PRODUCT ANALYSIS USING CUSTOMER REVIEWS

[1]Pranali Kosamkar, [2]Saurabh Marathe, [3]Akshay Jagtap, [4]Akash Jaiswal
[1,2.3.4]Dept. of Computer Engineering, University of Pune
MIT, Pune 411038, India
[1]pranali.kk@gmail.com, [2]saurabhmarathe1992@gmail.com, [3]akshayjagtap20@gmail.com,
[4]akashvsj@hotmail.com

***Abstract*—The web offers an overwhelming amount of textual data, containing traces of sentiment which may be published through, e.g. blogs, reviews or twitter. A field of Sentiment Analysis has sprung up in the past decade to analyze these textual data for extracting the information tailored to the needs of customers. In this paper, we aim to implement the algorithms Naïve Bayes and Rule-Based. The Rule-based approach identifies the sentiment first. In the second approach the Naive Bayes first identifies the sentiment and then on encountering a sentence which is difficult to analyze, it forwards the input to the Rule-based approach. The rule-based approach then analyze the sentence and then revert the result back to the Naive-Bayes approach, which it stores in it 'good' or 'bad' sentence training database. The entire process involves four steps - Review pre-processing (POS), feature-based sentiment extraction, word scoring based on the comparison with Wordnet library and polarity detection using algorithm. We test this algorithm on Amazon's customer reviews for one specific product. The precision and recall of classifier is used as a measure to determine accuracy of algorithm**

Keywords— *Sentiment, pre-processing, algorithms, feature, reviews, customer, product, analysis, data, extraction, Naïve-Bayes, Rule-Based.*

I. INTRODUCTION

Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents. In
recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The analysis of this data to extract latent public opinion and sentiment is a challenging task. The
analysis of sentiments may be document based where the sentiment in the entire document is summarized as positive, negative or objective. It can be sentence based where individual sentences, bearing sentiments, in the text are classified. SA can be phrase based where the phrases in a sentence are classified according to polarity. A text may contain many entities but it is necessary to find the entity towards which the sentiment is directed. It identifies the polarity and degree of the sentiment. Sentiments are classified as objective (facts), positive (denotes a state of happiness, bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow, dejection or disappointment on part of the writer). A novel approach for automatically classifying

the sentiment of Twitter messages. These messages are classifi1ed as either positive or negative with respect to a query term. This is useful for consumers who want to re-search the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brands. There is no previous research on classifying sentiment of messages on micro blogging services like Twitter. Here they present the results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision [10]. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are thus needed. In recent years, opinionated postings in social media have helped reshape businesses, and sway public sentiments and emotions, which have profoundly impacted on our social and political systems. Such postings have also mobilized masses for political changes such as those happened in some Arab countries in 2011. It has thus become a necessity to collect and study opinions on the Web. Of course, opinionated documents not only exist on the Web (called external data), many organizations also have their internal data, e.g., customer feedback collected from emails and call centers or results from surveys conducted by the organizations. Many companies are managing their resources in this field to know shortcomings of their products and for better sales.

II. RELATED WORK

Sentiment analysis is a kind of text classification that classifies texts based on the sentimental orientation (SO) of opinions they contain. Sentiment analysis of product reviews has recently become very popular in text mining and computational linguistics research. Bing Liu worked on extracting evaluating opinions in online discussions. The discussions can get highly emotional heated with many emotional statements and personal attacks. As a result, many of the postings and sentences do not express positive or negative opinions about the topic being discussed. To find people's opinions on a topic and its different aspects, the irrelevant sentences should be removed [1]. Sentiment analysis is also known as opinion mining, opinion extraction and affects analysis in the literature. Further, the terms sentiment analysis and sentiment classification have sometimes been used interchangeably. It is useful, however, to distinguish between two subtly different concepts. In this article, hence, sentiment analysis is defined as a complete process of extracting and understanding the sentiments being expressed in text documents, whereas sentiment classification is the task of assigning class labels to the documents, or segments of the documents, to indicate their SO. Sentiment analysis can be conducted at various levels. Word level analysis determines the SO of an opinion word or a phrase. Sentence level and document level analyses determine the dominant or overall SO of a sentence and a document respectively. The main essence of such analyses is that a sentence or a document may contain a mixture of positive and negative opinions. Some existing work involves analysis at different levels. Specifically, the SO of opinion words or phrases can be aggregated to determine the overall SO of a sentence (Hu and Liu, 2004a) or that of a review [11].Some sentiment analysis algorithms aim at summarizing the opinions expressed in reviews towards a given product or its features. Such sentiment summarization also involves the classification of opinions according to their SO as a subtask, and that it is different from classical document summarization, which is about identifying the key sentences in a document to summarize its major ideas. Sentiment analysis is closely related to subjectivity analysis. Subjectivity analysis determines whether a given text is subjective or objective in nature. It has been addressed using two methods in sentiment analysis algorithms. The first method considers subjectivity analysis a binary classification problem, for example, using Subjective and Objective as class labels [2]. Pang and Lee (2005) adopted this method to identify subjective sentences in movie reviews. The second method makes use of part-of-speech (POS) information

