Opinion Mining and Sentiment Analysis on Online Customer Review

Santhosh Kumar K L
Assistant Professor
Department of M Tech CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
santhosh.kumar.kadur@gmail.com

Jayanti Desai
Student
Department of M Tech CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
desaijayanti@gmail.com

Jharna Majumdar
Dean R&D, Prof & Head,
Department of M Tech CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
jharna.majumdar@gmail.com

Abstract— **The opinion mining is very much essential in e-commerce websites, furthermore advantageous with individual. An ever increasing amount of results are stored in the web as well as the amount of people would acquiring items from web are increasing. As a result, the users' reviews or posts are increasing day by day. The reviews toward shipper sites express their feeling. Any organization for example, web forums, discourse groups, blogs etc., there will be an extensive add up for information. Records identified with items on the Web, which are functional to both makers and clients. The process of finding user opinion about the topic or product or problem is called as opinion mining. It can also be defined as the process of automatic extraction of knowledge by means of opinions expressed by the user who is currently using the product about some product is called as opinion mining. Analyzing the emotions from the extracted opinions is defined as Sentiment Analysis. The goal of opinion mining and Sentiment Analysis is to make computer able to recognize and express emotion. This work concentrates on mining reviews from the websites like Amazon, which allows user to freely write the view. It automatically extracts the reviews from the website. It also uses algorithm such as Naïve Bayes classifier, Logistic Regression and SentiWordNet algorithm to classify the review as positive and negative review. At the end we have used quality metric parameters to measure the performance of each algorithm.**

Keywords— Opinion Mining, Sentiment Analysis, Naïve Bayes classifier, Logistic Regression, Senti Word Net

I. INTRODUCTION

Social media plays very important role in almost everybody's day to day life. It allows the people to convey what they think and feel about the products in E-commerce website. This is called as opinion or review. It aims to determine the mood of the writer or attitude of the speaker; it may be either positive or negative towards the product. These positive or negative emotions expressed by the people are known as sentiment. Opinion mining or sentiment analysis refers to the type of Natural Language Processing (NLP), Text-analysis and Computational-Linguistics to identify and extract subjective information in source material.

Different websites allows different review structure to be followed. Some websites allows user to explicitly write the advantageous and disadvantageous, in some cases along with the summary, in other cases there will not be any restrictions on user to write review, so that they can write however they want and express their feelings.

This paper focuses on Amazon product review mining which follows free review structure. There will not be any restrictions on user to write review.

II. RELATED WORK

The ongoing research work related to the Opinion mining and Sentiment Analysis are given in this section. In [1], focuses tools and techniques used in opinion mining. The process of opinion summarization has three main steps, such as "Opinion Retrieval, Opinion Classification and Opinion Summarization." User comments are retrieved from review websites. These comments contain subjective information and they are classified as positive or negative review. Depending upon the frequency of occurrences of features opinion summary is created.

In [2], focuses on review mining and sentiment analysis on Amazon website. Users of the online shopping site Amazon are encouraged to post reviews of the products that they purchase. Amazon employs a 1-to-5 scale for all products, regardless of their category, and it becomes challenging to determine the advantages and disadvantages to different parts of a product. In [3], presents how to mine product features in opinion sentences. It makes use of SentiWordNet based algorithm to find opinion of the sentence.

In [4], aims to provide summarized positive and negative features about products, laws or policies by mining reviews, discussions, forums etc. This approach diligently scans every line of data, and generates a cogent summary of every review (categorized by aspects) along with various graphical visualizations. In [5], proposes rule based hybrid approach. It finds sequential patterns and Normalized Google Distance (NGD) to obtain explicit and implicit aspects. "Aspect-based opinion mining focuses on extraction of aspects from customer reviews and ranking these aspects as positive or negative". In [6], aims is to automate the process of gathering online end user reviews for any given product or service and analysing those reviews in terms of the sentiments expressed

978-1-5090-0612-0/16/$31.00 ©2016 IEEE

about specific features. In [7], e-commerce websites, clients typically aggravate comments, which incorporate those properties of the product, those mentalities of the vendor, express conveyance majority of the data following purchasing the results. The majority of the data gives a critical reference to the point when others purchase results in the website. On assumption analysis and finer-grained idea mining approach concentrates for the resulting features. Past related exploration concentrates on the unequivocal target mining in any case neglects the understood ones. Whereas, those understood features, which need aid intimated toward a portion expressions or phrases, need aid thick, as huge and serious with express users' assumption.

The scientometric mapping of Opinion Mining and Sentiment Analysis (OMSA) is shown in [8] for the duration of 2000-2016. The research publication indexed in Web of Science (WoS) is used as input data. In [9], the Supervised learning approach and Dictionary based techniques are considered for the Sentiment analysis. The precision and recall measures are used to evaluate the accuracy of the algorithm. The survey paper in [10] discuss about the overview of different classification, clustering algorithms and challenges in sentimental analysis and opinion mining.

