Improving analysis and prediction of customer reviews using NLP and Bernoulli Classifier

[1]Priya Shahane, [2]G.P. Chakote
[1]ME Student CSE Dept., [2]Professor CSE Dept
MSS College of Engineering and Technology Jalna

***Abstract*— Customer feedback is important in improving the company's services, both in terms of intimacy and openness. Open-minded reviews mean comments, expressions, and direct comments from customers. However, companies have a variety of content or groups to evaluate with their scores and overall scores for the types of services that many customers are looking for. The problem is that some customers give points to reviews. Other reviewers must read the comments and provide feedback that is different from the rating. So this article offers analysis and forecasts from open customer reviews using the probability classifier. Classifiers will use case studies of hotels with customer reviews in open reviews for training data to group feedback on whether mining is a positive or negative feedback. Data mining, commenting or opinion analysis is a part of data mining. Data mining is a form of natural language processing used to record people's attitudes toward a particular subject or product. Most mining reviews give the category a positive, neutral or negative review. Recently, data mining reviews have been very successful due to the availability of enormous amounts of rich web resource reviews. Digital formats such as forums, discussion sites, blog reviews, etc. When using ecommerce websites, rudely increased the users not only. Instead of buying products on the site, but also providing feedback and suggestions that will benefit other users, compiled user reviews will be analyzed and organized to make better decisions.**

***Index Terms*— Opinion analysis, Sentiment analysis, Machine Learning Algorithm, Stanford Classifier, Reviews, E-commerce**

I. INTRODUCTION (HEADING 1)

Nowadays, people search for other people's opinions from the internet before making a purchase when they are unfamiliar with any product, with help from reviews, online ratings, etc. It provides useful information to customers to buy products and for manufacturers to improve. Product quality Mining is used to identify and select useful information from the text. Confidence can mean the perception, attitude or interpretation of a person. Classification of comments is a subset of the classification of messages related to opinions. Topics are expressed. [11] Opinion analysis also includes different names, including: mining, opinion, confidence analysis, retrieval. Confidence or Emotional Rating Analyze the confidence process to find the meaningful orientation of the criticism or opinion given. [3] Mining opinion plays an important role in social media in getting user feedback. It uses the mode of the form of online reviews, emails and social networking sites, such as Facebook, Twitter, LinkedIn, YouTube, blogs, etc. Feelings are usually given at the sentence level, document level, and skill level. It is very important to mine the comments. A feature or feature is a clear reference of an entity with a comment. Entity is a hierarchical representation of components and subcomponents. Each component is associated with a set of attributes. Comments may be positive, neutral, or negative. Comments can be evaluated in two ways. Direct and indirect comments In direct contrast, the review is given directly for the product, as the X camera is good. In an indirect review, it compares two products, such as an X camera, to a Y camera. The ML algorithms can be used for confidence analysis. For example, "The resolution of this phone is good." In this job, the owner of the comment is the user who gave the comment. Comment is The "resolution" of the phone and the opinion is that the word "good" is positive in terms of The main purpose of the Confidence Analysis is to classify the terminals. Orientation determines whether sentences are positive or neutral. Machine learning is a system that learns from observation, training, experience, etc. Learning supervision creates functions that generate data inputs to the desired results, or so-called labels, that they are examples of. Training labeled by human experts [4] can use any managed learning method, such as the Naïve Bayes classification and vector machine.

The sentiment analysis can be performed at one of the three levels: the document level, sentence, feature level [14].

1. Document Level Sentiment Classification: In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document.
2. Sentence Level Sentiment Classification: The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral.
3. Feature Level Sentiment Classification: Product features are defined as product attributes or components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models.

People post their views feedback experience about product then collect the corpus containing views given by people and then the corpus is processed. Pre-processing is done. Feature Extraction is performed to extract the relevant features. The review given by person is classified as positive negative or neutral by applying machine learning algorithm and the output is given. Because of the huge number of reviews in the form of unstructured text, it is impossible to summarize the information manually. Accordingly, efficient computational methods are needed for mining and summarizing the reviews from corpuses and Web documents. Different tools are available to track the sentiment of users which are Review Seer Tool, Red Opal, Web Fountain, Opinion Observer etc

II. RELATED WORK

The goal of Machine Learning is to develop algorithms to optimize the performance of the system using the information or experience of the past. Machine Learning has a solution to the two-stage classification problem.:

1. Learning model from the training data warehouse.

2. Invisible Classification of Training Patterns Confidence analysis of natural language text is an emerging field. Partial transformation of text into vector features is an important step in a data driven approach to sensory analysis. [13].

Lisette García-Moya, et al. [3] Uses summaries based on various aspects by offering new ways to retrieve product features from free customer feedback on products or services. Their proposal relied on a language modeling framework that included models of probability, comments, and stochastic mapping styles between words and product language variants. Their work expands the preliminary guidelines that focus on modeling the language of product features based on customer reviews. They provide an official method for retrieving product features from the approximate language model of the feature.

Bing Liu [12] uses a remote approach to extract comments, words, and phrases after pulling the facets. In calculating the polarity of each word, WordNet offers a set of mining techniques and product reviews based on data mining and natural language processing. The purpose is to provide a summary of the product features, many of which are customer reviews of products sold online.

Su Su Htay and Khin Thidar Lynn offer unique ways to help find words or phrases that are customer feedback for each feature in a useful way. When using adjectives, adverbs, verbs and nouns, they receive a form of comment, a word / phrase of the product attribute from a given review. Separated features and opinions help create useful summaries, which are important resources to help users. It is also useful for merchants to track the best choice of product. There are some tagger uses which specifies a phrase in the specified text, which has adjectives or adverbs, or a verb or noun as a comment phrase.

Richa Sharma, et al. [4] offer a view of the mining system to isolate positive, neutral, or negative comments for each feature. In denial, their systems are also managed. They use unattended techniques to carry out their tasks. The dictionaries used to define their opinions, words, and synonyms, and their opposite words, are WordNet. The first feature of all products that have been validated by customers is the first. The positioning of sentences for every feature is already given. Most of the commentary words help to find the terminology of the sentences given. Dim En Nyaung and Lai Lai Thein [5] mainly work on summing up the comments. For the summary of product reviews and commentary reviews, it is very important. To check the polarity and the numerical scores of all features of the Senti-Word Net Lexicon product, use the Find Feedback feature for positive and negative attributes. Their main purpose is to provide a summary of products from customers sold through an ecommerce site.

Madhavi Kulkarni and Mayuri Lingayat [6] offer techniques that can effectively rate products by reviewing genuine products. The system only allows users to write reviews about products purchased from the website. Other users are not allowed to comment. Reduce product errors and customers who receive reliable products. They offer a product ranking system that can use product data in questionnaires and systems to meet customer needs, as well as product ratings.

Arti Buche, et al. [7]. The paper provides information on data mining, feedback, machine learning, and associative analysis, classification, message classification, and tool usage. Let users know the product that uses the features. Product reviews are provided in a graphical format based on the features. Validation tools are used for automation and site integration. The Naïve Bayes discriminator is used to determine points for specific features made in the critique. A set of inputs such as clustering are used for learning that is not exempt during the labeling training. To use the static classification scheme used in the comments. To write a sentence pattern that is used as part of a speech tag. Semi-structured learning creates the right function or classifier, including labeled and unlabeled examples.

Aashutosh Bhatt, et al. [8] offers a system that performs the categorization of feedback defined by the customer, followed by the finding of confidence in the feedback received. Feature extraction as a rule for a given product is made. Their system creates a visual sense of the review, which is presented in the form of a chart. Critique of reviews, along with emotional analysis, improves system accuracy, providing accurate feedback to users. The main aim of the system is to ensure a fair outcome of the feelings and time savings that users spend reading the narration about the length of the review by summarizing the results in the chart.

Asmita Dhokrat, et al. [9] provides a review of mining techniques and tools used in mining. The paper presents the basic requirements of mining, the opinion to explore the current techniques used to develop a complete poultry farming system. It highlights the opportunities or uses and research of such systems. The tools available for creating such applications are presented with advantages and limitations.

Farhan Hassan Khan, et al. [10] Offers a new conceptual framework, Twitter Mining, to predict the positive or negative polarity of reviews. It improves classification accuracy. It was created using a different procedure.