about words to identify opinions because previous work on subjectivity analysis suggests that adjectives usually have significant correlation with subjectivity [3]. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data [12]. Data uncertainty is common in real-world applications due to various causes including imprecise measurement, network latency, and outdated sources and sampling errors. These kinds of uncertainty have to be handled cautiously, or else the mining results could be unreliable or even wrong. In this paper, we propose a new rule-based classification and prediction algorithm called uRule for classifying uncertain data. This algorithm introduces new measures for generating, pruning and optimizing rules. These new measures are computed considering uncertain data interval and probability distribution function. Based on the new measures, the optimal splitting attribute and splitting value can be identified and used for classification and prediction. The proposed uRule algorithm can process uncertainty in both numerical and categorical data. Eexperimental results show that uRule has excellent performance even when data is highly uncertain [14].

III. PROBLEM STATEMENT AND METHODOLOGY

The proposed system will take Amazon customer reviews and analyze those using Naïve-Bayes and Rule-Base approach to extract the sentiment of the product and sentiments of its features like cost, quality, reusability, popularity etc.

A. Preprocessing

The database contains files (customer reviews). The files are sent to preprocessor. Pre-processing includes three steps: Mapping each word with WorldNet 2.1 Performing part-of-speech tagging. Identifying entities Mapping each word with WordNet 2.1.Every word of the customer review file I mapped to WordNet 2.1 for enabling the identification of the category of the word. The category of words are: Noun, Pronoun, Adjective, Adverb, Verb

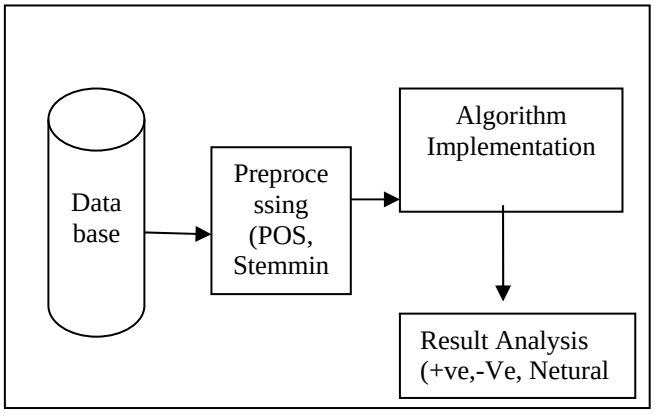

Fig. 1. High Level Design

B. Implementation of Naive Bayes Approach

- Find probability of word being positive or negative from training data
- Multiply all probabilities being positive
- Multiply all probabilities being negative.
- Find the maximum value of

C. Implementation of Rule Based Approach

- Match each word with positive database or negative database
- Assign weight to words accordingly
- Determine the overall sentiment
- Store the result to the database

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

We have used data from Amazon customer reviews for experiment and result analysis. This

data is in unstructured format. The sample file is shown in table I below. We have examined the precision and recall for Naïve based and Rule based Approach using the Amazon customer reviews. Table no.2 shows the precision and recall values for both the approaches. From the Fig. no. 2 and Fig.no.3 it is observed that Rule based approach gives good performance as compare to Naïve based approach for customer product reviews.

TABLE I. SAMPLE COMMENT FROM AMAZON CUSTOMER REVIEWS

Sr.No	Comment
1	excellent picture quality / color
2	i 'd highly recommend this camera for anyone who is looking for excellent quality pictures and a combination of ease of use and the flexibility to get advanced with many options to adjust if you like .
3	the camera is very easy to use , in fact on a recent trip this past week i was asked to take a picture of a vacationing elderly group .
4	i just told them , press halfway , wait for the box to turn green and press the rest of the way .
5	ensure you get a larger flash , 128 or 256 , some are selling with the larger flash , 32mb will do in a pinch but you 'll quickly want a larger flash card as with any of the 4mp cameras .
6	a few of my work constituants owned the g2 and highly recommended the canon for picture quality

TABLE II. PRECISON AND RECALL FOR NAÏVE BAYS AND RULE BASED APPROACH

Opinion	Naïve Bays Approach		Rule Based Approach	
	Precision	Recall	Precision	Recall
Good	80	128.57	80.48	166.66
Bad	50	133.33	44.44	120.00
Neutral	42.85	77.77	35.29	83.33

