ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Sentiment Analysis Using Support Vector Machine

Ms. Gaurangi Patil[1], Ms. Varsha Galande[2], Mr. Vedant Kekan[3], Ms. Kalpana Dange[4]

UG Student, Dept of Computer, VIIT Engineering College, Pune, India[1]
UG Student, Dept of Computer, VIIT Engineering College, Pune, India[2]
UG Student, Dept of Computer, VIIT Engineering College, Pune, India[3]
UG Student, Dept of Computer, VIIT Engineering College, Pune, India[4]

Abstract: Sentiment analysis is a subfield of NLP concerned with the determination of opinion and subjectivity in a text, which has many applications. In this paper we will be studying about classifiers for sentiment analysis of user opinion towards political candidates through comments and tweets sing Support Vector Machine (SVM),in the manner of the Pang, Lee and Vaithyanathan, which was the first research paper on this topic. The goal is to develop a classifier that performs sentiment analysis, by labeling the users comment to positive or negative. From which we can classify text into classes of interest.

Keywords**:** Opining mining, Support vectormachine, TF-IDF,Sentiment Classification.

I. INTRODUCTION

Sentiment is basically a thought, view based on emotion instead of reason. It is a kind of subjective impression and not facts, also termed as the expression of sensitive feeling in art and literature. Sentiment Analysis also referred as Opinion Mining is a *Natural Language Processing* and *Information Extraction* task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents Sentiment. Sentiment analysis is the computational technique for extracting, classifying, understanding and determining the opinions expressed in various contents. Itattempts to identify the opinion / sentiment that hold towards an object. It makes use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text. Generally speaking, sentiment analysis aims to determine the state of mind of a speaker or a writer with respect to some topic or the overall tonality of a document. Word of mouth (WOM) is the process of conveying information from person to person and plays a major role in customer decision making regarding services/products. In commercial situations, WOM involves consumers sharing attitudes, opinions, products, or services with other people.WOM communication functions based on social networking. In recent years, the massive increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is an immense repository of structured and unstructured data. The analysis of this data to extract dormant public opinion and sentiment is a challenging task. Sentiment analysis can be of use in online product reviews, recommendations, blogs, user opinion towards political candidates,etc.The next section covers therelated work done on sentiment analysis by various researchers. After that we propose a technique for sentiment analysis using SVM since SVM have been proven as one of the most powerful learning algorithms for text categorization [9].

II. RELATED WORK

Many researchers are trying to combine the text mining and sentiment analysis as next generation discipline [3] [6]. In sentiment analysis document – level classification is most promising topic [9]. In Sentiment classification there are four different levels of sentiment analysis - sentence level, document level, phrase level, word level. Subjectivity and sentiment are both relevant properties of language. Subjectivity refers to linguistic expression of somebody's opinion,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

beliefs, speculations. Main task of subjectivity is to classify the contents in objective or subjective. Figure 1 shows the sentiment analysis and subjectivity analysis classification.

