

A Comparative Study of Sentiments Analysis Using Rule Based and Support Vector Machine

Ankush Sharma¹, Aakanksha²

Assistant Professor, Department of C.S.E, Chandigarh University Gharuan, India^{1,2}

Abstract: Sentiment analysis (SA) in Natural language processing deals with sentiment and subjectivity in text. It has many names like subjectivity analysis, opinion mining, and appraisal extraction. In order to get information what other people thinks whether they have positive or negative opinion about a particular subject. For online reviews, like about any movie there will be favorable as well as unfavorable opinions and sentiment analysis is used to get information about the movie if it's good or not. This paper compares the various techniques used for sentiments analysis. The sentiment analysis for English product reviews has been achieved. For Chinese product reviews, feature-level based sentiment analysis is used.

Keywords: sentiments analysis, rule base, support vector machines, various techniques used.

I. INTRODUCTION

Natural language processing (NLP) is a field of computer science with the interactions between computers and human (natural) languages.

The following is a list of some of the most commonly researched tasks in NLP:-

- **Automatic summarization:** Produce a readable summary of a chunk of text.
- **Machine translation:** Automatically translate text from one human language to another.
- **Morphological segmentation:** Separate words into individual morphemes and identify the class of the morphemes.
- **Speech recognition:** Given a sound clip of a person or people speaking, determine the textual representation of the speech.
- **Speech segmentation:** Given a sound clip of a person or people speaking, separate it into words.
- **Topic segmentation and recognition:** Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

A. *Review segment on iPhone to introduce the problem*

“(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

B. *What to mine or extract from this review?*

- (1) Sentences (2), (3) and (4) express positive opinions, while sentences (5), (6) and (7) express negative opinions or emotions.
- (2) The opinion in sentence (2) is on the iPhone as a whole, and the opinions in sentences (3), (4) and (5) are on the “touch screen”, “voice”, “quality” and “battery life” of the iPhone respectively.

- (3) The opinion in sentence (7) is on the price of the iPhone, but the opinion/emotion in sentence (6) is on “me”, not iPhone. This is an important point.

- (4) In an application, the user may be interested in opinions on certain targets or objects, but not on all

The rest of the paper is organized as follows: section 2 describes related works, section 3 provides the process Sentiment and Subjectivity Classification, section 4 presents Rule Based Approach, and section 5 presents the Pre-Processing of text, section 6 presents the classifier (SVM), section 7 presents the techniques used for sentiment analysis, and section 8 concludes the comparisons result.

II. RELATED WORK

Sentiment analysis is how to extract the opinions or sentiments on the product features in the review, which is feature-level sentiment analysis [3].

III. SENTIMENT AND SUBJECTIVITY ANALYSIS

A. DOCUMENTATION-LEVEL SENTIMENT CLASSIFICATION

Given a set of opinionated documents D , it depicts whether each document $d \in D$ expresses a positive or negative opinion (or sentiment) on an object.

B. Sentence-Level Subjectivity and Sentiment Classification

Task: Given a sentence s , two sub-tasks are performed:

- (1) Subjectivity classification: It predicts whether s is a subjective sentence or an objective sentence,
- (2) Sentence-level sentiment classification: If s is subjective, then it expresses a positive or negative opinion.

C. Rule-Based approach

In the Rule-Based approach, rules are to be defined and in this methodology, certain rules are to be form and then the sentiments should be analyzed depending on it.

D. Feature-Based Sentiment Analysis

Document level or at the sentence level is useful in many cases, but they do not provide the necessary detail required for some other applications. To obtain such details, object feature level analysis is required.

The feature-level sentiment analysis task can be divided into two sub tasks:

- A. Identifying product features;
 - B. Determining whether the opinions on the features are positive, negative or neutral.
- E. *Terminology*
1. Sentiment analysis is more widely used in industry.
 2. Both are widely used in academia
 3. But they can be used interchangeably

IV. RULES BASED APPROACH

Rule based approach is used by defining various rules for getting the opinion or it uses lexical rules, created by tokenizing each sentence in every document and then testing each token, or word, for its presence. If the word is there and has with a positive sentiment, a +1 rating was applied to it. Each post starts with a neutral score of zero, and was considered positive if the final polarity score was greater than zero, or negative if the overall score was less than zero. In this, certain rules are to be form and then the sentiments should be analyzed depending on it [5].

The rule based approach result creates the rules by taking:

- Affecting words,
 - Inverted words, and
 - Negation words.
- a) After the output of rule based approach it will check or ask whether the output is correct or not. If the input sentence contains any word which is not present in the database which may help in the analysis of movie review, then such words are to be added to the database.
 - b) This is supervised learning in which the system is trained to learn if any new input is given.
 - c) This approach will always increase the efficiency of the system.

V. PREPROCESSING OF TEXT

In the pre-processing of text, meaningless words are to be removed which are not describing any sentiments. The words like this, it, is, am, are, who, or, etc does not give any idea for analysis of sentiments. So these words must be discarded from the input. E.g., this is a good movie. In this sentence, this, is, a, will get removed as they are not deriving any sentiments. After pre-processing of the text there are three tables for storing the refined input. The first table will be for the words that affect the analysis. These are the word that gives no meaning for deriving sentiments. The second table contains the general words and the third table will consist of the negative words [4].

