Unsupervised Opinion Mining From Text Reviews Using SentiWordNet

Vibha Soni[1], Meenakshi R Patel[2]

[1]M.Tech. Scholar, Department of Computer Science and Engineering RCET Bhilai (C.G.) INDIA

[2]Associate Professor, Department of Computer Science and Engineering RCET Bhilai (CG) INDIA

***Abstract*—Opinion mining is a discipline or area of text classification which continues gives contribution in research field. Sentiment analysis is one another name of Opinion mining. Opinion Mining analyse and classify the user generated data like reviews, blogs, comments, articles etc. Nowadays every people use web services and gives their opinions about every field, domain or peoples. The main objective of Opinion mining is Sentiment Classification i.e. to classify the opinion into positive or negative classes. There are basically two approaches first machine learning Or Supervised learning techniques and other unsupervised learning techniques. In this paper an unsupervised lexicon technique is used for Sentiment Classification.**

Keywords: Opinion Mining, Sentiment Classification, Sentiment Analysis, SentiWordNet, Lexicon, Sentiment Polarity.

I. INTRODUCTION

Today internet and worldwide web become the necessary part of the human being. It provides a vital resource for finding any information about anything. It can be about any domain like business, entertainment, politics, and social media. Today people share their idea with everyone through internet and also take opinion form it. Web becomes a decision making resource. Everyone takes opinion before taking any decision, or buying any products etc. In general opinion may be the result of a people's personal feelings, beliefs, sentiments and desires etc. For this reason Opinion mining become popular research topic.

Opinion mining is a type of text mining which classify the text into several classes. Sentiment analysis which also known as Opinion mining use some algorithm techniques to cauterize the user opinions into positive, negative and neutral classes .This categorization of text is called polarity of text. The main objective of Sentiment analysis is classification of sentiment. It classifies the given text into three level document level, sentence level, and entity/aspect level.

The document-level sentiment analysis problem is essentially as follows: given a set of documents D, a sentiment analysis algorithm classifies each document d ϵ D into one of the two classes 'positive' or 'negative'. Positive label denotes that the document d expresses an overall positive opinion and negative label means that d expresses an overall negative opinion of the user. The document-level sentiment analysis assumes that each document contains opinion of user about a single object.

The aspect-level sentiment analysis on the other hand assumes that a document contains opinion about multiple aspects/entities of one or more objects in a document. It is therefore necessary to identify about which entity an opinion is directed at.

After introduction section the paper is organized into following sections: Section II describes the background and related work. Section III describes methodology used for opinion mining. Section VI describes dataset and experimental setup. Section V describes generated result. Section VI discusses the conclusion part of the paper.

II. BACKGROUND AND RELATED WORK

Sentiment classification is the most widely studied and research topic nowadays. Basically there are two techniques for Sentiment analysis. First one is a supervised learning technique which is based on machine learning classifiers. It uses training on labelled data before they can be applied to the actual sentiment classification task. Naive Bayesian, Support Vector Machines (SVM), maximum entropy etc. are existing supervised learning methods can be readily applied to sentiment classification. Several numbers of papers mentioning "sentiment analysis" focus on the specific application of classifying review as to their polarity [1].

The Second technique is based on lexicon which is also called unsupervised learning technique. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions. It classifies the document using semantic orientation or sentiment dictionary for computing sentiment polarity of a text. SO-PMI-IR [Semantic Orientation - Pointwise Mutual Information-Information Retrieval] algorithm is another unsupervised approach, which uses the mutual occurrence frequency of selected words to compute the sentiment polarity [1].

Unsupervised learning is basically based on sentiment words. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions. The syntactic patterns are composed based on

part-of-speech (POS) tags. Part of speech defines that for given a sentence; determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Words of different POS may be treated differently. In Opinion mining adjectives are important indicators of opinions. The standard Penn Treebank POS tags are mostly used Standard POS tags. [1]. It computes semantic orientation of documents based on aggregated semantic orientation values of selected opinionated POS tags extracted from the document.

A number of unsupervised learning approaches take the credit of first creating a sentiment lexicon in an unsupervised manner, and then determining the degree of positivity (or subjectivity) of a text unit via some function based on the positive and negative (or simply subjective) indicators, as determined by the lexicon, within it.

There is lots of work had been done on unsupervised learning techniques. First Peter D. Turney presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs also introduces a simple algorithm for unsupervised learning of semantic orientation from extremely large corpora. The method involves issuing queries to a Web search engine and using point wise mutual information to analyse the results [2] [3].