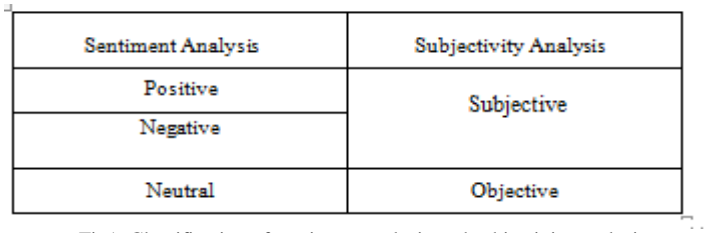

Fig1. Classification of sentiment analysis and subjectivity analysis

Phrase – level sentiment analysis is proposed by Theresa et al [4], which determines whether the given expression is neutral or polar, based on which the respective polarity of expression is decided .The approach automatically identify the contextual polarity for a huge subset. Sentiment expressions that are actually better than baseline but it require long time for calculations. Yi et al [8] proposed sentence – level polarity categorization which aims to classify positive and negative sentiment for each sentence. Phrase-level categorization can also be nested within sentence level classification in order to capture multiple sentiments that may be present within single sentence. But the accuracy of predicting the sentiment is not relevant [1]. Hence the new approach comes into picture. Pang & Lee [3] also came up with the approach based on sentence as being subjective or objective and then perform the sentiment classification on the subjective portion of sentence. But the results proved that it is not sufficient for predicting the sentiment of entities. Turney [5] proposed most challenging and effective model for sentiment classification which is based on document – level which involves two approaches: Term counting and machine learning approach [1].Term counting approach involves deriving a sentiment measure by calculating the positive and negative terms. In [3] authors propose machine learning approaches that molded again the sentiment classification problem as a statistical classification task. As compared to term-counting approaches, machine learning approaches usually achieve better performance, and have been adapted to more intricate scenarios, such as domain adaptation, multi-domain learning and semi-supervised learning for sentiment classification. Whitelaw et al [6] propose an approach considering adjectival expressions a crucial indication of the sentiment polarity in textual reviews. This approach is mainly based on extracting and analyzing the most appraisal words or group of words such as “very good” or “very bad” etc. Wang et al [2] proposed supervised learning methods have been popularly used and proven its effectiveness in sentiment classification. It is highly depend on large amount of labeled data which results in time consuming and also expensive one. Many semi-supervised learning methods are proposed to overcome the problem of supervised learning method .Semi-supervised methods requires small scale of labeled data along with larger amount of unlabelled data. Vapnik [6] proposed Support Vector Machine (SVM), that belongs to supervised learning method which classifythe data into two categories by constructing the N-dimensional hyper plane. SVM [7] uses g(x) as the discriminate function,

$g(x) = w^T f(x) + b$ (1)

wherew is the weights vector, b is the bias , and f(x) denotes nonlinear mapping from input space to high-dimensional feature space. The parameters w and b are learned automatically on the training dataset following the principle of maximized margin by

$\min \frac{1}{2} W^T W + C\sum_{i=1}^{N} ci$ (2)

where N denotes the slack variables and C denotes the penalty coefficient. Due to the dimension of feature space is quite large in text classification task, the classification problem is always linearly separable [1,4] and therefore linear kernel is commonly used.

III. PROPOSED WORK

An overview of sequential steps and techniques commonly used in sentiment classification approaches, as shown in Figure 1.Parts of speech is a model which aims to classify roles that means according to parts of speech has also been explored. In this model , information is used as part of a feature set which leads to sentiment classification on a dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

The model parts of speech is supposed to be the significant indicator of sentiment expression and which works on subjectivity detection that represents the close relationship between presence of adjectives and sentence subjectivity. But, many experimental results show that using only adjectives as features leads to worse performance.

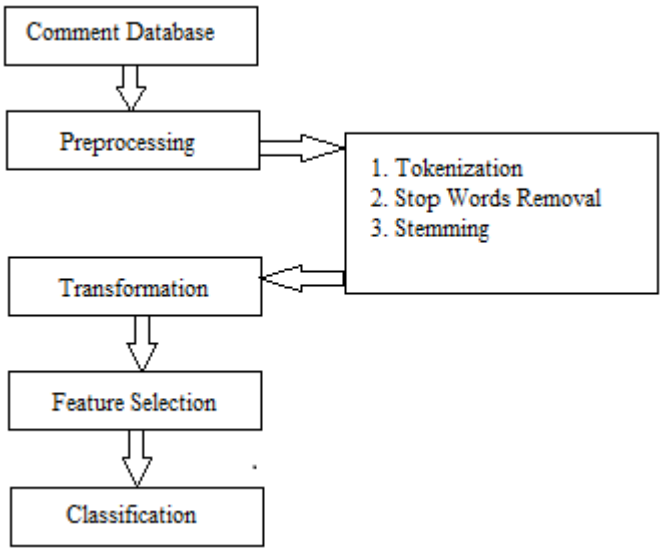

Fig.1 Steps and techniques used in sentiment classification.