III. The Proposed Method

This paper concentrates on mining reviews from the websites like amazon.com, which allows user to freely write the view. It automatically extracts the reviews from the website. It also uses algorithm such as Naïve Bayes classifier, Logistic Regression and SentiWordNet algorithm to classify the review as positive and negative review. The Fig.1, shows the data flow of the proposed system. The different processing components of the system are as follows:

A. Text Extraction

After the Login credentials, this module takes the amazon.com URL as the input and extracts all the text from the provide webpage.

B. Source Code Extractor

HTML source code of the webpage is extracted in this module.

C. List of Product

This module will display a list of products from which we have to select a products of our choice to extract review.

D. Display Review List

This module generates the dynamic link and displays all the reviews of the selected product.

E. Stop Word Dictionary

This function contains the stop word list which will be used to eliminate the stop words in the reviews.

F. Algorithm selection

This module allows the user to select any one algorithm among Naïve Bayes, Logistic Regression and SentiWordNet.

G. Calculate Performance

Once the algorithm is selected the training data is loaded and the performance of the algorithm is measured in terms of Recall, Precision and F-measure.

H. Display the Classification Result

This module displays two lists containing positive and negative review separately.

I. Positive and Negative Opinion Dictionary

This function contains the positive and negative word list which will be saved in the two separate text file and later it will be used for sentiment analysis.

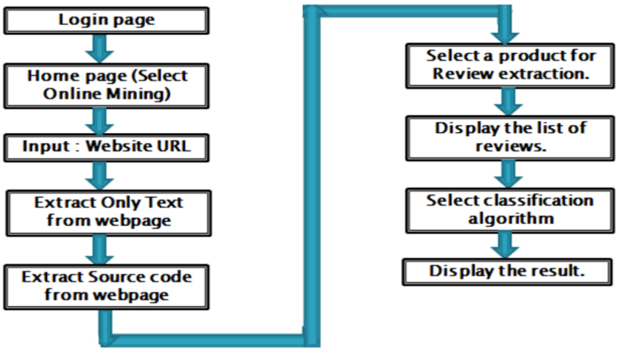
Login page
Home page (Select Online Mining)
Input : Website URL
Extract Only Text from webpage
Extract Source code from webpage
Select a product for Review extraction.
Display the list of reviews.
Select classification algorithm
Display the result.

Fig. 1. Data flow of the proposed system

IV. The Opinion Classification Methodologies

A. Naïve Bayes Text Classification

The Bayesian arrangement is utilized similarly as a probabilistic strategy (Naive Bayes content classification) [11]. Utilizing suitable samples which reflect nice, terrible or impartial sentiments, same should recognize the middle of them. Basic feeling demonstrating combines a statistically based classifier with a dynamical model. Those credulous bayes classifier utilizes single expressions also saying pairs concerning illustration Characteristics. It allocates the input under nice or terrible. The unbiased classes, marks +1, -1 what's more 0 individually. This numerical yield drives a basic first-order dynamical system, whose state speaks to the mimicked enthusiastic state of the experiment's representation.

B. Logistic Regression

Logistic regression has a place with the group of classifiers known as the exponential or log-linear classifiers [12]. Like innocent Bayes, it log-linear classifier works by extricating some set of weighted components from the information, taking logs, and joining them linearly (implying that every element is increased by a weight and afterward included). In fact, logistic regression alludes to a classifier that characterizes a perception into one of two classes, and multinomial logistic regression is utilized when arranging into more than two classes.

While logistic regression in this way varies in the way it calculates probabilities, it is still similar to naive Bayes in being a linear classifier. Logistic regression estimates P(y|x) by separating some set of elements from the input, consolidating them linearly (increasing every element by a weight and adding them up), and afterward applying a combination function.

C. *SentiWordNet*

The SentiWordNet[13] will give acceptable a development for Word Net, such-and-such constantly on systematic sets might make connected with a esteem concerning the negative, sure alternately target implication. SentiWordNet 3.0 [14] will be the progressed versify for SentiWordNet 1.0 Furthermore publicly uninhibitedly accessible to investigate end goal with an web interface. This development labels every synset with a worth for each classification between 0 and 1. Along these lines each synset could have An nonzero worth to each sentiment, as a result a portion synsets make positive, negative or target relying upon the setting done which they need aid utilized. Those web interface permits those clients should scan for whatever synset having a place should WordNet for its connected SentiWordNet scores.

Moreover the user has the capacity of visualization about the individual's scores. Each classification will be interfaced with a color, which may be red to negativity, blue for objectivity and green to positivity.

V. Results and Analysis

The performance of the classifications methods can be found out by using some of the following parameters:

A. *Recall*

The Recall is known as true positive function and defines as the ratio of correct instances classified as given class over the number of actual total in that class.

Recall = correctly classified / (correctly classified + Missed classified)

B. *Precision*

The Precision defines as the ratio of correctly classified over number of all experimental classifications.

Precision = correctly classified / (correctly classified + Errorly classified)

C. *F-Measure*

F-Measure is a combined measure for precision and recall. F-Measure = 2 * Precision * Recall / (Precision + Recall).

Note: The Recall and F-Measures has to be higher and Precision has to be lower for the good classification algorithm.

