Extracting Opinion for a Product based on User Reviews using R-Language

R. Rengaraj[1], Elavarsi[2], Raghavi[3], Raihan Parveen[4], Rajeshwari[5]
[1, 2, 3, 4, 5] Department of Information Technology
[1, 2, 3, 4, 5] Saranathan College of Engineering,Trichy-620012, India

Abstract- *In e-shopping many web domains are developed drastically in recent years. These websites provide plenty of products from different companies. Most of the companies are unknown for the buyer. This creates an ambiguity in choosing the product. The proposed system resolves this ambiguity by categorizing, ranking the extracted reviews through Natural Language Processing Technique and represented in R-language.*

Keywords- Opinion mining, Sentence analysis, Opinion extraction, Natural Language Processing.

I. INTRODUCTION

In World Wide Web e-shopping is a fast growing domain. It has influenced many buyers to switch from traditional to electronic. More options for a product create an ambiguity for the buyer. Now a days, buyers focus on reviews before choosing a product. Reviews are the experiences faced by previous buyers of the product. User reviews are used as feedback to the manufacturers in promoting the product. The survey of BrightLocal says that 88% of consumers buy the product with the help of user reviews. Situations may arise in which people may not be able to go through all the reviews as it is time consuming and also very difficult to understand the context.

To overcome this problem we have proposed a concept called opinion mining[1][5]. In this paper we use NLP as language processor.User's opinion about the product is generated from the user reviews which is extracted from the websites,blogs,etc…In opinion generation process we first preprocess the reviews and then Part-of-speech tagging is done for getting the component-feature pair. we find polarity for individual user's review by counting the positive and negative opinion.Then polarity for the component with respect to its feature is calculated. Finally we display the result in digramatical format which includes overall review polarity,positive and negative report and componenet-feature report.

II. LITERATURE SURVEY

Kang Liu proposed novel approach based on the partially supervised alignment model which helps in identifying opinion relations as an alignment. Then a graph based co-ranking algorithm is used to estimate the confidence of each candidate. Candidates with higher confidence are extracted as opinion word or opinion target. Compared to syntax based method

word alignment model effectively process the negative effect of parsing errors when dealing with informal online texts .This model obtains better precision for long-span relations when compared to unsupervised alignment model[8].

Angulakshmi signifies about the analysis tools and techniques used for opinion mining. Tools like Review seer tool, Web Fountain, Red Opal tool and opinion observer for analyzing and comparing opinions. This paper uses corpus based and dictionary based approach for sentiment analysis[6]. Double Propagation Algorithm is used for extracting product features and sentence words.

Diana Mynardsays ; rule based learning is used instead of machine learning for analyzing reviews which are entered by the user in social media. This paper deals with challenges in mining user reviews[7]. The rule based approach performs linguistic analysis and builds on a number of linguistic subcomponents to generate the final polarity and score.

Gaurav Dubey has done opinion mining based on three main steps 1) Parts of Speech Tagging: Identifying and categorizing words in a piece of text in a given language into defined parts of speech, 2) Rule mining and identifying opinion words: Inventing new methods, relations between two or more variables and categorizing words into positive and negative and 3) Summarizing and displaying the end result[9].

He has also proposed two techniques called supervised and unsupervised learning. Supervised learning technique can process large dataset with good performance. The limitation is, it has to be trained before processing. Unsupervised learning technique can process only small dataset with poor performance but it needs no training before processing [9].

Amiya Kumar Tripathy extracts opinion of the user using SentiWordNet dictionary and NLP as a language processor. Opinion of the product is extracted with its features and categorized into positive or negative response. Opinion words are classified into enhancers and reducers which provides positive or negative effect on the polarity of the opinion. Also uses conditional probability to find the correlation between the components and feature. Comparing the value with threshold value, the existence of relation is found [10].

III. METHODOLOGY

We have used Natural Language processing Technique to extract the user reviews and process it. Natural language processing has come a long way since its foundations were laid in the 1940s and 50s (for an introduction see, e.g., Jurafsky and Martin (2008): Speech and Language Processing, Pearson Prentice Hall). This CRAN task view collects relevant R packages that support computational linguists in conducting analysis of speech and language on a variety of levels - setting focus on words, syntax, semantics, and pragmatics[11].

The implementation is done using R language.R is a language and environment for statistical computing and graphics which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control[11].R is an integrated suite of software facilities for data manipulation, calculation and graphical display.[11] It includes an effective data handling and storage facility, calculation of arrays is done using operators and in particular matrices. A large, coherent, integrated collection of intermediate tools for data analysis. Graphical facilities for data analysis and display either on-screen or on hardcopy. It is a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN (Comprehensive R Archive Network) family of Internet sites covering a very wide range of modern statistics.

