2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

Featured Based Sentiment Classification for Hotel Reviews using NLP and Bayesian Classification

Tushar Ghorpade, Lata Ragha
Department of Computer Engineering, Ramrao Adik Institute of Technology
Mumbai University, India
tushar.ghorpade@gmail.com
lata.ragha@gmail.com

Abstract— **The internet revolution has brought about a new way of expressing an individual's opinion. It has become a medium through which people openly express their views on various subjects. These opinions contain useful information which can be utilized in many sectors which require constant customer feedback. Analysis of the opinion and it's classification into different sentiment classes is gradually emerging as a key factor in decision making. There has been extensive research on automatic text analysis for sentiments such as sentiment classifiers, affect analysis, automatic survey analysis, opinion extraction, or recommender systems. These methods typically try to extract the overall sentiment revealed in a sentence or document, either positive or negative, or somewhere in between. However, a drawback of these methods is that the information can be degraded, especially in texts where a loss of information can also occur. The proposed method attempts to overcome the problem of the loss of text information by using well trained training sets. Also, recommendation of a product or request for a product as per the user's requirements have achieved with the proposed method.**

Keywords— **sentiment analysis; natural language processing; ontology; machine learning; online traveller reviews; naive bayes classification.**

I. INTRODUCTION

There are two main categories in textual information, they are ***facts*** and ***opinions***. Facts are objective statements while opinions are subjective ones. A lot of research is being conducted to retrieve information from different sources to throw light on these two aspects of a statement. Opinions are the subjective statements and still rare in existing researches. Opinions reflect the people's sentiments or feelings about the product and events. Many of the existing research are based on mining and retrieval of factual information and not on opinions. Opinions are also important when someone wants to hear the other's viewpoint before they make a decision [1].

Sentiment Classification aims at mining the World Wide Web text of product reviews by customers to classify the reviews into positive or negative opinions. Automated opinion mining from the reviews is beneficial to both consumers and sellers. Examples of past work include mining reviews of automobiles, banks, movies, travel destinations, electronics and mobile devices. Potential applications include extracting opinions or reviews from discussion forums efficiently, and integrating automatic review mining with search engines to provide quick statistics of search results [2].

In the proposal we have focused more on how to improve the words extraction from the given reviews or opinions. Also, we would be able to design the well trained training sets by adding weights to the words. The trained training sets can easily filter the attributes as per the user requirements. The Jolly and Pleasant Exercise (JAPE) mathematical techniques can easily pre-process all the words from the given sentences. By using Bayesian algorithm we can easily classify the positive or negative classifications.

II. BACKGROUND

There are mainly two types of approaches for sentiment classification. One is machine learning method, the other, semantic orientation methods.

A. Semantic Orientation (SO)

Turney presented a Semantic Orientation (SO) mining method based on PMI-IR algorithm for sentiment classification by combining the Point Mutual Information (PMI) and the statistical data collected by Information Retrieval (IR) [3]. With two reference words pair (RWP) "excellent" and "poor" (presenting positive and negative opinions respectively), determine the semantic orientation of a Phrase' SO (*phrase*) according to Equation.

$$\boldsymbol{so}(\boldsymbol{phrase}) = \mathbf{log_2}[\frac{\boldsymbol{hits}(\boldsymbol{phrase\ NEAR}\ \mathbf{excellent})\boldsymbol{hits}(\mathbf{poor})}{\boldsymbol{hits}(\boldsymbol{phrase\ NEAR}\mathbf{poor})\boldsymbol{hits}(\mathbf{excellent})}]$$

A review's semantic orientation was calculated by averaging the SO values of all the extracted phrases in it. The opinion will be positive if its average semantic orientation exceeds a threshold and is negative if otherwise. In Turney's study the threshold was set to Zero [3].

978-1-4577-2078-9/12/$26.00©2011 IEEE

B. Machine learning methods

Classification algorithms are the core of text classification. At present, popular classification algorithms based on machine learning and statistics are k-neighbor classifier (KNN), Naïve Bayes classifier (NB) and Support Vector Machine (SVM) etc.

