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 8, August 2014

Three Level Feature Extraction for Sentiment Classification

G.Angulakshmi [1], Dr.R.Manicka Chezian [2]

Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[1].

Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[2].

ABSTRACT: Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to all human activities and key influencers of our behaviors. Product reviews written by on-line shoppers is a valuable source of information for potential new customers who desire to make an informed purchase decision. Identifying domain-dependent opinion words is a key problem in opinion mining. In this paper, the feature-based opinion mining model has been discussed. In many such cases, these nouns are not subjective but objective. The involved sentences are also objective and imply positive or negative opinions. Thus, this paper discuss about how reviews could be classified using naive bayes algorithm to produce effective result.

KEYWORDS: Auxiliary List, Opinion Mining, Preprocessing, Porter Algorithm.

I. INTRODUCTION

Opinion mining is the study of people opinions and emotions towards any entities, events and attributes [1]. In opinion mining review is to be evaluated at three levels [4] – document level, sentence level and feature level. Differently sentiment analysis at the word level focuses on sentiment polarities of opinion phrases. Rule-based semantic analysis approach is used to classify the documents based on text reviews as sentiments. Opinion mining involves analyzing user's opinion [10], attitude, and emotion towards particular topic. This consists of first categories text into subjective and objective information, and then finding polarity in subjective text [3]. Opinion mining and summarization process involve three main steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining.

II. RELATED WORK

G.Angulakshmi, Dr.R.ManickaChezian [1] compared the existing techniques and tools of Opinion Mining available earlier. Anjali Ganesh Jivani [2] propsed a work on how stemming works and comparison on different stemming algorithms. Ahmad Kamal [3] proposed a work on Opinion Indicator Seed Word and the analysis with feature attributes. M. Daiyan, A. Khan, A. Alam [4] proposed a work about classification on different techniques on machine learning approaches. Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang and Xiaoyan Zhu [5] propsed a work on Lexicons. Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen [6] had done a work on Sentiment Word Propagation and Polarity assignment. Lei Zhang, Bing Liu [7] propsed a method over Feature – Based Sentiment Analysis and also to prune Non –Opinated Features. Padmapani P. Tribhuvan, S.G. Bhirud, Amrapali P.Tribhuvan [8] has proposed an overview on feature based opinion mining and summarization. Richa Sharma, Shweta Nigam and Rekha Jain [8] proposed a work on extracting opinion words and Seed List Preparation and the polarity Detection & Classification. Richa Sharma, Shweta Nigam, Rekha Jain [10] proposed a work on basic of opinion mining and the measures used to evaluate them. S.L. Ting, W.H. Ip, Albert H.C.

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 8, August 2014

Tsang [11] proposed a methodology on Preprocessing, Feature Selection and a model evaluation on how naive Bayes algorithm works on text. Wahiba Ben Abdessalem Karaa [12] proposed a work on Porter Stemmer and the error that occur in the algorithm. Henrique Siquerira and Flavia Barros [13] proposed a work on the corpora and the evaluation measures.

III. THE PROPOSED METHOD

The approach comprises of a series of steps that gradually detect opinion words. Each step creates a pool of opinion words that will constitute the feed for the next detection step. More specifically the construction of the proposed algorithm can be summarized in the following steps: 1) Preprocessing 2) First Level Feature Extraction 2) Second Level Feature Extraction 3) Third Level Feature Extraction 4) Classifier. Figure 1 shows the overview of approach.

