# Feature Extraction Method for Aspect-Based Sentiment Analysis

Win Lei Kay Khine, Nyein Thwet Thwet Aung, Thet Thet Zin
*University of Information Technology, Yangon, Myanmar*
*winleikkhine@uit.edu.mm, nyeinthwet@uit.edu.mm, thetthetzin@uit.edu.mm*

## Abstract

*In our daily life, we take opinions of our friends and we are influenced in decision making process. Opinion is the view or the judgment about something. Opinion Mining (OM) or Sentiment Analysis (SA) is the computational analysis of public's opinion, emotion, sentiments, and attitude toward entities and their attributes expressed in written text. These entities may be products, services, organizations, individuals, events, issues, or topics. In sentiment analysis, formal and informal opinion text like product reviews, news articles, tweets, forum discussions, blogs, and Facebook posts are also applicable to all domains. The main purpose of sentiment analysis is to extract the main opinions, on which the decision can be made very right. Paper intends to classify sentiment polarity on product review datasets by using Mutual Information as a feature selection method. Because product reviews are highly focused and they are opinion rich. After the feature selection, we aim to classify the extracted features with Naïve Bayes, SVM and Maximum Entropy to get the accurate sentiment polarity.*

**Key Words-** Sentiment Analysis, Opinion Mining, Feature Selection, Feature Extraction, Aspect-Based Sentiment Analysis

## 1. Introduction

Sentiment analysis has been an active research field of natural language processing (NLP) for finding sentiments of people for a specific product, services, films, news, organizations and so on. It aims to mine this information to find out the popular sentiment about any product and its associated features. The main goal is to determine whether the expressed opinion in the text is positive, negative or neutral. Neutral opinion usually means "no opinion". The Sentiment analysis can be done by classifying the text on document level, sentence level and feature or aspect level. Document level only gives the whole document polarity; it considers the opinion of only one opinion holder. Sentence level deals with multiple sentences, but can distinguish only subjectivity or objectivity of sentences which is not equivalent to derive sentiment of particular feature. It can only determine sentence's polarity that is positive, negative and neutral. If there are more than one object and feature with complex sentences then the above levels failed to get accurate sentiment analysis. Neither document-level nor sentence-level analyses discover what people like and dislike exactly. For it, Feature level sentiment analysis is adopted not only dealing with the polarity classification of particular features but also handling many problems as coordinating conjunctive sentences and comparative sentences which are quite tough to extract the opinion. This level of analysis was earlier called feature-based, which is now called aspect-based sentiment analysis. The goal of this level is to discover sentiment on entities and/or their aspects. In this paper, we intend to use aspect-based sentiment analysis because sentiment classification on both document and sentence levels are not sufficient: they do not tell what people like or dislike. The main objectives are to extract people's opinion about a particular product and produce a feature-based opinion summary of multiple reviews.

## 2. Related Works

In the literature [4] proposed a novel feature reduction method using standard deviation based on more variation or dispersion of features in feature space. They used three popular classifiers, namely: Naïve Bayes, Maximum Entropy and Support Vector Machine for sentiment classification and ensemble of these classifiers. They then compared their proposed method with other feature reduction methods used on book and music reviews.

According to Tuba Parlar, Selma Ayse Özel [5], they proposed a new feature selection method called "Query Expansion Ranking" that is based on query expansion term weighting methods. They compared their Query Expansion Ranking with Chi Square method and Document Frequency Difference (DFD). Experiments are conducted on four Turkish product review datasets that are Book, DVDs, electronics and kitchen appliances reviews by using a supervised machine learning classification method, namely Naïve Bayes Multinomial classifer. Finally, their new feature selector improves classification accuracy better than Chi Square and Document Frequency Difference.

Bagheri et al. [7] proposed a sentiment classification model for the document level of cell phone reviews. They presented a new feature selection approach MMI (Modified Mutual Information) based on the Mutual Information method and applied three feature selection approaches, MI, Term Frequency Variance (TFV) and MMI with the Naïve Bayes learning algorithm. To test their method, they compiled a dataset of 829 online customer reviews in Persian language from different brands of cell phone products including Nokia, Apple, Samsung, Sony, LG, Motorola, Huawei and HTC. Their proposed approach, MMI in overall can reach to 85% of F-score classification correctly and it overcomes other techniques. (MI-58% and TFV-84%)

J. Ashok Kumar and S. Abirami [8] implemented the OMSA approach and analysed the results by using a single dataset for different feature extraction or selection techniques namely single word, Multiword, Document Level, Phrase Level, Tf-idf single word and Tf-idf Multiword. Experimental procedure has been carried out with an extension of the OMSA approach. In this approach, the Polarity Classification Algorithm (PCA) and evaluation procedure is applied to verify the accuracy. The evaluation procedure is tested with four different datasets.

According to my review, many researchers did sentiment analysis on many fields using various methods. This paper intends to do sentiment analysis with high accuracy. Accuracy will increase with excellent features for relative domain and classification methods. Therefore, this paper aims feature selection and classification for sentiment analysis.