C. Ontology

The ontology learning framework consists of several steps, as depicted in Fig. 1. We briefly describe these steps and show how the characteristics of the Web services context influenced their design [4], [5].

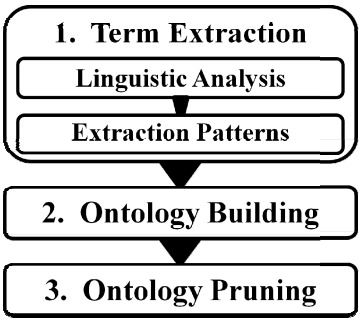

Fig. 1. The Ontology Learning Framework [4]

1. *Term Extraction:* A word or a set of words that are identified as useful for ontology building form a "term". Term extraction is done in two steps. First, in a linguistic analysis phase, the corpus is annotated with linguistic information. Then, a set of extraction rules are applied on this linguistic information to identify the potentially interesting terms. The characteristics of the Web services domain influenced our design choices in several ways. First, to overcome the limitations of the poor grammatical quality of the texts, we employed linguistic analysis of different complexity. Then, the small size of the corpus and its sublanguage features facilitated the use of a rule-based solution.

2. *Ontology Building:* The ontology building phase derives both static and procedural knowledge in the form of a hierarchy of frequent domain concepts and a hierarchy of Web service functionalities. The strong sublanguage features of the analysed corpora allow extracting terms that are highly relevant for ontology building. Therefore, it suffices to use simple ontology learning techniques and to adapt them to the requirements of the domain (e.g., extract procedural knowledge).

3. *Ontology Pruning:* The low grammatical quality of the corpus and its sublanguage characteristics cause a suboptimal functioning of the used linguistic tools. Therefore, some of the derived concepts do not have any domain relevance. The pruning stage filters out these potentially uninteresting concepts

D. Natural Language Processing (NLP)

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages; it began as a branch of artificial intelligence. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

Modern NLP algorithms are grounded in machine learning, especially statistical machine learning. Research into modern statistical NLP algorithms requires an understanding of a number of disparate fields, including linguistics, computer science, and statistics [6].

The lists of some of the most commonly researched tasks in NLP [6] are Automatic summarization, Named entity recognition (NER), Part-of-speech tagging, Parsing, Sentiment analysis.

E. JAPE Rule

A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The grammar always has two sides: Left and Right. The LHS of the rule contains the identified annotation pattern that may contain regular expression operators (e.g. *, ?, +). The RHS outlines the action to be taken on the detected pattern and consists of annotation manipulation statements. Annotations matched on the LHS of a rule are referred on the RHS by means of labels. For example,

LHS (regular expression for annotation pattern): i.e., Lookup for annotation player and label it *player*

RHS (manipulation of the annotation patter from LHS): Annotation is the act of tagging, commenting or marking the media with some metadata information to identify type or the content of the media. In the context of GATE annotation it means creating new or identifying existing metadata information, for example, identifying a person's name from the text.

i.e., Get the gender of the *player*, if gender=male re-label *player* as *Male-Player,* else re-label as *Female-Player* [4], [7].

III. SYSTEM FRAMEWORK DESIGN

A. Working and Component Details of the Proposed Method Architecture

Step 1, Input: To collect the inputs from the different websites through crawler or manually do it at the initial stage. Also user can submit review through our web-site.

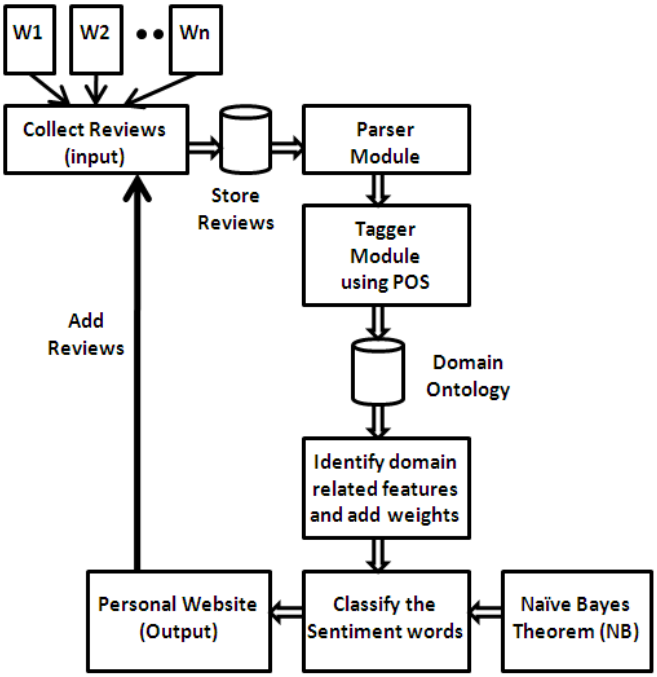

Fig. 2 The Proposed Architecture of Featured based Sentiment Classification.

Step 2, Parser Module: A natural language parser works out the grammatical structure of sentences, for instance, words which are grouped together are "phrases" and the Stanford parser identifies words as subjects or objects of a verb [8], [9].

Step 3, Tagger module: We use Part Of Speech (POS) tagger to assign POS tags to words in a sentence (such as: tags for nouns, verbs, and adjective). To implement this process we use Stanford Tagger. This tagger is based on a technique that has been effective in a number of natural language applications which include part of speech and word sense tagging, prepositional phrase attachment, and syntactic parsing [10].

Step 4, Apply Domain Ontology: In this step we present some modules indicating how to separate the sentences, then how to extract the nouns and respective Verbs, Adverbs and Adjectives. Also, how to create and upgrade the word dictionary. It is important to find out how to generate a well trained training set which will easily support the Bayesian algorithm to classify the reviews.

Module 1, Separate the sentences by using maximum delimiters.

If (S.Delimeters= "." + "," + "!" + "&")
Then, Separate the sentences.

Where, S= Sentences or reviews

Module 2, Extract feature based information.

If (S.Tokens.Nouns = N_1, N_2, N_n)
Then, extract Verb, Adverb & Adjective from the sentences to the respective nouns.

Else If (S.Tokens.Nouns! = N_1, N_2, N_n)
Then, convert all nouns into **OTHER** as keyword and extract the Verb, Adverb & Adjectives from the sentences to the respective nouns.

Where, S=Sentences.
N_1, N2, Nn = Specified Nouns.
OTHER = collection of all other nouns.

Module 3, Create Word dictionary for good and bad words by assigning a weight for every word.

If (Word=Positive)
Then, Assign a weight between the ranges of 1 to 5.

Else If (Word= Negative)
Then, Assign a weight between the ranges of -5 to -1.

Module 4, Upgrade the Word dictionary.

If (Word= "New Word")
Then, Add the New Word into the Word dictionary with Zero weight.

Module 5, summarize the Noun weights and submit it into the database.

If (Nouns = N_1, N_2, N_n) then,

$$w_{N1} = \sum_{k=1}^{n} W_K \quad (1)$$

Else If (Nouns! = N_1, N_2, N_n) then,

$$W_{OTR} = \frac{\sum_{k=1}^{n} W_K}{n} \quad (2)$$

Where, W_{N1} = Total weight of noun1
W_{OTR} =Average weight of **OTHER** nouns
W_k = K^{th} Weight
n = Total no of reviews.

Module 6, Set the parameters or Nouns for Classification and then we can apply Bayesian Theorem for the classifying reviews.

$$\sum_{j=1}^{n} W_{NJ+}\ W_{OTR}\ \geq 0\ then\ Y = Positive \quad (3)$$

$$\sum_{j=1}^{n} W_{NJ+}\ W_{OTR}\ < 0\ then\ Y = Negative \quad (4)$$

Where, W_{NJ} = Total weights of specified Nouns.
W_{OTR} = Average weights of other Nouns.
Y = Define a Class

Module 7, finally summarize the data into one table for presenting information through personal site.

