Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.425 – 430

RESEARCH ARTICLE

Selecting Best Features Using Combined Approach in POS Tagging for Sentiment Analysis

Nikita D. Patel [1], Chetana Chand [2]

[1] [2] Department of Computer Science and engineering & Gujarat Technological University, India

[1] emailtoniki@gmail.com; [2] chetnachand87@gmail.com

Abstract— ***Today E-commerce popularity has made web an excellent source of gathering customer reviews / opinions about a product that they have purchased. The number of customer reviews that a product receives is growing at a very fast rate. Opinion mining from product reviews, forum posts and blogs is an important research topic today with many applications. Customers use the reviews for deciding quality of product to buy. So, opinion mining may be a Decision making process. It means reviews are given to promote or to demote the product. There is need to find how many reviews are positive and how many are negative. So, to find out it features for which classification is going to be performed should be best or optimal. This Paper presents various approaches of classification for sentiment analysis and proposed work is selecting best feature set such as pos tags from reviews which we can easily classify the review of customer. Only features which are giving best decision for analysis have been selected in pre-processing task and Combination of best feature set will be used to classify reviews.***

Keywords— Include Customer Reviews; Opinion mining; Decision making; Pos tags

I. INTRODUCTION

Text classification has been one of the key tools to automatically handle and organize text information for decades. In recent years, with more and more subjective information appearing on the internet, sentiment classification, as a special case of text classification for subjective texts, is becoming a hotspot in many research fields, including natural language processing (NLP), data mining (DM) and information retrieval (IR) [12].The main objective of opinion mining technique is to extract opinion from large amount of data.

Two main research directions are there: subjectivity classification and sentiment classification. Sentiment classification is almost same as text classification. Bag of words (BOW) are used as a input features to classification algorithm, which are Naïve Bays (NB), Support Vector Machine (SVM), Maximum Entropy (ME).These algorithms are machine learning . The effectiveness of machine learning techniques when applied to sentiment classification tasks is evaluated in the pioneering research by Pang et al. [2]. The experimental results on the movie-review dataset produced via NB, ME, and SVM are substantially better than those results

obtained through human generated baselines. But their performance is not as remarkable as when they are used in topical text classification.

Only Bag of word features are not sufficient for sentiment classification, so higher-order n-gram features, POS based features, word pairs and dependency relations, have been exploited to improve sentiment classification performance. In proposed paper best features, which are only affecting to opinion are being considered

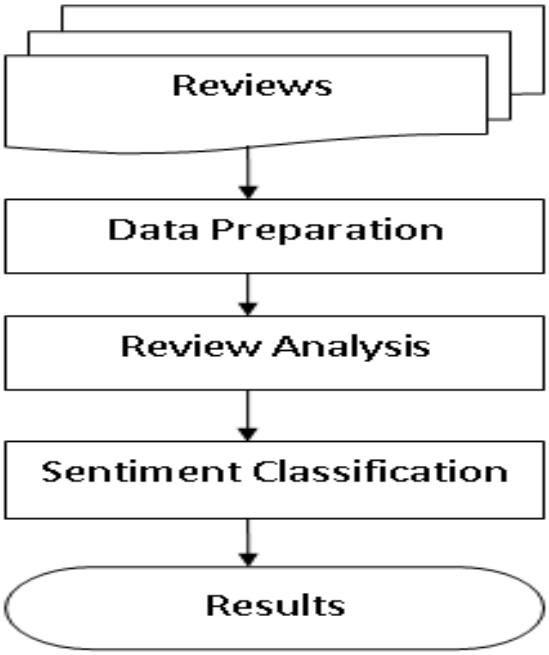

Fig. 1 Sentiment analysis Model[14]

In this era of electronic almost all people come in contact of e-commerce. People are buying products on the web. For making people to do so online, merchant s have to enhance (increase) customer satisfaction and their experience in online shopping. For that if mostly online customers give their review for particular product. These reviews help us for decision making process .for example, if for any customers give positive reviews, then product is good so new customers can be satisfied by this positive review and can buy it trustfully.