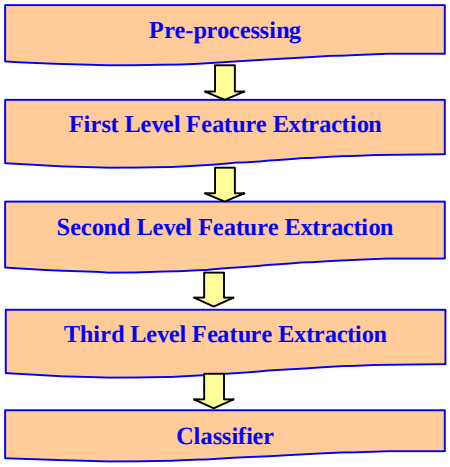

Figure 1: Overview of Proposed Approach

A. PREPROCESSING

Algorithm receives user opinions in raw form. We implement some form of preprocessing in order to filter-out noise. Sentence splitting is a critical step in this module (opinion delimitation) since double propagation takes into account neighborhood sentences in order to propagate sentiment. Additionally in order to increase the efficiency of the extraction process we have adopted an on-line stemmer engine.

Porter Stemming

Stemming is a technique that reduces words to their common root, or *stem*. The Porter algorithm [12] differs from Lovins-type stemmers in two major ways. The first difference is a significant reduction in the complexity of the rules associated with suffix removal [2]. The second difference is the use of a single, unified approach to the handling of context. Porter uses a minimal length based on the number of consonant-vowel consonant strings (the *measure*) remains after removal of a suffix. This idea, which may be regarded as an easily computable representation of a syllable. A typical rule is thus as follows:

$$(M > 0) * FULNESS \rightarrow * FUL$$

This means that the suffix *FULNESS should be replaced by the suffix *FUL if, and only if, the resulting stem has a non-zero measure (*m*). Porter's algorithm [12] is iterative in nature, i.e., it allows a long, multi-component suffix to be removed in stages.

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 8, August 2014

Auxiliary List Preparation

Auxiliary word list comprises a series of word sets like articles, verbs, comparatives, conjunctions, decreases (e.g. “less”), increasers (e.g. “extra”), negations (e.g. “not”) and pro-nouns. These words will constitute a main feed of the algorithm. The proposed approach utilizes this seed in the construction of all extraction patterns.

IV.FEATURE EXTRACTION

A major information loss of this word level dependency tree compared with constituent tree is that it does not explicitly provide local structures and syntactic categories (i.e. NP, VP labels) of phrases [8]. On the other hand, dependency tree provides connections between distant words, which are useful in extracting long distance relations. Formally, we define the dependency parsing with phrase nodes as phrase dependency parsing. A dependency relationship which is an asymmetric binary relationship holds between two phrases. One is called head, which is the central phrase in the relation. The other phrase is called dependent, which modifies the head Pruning [7] away local dependency relations by additional phrase structure information, phrase dependency parsing accelerates following processing of opinion relation extraction.

A. First Level Filtered Seed Extraction

The Filtered Seed S0 is the set of Seed words that also appeared in the collection of opinion documents$S' = S \cap V$. In other words we filter-out words from the Seed [3] that don't appear in the corpus. In Seed list, the polarity of each word is provided. However, depending on the way the word is used, it might alter its polarity (“this phone is definitely not lightweight”). Hence, we apply a step of polarity disambiguation using a set of language patterns [10]. The end of this process we have a pool of newly discovered opinion words along with their polarity.

B. Second Level Conjunction-Based Extraction

At this level we exploit the assumption of sentiment consistency that applies in conjunct words (e.g. “lightweight and well-built device”). That way algorithm discovers new opinion words by making use of certain conjunction patterns that have been selected to the sentiment consistency theories. For this level we have utilized 6 positive and 4 negative extraction patterns. Candidate opinion words of this step also go through a polarity [6] disambiguation process like the previous step. At the end of this process we have an extended list of opinion words $S' \cup C$ where C is the list of opinion words extracted from this level.