## 3. Theory Background

Sentiment Analysis has been a good research area from a long time. It is a very challenging task in these days, which is a great requirement in every field as in Political field, Marketing field and in social field mainly. Many techniques can be found for sentiment analysis. It can be done based on Document level, Sentence level and Aspect or Feature level. But the first two levels didn’t consider object features that have been commented in a sentence. So the aspect/feature level sentiment analysis is more appropriate compared to both levels. In this paper, feature extraction method for product review dataset will be proposed. There are so many feature selection techniques for Feature Extraction. These are document frequency, Information Gain, Chi-Square, TF-IDF, Information Mutual, Standard deviation, etc... . Here are some of the feature selection methods:

### 3.1. Feature Extraction

There are so many feature selection methods like uni-gram, bigrams, Information Gain (IG), Chi-square (CHI), Document Frequency (DF), Information Mutual (IM) and so on. The main task of feature selection and feature reduction/extraction is reduction dimension in feature space. It causes the removal of irrelevant features and results in the following outcomes: more efficient categories; easier analysis of sentiment after reduction; visualization of results; and there may be a better perception of low dimension.

**3.1.1. Document Frequency.** In the Document Frequency (DF) method, features are ordered by document frequency for each feature in a whole document. This method is the simplest measure for feature reduction and has a linear time complexity capable of scaling a large dataset.

**3.1.2. Information gain (IG).** Information gain is the most commonly used feature selection method in the field of machine learning. It calculates the relevance of a feature for prediction of sentiment of review by analyzing the presence or absence of a feature in a document.

**3.1.3. Chi-square (CHI).** Chi square measures the lack of independence between a feature and a class. If a feature *f* in the related class *c* has a low score, it can be less informative, so it can be removed. [5]

The challenges in feature extraction in sentiment analysis are facing different issues like large feature space problems, redundancy, domain dependency, difficulty in implicit feature identification. After the feature extraction, we will use classifiers to evaluate the results. For the classification, the most commonly used algorithms are Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees and Maximum Entropy (ME) classifiers are used to classify the polarity of a given text of the feature.

### 3.2. Classification Techniques

**3.2.1. Naïve Bayes.** The Bayesian classification is a statistical method underlying a probabilistic model and supervised learning algorithms. Naive Bayes (NB) uses a features vector matrix to determine a document depending upon polarity classes (i.e. positive and negative classes) by probability. It attaches a document to the relevant class with the highest probability.

**3.2.2. Support Vector Machine.** Support Vector machine (SVM) is a most popular algorithm that can classify data as either linear or nonlinear. It can also map input data to high dimensional feature spaces, in addition to classifiers' SVM support regression, binary and multiclass classification respectively. The support vector can be either linear or nonlinear.

**3.2.3. Decision Tree.** It is a tree in which internal nodes are represented by features, edges represent tests to be done at feature weights and leaf nodes represent categories which results from above tests. It categorizes a document by starting at the tree root and moving successfully downward via the branches (whose conditions are satisfied by the document) until a leaf node is reached. The document is then classified in the category that labels the leaf node. Decision Trees have been used in many applications in speech and language processing [6].

**3.2.4. Maximum Entropy.** Maximum Entropy (ME) classifier is one of the machine learning methods used for natural language processing applications, as it is implemented using a multinomial logit model as the classifier rule. ME is a kind of statistical inference that can be used to estimate any probability distributions on the partial knowledge.

## 4. Proposed System

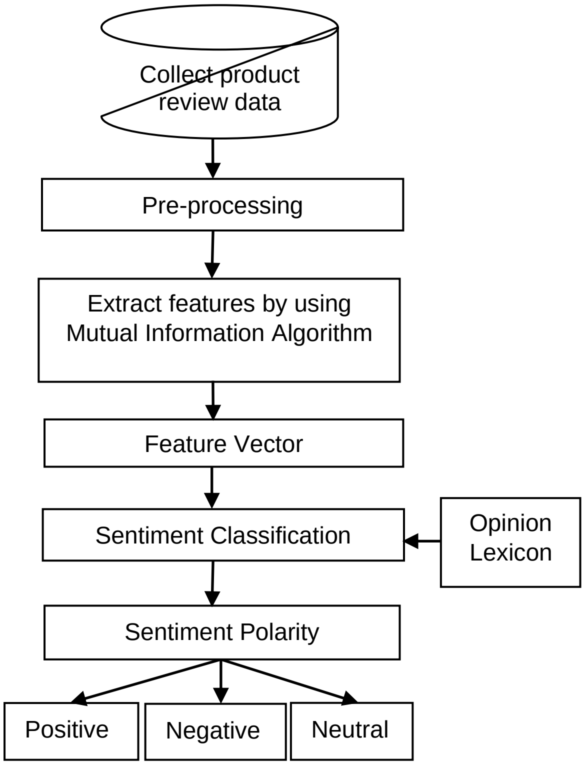

**Figure 1. Proposed system architecture**

Figure 1 gives overall system architecture for sentiment analysis. Firstly, we collect the products' review data from Amazon.com. This data may be noisy and hence we need to be pre-processed. Secondly, the pre-processing step is used for the removal of stop words and special characters. After the pre-processing phase, we extract the relevance features by using the "Mutual Information" feature selection method. This paper intends to modify the Mutual Information Algorithm to get the right features for classification phase. Feature selection tries to find a subset of features that are important and valuable for the classification task. Finally, the classifier is applied to test whether it is able to detect the right features or give the right classification. This paper intends to get the higher accuracy for selected features that are extracted by Mutual Information than by using other methods.