Example:

For each feature, we identify how many customer reviews have positive or negative opinions. The specific reviews that express these opinions are attached to the feature. This facilitates browsing of the reviews by potential customers. We give a simple example to illustrate. Assume that we summarize the reviews of a particular digital camera, digital_camera_1. Our summary looks like the following:

Digital_camera_1:
picture quality:
  Positive: 253 <individual reviews>
  Negative: 6 <individual reviews>
size:
  Positive: 134 <individual reviews>
  Negative: 10 <individual reviews>
…

picture quality and size are opinion features. There are 253 customer reviews that express positive opinions about the picture quality, and only 6 that express negative opinions. <individual reviews> points to the specific reviews that give positive (or negative) comments about the feature. With such a feature-based opinion summary, a potential customer can easily see how the existing customers feel about the digital camera. If he/she is very interested in a particular feature, he/she can drill down by following the <individual reviews> link to see why existing customers like it or what they complain about.

Task is performed by ensemble (combining) the best features of words (pos based tags) which we select for opinion mining. Picture quality and size are opinion features, so in above example 253 reviews are positive about picture quality and only 6 reviews are negative.

In this research we propose to study the problem of selecting best feature which are most useful for finding opinion whether the product is good or not.

Automatic detection of emotions in text is becoming increasingly important from an applicative point of view. Websites are used for getting knowledge about particular product by customer's opinion (by review, blog, survey) which is useful for deciding company's place in the market. Opinion mining is based on feature of product on which they have been purchased by customers and these feature help to decide whether opinions are positive or negative.

Moreover opinions mining can be used for recommendation system, government intelligence, citation analysis, human computer interaction and its computer assisted creativity.

The rest of paper is organized as follows. In section 2, literature survey is discussed. Section 3 gives an overview on proposed system.

II. LITERATURE SURVEY

(A) Data Pre-processing[13]

Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data. Data must be pre-processed in order to perform any data mining functionality. Data Pre-processing involves the following tasks

(1) Removing URLs
In general URLs does not contribute to analyze the sentiment in the informal text. For example consider the sentence "I have logged in to www.Ecstasy.com as I'm bored" actually the above sentence is negative but because of the presence of the word ecstasy it may become neutral and it's a false prediction. In order to avoid this sort of failures we must employ a technique to remove URLs.

(2) Filtering
Usually people use repeated letters in words like happyyyyy to show their intensity of expression. But, these word are not present in the sentiwordnet hence the extra letters in the word must be eliminated. This elimination follows the rule that a letter can't repeat more than three times hence can eliminate such letter.

(3) Questions
The question words like what, which, how etc., are not going to contribute to polarity hence in order to reduce the complexity such words are removed.

(4) Removing Special Characters
Special characters like.,[]{}()/' should be removed in order to remove discrepancies during the assignment of polarity. For example "it's good:" if the special characters are not removed sometimes the special characters may concatenate with the words and make those words unavailable in the dictionary. In order to overcome this we remove special characters.

(5) Removal of Retweets.
Retweeting is the process of copying another user's tweet and posting to another account. This usually happens if a user likes another user's tweet. Retweets are commonly abbreviated with \RT." For example, consider the Following tweet: Awesome! RT @rupertgrintnet Harry Potter Marks Place in Film History http://bit.ly/Eusxi :).

Where it has to be done before applying any classification algorithm. In [13, 11] performed three pre-processing tasks. One task to remove URLs from the input file next one to remove special characters, and also removed repeated letters from a word, the last task is to remove question words. The pre-processed document can be given as input to any Machine Learning algorithms.

(B) Review analysis

In [4] , author has proposed a number of techniques for mining opinion features from product reviews based on data mining and natural language processing methods. The objective is to produce a feature-based summary of a large number of customer reviews of a product sold online. Feature based summary is produced by generating frequent pattern features (Association rule mining), opinion words extraction, infrequent feature identification, Opinion Sentence Orientation Identification, etc...