IV METHDOLOGY

A. Experimental Data

We selected hotel reviews from the links mentioned in [11], [12], [13] as the experiment corpus. We ultimately finalized 11 hotels from Mahableshwar City, Maharashtra State, India. Customer reviews for these 11 hotels were obtained as the data of our study. In this research, we manually retrieved reviews on the web and stored them into our database. We have considered only the reviews written in English. Reviews were selected on the basis of the content. We got a total of 128 reviews. The distribution of the scores of the 128 reviews is showed in Table. 1.

Table 1. Results of Manual Classification

City	No of Hotels	Positive Reviews	Negative Reviews
Mahabaleshwar (Maharashtra)	**11**	**93**	**35**

B. Test Data

We selected 36 reviews as training data set, out of which 21 were positive and 15 were negative. A Bayesian sentiment classification model was trained by this training data set. Then the sentiment classification model was applied on the test data set that has 128 reviews with 93 positive and 35 negative samples. The classification result is shown in the Table 2. Fig. 3 gives the Bayesian Sentiment classification approach.

C. Evaluation method to the performance.

There are three indexes generally used in text categorization: Recall, Precision, and Accuracy. So we adopted these indexes to evaluate the performance of sentiment classification in our study [14].

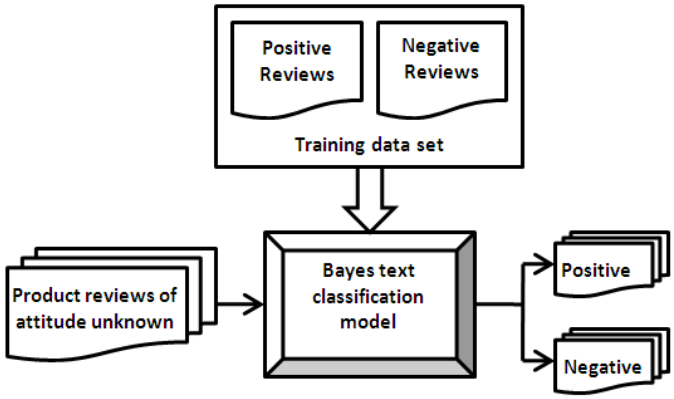

Fig 3. Bayes Sentiment classification approach [2]

Table 2. Contingency Table for Performance Evaluation [14]

	Actual positive	**Actual negative**
Predict positive	**a (Tp)**	**b (Fp)**
Predict negative	**c (Fn)**	**d (Tn)**

These indexes can be calculated according to the figures in Table 2. and their formulas are as follows.

$$Accuracy\ (A) = \frac{a+d}{a+b+c+d} \quad (5)$$

$$Precision(p) = \frac{a}{a+b} \quad (6)$$

$$Recall(p) = \frac{a}{a+c} \quad (7)$$

$$Precision(n) = \frac{d}{c+d} \quad (8)$$

$$Recall(n) = \frac{d}{b+d} \quad (9)$$

Here, Accuracy is the overall accuracy of sentiment classification. Recall (p) and Precision (p) are the recall ratio and precision ratio for actual positive reviews. Recall (n) and Precision (n) are the recall ratio and precision ratio for actual negative reviews.

V EXPERIMENT RESULTS

Based on the Bayesian classification model, according to the training set, the classification result of 128 reviews in the testing set is shown in Table 3 and its graphical representation is shown in Fig. 4.

Table 3. Performance Evaluation

	Actual Positive Reviews	Actual Negative Reviews	Total
Predict Positive	**89**	**04**	**93**
Predict Negative	**01**	**34**	**35**
Total	**90**	**38**	**128**