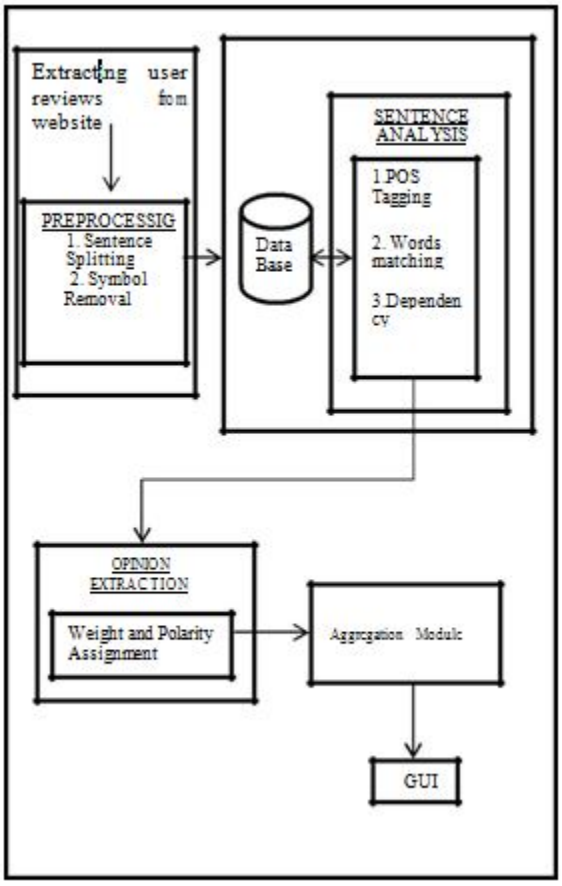
In recent years, R teams have elaborated a framework to be used in packages dealing with the processing of written material: the package tm. Extension packages in this area are highly recommended to interface with tm's basic routines and users are cordially invited to join in the discussion on further developments of this framework package. To get into natural language processing, the cRunchservice and tutorials may be helpful.we use packages like RMySQL, RCurl, XML, plotrix, OpenNLP, NLP, stringr,etc…RMySQL is a database interface and MySQL driver for R. This version complies with the database interface definition as implemented in the package DBI 0.2-2. XML: Tools for Parsing and Generating XML. Within R and S-Plus.Many approaches for both reading and creating XML (and HTML) documents (including DTDs), both local and accessible via HTTP or FTP. Also offers access to an 'XPath' "interpreter".RCurl: General Network (HTTP/FTP/...) Client Interface for R.A wrapper for 'libcurl' <http://curl.haxx.se/libcurl/> Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server. stringr: Simple, Consistent Wrappers for Common String Operations. A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with "NA"'s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another. openNLP: An interface to the Apache OpenNLP tools (version 1.5.3). The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. See <http://opennlp.apache.org/> for more information. plotrix: Various Plotting Functions,Lots of plots, various labeling, axis and color scaling functions.

A) Abbreviations and Acronyms

– -NLP(Natural Language Processing) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages.NLP is related to the area of human– computer interaction[1].

– -POS (Part of Speech) tagging is identifying and categorizing words in a piece of text in a given language into defined part of speech

a) System Architecture



B) Preprocessing user reviews

The user reviews are obtained through blogs, Websites, forums, etc. For our work, we have targeted these reviews which are readily available to the companies. The extraction of reviews present in blogs, forums requires customized crawlers which will go through the suggested site and extract the reviews for the system. The extracted reviews are then cleaned to remove symbols, split the sentence. Then the preprocessed reviews are then stored in the database.

C) Sentence Analysis

The main function of Sentence Analysis is to find out the Individual part of speech for each word of the sentence. The POS Tagging is done to find the component-feature pair. The positive and negative words that are stored in data base is matched with the generated user reviews to find the negative and positive opinion. The component-feature pair is found using POS Tagging in which the noun and adjective is compared. Then the component along with feature is stored in database.

D) Opnion Extraction

After storing the component-feature pair weight and polarity is assigned to them. Firstly it is checked whether an entry for the given pair is available or not. If it is available then the respective polarity and weights are assigned to the pair. If it is not available then a general polarity is assigned to the pair. Next, the weight and polarity of the modifiers is obtained. The modifiers as stated earlier, are the words which either enhance or reduce the polarity of opinion words. Based on the final weight polarity is obtained for the sentence.

E) Aggregation

This module gets the extracted data and then aggregates the positive and negative reviews. It checks whether the product gives positive or negative opinion. Then the result is displayed in graphical format.

IV. IMPLEMENTATION

Our user interface is framed using php and back end is MySQL. We have implemented our project in R language and displayed the analysis result using php which consist of piechart, barcharts..etc.