Improvement can be done in these techniques by grouping features according to the strength of the opinions that have been expressed on them, e.g., to determine which features customers strongly like and dislike. This will further improve the feature extraction and the subsequent summarization.

The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure [12]. First, two types of feature sets are designed for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets.

(C) Sentiment Classification

Considering the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, it is found that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods are Naive Bayes, maximum entropy classification, and support vector machines. In terms of relative performance, Naive Bayes tends to do the worst and SVMs tend to do the best, although the differences aren't very large [2].
These three text classification algorithms, namely naı¨ve Bayes, maximum entropy and support vector machines, are employed as base-classifiers for each of the feature sets. [12]

SVMs are popular in text classification tasks since they scale to the large amount of features often incurred in domain [5]

(D) Motivation from Literature

In this world of internet mostly people are purchasing products online. So, review of customer matters a lot. To mine customer reviews accurately combined best features can be extracted and decision can be positive or negative review. The methods found in literature can select only best features or frequent features for classification. It motivates to propose the method combined approach in POS tagging in sentiment analysis.

III. PROPOSED APPROACH

Proposed system find out best features of POS tags and these features will be classified using Naïve Bayes classifier. Accuracy of this process will be increased by applying this system.

The Steps for proposed Algorithm are as follows:

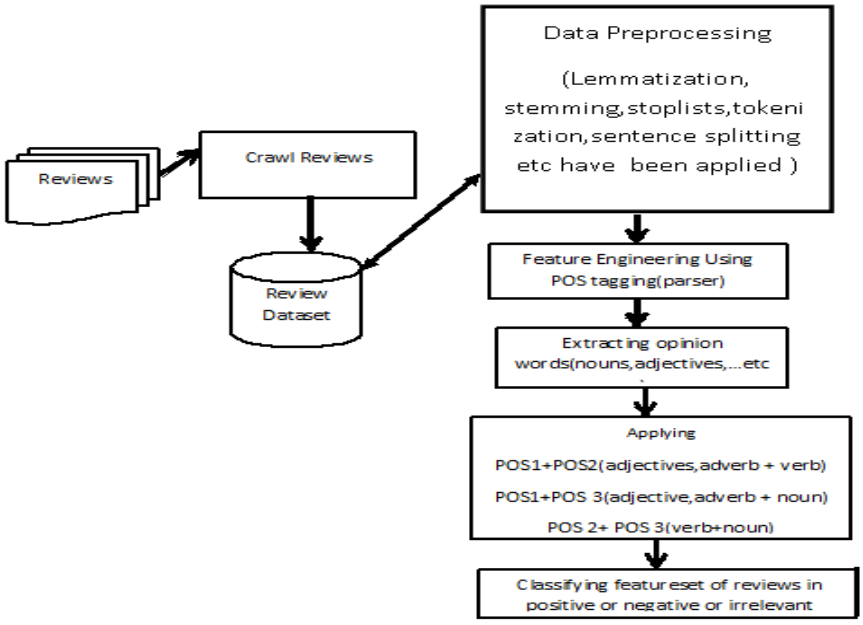

Fig. 2 Flow chart of proposed system

Figure 2 gives an architectural overview for our opinion mining system. The system performs the feature extraction and opinion orientation identification. Given the inputs, the system first downloads (or crawls) all the reviews, and puts them in the review database. After getting review dataset which are already labelled, data pre-processing steps are being applied. In this step only stop words, sentence splitting and tokenization are applied.

E.g. I definitely recommend this film.
For above sentence tokens are I, definitely, recommend, this, film

Next step describe feature engineering, in which all words are given specific tags, according to their category.
POS tagging: Tags are given to tokens
Like I –PRP, definitely-RB, recommend-VBP, this-DT, Film-NN

Feature extraction function, first features that a lot of people have expressed their opinions on in their reviews, and then finds only three groups of POS tags are
1) POS-1 adjectives and adverb,
2) POS-2 verbs
3) POS-3 Nouns.