From table 1 to 3, we can see the sample performance measure values obtained for three different algorithms for three different products such as Apple Iphone 5S, Samsung J7 and Redmi Note 3. Fig. 2 shows the graphical representation of overall analysis of performance measures obtained from the tabular values. Among three algorithm Naïve Bayes algorithm has comparatively better results over the Logistic Regression and SentiWordNet techniques.

TABLE I. Performance Measure on - Apple iphone 5S

Classifier	Recall	Precision	F-Measure
Naïve Baye's	0.870	0.675	0.760
Logistic Regression	0.778	0.713	0.744
SentiWordNet	0.570	0.820	0.672

TABLE II. Performance Measure on - Samsung J7

Classifier	Recall	Precision	F-Measure
Naïve Baye's	0.770	0.557	0.857
Logistic Regression	0.678	0.645	0.661
SentiWordNet	0.670	0.651	0.660

TABLE III. Performance Measure on Redmi Note 3

Classifier	Recall	Precision	F-Measure
Naïve Baye's	0.796	0.504	0.802
Logistic Regression	0.596	0.590	0.593
SentiWordNet	0.590	0.620	0.605

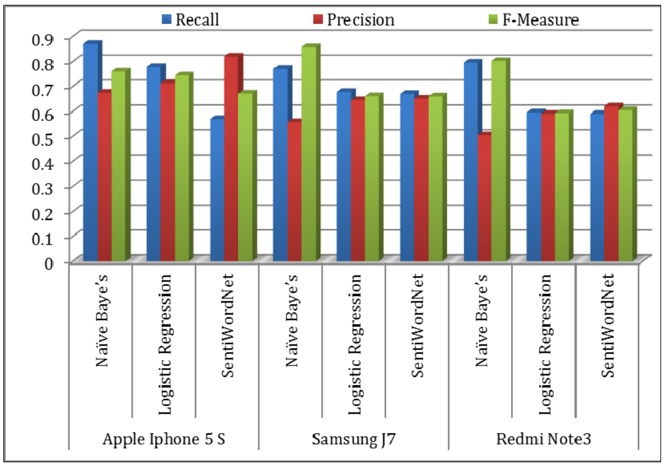

Fig. 2. Analysis of Performance measures on Opinion classification techniques

VI. Conclusion and Future Work

According to our experiment, the Naïve Bayes classification proves to be the most efficient among three algorithms for text classification of opinion mining. This work focuses only on the reviews taken from Amazon website using 3 different algorithms. The work can be extended on mining reviews from multiple website such as Flip kart, Snap deal etc. Further, to incorporate more classification algorithms to analyze their efficiency. This will help us in deciding the best text classifier in opinion mining and sentiment analysis.

Acknowledgment

The authors express their sincere gratitude to Prof. N. R Shetty, Advisor and Dr. H C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT.

The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

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.

[2] Callen Rain, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning", Swarthmore College Compter Society, November 2013

[3] Weishu Hu, Zhiguo Gong, JingzhiGuo, "Mining Product Features from Online Reviews" IEEE International Conference on E-Business Engineering, 2010.

[4] PoojaKherwa, ArjitSachdeva, Dhruv Mahajan, "An approach towards comprehensive sentimental data analysis and opinion mining". IEEE International Advance Computing Conference (IACC), 2014.

[5] Toqir Ahmad Rana, Yu-N Cheah, "Hybrid Rule-Based Approach for Aspect Extraction and Categorization from Customer Reviews", 9th International Conference on IT in Asia (CITA'15), At Kuching, Sarawak, Malaysia, August 2015

[6] Prashast Kumar, ArjitSachdeva, "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites". 5th International Conference- Confluence The Next Generation Information Technology Summit (Confluence), 2014.

[7] Hui Song, Jianfeng Chu, Yun Hu, Xiaoqiang Liu, "Semantic Analysis and Implicit Target Extraction of Comments from E-commerce Websites". Fourth World Congress on Software Engineering, 2013.

[8] R. Piryani, D. Madhavi, V.K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000-2015", Information Processing and Management, ScienceDirect, Elsevier 2016

[9] Vidisha M. Pradhan, Jay Vala, Prem Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining", International Journal of Computer Applications, Volume 133, No.9, January 2016

[10] G. Sneka, CT. Vidhya, "Algorithms for Opinion Mining and Sentiment Analysis: An Overview", International Journal of Advanced Research in Computer Scinece and Software Engineering, Volume 6, Issue 2, February 2016

[11] Neetu, "Hierarchical classification of web content using Naïve Bayes approach", International Jouranal on Computer Science and Engineering (IJCSE), Vol. 5, No. 05, May 2013

[12] Daniel Jurafsky, James H Martin, Chapter on "Logistic Regression – Speech and Language Processing", Standford University, November 2016

[13] Andrea Esuli, Fabrizio Sebastiani, "SentiWordNet: A Publicly Available Lexial Resource for Opinion Mining", 5th Conference on Lanuage Resources and Evaluation (LREC 2006), Genova, IT, 2006

[14] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), Valletta, MT, 2010