A. Text Preprocessing

Pre processing of data is the process of preparing and cleaning the data of dataset for classification. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

1. Tokenization: Given input as character sequence, tokenization is a task of chopping it up into pieces called tokens and at the same time removing certain characters such as punctuation marks . A token is an instance of sequence of characters that are grouped together as a useful semantic unit for processing.

2. Stop word removal: A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include "a", "of", "the", "I", "it", "you", and ”and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words .Hence it is practical to remove those words which appear too often that support no information for the task.

3. Stemming: It is the process for reducing derived words to their stem, or root form. Stemming programs are commonly referred to as stemmers or stemming algorithms. A simple stemmer looks up the inflected form in a lookup table, this kind of approach is simple and fast. The disadvantage is that all inflected forms must be explicitly listed in table.eg. “developed”, “development” , ”developing” are reduced to the stem “develop”.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

B.Transformation

The weight of each word in the corpus is calculated with the help of TF-IDF, so that it is easy to determine what words in the corpus of documents might be more favorable to use in a further processing. TF-IDF calculates [9] values for each word in a document defined as below –

$w_{d} = f_{w,d} * \log(|D|f_{w,D})$

D is collection of documents ,w represents words, d is individual document belongs to D,|D| is size of corpus, $f_{w,d}$is number of times w appears in d,$f_{w,D}$is number of documents in which w occurs in D.

C.Feature Selection

Feature Selection is used to make classifiers more efficient by reducing the amount of data to be analyzed as well as identifying relevant features to be considered in classification process. Ideally, feature selection stage will refine features, which are input into a classification / learning process.

- Identify the parts of corpus to contribute to positive and negative sentiment.
- Join these parts of corpus in such a way that the document falls into one of these polar categories.

D. Classification

Goal of text classification is to classify data into predefined classes. Here they are positive and negative classes. Text classification is supervised learning problem.

First step in text classification is transforming document which is in string format into format suitable for learning algorithm and classification task. In information retrieval it is found that word stem works well as representation unit. This leads to attributed value representation of text. Each word corresponds to feature with, number of times word occurs in document, as its value. Words are considered as features only if they are not stop words (like “and”, “or”, etc). Scaling the dimension of feature with IDF improves the performance[12].

SVM- Support vector machines are universal learners[12]. Remarkable property of SVM is that their ability to learn can be independent of dimensionality of feature space. SVM measures the complexity of Hypothesis based on margin that separates the plane and not number of features[12].

SVM learning Algorithms for Text Categorization -

SVM has defined input and output format. Input is a vector space and output is 0 or 1 (positive/negative).

Text document in original form are not suitable for learning. They are transformed into format which matches into input of machine learning algorithm input. For this preprocessing on text documents is carried out.Then we carryout transformation.Each word will correspond to one dimension and identical words to same dimension. As mentioned before we will see TF-IDF for this purpose. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings.SVM has been proved one of the powerful learning algorithm for text categorization[12].

SVM Benefits]-

1. High Dimension Input Space - while text classification we have to deal with many features (may be more then 1000). Since SVM uses over fitting protection[12], which does not depend on number of features so they have ability to handle large number of features.
2. Document Vector Space - despite the high dimensionality of the representation, each of the document vectors contain only a few non-zero element[12].More Text Categorization problems are linearly separable[12].