III. PROPOSED ARCHITECTURE

Before Mining is one of the sub-fields related to information retrieval and knowledge discovery. Many researches can be done in this area. Mining is considered an interesting area of research because of its various uses. Over the years, special attention has been paid to reviewing customer reviews for specific products.

Polarization or classification of feelings is classified as a positive, negative or neutral document. It is a technique to analyze useful information in a large number of messages. A common way to classify confidence is to use a learning algorithm such as Stanford classifier , Naive Bayes Classifier , SVM, C4.5 etc

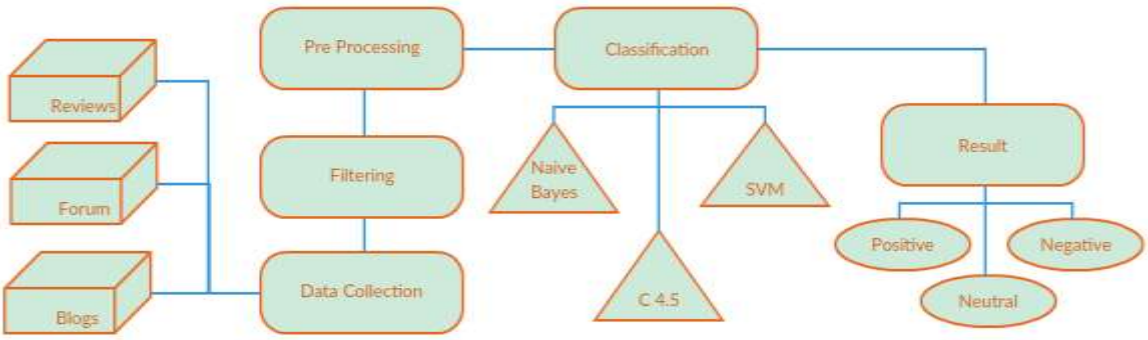

Figure 1: Basic Steps in preprocessing and classification

Based on the review of literature, typical steps involved in opinion mining are shown in Figure 1. The different steps are data selection, data collection, preprocessing, classification and result or output.

1. Data-set Selection: Data selection can be done from various data sources such as blogs, online reviews e.g. amazon, yelp, etc., forums and micro-blogging sites e.g. twitter.
2. Data Collection: The reviews are collected from the chosen data-set.
3. Pre-Processing: The collected data is pre-processed or cleaned for analysis to get fair text review. Cleaning of data is done by removal of special characters (such as :":/.,'#$*^&-) to retrieve best results. Stop words are also removed.
4. Classification: The given Reviews are classified as positive, neutral or negative. ML Algorithms are used for classification. Different ML algorithms which can be used for this step are SVM, Naïve Bayes, Maximum Likelihood, ANN and Decission Tree. By applying any of these algorithms the document, sentence or aspect level can be oriented positively, or negatively.

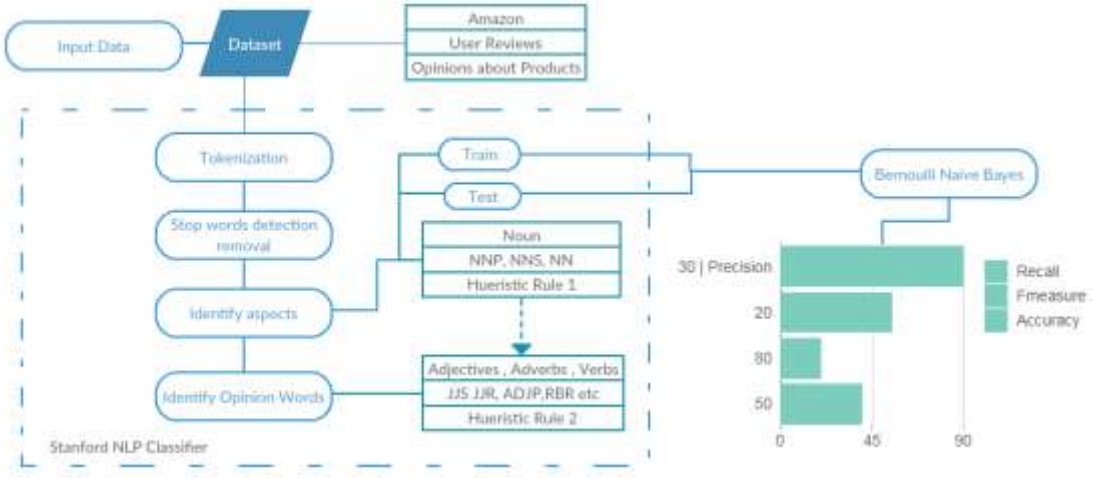

Figure 2. Proposed Architecture

As suggested in figure , our proposed model will allow users to take reviews at runtime or from datasets and perform tokenization with parts of speech tagging using Stanford Classifier and further train the system to classify using Bernoulli Naive Bayesian Classifier.

We use heuristic rules to identify different types of parts of speech and tag our reviews accordingly for further classification.

Heuristic rules are pattern related features that allow us to identify and process each word or token in a sentence to depict certain feature and dig further to analyze their patterns.

Pattern-Related Features

The models selected are called "sentiment expression" [9] are very common, even in spoken language. However, their number is small, they are not unique and most reviews in our workout and test sets do not contain them. However, we dig deeper and extract another set of features.

We offer more efficient and reliable models. We divide the words into several classes [12]: containing the words whose grammatical function is more important. If a word belongs to the first category, it is lemmatised; Otherwise, it is replaced by a certain expression. The expressions [13] used to replace these words are presented in

TABLE 2. The classification is done according to the part of voice tag of the word in the opinion or review sentence.

POS Tag	Expression
CD	CARDINAL
FW	FOREIGNWORD
UH	INTERJECTION
LS	LISTMARKER
NN, NNS, NNP, NNPS	NOUN
PRP	INTERJECTION
MD	MODAL
PB, RBR, RBS	ADVERBS

Bernoulli Naive Bayes Classifier

Bernoulli Naive Bayes is a probabilistic classifier, which means that for a document d, all classes c ∈ C the classifier returns the class c which has the posterior maximum probability given the document.

The Bernoulli Naive Bayes classifier is a simple probabilistic classifier that is based on the Bayes theorem with strong and naive assumptions of independence. This is one of the most basic text classification techniques with various applications in the detection of e-mail, personal message sorting, categorization of documents, detection of sexually explicit content, detection of languages and The detection of feelings. Despite the naive design and simplified assumptions that this technique uses, Bernoulli Naive Bayes works well in many complex real world problems.

Even though it is often surpassed by other techniques such as boosted trees, random forests, Max Entropy, Support Vector Machines etc., Bernoulli Naive Bayes classifier is very effective as it is less costly in computing (both CPU and memory) And it requires a small amount of training data. In addition, the training time with Naive Bayes is much lower as opposed to the alternative methods.

The Bernoulli Naive Bayes classifier is superior in terms of CPU and memory consumption as shown by Huang, J. (2003), and in many cases its performance is very similar to more complicated and slower techniques.

A Bernoulli naive bayes classifier[15] is a simple probabilistic model based on the Bayes rule with a strong hypothesis of independence. The Naïve Bayes model implies a simplified conditional independence hypothesis. This is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not significantly affect the accuracy of the text classification, but makes the classification algorithms very fast applicable to the problem. In our case, the probability of maximum likelihood of a word belonging to a given class is given by the expression:

$$p(a|b) = \frac{countoffeaturesinareview}{totalnumbersofwordinareview}$$

Here, the xi s are the individual words of the post review. The classifier delivers the class with the maximum a posteriori probability. We also remove duplicate words from reviews, they do not add any additional information; This type of Naive Bayes algorithm is called Bernoulli Naïve Bayes. The inclusion of the presence of a word instead of the count has been found to improve performance marginally, when there are a large number of training examples.

IV. Key Index Parameters

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance.

In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are:

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp + tn/tp + tn + fp + fn$

V. Conclusion

Mining in customer feedback is very important in improving the service, which forms will be compared between the decision tree and naIve Bayes. The advantage of the classification model is calculated from the probability projected in Level label However, the advantage of the decision tree structure is that it shows the hierarchical order of trees to help analyze service improvement and key factors. In addition, the Bernoulli Naive Bayes model can use a probability that is similarly estimated, calculated automatically. Even customers will read the comments. However, the system can conclude all classification of the comments. So customers can make quick decisions.

References

List and number all bibliographical references in 10-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example: [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer

simply to the reference number, as in "[3]"—do not use "Ref. [3]" or "reference [3]". Do not use reference citations as nouns of a sentence (e.g., not: "as the writer explains in [1]").
Unless there are six authors or more give all authors' names and do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.
For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] S. 1. Wu, R.D. Chiang and Z.H. Ji, Development of a Chinese opinionmining system for application to Internet online forum, The Journal of Supercomputing, Springer US[Online], 2016.
[2] Z. Li, L.Liu and C.Li, Analysis of customer satisfaction from chinese reviews using opinion mining, Proceeding of the 6th IEEE International Conference on Software Engineering and Service Science(ICSESS). 2015, pp.95-99.
[3] Q.Su, X.Xu, H.Guo, Z.Guo, X. Wu, X. Zhang and B.Swen. Hidden Sentiment association in Chinese web opinion mining. Proceeding ofthe 17th International Conference on World Wide Web, 2008, pp.959-968.
[4] S.Atia and K. Shaalan, Increasing the accuracy of opinion mining in Arabic. Proceeding of the 1st International conference on Arabic computing linguistics, 2015, pp.l 06-113.
[5] R.M. Duwairi and I. Qarqaz, Arabic Sentiment Analysis using Supervised Classification. Proceeding of 2014 International Conference on Future Internet of Things and Cloud. 2014, pp. 579-583.
[6] H.S. Le, T.V. Le and T.V. Pham, Aspect Analysis for Opinion Mining of Vietnamese Text. Proceeding of Internation Conference on Advance Computing and Application, 2015, pp.118-123.
[7] T. Chumwatana, Using sentiment analysis technique for analyzing Thai customer satisfaction from social media. Proceeding of the 5th International Conference on Computing and Informatics, 2015, pp.659- 664.
[8] S.Ahmed and A.Danti, A novel Approach for sentimental analysis and opinion mining based on sentiwordnet using web data. Proceeding of International Conference on Trends in Automation, Communications and Computing Technology, 2015, pp.1-5.
[9] R.K. Bakshi, N. Kaur, R. Kaur and G.Kaur, Opinion mining and sentiment analaysis, Proceeding of the 3rd International Conference on Computing for Sustainable Global Development, 2016, pp. 452-455.
[10] P.Barnaghim, 1.G. Breslin and P. Ghaffari, Opinion mining and sentiment polarity on Twiiter and correlation between events and 77 .
[11] N. Kumari and S. N. Singh, Sentiment analysis on E-commerce application by using opinion mining, Proceeding of the 6th International Conference-Cloud System and Big Data Engineering(Confluence), 2016, pp. 320-325
[12] V.B. Raut and D.D. Londhe, "Survey on opinion mining and summarization of user review on web", International Journal of Computer Science and Information Technology, Vol. 5(2), 2014, pp. 1026-1030
[13] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T.H, Kim, Opinion Mining over twiiterspace: Classifying tweets programmatically using the R approach. Proceeding of the 7th International Conference on Digital Information Management, 2012, pp. 313-319.
[14] M. R. Islam and Minhaz F. Zibran, "Exploration and Exploitation of Developers' Sentimental Variations in Software Engineering", Internation Journal of Software Innovation, Vo1.4(4), 2016, pp.35-55.
[15] Y.Yokoyama, T. Hochin and H. Nomiya, "Estimation of Factor Scores of Impressions of Question and Answer Statements", Internation Journal of Software Innovation, Vol. 1(4), 2013, pp.53-66.
[16] L. Lin, 1. Li, R. Zhang, W. Yu and C. Sun, Opinion mInIng and sentiment analysis in social networks: A retweeting structure-aware approach. Proceeding of the 7th International Confernece on Utility and Cloud Computing, 2014, pp.890-895
[17] A.H. Al-hamaami and S. H. Shahrour, Development of an opinion blog mining system, Proceeding of the 4th International Conference on Advanced Computer Science Application and Technologies, 2015, pp. 74-79.