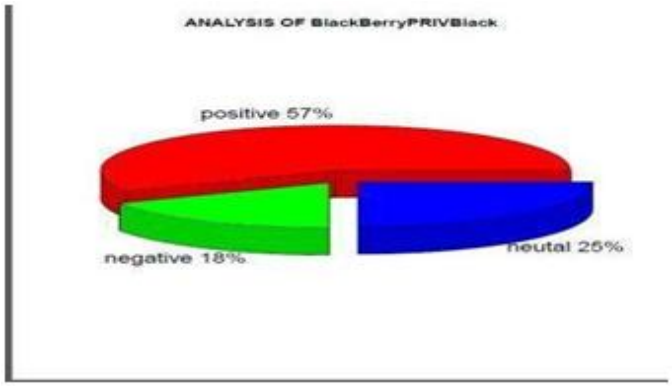
b) Implementation in R studio(IDE)

c) PHP User interface

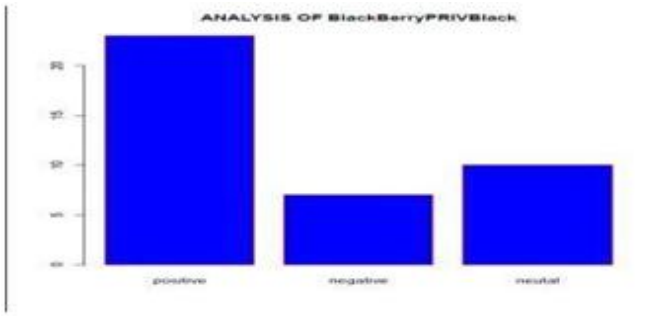
V. CONCLUSION

User reviews are first extracted from the website. The reviews are first preprocessed to remove the unwanted symbols. Then sentence splitting is done for the reviews having multiple sentences. Then every sentence goes through sentence analysis for POS tagging, positive and matching the positive and negative words. In opinion extraction module component-feature pair is obtained and then weights are assigned to these pairs. Finally polarity is assigned. At last the final result is displayed in graphical format. We processed those user reviews by relating component-feature pair and Weightage is given to those pairs. Finally based upon these weights the polarity is assigned and then displayed in graphical format. The report for positive and negative opinion of a product and component-feature is also displayed separately.

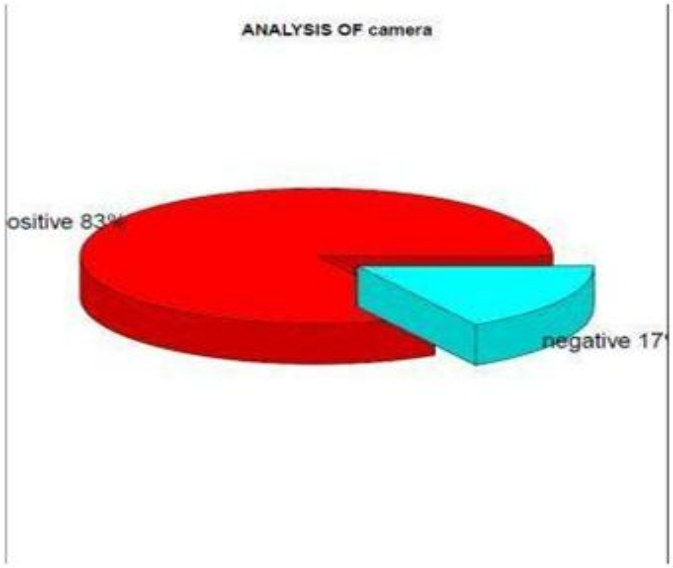
d) Pie chart representation of overall product



e) Bar Chart repesentation of overall product



f) Pie chart representation of particular component of a product



g) Review analysis of products

Product Name	*No.Of.Reviews*	*No.Of.Positive Count*	*No.of.Negative Count*	*No.Of.Positive Reviews*	*No.Of.Negative Reviews*	*No.Of.Neutral Reviews*	Sample Component	
							Camera PositiveCount	*Camera NegativeCount*
Apple iPhone 5s Space Grey16GB	450	128	30	26	4	0	1	0
BlackBerry PRIV Black	40	293	125	18	5	7	4	1
Honor 5X Gold	120	186	53	25	1	4	6	0
Lenovo A7000 Black	410	54	16	23	4	3	2	0
Microsoft Lumia 535 Black 8GB	420	1313	604	260	102	58	65	1

REFERENCES

[1] Lipika Dey, Sk. Mirajul Haque, "Opinion mining from noisy text data",International Journal on Document Analysis and Recognition (IJDAR),Vol. 12, Issue 3, pp. 205-226, September 2009.

[2] Subhabrata Mukherjee, "Sentiment Analysis, A LiteratureSurvey",Indian Institute of Technology, Bombay, India, 29 June 2012

[3] https://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf

[4] http://www.sersc.org/journals/IJAST/vol60/1.pdf

[5] Kavita Ganesan, Hun Duk Kim, "Opinion Mining Tutorial", University of Illinios.

[6] G.Angulakshmi1, 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

[7] Diana Maynard, Kalina Bontcheva, Dominic Rout" Challenges in developing opinion mining tools for social media"Department of Computer Science University of Sheffield Regent Court, Sheffield, S1 4DP, UK diana@dcs.shef.ac.uk

[8] Kang Liu, Liheng Xu, and Jun Zhao"Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model"ieee transactions on knowledge and data engineering, vol. 27, no. 3, march 2015

[9] Gaurav Dubey,Ajay Rana"User Reviews Data Analysis using Opinion Mining on Web"2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015)

[10] Amiya Kumar Tripathy1, Revathy Sundararajan" Opinion Mining from User Reviews"2015 International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India

[11] https://cran.r-project.org