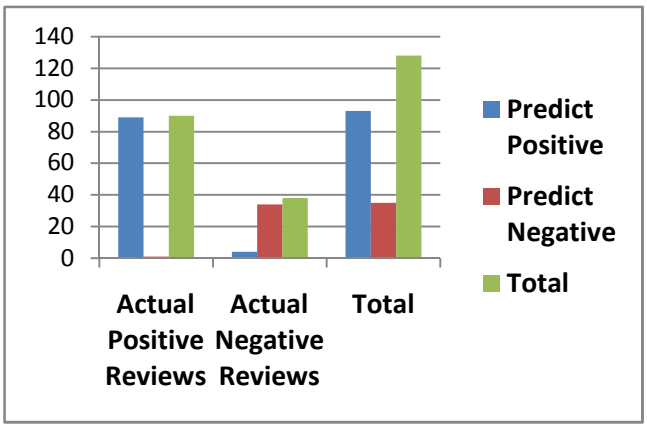

Fig 4. Distribution Of Positive & Negative Reviews

$$Accuracy\ (A) = \frac{89 + 34}{89 + 04 + 01 + 34} = 96.09\% \quad (10)$$

$$Precision\ (p) = \frac{89}{89 + 04} = 95.69\% \quad (11)$$

$$Precision\ (n) = \frac{34}{01 + 34} = 97.14\% \quad (12)$$

$$Recall\ (p) = \frac{89}{89 + 01} = 98.88\% \quad (13)$$

$$Recall\ (n) = \frac{34}{04 + 34} = 89.47\% \quad (14)$$

IV. Conclusions and Future

In the proposed method, we have focused more on how to improve the words extraction from the given reviews or opinions. Also, we would be able to design the well trained training sets by adding weights to the words. The trained training sets can easily filter the attributes as per the user requirements. The JAPE mathematical techniques can easily pre-process all the words from the given sentences. By using the machine learning algorithm we can easily classify the positive or negative reviews.

The proposed method attempts to overcome the problem of the loss of text information by using the well trained training sets. The trained training sets can easily filter the attributes as per the user requirements. Also, the proposed method allows the user to make recommendations for a product or make request for a product as per the user's requirements.

We can focus more on, to design intelligence system for word dictionary. Also, we can consider to Support Vector Machine algorithm for classification for future studies.

REFERENCES

[1] Khin Phyu Phyu Shein, *"Ontology Based Combined Approach For Sentiment Classification"*, Proceedings Of The 3RD International Conference On Communications And Information Technology, Manuscript Received © October 9, 2001.

[2] Qiang Ye, Bin Lin, Yi-Jun Li, *"Sentiment Classification For Chinese Reviews: A Comparison Between Svm And Semantic Approaches"*, Proceedings Of The Fourth International Conference On Machine Learning And Cybernetics, Guangzhou, © 18-21 August 2005 IEEE. pp. 2341-2346.

[3] Qiang Ye, Wen Shi, Yijun Li, "*Sentiment Classification for Movie Reviews In Chinese By Improved Semantic Oriented Approach"*, Proceedings Of The 39th Hawaii International Conference On System Sciences ©2006 IEEE. pp. 53b.

[4] Vladimir Oleshchuk, Asle Pedersen, "*Ontology Based Semantic Similarity Comparison of Documents",* Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03)" © 2003 IEEE. pp.735-738.

[5] Wu Di1, Li Xiaojing2, Zhang Chengwei3 "*The Design of Ontology-based Semantic Label and Classification System of Knowledge Elements*", 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering, 978-1-4244-9983-0/11 ©2011 IEEE. pp. 95-98.

[6] www.wikipedia.com, Natural language processing Techniques.

[7] Dhaval Thakker, PA Photos, UK, Taha Osman,"GATE JAPE Grammar Tutorial", Nottingham Trent University, UK, Phil Lakin, UK, Version 1.0.

[8] Alexander O'Neill, "*Sentiment Mining for Natural Language Documents*", Comp 3006 Project Report, Department of Computer Science Australian National University, © November 2009.

[9] Marie-Catherine de Marne and Christopher D. Manning, "*Stanford typed dependencies manual*", September 2008, revised for Stanford Parser v. 1.6.5 in November 2010.

[10] Beatrice Santorin "*Part-Of-Speech Tagging Guidelines For The Penn Treebank Project (3rd Revision)*", University Of Pennsylvania.

[11] http://travel.yahoo.com/p-hotel-363551, September, 2011.

[12] http://travel.yahoo.com/p-hotel-329920, September, 2011.

[13] http://travel.yahoo.com/p-hotel-1320343, September, 2011.

[14] Wenying Zheng, Qiang YE "*Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm*", Third International Symposium on Intelligent Information Technology Application. ©2009 pp.335-338.