C. Third Level Double Propagation Extracted

. The newly extracted sentiment words and features are utilized to extract new sentiment words and new features which are used again to extract more sentiment words and features. The propagation [8] ends until no more sentiment words or features can be identified. As the process involves propagation through both sentiment words and features, we call the method *double propagation*. The process of detecting new opinion words follows the theory of double propagation. The assumption is that each opinion word has an opinion target attached to it (e.g."nice screen"). Here, there is a direct connection (opinion word, `nice')! (Opinion target, `screen'). Based on double propagation [6] and using the current list of opinion words, we are able to identify opinion targets. Using this set of opinion targets we are able to extract new opinion words following the same logic. So this process is repetitive. It iterates i times or as long as new opinion words are discovered. After parsing, words in a sentence are linked to each other by certain relations. In dependency grammar, the relation between two words *A* and *B* can be described as *A*(or *B*) depends on *B*(or *A*).

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 8, August 2014

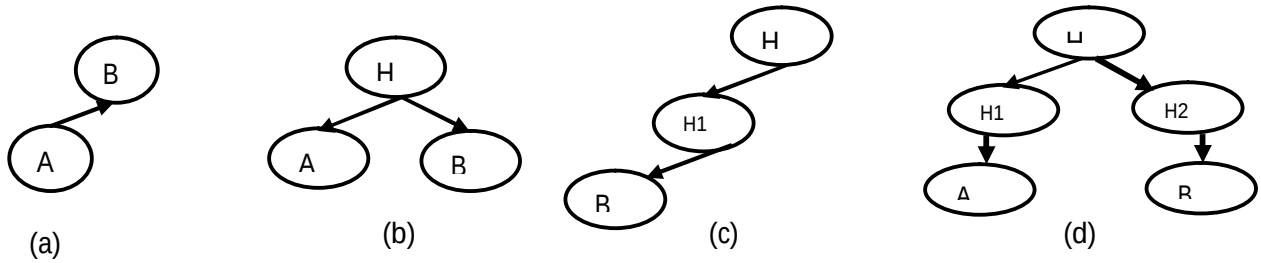

Figure 2. Different relations between words *A* and *B*. (a) and (b) are two direct relations; (c) and (d) are two indirect relations.

Direct Relation (DR): A *direct relation* means that one word depends on the other word directly [6] or they both depend on a third word directly. Some examples are shown in Figure 2 (a) and (b). In (a), *A* depends on *B* directly while they both directly depend on *H* in (b).

Indirect Relation (IDR): An *indirect relation* means that one word depends [6] on the other word through other words or they both depend on a third word indirectly. Some examples are shown in Figure 2 (c) and (d). In (c), *A* depends on *B* through *H1*; in (d), *A* depends on *H* through *H1* while *B* depends on *H* through *H2*. In more complicated situations, there can be more than one *H1* or *H2*. **DR** can be regarded as a special case with no *H1* or *H2* in the dependency path. Note that in (d), there are cases that no *H1*(or *H2*) between *A*(or *B*) and *H*, but more than one *H2*(or *H1*) between *B*(or *A*) and *H*.

However, complex relations can make the algorithm vulnerable to parsing errors. Parsing is considerable more difficult and error prone with informal expressions used in the Web environment. At each step Pi opinion words are discovered. In the end of this step we end up with list $D = S' \cup C \bigcup_i P_i$. For the reverse step of double propagation. The newly discovered opinion words are going through polarity disambiguation [5]. At this step we take advantage of intra-sentential and inter-sentential sentiment consistency. The intra-sentential consistency suggests that if there are other opinion words in a sentence with known orientation, then, the newly found word will get the accumulated sentiment of these words. When there are no other known opinion words in the sentence, the inter-sentential assumption is applied. At the end of this process we have a pool of new opinion words and their orientation, the double propagation opinion words.

IV. CLASSIFICATION

There are many possible approaches to identifying the actual polarity of a dataset. Our analysis uses statistical methods, namely supervised machine learning, to identify the likelihood of reviews having "positive" or "negative" polarity with respect to previously hand-classified training data.

A. Naive Bayes Classifier

The Naive Bayes classifier is a well-known supervised machine learning approach. In this paper the "features" used to develop Naïve Bayes are referred to as "attributes" to avoid confusion with text "features"[11]. In our approach, double propagation word that appears in the training data are collected and used as attributes. The formula of our Naïve Bayes classifier is defined as

$$Pr(C|rv) = \prod_{w \in W} Pr(app_w|C) \prod_{w \notin W} Pr(app_w|C)$$

$$\hat{C} = argmax Pr(c|rv)$$

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 8, August 2014

Where **rv** is the review under consideration, **w** is a feature extraction words pair that appears in the given document, ***Pr*** (app_w|class) is the probability that a feature extraction words pair appears in a document of the given class in training data, and ***bc*** is an estimated class. The probabilistic models computed by the Naïve Bayes classifiers were sorted by log posterior odds on positive and negative orientations for the purpose of ranking, i.e. by a "score" computed as follows

score = log Pr(+|rv) - log Pr(-|rv)

Where ***rv*** is the review under consideration, ***Pr (+|rv)*** is the probability of rv being a review of positive polarity, **Pr*(-|rv)*** analogously is the probability of the review being of negative polarity.

V. RESULTS AND DISCUSSIONS

For the evaluation, the consideration of initial seed as the ground truth set of opinion words. Since the initial seed is generic, at this step the evaluations of ability in the approach to extract opinion words that are not domain specific. The set of results focuses on the impact of each step at opinion-word discovery. In brief, Conjunction-based extraction is more conservative at finding new opinion words while Double Propagation tends to discover more words. Note that the evaluation is only indicative since the approach is evaluated in terms of ability to identify opinion words from the original seed which is not domain-specific and only a few of them appear in the extracted opinions. That language patterns are not always followed by the users. At this point the discussion of the contribution on each extraction step to the sentiment classification task. Review text classified as positive or negative by opinion classification method. Effectiveness of method is determined by precision or recall values [10].

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

- True positives [9] (TP) - number of reviews correctly labeled as belonging to particular class (positive/negative).
- False positives (FP) - number of reviews incorrectly labeled as belonging to particular class.
- False negatives [11] (FN) - number of reviews were not labeled as belonging to the particular class but should have been labeled.

Table1: Precision Vs corpus sizes

Classifier	Corpus size			
	500	**1000**	**1500**	**2000**
Naive Bayes	48	55	59	**61**
J48	43	49	46	49
SVM	42	46	45	**55**

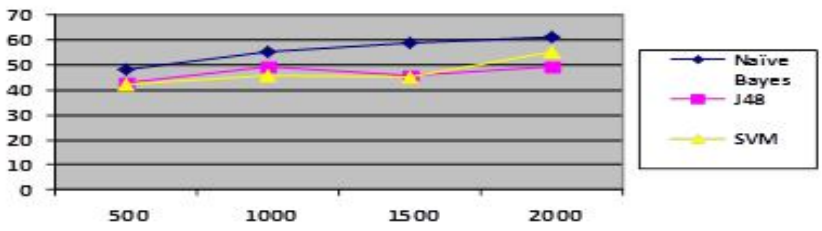

Figure. 3 Different classifier with dataset size

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 8, August 2014

From the above figure 3 there is a comparison with different classifiers on different dataset size; Naïve Bayes have the highest value. The graph is calculated using Java.

Table 2: Performance Measure

Classifiers	**Avg. Accuracy**	**Max. Accuracy**	**Avg F**
SVM	69.82%	71.33%	0.688
J48	73.25%	77.60%	0.728
Double propagation Naive Bayes (NB)	**91.86%**	**94.82%**	**0.920**

From Table 2 the calculations on measures are used and when using Double Propagation Naïve Bayes Classifier the values are increased when compared to others.

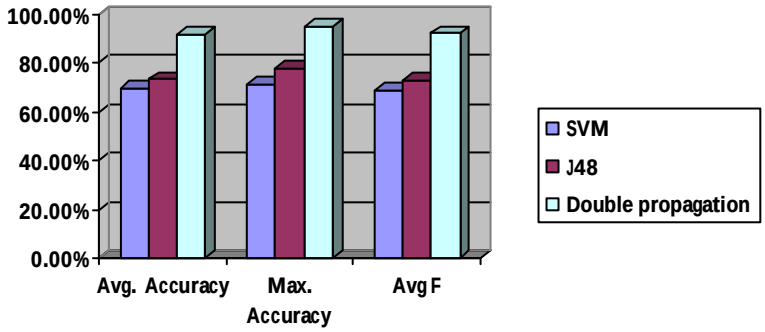

Figure 4: Comparison of classification algorithm

From the above figure 4 the process using Double Propagation has achieved the highest precision values compared to Support Vector Machine and J48.

VI. CONCLUSION

In this paper an approach on the method for domain-specific opinion word discovery was presented. Word polarity is calculated automatically by following a set of polarity disambiguation procedures. The experimental evaluation suggests that we can achieve satisfactory sentiment classification using this completely unsupervised approach. Naïve Bayesian which summarizes review depending on features and technical feature value extracted from the reviews. The purpose of a domain sentiment word extraction approach based on the propagation of both known sentiment lexicon and extracted product features, which we call *double propagation,* algorithm exploits dependency relations to capture the association between features and sentiment words.

REFERENCES

[1] G.Angulakshmi, Dr.R.ManickaChezian, "An Analysis on Opinion Mining: Techniques and Tools", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014. Page No: 7483 -7487.
[2] Ms. Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms", International. Journal. Computer. Technology. Applications., Vol 2 (6), 1930-1938, ISSN:2229-6093.
[3] Ahmad Kamal, "Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources" Department of Mathematics, Jamia Millia Islamia (A Central University), New Delhi.

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 8, August 2014

[4] M. Daiyan, A. Khan, A. Alam, " To Classify Opinion of Different Domain Using Machine Learning Techniques", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 5, May 2013.
[5] Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang and Xiaoyan Zhu, "Cross-Domain Co-Extraction of Sentiment and Topic Lexicons".
[6] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, "Expanding Domain Sentiment Lexicon through Double Propagation", Department of Computer Science, University of Illinos at Chicago. Page No: 1199-1203.
[7] Lei Zhang, Bing Liu, "Identifying Noun Product Features that Imply Opinions".
[8] Padmapani P. Tribhuvan, S.G. Bhirud, Amrapali P.Tribhuvan, "A Peer Review of Feature Based Opinion Mining and Summarization", International Journal of Computer Science and Information Technologies, Vol 5(1), 2014, 247-250. ISSN: 0975-9646. Page No: 247-250.
[9] Richa Sharma, Shweta Nigam and Rekha Jain, "MINING OF PRODUCT REVIEWS AT ASPECT LEVEL", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014. Page No: 87-95.
[10] Richa Sharma, Shweta Nigam, Rekha Jain, "Determination of Polarity of Sentences using Sentiment Orientation System", International Journal of Advances in Computer Science and Technology. Volume 3, No.3, March 2014. ISSN 2320-2602. Page No: 182-187.
[11] S.L. Ting, W.H. Ip, Albert H.C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification", International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011.
[12] Wahiba Ben Abdessalem Karaa, "A NEW STEMMER TO IMPROVE INFORMATION RETRIEVAL", International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013
[13] Henrique Siquerira and Flavia Barros, "A Feature Extraction Process for Sentiment Analysis of Opinions on Services", Universidade Federal de Pernambuco, Brazil.

BIOGRAPHY

G.Angulakshmi is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi. She received her Master of Computer Science (M.Sc) in 2013 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has published a paper, presented papers in International/National conferences and attended Workshop, Seminars. Her research focuses on Data Mining.

Dr. R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published hundred papers in International/National journals and conferences. He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.