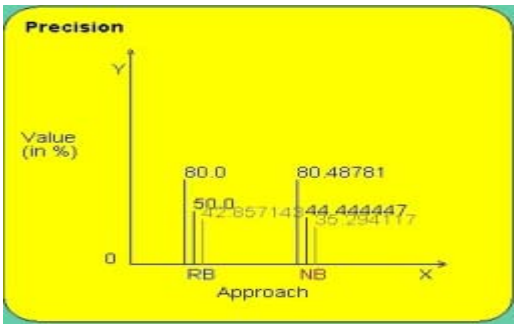

Fig. 2. Precision graph for Rule based and Naïve based Approach

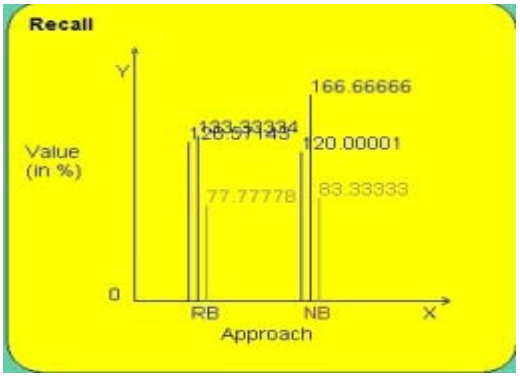

Fig. 3. Recall graph for Rule based and Naïve based Approach

V. CONCLUSION

Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others' opinions. This project deals with identifying the sentiment of the product expressed in the reviews and it also involves identifying feature-by-feature sentiment of the product. This is achieved by implementing two different approaches. The Naïve-Bayes approach focuses on computing the relative probability of each class (positive or negative) of word so as to compute total sentiment expressed in the review. A similar approach also computes sentiment of features of the product identified in the previous pre-processing stage. The Rule-based approach involves multiple parsing stages with each stage finds and refines the sentiment expressed. This approach uses similar steps to compute pre-identified feature-based sentiment.

Finally, after experiment and from precision and recall graph it is observed that Rule based approach performs better than Naïve based approach for sentiment analysis for customer product reviews.

REFERENCES

[1] Zhongwu Zhai, Bing Liu, Lei Zhang, Hua Xu, Peifa Jia Identifying ,"Evaluative Sentences in Online Discussions".

[2] Bo Pang and Lillian Lee, "*Thumbs up? Sentiment Classification using Machine Learning Technique*".*s* Department of Computer Science Cornell University Ithaca, NY 14853 USA

[3] Peter D. Turney, "*Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*". Institute for Information Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6

[4] Alec Go, Richa Bhayani,Lei Huang, "*Twitter Sentiment Classification using Distant Supervision*".. Stanford University Stanford, CA 94305 .

[5] Cane W. K. Leung, Stephen C. F. Chan , "*Sentiment Analysis of Product Reviews*". Department of Computing The Hong Kong Polytechnic University Hung Hom, Kowloon Hong Kong SAR

[6] Naive-Bayes Classification Algorithm [PDF].

[7] Biao Qin, Yuni Xia ,*"A Rule-Based Classification Algorithm for Uncertain Data* ".Department of Computer Science Indiana University -Purdue University Indianapolis, USA Sunil Prabhakar Department of Computer Science Purdue University Yicheng Tu Department of Computer Science and Engineering University of South Florida

[8] Dmitry Kan-, "Rule-based approach to sentiment analysis". at ROMIP 2011 AlphaSense Inc.

[9] Benjamin Paterson, Weifeng Zhang, Tim Mwangi, "*Recommending movies and TV shows based on Facebook probe data* ".CS229 Project Report December 13, 2012.

[10] Zakaria Elberrichi, Abdelattif Rahmoun, and Mohamed Amine Bentaalah 1EEDIS Laboratory, "*Using Word Net for Text Categorization*", Department of Computer Science, University Djilali Liabès, Algeria King Faisal University, Saudi Arabia

[11] Sasha Blair, Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, Jeff Reynar, "*a Sentiment Summarizer for Local Service Reviews*". NLPIX 2008.

[12] Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Objectivity Summarization Based on Minimum Cuts".

[13] Zhongwu Zhai,Bing Liu, Lei Zhang, Hua Xu,Peifa Jia, "Identifying Evaluative Sentences in Online Discussions"

[14] Shoushan Li, Rui Xia, Chengqing Zong, Chu-Ren Huang, "*A framework of feature selection methods for text categorization.*" 2009 ACL and AFNLP.

[15] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis" (Early Text)

[16] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena., "*Large-scale Sentiment Analysis for News and Blogs*". New York 11794-4400 USA

[17] Tony Mullen and Robert Malouf ,"A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse".