LOCALIZED SENTIMENT ANALYSIS USING RANDOM WALK ALGORITHM IN HINDI

Surbhi Maheshwari, Rajasthan Technical University, India (subhi.kabra@gmail.com)
Pallavi Gupta, Rajasthan Technical University, India (pallavigupta9509@gmail.com)
Ritu Dhabhai, Rajasthan Technical University, India (ritudhabhai7@gmail.com)

ABSTRACT

In the dissertation, we have extended the previous work of the review analysis which is capable of only scoring the reviews by calculating the scores using the SentiWordNet. The features we have added in our dissertation are feature level debate analysis , in which the features related to debates, negation handling, intensifiers handling, conjunction handling and multi-document review analysis. The most important part of our dissertation is that we have worked on HINDI, for this purpose we have use the SentiWordNet Tool. Sentiment analysis is the process of extracting knowledge from the people's opinions, appraisals and emotions toward entities, events and their attributes. This dissertation presents the limitations of existing feature and sentence level sentiment analysis approaches and recommends the use of SentiWordNet Tool for better scoring the sentiment for feature level.

Keywords - Abstractive, Extractive, Information Retrieval, Summary Generation, Text Summarization.

INTRODUCTION

Analyzing sentiment of text is a challenging problem in Natural Language Processing. This is a wide range task in extracting sentiment from text. Previous work is focused on polarity classification, opinion extraction and opinion source assignment. These systems has solved classification sentiment at different level of granularity, form lexical level, phrased level, sentence level, as well as document level. Different application has different needs, so research of sentiment classification on multiple levels is very important. For example, a question answering system might require sentiment of sentence level; a system of product review would mostly require polarity classification at phrase level or sentence level; a system which determines articles from online news source would require document level. This work focuses on constructing a model on multiple level and analyzes global sentiment and local sentiment of document at different granularity. There are several kinds of models used in analysis of sentiments. Structured model has been used for sentiment analysis. Choietal used CRFs to learn a global sequence model to classify and assign sources to opinions. Mao and Lebanon has measured sentiment on sentence level and constructed sentiment flow of authors in news reviews by using a CRFs regressing model. Cascaded models for sentiment analysis were studied by Pang and Lee. In their work, an initial model classified each sentence as being subjective or objective using a global min-cut inference algorithm that considered local labeling consistencies. The top subjective sentences are input into a standard document level polarity classifier with improved results. The current work differs from that in Pang is that we introduce global sentiment and local sentiment based on semantic analysis and construct a multiple level model which determines sentiment of document through local sentiment. However, sentiment of a text is complicated. Sentiment of different granularity co-exists and influences each other. In order to identify the sentiment of the whole text, we propose a model of multiple levels which analyzes polarity of text from local sentiment to global sentiment. [1]

The use of an entity or aspect based sentiment analysis is to compute the document level sentiment; some of the problems of document level sentiment analysis, such as entity identification, subjectivity detection and negation are automatically taken care of. Document may contain multiple entities. It is very important to find out what the sentiment is towards a particular entity. For example, "Sir has better presentation skill but knowledge is poor." "Camera has better battery but display is bad." Here the, positive words are for presentation and battery. But these sentences are negative for knowledge and display, respectively. In our approach, we first evaluate each document based on multiple aspects such as knowledge, presentation, communication and regularity, and then we classify the document into multiple classes such as strong positive, positive, negative and strong negative for every aspect. This allows us to have wider range of analysis of sentiment towards faculty performance by student. We further analyze

the text at sentence level for subjectivity detection. Subjectivity detection is used for finding out whether the sentence is opinionated or non-opinionated text. For example, "I hate theory classes." "I like the way of teaching." Here when we analyze the sentence for particular teacher, then the first example is objective type of sentence, whereas the second example has the sentiment of a particular teacher. As we can see the first sentence has sentiment bearing word "hate". This word has a negative meaning for theory courses, but not for the teacher. Our approach allows us to remove objective type of sentences from an entity perspective, for computation of the score of that entity. This helps us to improve the accuracy of system.

Negation problems are not solvable at document level or sentence level. Handling of negation is very important task in sentiment analysis. For example "I do not like the way he teaches." "The battery is not good, but I like the screen of mobile." In first example, 'like' word has a positive meaning, but overall sentence has a negative meaning. In second example, the negation word has an impact on first half of the sentence, but in second part of sentence has positive sentiment towards the screen of a mobile. This has to be handled at entity level. To take care of negation, we use bi-gram as a feature. The TF-IDF value of the negating feature is replaced by the anonym feature TF-IDF value. [2]

- **SOLUTIONS APPROACHES USED**

After reviewing 35 research papers on sentiment analysis, following issues were found, which has to be addressed, while the designing and implementation of Movie Reviews analysis Using SentiWordNet.

Issue 1: Document level sentiment classification
Issue 2: Sentence level sentiment classification
Issue 3: Feature level sentiment classification

The discussion of various solution approaches under specific issues has been presented below.

1. Solution Approaches to the issue "Document level sentiment classification".

Document level sentiment classification means to determine the overall sentiment orientation of the document depends on classes which can be positive, negative or neutral. In order to find out the document level sentiment analysis score, approaches are used based on the underlying entity or aspect based scores [3] and in such approaches the document is split into multiple classes with multiple aspects which are taken into consideration. The approach will allow us to automatically take care of the current problems of document level sentiment analysis, such as, the entity identification, the subjectivity detection and the negation.

Another approach is to find and record the statistical information such as the words and their frequency (TF) in every paragraph. In this approach, the Co-existent frequency of the same words in sequential paragraphs indicates the similarity of these paragraphs.

2. Solution Approaches to the issue "Sentence level sentiment classification".

Sentence level classification means each sentence as a separate unit and assumes that sentence should contain only one opinion. An approach for Sentence-level Sentiment Classification is Dependency Tree based Approach [5]. In contrast to document, a sentence just contained little information and a small set of features which could be used for the sentiment classification. The Dependency Tree based Approach introduced the dependency tree into the sentence-level sentiment classification and then combined flat features with structured features to form a novel feature representation. Another proposed approach is a rule based domain independent sentiment analysis method [6]. The method classified the subjective and the objective sentences from reviews and blog comments. In the first step, sentences were split into subjective and objective ones based on lexical dictionary. Subjective sentences were further processed for extraction to classify as positive, negative or neutral opinions. A rule based lexicon method was used for the classification of subjective and objective sentences.

Another approach is statistical learning based computational method [7] for the automatic construction of domain-specific sentiment lexicons to enhance cross domain sentiment analysis.

3. Solution Approaches to the issue "Feature level sentiment classification".

Feature level classification means to produce a feature-based opinion summary of multiple reviews.

One approach of feature level classification is, "SENTIMENT FUZZY CLASSIFICATION"[8]. Sentiment polarity is vague with regard to its conceptual extension. There is not a clear boundary between the concepts of "positive", "neutral" and "negative". To better handle such intrinsic fuzziness in sentiment polarity, in this approach author apply the fuzzy set theory to sentiment classification. To do so, they first redefine sentiment classes as three fuzzy sets, and then apply existing fuzzy distributions to construct membership functions for the three sentiment fuzzy sets. A fuzzy set is defined by a membership, function. And the simplest statistical approach for feature selection is to use the most frequently occurring words in the corpus as polarity indicators. The majority of the approaches for sentiment analysis involve a two-step process:

- Identify the parts of the document to contribute the positive or negative sentiments.
- Join these parts of the document in ways that increases the odds of the document falling into one of these two polar categories.

Another approach is based on an unsupervised linguistic approach [9]. In this, author has used a pattern-based method which applies a classification rule according to which each review is classified as positive or negative. In this approach, they used SentiWordNet to calculate overall sentiment score of each sentence. In this work, they proposed a domain dependent rule based method for semantically classifying sentiment from online customer reviews and comments.

Another approach is automated opinion mining approach. This task of automatic opinion mining can be done mainly at three different levels, which are document level, sentence level and aspect level. Most of the previous work is done in the field of document or sentence level opinion mining. In this, author has proposed a new syntactic based approach for it, which uses syntactic dependency, aggregate score of opinion words, SentiWordNet and aspect table together for opinion mining.

- **TECHNOLOGY USED**

In order to simulate our proposed work of movies review analysis, we have to make use of SentiWordNet and the software which we have used in our proposed work is,

- *Eclipse Keepler*
- *Java SDK*

Eclipse Keepler

In computer programming, Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in Java and its primary use is for developing Java applications, but it may also be used to develop applications in other programming languages through the use of plug-ins, including: ADA, ABAP, C, C++, COBOL, Fortran, Haskell, JavaScript, Lasso, Lua, NATURAL, Perl, PHP, Prolog, Python, R, Ruby (including Ruby on Rails framework), Scala, Clojure, Groovy, Scheme, and Erlang.

Java SDKs

The Java Development Kit (JDK) is an implementation of either one of the Java SE, Java EE or Java ME platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris, Linux, Mac OS X or Windows. The JDK includes a private JVM and a few other resources to finish the development of a Java

Application. Since the introduction of the Java platform, it has been by far the most widely used Software Development Kit (SDK).On 17 November 2006, Sun announced that it would be released under the GNU General Public License (GPL), thus making it free software. This happened in large part on 8 May 2007, when Sun contributed the source code to the OpenJDK.

To perform Data Pre-processing of movie reviews by doing,

-Tokenization
-Stop Word Removal
-Stemming
-POS Tagging using Stanford Parser

- Web Users: - To determine the polarity of the sentences, based on aspects, large numbers of reviews are collected from the Web.
- Blogs and Reviews: - There are a lot of websites on the Internet where the large numbers of reviews are available. For Example: RottenTomatoes.com, reviews.imdb.com, twitter.com etc. is used to collect the reviews.
- Input Documents: - After collecting the reviews, they are sent to the POS tagging module where POS tagger tags all the words of the sentences to their appropriate part of speech tag. POS Tagging is an important phase of opinion mining. It is necessary to determine the features, POS tagging can be done manually or with the help of POS tagger. Manual POS tagging of the reviews take lots of time. Here, POS tagger is used to tag all the words of reviews.
- POS tagged output: - The processed input documents are collected after tagging process, the part of speech tagger is applied in input documents and output is collected as tagged output in noun, pronoun, adjective, adverb etc.
- List of opinions: - From tagged output, the list of opinion words is extracted and stored in the database for further process of analysis like good, bad etc.
- List of feature words:- From the tagged output, the list of feature words like story, casting, music etc. are stored in another database from the sentences .
- Feature Based Polarity: - The polarity of the sentences is determined for each feature. Polarity is determined on the basis of majority of opinion words. If the number of positive words is more, then the polarity of the sentence is positive otherwise the polarity is negative and if the number of positive and negative words is equal then the sentence shows the neutral polarity. The scores of word are calculated by using SentiWordNet Tool.

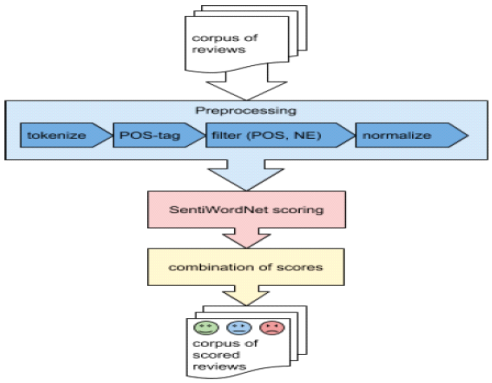

Fig.1.Flowchart for opinion mining process applied on reviews using SentiWordNet

- **DATA AND RESULTS**

This section includes the comparative analysis of the solution approaches based on the variable and parameters used by the researchers in their proposed solution and the results which they have obtained, along with variations in the results. The issue-wise Tables 1 and 2 have been included, which describes the solution approach, its reference, input and output variables used, and the result of the solution approach.

Table 1: Solution Approaches for Different Domains.

Authors Name	**Solution Approach**	**Input Parameters**		**Results**	**Tools**
		Domain	**Reviews**		
Zhongchao Fei, Jian Liu, Gengfeng Wu [02]	Phrase patterns based on unsupervised learning method , POS tagging	Sports	Total-320 Pos-170 Neg-150	Sentiment classification-86%	
Ethan Zhang [17]	Bayesian logistic regression method, Pearson's correlation coefficient method	Movies	82,437 - documents 6,511-features (5,604 adjective)	UCSCAUTO (The product of the retrieval score and the predicted probability)gave better performance with 61%	SentiWordNet 1.0, Lemur Toolkit
Farah Benamara, Carmine Cesarano, Diego Reforgiato [03]	adverb-adjective Combinations (AAC)sentiment analysis technique	News	200 documents, 400 blog posts	Showed adjectives as more important than adverbs , AAC was better than adjective and adverb based scoring.	OASYS (An Opinion analysis system)
AnnDevitt, Khurshid Ahmad [04]	Cohesion-based Text Representation algorithm	News	Total- 30 texts	Showed modifiers as better indicators of text polarity than other word classes (non modifiers) Postive polarity-70% Negative polarity-50%	Wordnet
Kerstin Denecke [12]	LingPipe Language Identification Classifier, SentiWordNet Classifier with classification rule, SentiWordNet Classifier with machine learning,	News Movies	535 articals Neg-1000 Pos-1000 in English Neg-100,Pos-100 in German	Sentiment classification-(For English) (a) 54% (b) 51% (c) 62% (for German) (a)59% (b)58% (c) 66%	PROMT eXcellent Translation (XT) Tool

Table 2: Solution Approaches for Different Domains Using SentiWordNet

Author's Name	**Solution Approach**	**Input Parameters**		**Results**	**Tools**
		Domains	**Reviews**		
Bruno Ohana , Brendan Tierney [05]	Stanford Part of Speech Tagger,SVM Algorithm	Movie		Sentiment classification-65.85% by sentiwordnet	SentiWordNet
Ankit Ramteke , Pushpak Bhattacharyya [27]	Naïve approach for thwarting detection ,SVM, domain ontology parsing technique		Total-1196	non-thwarted-98% thwarted-73% Sentiment classification-56.3%	SentiWordNet
Takashi Inui , Mikio Yamamoto [06]	sentiment-oriented sentence filtering method, SVM	Movie	1392-English (609 pos, 783 neg) 724 - Japanese (340 pos, 384 neg)	Removed the translation errors in multilingual review classification with 87% classification accuracy	Wordnet, Japanese-language dictionary
Si Li, Hao Zhang, Weiran Xu, Guang Chen and Jun Guo [13]	Left-Middle-Right template and CRF(Conditional random fields) , TF-IDF	Movies	829 website texts(164 pos , 108 neg and 557 neutral) 1741 tagged sentiment words, 10000 sentences without emotion	Sentiment classification-96.4% compared to sentence level SA, removed the transitional words problems	
Raymond Y.K. Lau ,Wenping Zhang, Peter D. Bruza [28]	Kullback-Leibler (KL) divergence statistical learning method,TF-IDF	Product	996,167 reviews	Sentiment classification - 68.5%	OpinionFinder
Aurangzeb khan, Baharum Baharudin [29]	Rule based approach, POS tagging	Sports	1500 comments	Sentiment classification-83% Word Sense Disambiguation was handeled	SentiWordNet,

REFERENCES

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka," SentiFul: A Lexicon for Sentiment Analysis" IEEE transactions on affecting computing, vol. 2, no. 1, january-march 2011 1949-3045/11/$26.00 _ 2011 IEEE

Alexandra Balahur, Andrés Montoyo," A Feature Dependent Method for Opinion Mining and Classification" 978-1-4244-2780-2/08/$25.00 ©2008 IEEE

Alexandre Trilla and Francesc Alías.” Sentence-Based Sentiment Analysis for Expressive Text-to-Speech1” IEEE transactions on audio, speech, and language processing, vol. 21, no. 2, February 2013 1558-7916/$31.00 © 2012 IEEE

Andrea Esuli and Fabrizio Sebastiani. 2006.“Determining term subjectivity and term orientation for opinion mining” In Proceedings of eacl-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT. Forthcoming.

Ann Devitt, Khurshid Ahmad “Sentiment Polarity Identification in Financial News: A Cohesion-based Approach” Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 984–991, Prague, Czech Republic, June 2007. c 2007 Association for Computational Linguistics

Antonis Koukourikos1, Giannis Stoitsis2,3, Pythagoras Karampiperis1” Sentiment Analysis: A tool for Rating Attribution to Content in Recommender Systems” 2012.

Aniket Dalal, Kumar Nagaraj, Uma Sawant” Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach”

Ankit Ramteke, Akshat Malu, Pushpak Bhattacharyya “Detecting Turnarounds in Sentiment Analysis: Thwarting” 2012

Aurangzeb khan Baharum Baharudin,"Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs", 2011 IEEE.

(A complete list of references is available upon request from Surbhi Maheshwari at subhi.kabra@gmail.com)