### 4.1. Datasets

The data collection process is the first stage in this step. In this stage, a freely available dataset will be used for preprocessing the data. The collected dataset serves as input to the pre-processing stage and further the feature selection or extraction method has been applied over it to classify the polarity into positive, negative, and neutral. In this step, the datasets were downloaded from [9]. It contains 28000 reviews for book information, for example. The attributes in this book dataset include RowID, ProductID, Publisher, ReleaseDate, ProductDimensions, ShippingWeight, Language, NumPages, Type Edition and FullDesc. Other datasets are from [10].

### 4.2. Preprocessing phase

The preprocessing phase involves the following steps:

**4.2.1. Tokenizing.** In the pre-processing phase, reviews are scanned to extract tokens consisting of words and numbers.

**4.2.2. Removal of Stop Words.** Stop words do not have any sentiment information, so we need to remove stop words in the pre-processing step i.e. words such as she, he, at, about, of, in, on, the, etc…

**4.2.3. POS Tagging.** POS tagging assigns a tag to each word in a text and classifies a word to a specific morphological category such as noun, verb, adjective, adverb, etc.

**4.2.4. Stemming.** The stemming process converts all the inflected words present in the text into a root form called a stem. For example, ‘automatic,’ ‘automate,’ and ‘automation’ are each converted into the stem ‘automat.’

### 4.3. Feature Selection and Extraction

The aim of feature selection is to remove the irrelevant and redundant features, so it can produce better prediction accuracy and better efficiency. In this paper, we will use the Mutual Information (MI) as a feature extraction. MI is one of the most effective approaches for optimal feature extraction. It measures the mutual dependence of two or more variables. In this context, the feature extraction processing a feature vector from the data which have the largest dependency. MI selects features that are not uniformly distributed among the sentiment classes because they are informative of their classes. And we can see that MI giving more importance to the rare term.

### 4.4. Sentiment Classification

In this paper, Naïve Bayes, Support Vector Machine and Maximum Entropy will be used as a classifier to get the accurate the sentiment polarity with high accuracy. The experimental will be evaluated by precision, recall and F-Score methods.

## 5. Conclusion

Sentiment analysis is one of the widest areas of research and improvement with techniques and classification approaches. In this paper, feature extraction in sentiment analysis will be proposed. Feature extraction in sentiment analysis is now becoming an active area of research. Feature selection methods can help to improve the classification performance of sentiment analysis in terms of both accuracy and run time. For the future work, we will try to use the Mutual Information (MI) as a feature selection. After that, Naïve Bayes, Support Vector Machine and Maximum Entropy are used to compare the results.

## 6. References

[1] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, May 2012.

[2] R.Rajput, A.K.Solanki, “Review of Sentiment Analysis Methods using Lexicon Based Approach”, International Journal of Computer Science and Mobile Computing, India, 2016, pp. 159-166.

[3] T.Shaikh, Dr.D.Deshpande, “Feature Selection Methods in Sentiment Classification of Amazon Product Reviews”, International Journal of Computer Trends and Technology (IJCTT)-Volume 36 Number 4, ISSN:2331-2803, India, June 2016, pp 225-230.

[4] A.Yousefpour, R.Ibrahim, H.Nuzly, A.Hamed, “A Novel Feature Reduction Method in Sentiment Analysis”, International Journal of Innovative Computing 4, 2014, pp 34-40.

[5] T.Parlar, S.A.Ozel, “A New Feature Selection Method for Sentiment Analysis of Turkish Reviews”, IEEE, 2016.

[6] N.S.Joshi, S.A.Itkat, “A Survey on Feature Level Sentiment Analysis”, International Journal of Computer Science and Information Technologies, Vol.5(4) India, 2014, pp 5422-5425.

[7] A.Bagheri, M.Saraee, F.de Jong, “Sentiment Classification in Persian: Introducing a Mutual Information-based Method for Feature Selection”, IEEE, 2013.

[8] J. A. Kumar, S. Abirami, “An Experimental Study of Feature Extraction Techniques in Opinion Mining”, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.4, No.1, India, February 2015, pp 15-21.

[9] http://liu.cs.uic.edu/download/data/.

[10] http://jmcauley.ucsd.edu/data/amazon/links.html.