After finding this opinion words or POS tags now the main focus of this paper is to combine the three POS tags. Now feature set contains following types of word features

1 POS1 and POS 2 (adjective and verb)
2. POS2 and POS 3(verb and noun)
3. POS 1 and POS3 (adjective and noun)

These feature sets are applied to classifier to classify it into positive or negative.

As we know Data mining's best application is classification through which we can classify feature set. In proposed system naïve Bayesian classifier is used. Using Bayesian classifier input is given as feature set, Class values are positive or negative.
In the Bag Of Words framework, a document x is represented by [w1… wk] where wk denotes the k^{th} word appearing in the document. Naıve Bayes assumes that words are mutually independent. Under this assumption, the conditional probabilities can be simplified as

$$P(\mathbf{x}|y_j) = P([w_1, \ldots, w_m]|y_j) \approx \prod_{k=1}^{m} P(w_k|y_j).$$

The naıve Bayes decision can be described as

$$\omega^* = \arg\max_{j=1,\ldots,c} \prod_{k=1}^{m} P(w_k|y_j)P(y_j).$$

The probabilities P(wk/yj) and P(yj) can simply be estimated. From value of W* the review is classified in positive or negative.

IV. CONCLUSIONS

Sentiment analysis is an emerging field and deals with opinion extraction from online reviews, blogs, posts etc... This paper presents several techniques for sentiment analysis such as pre-processing, review analysis, sentiment classification. The proposed work is used for selecting only best features which affect reviews for accurate classification then more accurate result can be get.blem.

REFERENCES

[1] Liu , Lei Zhang "A SURVEY OF OPINION MINING AND SENTIMENT ANALYSIS ".
[2] Bo Pang and Lillian Lee Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classi¯cation using Machine Learning Techniques", Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.
[3] Opinion Mining with the SentiWordNet Lexical Resource", Bruno Ohana A.
[4] Minqing Hu and Bing Liu," Mining Opinion Features in Customer Reviews", American Association for Artificial Intelligence -2004,PP-755-760
[5] Gilad Mishne,"Experiments with Mood Classification in Blog Posts", Style2005, Stylistic Analysis Of Text For Information Access , 2005.
[6] Pang and L. Lee,"Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) PP: 1–135.
[7] Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik ,"Mining Opinion from Text Documents: A Survey", 3rd IEEE International Conference on Digital Ecosystems and Technologies PP: 217-222,2009.
[8] Michael Wiegand and Dietrich Klakow ,"Predictive Features in Semi-Supervised Learning for Polarity Classification and the Role of Adjectives" ,NODALIDA 2009 Conference Proceedings, pp. 198–205
[9] Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik,"Mining Opinion from Text Documents: A Survey", 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies PP: 217-222.
[10] Khin Phyu Phyu Shein and Thi Thi Soe Nyunt" Sentiment Classification based on Ontology and SVM Classifier", 2010 Second International Conference on Communication Software and Networks PP:169-172.
[11] Krzysztof Jędrzejewski, Mikołaj Morzy, "Opinion Mining and Social Networks: a Promising Match", International Conference on Advances in Social Networks Analysis and Mining,IEEE 2011.
[12] Rui Xia, cheng quing zong,Shoushan Li, " Ensemble of feature sets and classification algorithms for sentiment classification" information Sciences 181(2011),PP:.1138-1152.
[13] I.Hemalatha ,Dr. G. P Saradhi Varma, Dr. A.Govardhan," Preprocessing the Informal Text for efficient Sentiment Analysis", IJETTCS, Volume 1, Issue 2, July – August 2012, PP: 58-60
[14] V. S. Jagtap, Karishma Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) ,Volume 2 Issue 3, PP : 164-170 1 April 2013
[15] Adwait Ratnaparkhi,"A maximum entropy model for Part of Speech Tagging"
[16] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar,"Opinion Mining and Analysis: A Survey", IJNLC Vol. 2, No.3, June 2013, PP: 39-49.