VI. SUPPORT VECTOR MACHINE(SVM)

Support Vector Machines (SVM) A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task. High dimensional input space: When learning text classifiers, one has to deal with large number of features. Since SVM use over fitting protection, which does not necessarily depend on the number of features, they

have the potential to handle these large feature spaces. Most text categorization problems are linearly separable: All categories are linearly separable and so are many of the Reuters Tasks. The idea of SVMs is to find such linear separators [3].

The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. With increasing amounts of data being generated by businesses and researchers there is a need for fast, accurate and robust algorithms for data analysis. Improvements in databases technology, computing performance and artificial intelligence have contributed to the development of intelligent data analysis. The primary aim of data mining is to discover patterns in the data that lead to better understanding of the data generating process and to useful predictions. Examples of applications of data mining include detecting fraudulent credit card transactions, character recognition in automated zip code reading, and predicting compound activity in drug discovery.

A. Benefits of SVM

1. High-dimensional input space.
2. Few irrelevant features: almost all features contain considerable information. He conjectures that a good classifier should combine many features and that aggressive feature selection may result in a loss of information.
3. Document vectors are sparse: despite the high dimensionality of the representation, each of the document vectors contains only a few non-zero elements.
4. Most text categorization problems are linearly separable.

A support vector machine (SVM) is for a set of related supervised learning methods, it takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which class. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

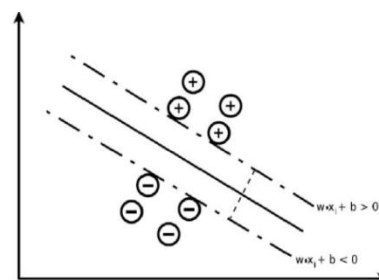


Fig.1. Support Vector Machine [4]

SVM is a machine learning approach. It is used to search a usually very large space of potential hypothesis to determine the one that will best fit the data. SVM generates training models for text classification.

VII. CONSOLIDATED REPORT OF TECHNIQUES USED FOR SENTIMENT ANALYSIS

This is the comparison table which shows the technique used for Sentiment Analysis.

1. In 2010, SentiWS – a Publicly Available German-language Resource for Sentiment Analysis, the approach used is SentiWS(v1.8b) And Approach is PMI.
2. In 2011, Feature-level Sentiment Analysis for Chinese Product Reviews, the approach used is Feature Level Sentiment Analysis And Hybrid Approach.
3. In 2011, Subjectivity and Sentiment Analysis of Modern Standard Arabic, the approach used is Sentence Level SSA.
4. In Jan 2012, Sentiments analysis using Hybrid Approach involving Rule-Based & Support Vector Machines methods, the approach used is Rule Based and Support Vector Machines.
5. In March 2012, “Movie Review analysis using Rule-Based & Support Vector Machines methods, the approach used is Rule Based and Support Vector Machines.

TABLE I. Consolidated Report

S. No	Comparison Technique	Year
1.	SentiWS(v1.8b) And Approach is PMI	2010
2.	Feature Level Sentiment Analysis And Hybrid Approach	2011
3.	Sentence Level SSA	2011
4.	Rule Based and Support Vector Machines	Jan,2012
5.	Rule Based and Support Vector Machines	March,2012

IX. CONCLUSION

In contrast to the rule-based approach, the results of the machine-learning based classifier are significantly better. The benefit of the rule-based approach is that no training material is required. But, a problem for the rule-based approach is to decide for a polarity when the number of positive words equals the number of negative words.

FUTURE SCOPE

SentiWS is work in progress and it will be error prone and will be continually processed by adding missing words and word forms and it removes the words which are ambiguous. It will be extended by introducing a new dimension indicating subjectivity. Also it is necessary to use weighting schemes.

ACKNOWLEDGMENT

The success of this Paper would have been uncertain without the help and guidance of dedicated person. Thus as a token of appreciation of their effort in making it a success, With a deep sense of gratitude I would like to extend thanks to Mr. Ankush Sharma for their kind cooperation in the preparation of this Paper. Last but not the least; we would like to convey our heartfelt gratitude to our beloved parents, without their constant inspiration throughout our career we would not have reached this stage.

REFERENCES

- [1] Robert Remus, Uwe Quasthoff and Gerhard Heyer, “SentiWS – a Publicly Available German-language Resource for Sentiment Analysis”, 2010
- [2] Haiping Zhang, Zhengang Yu, Ming Xu and Yueling Shi, “Feature-level Sentiment Analysis for Chinese Product Reviews”, 2011
- [3] Muhammad Abdul Mageed, Mona T.Diab, Mohammed Korayem, “Subjectivity and Sentiment Analysis of Modern Standard Arabic”, June 19, 2011
- [4] Swati A. Kawathekar, Dr. Manali M. Kshirsagar, “Sentiments analysis using Hybrid Approach involving Rule-Based & Support Vector Machines methods”, IOSR Journal of Engineering (IOSRJEN) Vol. 2 Issue 1, Jan,2012, pp. 055-058.
- [5] Swati A. Kawathekar1, Dr. Manali M. Kshirsagar2, “Movie Review analysis using Rule-Based & Support Vector Machines methods”, IOSR Journal of Engineering Mar. 2012, Vol. 2(3), March. 2012, pp: 389-391.