Arzu Baloglu, Mehmet S. Aktas introduce an architecture, implementation, and evaluation of a Weblog mining application, called the Blog Miner, which extracts and classifies opinions and emotions (or sentiment) from the contents of weblogs. They use SentiWordNet for finding Sentiment Score [4].

Vivek Kumar Singh, Mousumi Mukherjee, Ghanshyam Kumar Mehta, Shekhar Garg, and Nisha Tiwari uses machine learning technique and unsupervised learning technique. The paper collects blog data on three interesting topics, transforming the collected blog data into vector space representation, and then performing opinion mining [5].

Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar surveyed and analysed various techniques that have been developed for the key tasks of opinion mining. On the basis of survey and analysis, it provides an overall picture of what is involved in developing a software system for opinion mining [6].

V.K. Singh, M.K. Singh, P. Waila, uses machine learning approach with unsupervised learning. The unsupervised learning use Semantic Orientation Pointwise Mutual Information- Information-Retrieval known as SO-PMI-IR algorithm for sentiment analysis. The paper uses both pre-existing data sets and own dataset collection comprising of a large number of user reviews for Hindi movies. The Naive Bayes and SVM approaches were implemented in multiple folds [7].

Other than this several paper were published which is based on unsupervised opinion mining using semantic orientation pointwise mutual information [8], [9], [10], [11].

Hussam Hamdan, Frederic Bechet and Patrice Bellot propose to use many features in order to improve a trained classifier of Twitter messages, these features extend the feature vector of unigram model by the concepts extracted from DBpedia, the verb groups and the similar adjectives extracted from WordNet, the Senti-features extracted using SentiWordNet and some useful domain specific features. They built a dictionary for emotion icons, abbreviation and slang words in tweets which is useful before extending the tweets with different features [12].

Vivek Kumar Singh, Rajesh Piryani, Pranav Waila, and Madhavi Devaraj presents evaluation of the performance of different machine learning as well as lexicon based methods for sentiment analysis of texts obtained from variety of sources two aspects of sentiment analysis [13].

Bruno Ohana, Brendan Tierney presents the results of applying the SentiWordNet lexical resource to the problem of automatic sentiment classification of film reviews. They counts positive and negative term scores to determine sentiment orientation, and an improvement is presented by building a dataset of relevant features using SentiWordNet as source, and applied to a machine learning classifier [14].

Monalisa Ghosh, Animesh Kar describes a simple technique to perform sentiment classification based on an unsupervised linguistic approach. They use SentiWordNet to calculate overall sentiment score of each sentence [15].

Marco Guerini, Lorenzo Gatti, and Marco Turchi present various approaches based on SentiWordNet. They compare the most often used techniques together with newly proposed ones and incorporate all of them in a learning framework to see whether blending them can further improve the estimation of prior polarity scores.

III. METHODOLOGY

In this paper we are using unsupervised lexicon based method based on the SentiWordNet lexicon. In introduction section we had defined that sentiment analysis

can be done either document level or aspect level. Here we are using aspect level sentiment analysis. No one can judge any review or blog using one single entity. There are lots of aspects through which one can identify it. These aspects may be positive or negative. Thus Aspect level analysis finds out the positive and negative aspects of any object. The main objective of aspect level analysis is to identify the features which are to be analysed, extract this features and calculate its polarity. For this work we perform the following steps.

DATA COLLECTION - The first step of opinion mining is to design a dataset. Here collect opinions from various sources. Taking reviews, blogs etc. of particular domain which is selected for analysis. Since aspect level analysis is based on aspects so we had search different websites, magazines for the identification of aspects.

DATA PREPROCESSING - The data preparation step performs necessary data preprocessing and cleaning on the dataset for the subsequent analysis. Some commonly used preprocessing steps include removing non-textual contents and markup tags (for HTML pages), and removing information about the reviews that are not required for sentiment analysis, such as review dates and reviewers' names.

FEATURE EXTRACTION - Feature identification and selection is most important task of opinion mining. There is more than one name for the same aspects. For example someone use "story of the book is good" or someone use "the storyline of the book is fantastic" but meaning of story and storyline is same. Hence also identify same synonym of the different aspects and design an aspect matrix. The following Table I describes an example of Aspect Matrix of mobile phone

TABLE I
ASPECT MATRIX

Aspects	**Synonyms**
phone	handset, cordless, wireless
camera	megapixels, pixels, lens
screen	Display
speaker	speakerphones, sound, volume
battery	batteries, battery life, ultra-power
resistance	shock-resistant, water-resistant, waterproof, dust-resistant
apps	Software
weight	lightweight, thinner, slim
design	plastic, rubber
memory	storage, space
charging port	charger, charging
internet	wi-fi, bluetooth, infrared

POS TAGGING – After designing the aspect matrix parse the data using parser. POS tagger parses a sentence or document and tags each term with its part of speech. For POS tagging we used the Stanford POS tagger. This tagger used by splitting text data into sentences and to produce the part-of-speech tag for each word (whether the word is a noun, verb , adjective, etc.). The following shows a sentence with POS tags.

"The feel of the phone is the best of the series."

When we apply the POS-tagger then it generates the following part of speech of the sentence.

"The_DT feel_NN of_IN the_DT phone_NN is_VBZ the_DT best_JJS of_IN the_DT series_NN."

CALCULATE SENTIMENT POLARITY – Now find out sentiment polarity using SentiWordNet. SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment numerical scores: Obj(s), Pos(s) and Neg(s) describing how Objective, Positive and Negative the terms contained in the synset are. Each of the three scores ranges from 0.0 to 1.0, and their sum is 1.0 for each synset and the entries contain the parts of speech category of the displayed entry, its positivity, its negativity, and the list of synonyms. The word or lemma present in the form lemma #sense-number, where the first sense corresponds to the most frequent and different word senses can have different polarities. The following two tables describe the SentiWordNet where Table II defines the SentiWordNet database record structure and Table III shows some sentiment scores associated to SentiWordNet entries.

TABLE II
SENTIWORDNET DATABASE STRUCTURE

Fields	Descriptions
POS	Part of speech linked with synset. This can take four possible values: a = adjective n = noun v = verb r = adverb
Offset	Numerical ID which associated with part of speech uniquely Identifies a synset in the database.
PosScore	Positive score for this synset. This is a numerical value ranging from 0 to 1.
NegScore	Negative score for this synset. This is a numerical value ranging from 0 to 1.
SynsetTerms	List of all terms included in this synset

TABLE II
SENTIWORDNET LIBRARIES

POS	Offset	PosScore	NegScore	SynsetTerms
a	1740	0.125	0	able#1
a	2098	0	0.75	unable#1

n	388959	0	0	divarication#1
n	389043	0	0	fibrillation#2
r	76948	0.625	0	brazenly#1
r	77042	0.125	0.5	brilliantly#2
v	1827745	0	0	slobber_over#1
v	1827858	0.625	0.125	look_up_to#1

There are mainly two methods of SentiWordNet. First one is SWN(AAC) which is Adverb+Adjective combination and another is SWN(AAAVC) which is Adverb+Adjective, Adverb+Verb combination. We apply both this methods for finding out of sentiment polarity of all aspects in one review. Similarly find out polarity for all reviews. Then aggregate score for a particular aspect from all the reviews and find out a net score of that aspect. At the end generate a sentiment profile with respect to selected aspects and their sentiment score.

IV. Dataset And Experimental Setup

In this paper we are using multiple dataset where some of them are existing dataset and some of them are new dataset. For the implementation of SentiWordNet we use java Netbeans IDE.

V. Result

Here we are using multiple dataset. Aspect-level sentiment analysis generates sentiment profile. We implemented all the existing dataset. Also we have design new dataset for mobile. For this we collect reviews from Amazon.com websites and implemented it with both scheme SWN(AAC) and SWN(AAAVC). Here we present the example of mobile Samsung Galaxy S5- Wireless on selected aspects of a review. The figures 1 and 2 present sentiment profiles of this mobile phone using both methods. Both methods present similar methods. Both sentiment profiles of this phone defines positive results with many aspects.

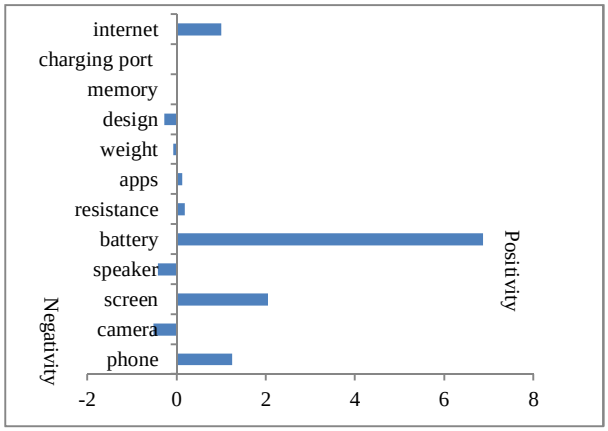

Fig. 1 shows the sentiment profile of Samsung Galaxy S5 with SWN(AAC) method.

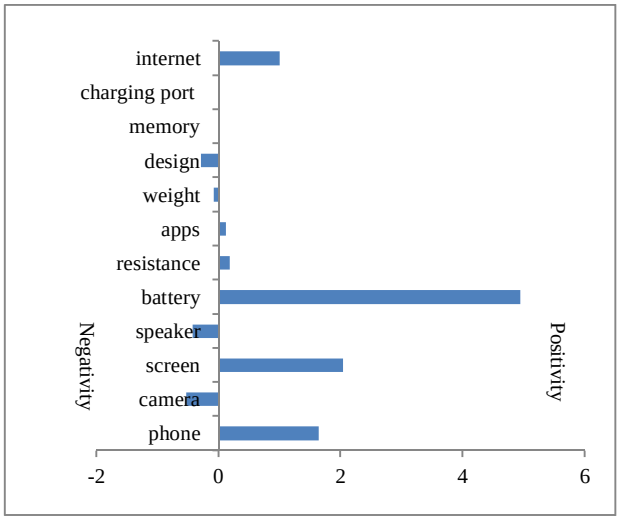

Fig. 2 shows the sentiment profile of Samsung Galaxy S5 with SWN(AAAVC) method.

VI. Conclusion

Opinion mining become popular research area due to the increasing number of internet users, social media etc. Here we have work on aspect level analysis using SentiWordNet. The method used here is very simple and domain independent. In this paper we present our experiment with reviews which generate great result. In future we will work it with the blogs. Also improve the accuracy of the method.

Acknowledgement

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them. I am highly indebted to Prof. Meenakshi R Patel for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. I would like to express my gratitude towards my parents & professors of my college for their kind cooperation and encouragement which help me in completion of this project. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

References

[1] Bing Liu, "Sentiment Analysis and Opinion Mining, " Morgan & Claypool Publishers, May 2012.

[2] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation applied to Unsupervised Classification of Reviews," *National Research Council of Canada* , pp. 417-424, 2002.

[3] P. Turney and M. L. Littman , "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus," *National Research Council of Canada*, pp. 9, 2002.

[4] A. Baloglu, Mehmat A. Aktas, "An Automated Framework for Mining Reviews from Blogosphere," *International Journal on Advances in Internet Technology*, vol. 3, 2010.

[5] V. K. Singh, M. Mukherjee and G. k. Mehta, "Opinion Mining from Weblogs and its Relevance for socio-political Research," *Institute for computer science, Social Informatics, and Telecommunications Engineering*, vol. II, pp. 134-145, 2012.

[6] A. Buche, Dr. M. B. Chandak and A. Zadgaonkar, "Opinion Mining and Analysis: Survey," *International Journal on Natural Language Computing*, vol. 2, June 2013.

[7] V. K. Singh, M. K. Singh and P. Walia, "Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews," *IEEE International Conference.*, 2012.

[8] W. Guangwei, A. Kenji "An Unsupervised Opinion Mining Approach for Japanese Weblog Reputation Information Using an Improved SO-PMI Algorithm," *IEICE Transactions on Information and Systems*, 2010.

[9] J. Rothfels, J. Tibshirani, "Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items," 2010.

[10] Wei Jin, H. Hay Ho and R. K. Shrihari, "OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction," *ACM*, 2009.

[11] D. Shaw, "Opinion Mining of Movie Reviews," *ENEE752*, 2009.

[12] H. Hamdan, F. Bechet and P. Bellot, " Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging," *Association For Computational Linguistic*, pp. 455-459, June -2013.

[13] V. K. Singh, R. Piryani, P. Walia and M. Devaraj, "Computing Sentimen Polarity of Texts at Document and Aspect Levels," *ECTI Transaction On computer and Information Technology*,vol. 8, no. 1, May 2014.

[14] B. Ohana, B. Tierney, "Sentiment Classification of Reviews Using SentiWordNet," *9th. IT&T Conference, Dublin Institute of Technology*, October, 2009.

[15] M. Ghosh, A. Kar "Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using Sentiwordnet 3.0," *International Journal of Engineering Research & Technology (IJERT)* , vol. 2, no. 9, September -2013.