SVM Characteristics-

1. ML algorithms typically use a vector-space (attribute-value) [10]representation of examples, mostly the attributes correspond to words. However word-pairs or the position of a word in the text may have considerable information, and practically infinitely many features can be constructed which can enhance classification accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

2. Categories are binary, but generally documents are not assigned so precisely. Often a document D is said to belong a little to category X1 and a bit to category X2, but it does not fit well into any of the two. It probably would require a new category, as it is not similar to any of the documents seen before.
3. Number of words increase if we increase the number of documents. Heap's law[11] describes how the number of distinct words increase if number of document increases.
4. Representations use words as they are in texts. However, words may have different meanings, and different words may have the same meaning. The proper meaning of a word can be determined by its context i.e. each word influences the meaning of its context. However, the usual (computationally practical) representation neglect the order of the words. Task of SVM is to learn and generalize the input-output mapping. In case of text categorization input is set of documents and output is their respective class. Consider spam filter as example input is an email and output is 0 or 1(either spam or no spam)[10].

SVM Evaluation-

Text categorization systems may make mistakes. To compare different text classifiers for deciding which one is better, performance measures are used. Some of these measures the performance on one binary category, others aggregate per-category measures, to give an overall performance. TP, FP, TN, FN are the number of true/false positives/ negatives[13]. The most important per-category measures for binary categories are [13]

- Precision: p=TP/(TP+FP)
- Recall: r= TP/(TP+FN)

The most important averages are: micro-average[13], which counts each document equally important, and macro-average, which counts each category equally important.

IV. CONCLUSION AND FUTURE WORKS

In this paper we discussed techniques for preprocessing and information retrieval with help of TF-IDF, SVM. Also we study Support Vector Machine for text categorization which can be used to find out the polarity of textual comment. From study we can conclude that SVM acknowledge some properties of text like a)High Dimensional feature space b)few irrelevant feature c) sparse instance vector. Performance evaluation is for SVM is also stated in paper which is done using Recall and Precision. Different results show that SVM gives good performance on text categorization as compared with ANN. With ability to generalize high dimensional feature space, SVM eliminates need of feature selection.

REFERENCES

1. Ms. K. Nirmala Devi, Ms. K. Mouthami,Dr. V. MuraliBhaskaran *'Sentiment Analysis and Classification Based on Textual Reviews'*,2012.
2. Li, S., Wang, Z., Zhou, G., & Lee, S.Y.M. , *'Semi-supervised learning for imbalanced sentiment classification'*, In Proceedings of international joint conference on artificial intelligence, pp. 1826–1831,2012.
3. Pang, B., & Lee, L., *'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts'*, In Proceedings of the association for computational linguistics, pp. 271–278, 2004.
4. Theresa Wilson, JanyceWiebe, Paul Hoffmann, *'Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis'*, In Proceedings Of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, pp. 347–354 2004.
5. Turney, *'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews'*, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, 2002.
6. Vapnik, V., *'The Nature of Statistical Learning Theory'*, Springer- Verlag, pp. 863-884, 2000.
7. Yang. Y, X. Liu, A re-examination of text categorization methods, in: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), ACM, New York, NY, USA, pp. 42–49,1999.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

8. Yi, J., Nasukawa, T., Niblack,W., &Bunescu, R. , *'Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques'*, In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), USA, pp. 427– 434,2003.
9. Rodrigo Moraes, Joao Francisco Valiati,Wilson P. GaviaoNeto,*'Document-level sentiment classification : An empirical comparison between SVM and ANN'*,Expert Systems with Applications 40 621-633,2013.
10. HumaLodhi, Craig Saunders, John Shawe-Taylor, NelloCristianini, Chris Watkins: *Text Classification using String Kernels*, The Journal of MachineLearning Research, Volume 2, pp. 419-444,2002.
11. Heaps' law: http://en.wikipedia.org/wiki/Heaps'_law
12. Thorsten Joachims: *Text categorization with support vector machines: learningwith many relevant features*, Proc. of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE,pp. 137-142,1998.
13. FabrizioSebastiani: *Machine learning in automated text categorization*, ACM Computing Surveys (CSUR), Vol. 34 Issue 1, ACM Press, New York, NY, USA, pp. 1-47,2002.