT @business: U.S. engagement in Afghanistan has deteriorated to a new low https://t.c
"RT @gerrystone: Women from Rwanda and Afghanistan learn how to achieve business su
"RT @business: U.S. engagement in Afghanistan has deteriorated to a new low https://t.c
"RT @gerrystone: Women from Rwanda and Afghanistan learn how to achieve business su
"RT @business: Pentagon spokesman says Afghanistan bombing was the first time the GE
"RT @Jamal_Mousavi: 2 local entrepreneurs in #Helmand turned their homemade \"cooler
"RT @business: U.S. engagement in Afghanistan has deteriorated to a new low https://t.c
"RT @gerrystone: Women from Rwanda and Afghanistan learn how to achieve business su
"RT @business: U.S. engagement in Afghanistan has deteriorated to a new low https://t.c
"RT @gerrystone: Women from Rwanda and Afghanistan learn how to achieve business su
"RT @business: Pentagon spokesman says Afghanistan bombing was the first time the GE
```

Fig.5 Extracting Tweets from Twitter

```
"rt AT_USER u.s. engagement in afghanistan has deteriorated to a new low URL URL '
"rt AT_USER the bomb dropped today in the middle of nowhere, afghanistan, cost $314,000,000. URL "
"rt AT_USER women from rwanda and afghanistan learn how to achieve business success with bofa partner ieew. URL "
"rt AT_USER i think afghanistan is a bit closer, given tribalism (prans or gangs in crime and/ or business), na\x85 URL "
"rt AT_USER u.s. engagement in afghanistan has deteriorated to a new low URL URL '
"rt AT_USER the bomb dropped today in the middle of nowhere, afghanistan, cost $314,000,000. URL "
"rt AT_USER women from rwanda and afghanistan learn how to achieve business success with bofa partner ieew. URL "
"rt AT_USER i think afghanistan is a bit closer, given tribalism (prans or gangs in crime and/ or business), na\x85 URL "
"rt AT_USER u.s. engagement in afghanistan has deteriorated to a new low URL URL '
"rt AT_USER the bomb dropped today in the middle of nowhere, afghanistan, cost $314,000,000. URL "
"rt AT_USER women from rwanda and afghanistan learn how to achieve business success with bofa partner ieew. URL "
"rt AT_USER i think afghanistan is a bit closer, given tribalism (prans or gangs in crime and/ or business), na\x85 URL "
"rt AT_USER u.s. engagement in afghanistan has deteriorated to a new low URL URL '
"rt AT_USER the bomb dropped today in the middle of nowhere, afghanistan, cost $314,000,000. URL "
"rt AT_USER women from rwanda and afghanistan learn how to achieve business success with bofa partner ieew. URL "
"rt AT_USER i think afghanistan is a bit closer, given tribalism (prans or gangs in crime and/ or business), na\x85 URL "
"rt AT_USER u.s. engagement in afghanistan has deteriorated to a new low URL URL '
```

Fig.6 Preprocessed Data of Tweets

```
Most Informative Features
        contains(setup) = True          negati : neutra =      9.7 : 1.0
         contains(firm) = True          negati : positi =      9.3 : 1.0
      contains(seeking) = True          negati : neutra =      6.9 : 1.0
        contains(jafza) = True          negati : neutra =      6.9 : 1.0
   contains(assistance) = True          negati : neutra =      6.9 : 1.0
    contains(formation) = True          negati : neutra =      5.8 : 1.0
      contains(company) = True          negati : neutra =      5.3 : 1.0
      contains(bahrain) = True          neutra : positi =      5.3 : 1.0
      contains(support) = True          positi : neutra =      5.2 : 1.0
```

Fig.7 Extracting Features of Tweets

The raw data of the twitter are preprocessed using NLTK library techniques as shown in Fig. 6. The preprocessed tweets are appended for feature extraction process. The score words are evaluated based on Term Frequency-Inverse Document Frequency (TF-IDF). The features which contain highest score that is the highest feature of the tweets. It indicates that these features are the repeated words in the system. After feature extraction, Naïve Bayes Classifier is trained.

In Fig.7, the experiments are conducted using all words as features in the tweets. For Business, the input data 8000 instances that is trained and tested on 2000 instances. For Technology, the input data is 7864 instances and tested on 2000 instances.

Fig.8, Fig.9 and Fig.10 shows the rate of change of Business and Technology in Korea, Vietnam and Myanmar on real testing twitter data. The input data is 8000 instances that are trained and testing data is real-time twitter data.

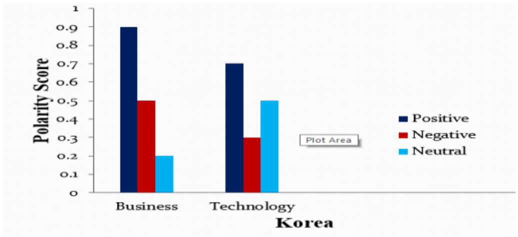

Fig.8 Graphical Analysis of Korea

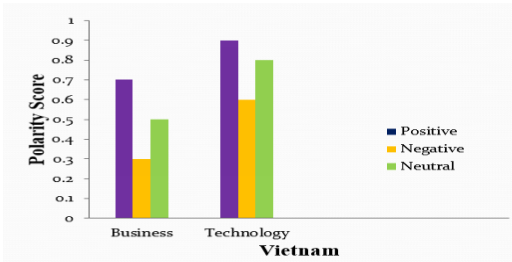

Fig.9 Graphical Analysis of Vietnam

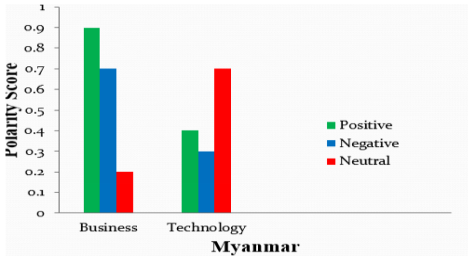

Fig.10 Graphical Analysis of Myanmar

VII. RELATED WORK

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification (Pang and Lee 2008) to learning the polarity of words and phrases (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)).

Text Classification using machine learning is a well studied field (Manning and Schuetze 1999). (Pang and Lee 2002) researched the effects of various machine learning techniques (Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM)) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model. (Pang and Lee 2004) present a hierarchical scheme in which text

is first classified as containing sentiment, and then classified as positive or negative.

The system uses machine learning algorithm such as Naïve Bayes. The system also uses NLTK metrics module to perform the accuracy, precision and recall. In the system, the input data is 8000 for business and 7864 for technology that are trained. Testing data can be crawled from real time twitter data of business and technology.

VIII. CONCLUSION

In this paper, the study of Transformation, Tokenization, Filtering, Normalization and Feature Extraction have been presented. The system design also presented. We have classified tweets by using Naïve Bayes Classifier. Twitter is an excellent initial point for social media analysis. People share their opinions through Twitter to the general public. One of the very common analyses which can perform on a large number of tweets is sentiment analysis. In the proposed system, tweets are crawled using Twitter streaming API from twitter. The collected tweets are preprocessed using Natural Language Toolkit techniques. The features of the tweets are selected based on TF-IDF and Naïve Bayes Classifier is used to classify the tweets as positive or negative or neutral. The proposed system is implemented using Python. The proposed system would be useful for the selected ASEAN countries to know their citizens' social media usage behavior. The proposed system is to study how to perform machine learning model in the case of social media mining mainly on sentiment analysis.

The system can easily check the selected ASEAN countries' conditions in these business and technology fields according the results. The system provides analytical results of Business and Technology data for Economists and Technology Department' needs.

ACKNOWLEDGMENT

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 grateful to Dr. Yi YiMyint, Assistant Lecturer, University of Technology (Yatanarpon Cyber City), for giving me valuable advices. I am also grateful to Dr. PhyuThwe, Assistant Lecturer, University of Technology (Yatanarpon Cyber City), for giving me valuable advices. I am also grateful to Dr. NawNaw, Assistant Lecturer, 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] M. Vadivukarassi, N. Puviarasam and P. Aruna, "Sentiment Analysis of Tweets Using Naïve Bayes Algorithm", World Applied Sciences Journal 35 (1):54-59, 2017.

[2] R.A.S.C. Jayasanka,M.D.T.Madhushani, E.R. Marcus, I.A.A.U. Aberathne and S.C. Premaratne, "Sentiment Analysis for Social Media".

[3] Daniel Jurafsky, James H. Martin, "Naïve Bayes and Sentiment Classification".

[4] GiulioAngiani, Laura Ferrari, TomasoFontanini, Paolo Fornacciari, EleonoraIotto, Federico Magliani, and Stefano Manicardi, "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter".

[5] David Osimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis".

[6] http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html

[7] Lucas C., "Sentiment Analysis a Multimodal Approach", Department of Computing, Imperial College London, September 2011.

[8] Haddi, E., Liu, X.,Shi, Y.,"The role of text pre-processing in sentiment analysis", Procedia Computer Science 17, 26-32 (2013).

[9] Neethu, M.S., Rajasree, R.,"Sentiment Analysis in Twitter Using Machine Learning Techniques" Department of Computer Science, 2013.

[10] https:// streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall