SENTIMENT ANALYSIS OF PRODUCT REVIEWS FOR E-COMMERCE RECOMMENDATION

[1]D. MALI, [2]M. ABHYANKAR, [3]P. BHAVARTHI, [4]K. GAIDHAR, [5]M.BANGARE

[1,2,3,4,5]Department of: Information Technology, Smt. KashibaiNavale College of Engineering, Pune.
E-mail: [1]digvijaymali3694@gmail.com, [2]madhura20a@gmail.com, [3]paras.bhavarthi@gmail.com, [4]krunalgaidhar1@gmail.com, [5]manoj.bangare@gmail.com

Abstract— The social web has made enormous amounts of information available to users globally at just the click of a button. Consumers often tend to rely on such text, especially those in the form of opinions or experiences regarding a particular product which makes it essential that this information should be available in a systematic manner. Sentiment analysis studies these opinions. This paper explains different methods for sentiment analysis and showcases an efficient methodology. It also highlights the importance of Naïve Bayes classifier over other classification algorithms.

Keywords— Sentiment analysis, E-commerce, Machine learning, NaïveBayes, WordNet.

I. INTRODUCTION

Others' opinions can be crucial when it is time to make a decision, especially when those choices involve valuable resources like time or money. In such cases, people often rely on their peers' past experiences. Social media allows us to efficiently create and share ideas with everyone connected to the World Wide Web via forums, blogs, social networks, and content-sharing services. This information is unstructured and thus capturing public opinion about a variety of topics results in the emergence of the fields, opinion mining and sentiment analysis. [1]

When an individual wants to make a decision about buying a product or using a service, they have access to a huge number of user reviews, but reading and analyzing all of them is a tedious task. Also when an organization wants to benefit by obtaining the public opinion or to market its products, even to identify new opportunities, predict sales trends, or manage its reputation, it needs to deal with an overwhelming number of available customer comments. With sentiment analysis techniques, it is possible to analyze a large amount of available data, and extract opinions from them that may help both customers and organization to achieve their goals.

Sentiment analysis,also opinion mining is the field of computational study that analyzes people's opinionsexpressed in written language, where focus of research is on the processing of text in order to identify opinionated information. This differs from mining and retrieval of factual information which is the target of much of the existing research in natural language processing and text analysis.

II. RELATED WORK

Current research focuses on sentiment analysis of information gathered from social networking websiteslike Twitter, Facebook, MySpace to conclude viewers' response to a particular social event or issue. Sentiment analysis has endless applications like forecasting market movement based on news, blogs and social media. Currently, sentiment analysis is a very lucrative approach for hefty applications like 'Smart Cities'. These applications use methods based on document level and sentence level classification which use purely supervised or unsupervised classification algorithms. These algorithms are advanced by Fuzzy Formal Concept, Genetic Algorithms or Neural Networks by making them semi-supervised. Research also focused on sentiment analysis with networking to give a degree of parallelism [2]. It focused on online accrued utility scheduling algorithm which gave them high speed on multiple processors. But this made the system much more complex. Research was also focused on Twitter sentiment analysis for security-related information gathering using normalized lexicon based sentiment analysis [3]. While it provided a positive outcome, a universal dataset was not used.

Current online product recommendation applications are comparing parameters like price, ratings and special offers on the product on different e-commerce websites and are not focusing on customers' personal experience by analyzing their reviews. Hence there is a need to develop a comprehensive application based on sentiment analysis which will give more importance to customer reviews.

III. EXISTING MODELS: COMPARISON

As per the thorough literature survey, the major identified techniques are as follows:

(A) Lexicon Based Model –It employs frequent and explicit product features extraction involving Syntax Tree Based Classification-Design Syntactic Patterns. [4]

(B) Word Alignment Model (Unsupervised)-It concerns with Word Co-occurrence Frequencies and Position of Words. [5]

(C) Word Alignment Model (Semi-supervised) - It involves analysis of Formal and Informal Text Separately [5].

Based on the above techniques, these are the various models that have been identified and some of the related models are discussed below.

It is possible to combine some features from Word Alignment Model and Lexicon Based Model to design a new semi-supervised lexicon based model so that it is possible to use lexical databases like WordNet, SentiWordNet and Attempto Controlled English Lexicon [ACE] [4][6]. Among these lexical databases WordNet groups English words into sets of synonyms called synsets. SentiWordNet processes unstructured information and extracts meaningful numeric indices from the text and aims to provide an extension for WordNet such that all synsets can be associated with a value concerning the negative, positive or objective connotation. ACE provides deep classification of parts of speech but it is better to use ACE along with WordNet to increase recognition rate of lexemes [4][6] .

IV. NATURAL LANGUAGE PROCESSING [NLP]

NLP is the set of methodologies and techniques which allows computer to make sense of human speech as it is spoken. Common NLP tasks in machine learning include: sentence segmentation, parts of speech tagging, parsing text results, deep analytics and named entity extraction [4][7].Controlled Natural Language [CNL] processing is a subset of NLP which restricts the grammar to increase the simplicity for tagging [4]. Basic units of sentiment analysis are features which can be extracted form product reviews. These features can be lemmas, multi-words and valance shifters [8].

V. PROPOSED WORK

A proposed test model is discussed below which will give direct result of sentiments. It will help users gather the best information about any particular product they desire, based on other customers reviews, and help in making a decision about any product.

A. Information Retrieval

Information retrieval (IR) deals with the storage, representation, organization of, and access to information items, the representation and organization of which provides the user with easy access to the information in which he is interested.[9]In other words, IR is finding material of an unstructured nature that satisfies an information need from within large collections. [3]IR systems identify the documents in a collection which matches a user's query and thus narrow down the set of documents that are relevant to a particular problem thereby speeding up the analysis considerably by reducing the number of documents to be analyzed. [2]

There are three major information retrieval techniques:

1. Scraping reviews from URL's using RE
2. Collecting data sets
3. By web API's

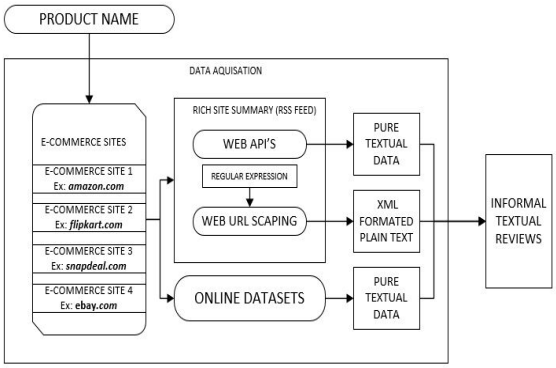

Figure 1: Information Retrieval

B. Preprocessing and Cleaning

Data preprocessing is a technique that involves transforming raw data into understandable format by eliminating incomplete, noisy and inconsistent data. Online informal text requires more sophisticated methods to clean noise in raw text to perform sentiment analysis. Therefore equal importance should be given to preprocessing along with classification [10]. 'Bag of words' is required to identify opinion targets (features) from this pure textual information. This is also known as product feature extraction. [11][12].

It involves the following tasks:

- Removal of URLs and Hyperlinks
- Removing HTML tags and special characters
- Abbreviations extending
- Parts of speech tagging

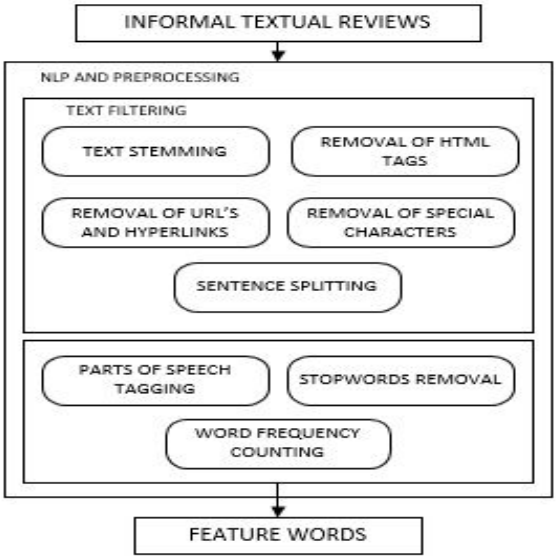

Figure 2: Preprocessing and Cleaning

C. Sentiment Analysis

1. Methods of Machine Learning for Sentiment Analysis

Following are the approaches for machine learning:

(A) Machine learning based approach includes: (i) supervised learning (ii) unsupervised learning and (iii) semi-supervised learning (iv) lexicon based [13].

(B) Using Semantic Orientation scheme of extracting relevant n-grams of the text and then labeling them either as positive or negative and consequentially the document [6].

(C) SentiWordNet approach -It is based on quantitative analysis of the glosses associated to synsets and on the use of the resulting vectorial term representations of semi-supervised synset classification. SentiWordNet is computationally a favorable algorithm but achieves relatively lower accuracy. [6][14]

The method discussed in this paper is based on semi-supervised learning which uses WordNet as lexicon based data dictionary to convert features into target words.

Figure 3: Target Words Extraction

2. *Classification*

In machine learning terms, classification is the problem of identifying to which of a set of categories a new observation belongs. This is decided on the basis of a training set of data containing observations whose category membership is known.

Features	Naive bayes [19]	Max Entropy [19]	Boosted trees[19]	SVM[6]	Random forest[19]	KNN[15]
Based on	Bayes theorem	Feature based classifier	Decision tree Learning	Distance vector	Decision tree Aggregation	Nearest neighbor
Simplicity	Very Simple	Hard	Moderate	Moderate	Simple	Simple
Performance	Better	Good	Good	Better	Excellent	Poor
Accuracy	Good	High	Poor	Good	Excellent	Good
Memory requirement	Low	High	Low	High	High	Low
Time Required	Low	Moderate	High	Moderate	High	Very low

Figure 4: Most Frequently used classification Algorithms for Sentiment Analysis

a) Deterministic classification

1. Rocchio Classifier (Nearest Centroid Classifier)

It computes the centroid of the document for positive and negative classes. The Rocchio algorithm often fails to classify multimodal classes and relationships. For instance, review regarding Holland and Netherlands should fall in the same class.

3. Margin Classifiers

(i) Passive Aggressive Method (PA)

For massive streams of data like documents are coming one by one, for example twitter data. It suffers the problem of 'Hinge Loss' if the incoming format of stream changes and it is not good for small streams.

(ii) Support Vector Machine (SVM) based on Vector Space Model (VSM)

This has high accuracy and almost always gives right predictions. There is always an element of chance or uncertainty involved in textual reviews with alternate solutions which is considered only in probabilistic approach. 'Contradict Optimization Problem' may occur in SVM but it is removable. Overall performance of SVM is good and it is way better than PA [4].

4. Nearest Neighbor Classifier: k-Nearest Neighbor Classifier (KNN) and its modifications

It is a majority of class theorem for the newly came unclassified document where k denotes the number of already classified documents and k is not the multiple of number of classes. (i) Standard KNN- k is fixed. Weight factor is not considered. Time consuming.(ii) k-variable KNN- Improved k-variable KNN, Basic k-variable KNN, Weighting KNN are good if they are combined into one 'Flexible KNN' algorithm which switches the algorithms according to k value available but again it is somewhat complex also not feasible real time sentiment analysis [15][16].

b) Probabilistic classification

These classifiers are good for 'Sentence Level Classification'. It does not allow you to say exactly what outcome will be, especially since there might be an element of uncertainty present. It includes both 'Deterministic Component' and 'Random Error Component'. Very good for small stream like sentence or segment level data.

Naïve Bayes Algorithm

This is the simplest classifier based on 'Bayes Theorem'. Because of its simplicity it is not only has variety of applications but also considered as baseline algorithm for research in decision level classification problems. It directly considers the probability of positivity and negativity of the text with respect to that class to which it belongs. Here, relative probability is considered. Naïve Bayes is the fastest algorithm for sentiment analysis. However, it may give low accuracy as compared to SentiWordNet approach or SVM if the feature is not well defined. For ex: classification of that word which has no meaning [6]. It makes independent assumptions for its features which cause 'overfitting'. Overfitting also occurs when trained data is quite large hence good for sentence level classification [8][17][18].

This paper considers Naïve Bayes as multiclass classifier where classes are divided for each e-commerce websites. The probability for each class

($\Pr(ci|x)$) can be calculated by using following formula:

$$\Pr(ci|x) = \frac{\Pr(x|ci)\Pr(ci)}{\Pr(x)}$$

where,

ci ∈ C = {c1,..,cn} (n is no. of e-commerce websites)

x ∈ X =(a1, a2,...,am) (m is number of target words)

$\Pr(x|ci)$is probability or likelihood of target word x over class ci.

Class with maximum probability can be considered as resulting class (cr∈ C) for that target word.

The question arises about what to choose in case of large number of small documents or small number of large documents. For sentiment analysis, if you are scrapping the data review by review during runtime, 'Sentence Level Classification' is a better approach as each review is examined independently. Most of the models of sentiment analysis use SVM and Naïve Bayes algorithm for classification. In order to increase their accuracy and efficiency, these classification algorithms are advanced by Artificial Neural Network (ANN with multilayer perceptron) or Back Propagation Network (BPN) [19] by making the approach non-deterministic. Also Naïve Bayes with modified k-Means clustering is more efficient than Naïve Bayes or SVM alone. Also, SVM and KNN algorithms are do not consider the uncertainty in the features. For example, 'That movie is formidably horror ' or 'The movie with full of conspiracy' doesn't mean that these movies belong to 'Bad Movies' category. Hence relative probability should be considered along with variable randomness. Even though other algorithms have better accuracy and performance, Naïve Bayes takes far less time in terms of run time sentiment analysis so this classifier should be used when training and testing time are crucial factors.

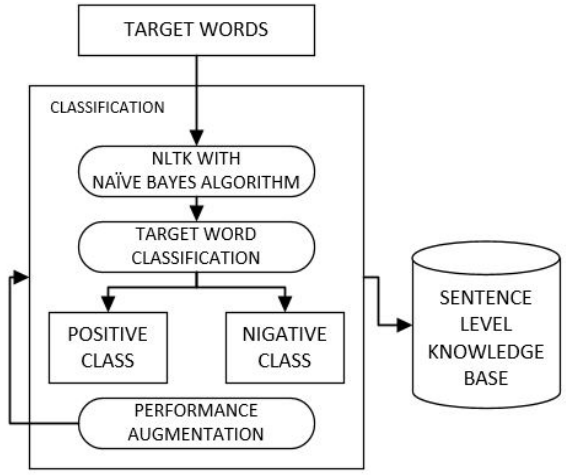

Figure 5: Classification of words

If the features or target words are well-defined by using lexical databases like WordNet (for synonyms only) and then sent for the classification by using toolkits like NLTK with Naïve Bayes algorithm, the chances of misclassification and classification of unnecessary 'stopwords' get reduced [20].Further the intelligence of this approach can be increased by advancing them using adaptive heuristic approach like Genetic Algorithms or Back Propagation Algorithms (BPN).

5. Aggregation and Evaluation

A customer review comprises of a number of sentences. To calculate the polarity of the review, the polarity of each individual sentence needs to be calculated. Aggregation is finding out the polarity of each review to conclude if it falls in the positive class or negative class. However, to find out the overall response about the product, an evaluation of all the reviews is required. Evaluation can also be further used for comparison amongst various e-commerce websites.

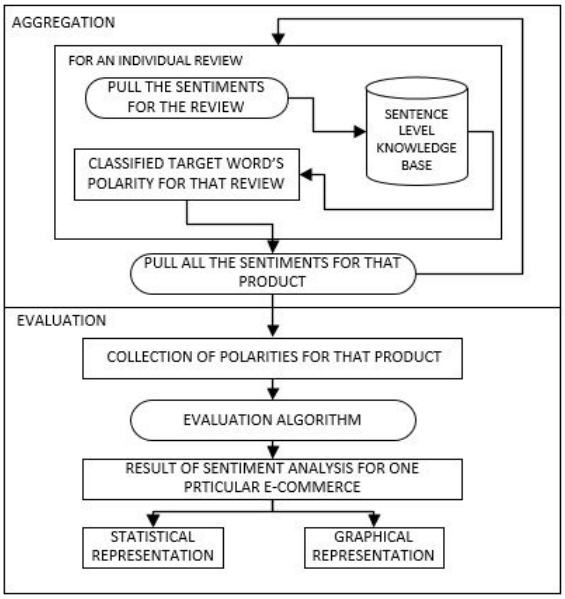

Figure 6: Aggregation and Evaluation

CONCLUSION

This paper depicts the available methods for carrying out sentiment analysis of reviews and have showcased the methods which the survey has shown to be the most efficient. Instead of using purely supervised or unsupervised learning algorithms advanced by heuristic approaches like ANN to increase accuracy, a semi-supervised approach is proposed in which emphasis is given to meaningful opinion words identified using WordNet. Sentiment analysis will be carried out at sentence level using NLTK with Naïve Bayes probabilistic model. Representation of results will be done graphically and statistically.

ACKNOWLEDGMENTS

This project was sponsored by Smt. KashibaiNavale College of Engineering Pune.

REFERENCES

[1] Erik Cambria, Björn Schuller, Yunqing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Computer Society, March/April 2013, Pages 15-21

[2] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K., "A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Pages: 223-233, Vol. 1, Issue 4, June 2012

[3] Christopher D. Manning, PrabhakarRaghavan, HinrichSch¨utze, "Introduction to Information Retrieval", ISBN-13 978-0-511-41405-3, 2013

[4] HuLi, Yong Shi"WordNet based lexicon model for CNL" 2009 IEEE proceeding at 2009 International Conference on Test and Measurement

[5] Kang Liu, Liheng Xu, and Jun Zhao "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model"2014 IEEE proceeding at IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[6] V.K. Singh, R. Piryani, A. Uddin P. Waila, Marisha "Sentiment Analysis of Textual Reviews ,Evaluating Machine Learning, Unsupervised and SentiWordNet Approaches" proceeding in IEEE 2013 5th International Conference on Knowledge and Smart Technology (KST)

[7] Mrs. Sayantani Ghosh, Mr. sudipta Roy, Prof. Samir K. Bandyopadhyay "A tutorial review on Text Mining Algorithm" Proceeding in International Journal of Advanced Research in Computer and Communication Engineering 2012.

[8] Pablo Gamallo,MarcosGarcia"Citius: A Naïve Bayes Strategy for Sentiment Analysis on English Tweets" Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.

[9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto,"Modern Information Retrieval", ISBN-13: 978-0201398298, 2013

[10] Fan Sun, Ammar Belatreche, Sonya Coleman,T. M. McGinnity ,Yuhua Li "Pre-processing Online Financial Text for Sentiment Classification: A Natural Language Processing Approach"

[11] HarunaIsah, Paul Trundle, Daniel Neagu "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis"2014 IEEE.

[12] I.Hemalatha, Dr. G. P Saradhi Varma, Dr. A.Govardhan "Preprocessing the Informal Text for efficient Sentiment Analysis" proceeding in International Journal of Emerging Trends & Technology in Computer Science 2012.

[13] ZohrehMadhoushi, Abdul RazakHamdan, SuhailaZainudin"Sentiment Analysis Techniques in Recent Works" proceeding in Science and Information Conference 2015.

[14] Kerstin Denecke "Using SentiWordNet for Multilingual Sentiment Analysis" 2008 IEEE.

[15] Zhang Yunliang, Zhu Lijun, QiaoXiaodong, Zhang Quan "Flexible KNN Algorithm for Text Categorization by Authorship based on Features of Lingual Conceptual Expression" 2008 IEEE ,World Congress on Computer Science and Information Engineering.

[16] Federica Bision, Paolo Gastaldo, Chiara Peretti, Rodolfo Zunino and Erik Cambria "Data Intensive Review Mining for Sentiment Classification across Heterogeneous Domains" 2013 IEEE at ACM International Conference on Adavances in Social Networks Analysis and Mining.

[17] Chris Tseng, NishantPateli, HrishikeshParanjape, T Y Lin, SooTee Teoh "Classifying Twitter Data with Naïve Bayes Classifier" IEEE 2012.

[18] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam "Comparative Study of Classification Algorithms used in Sentiment Analysis" proceeding in International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014.

[19] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam "Comparative Study of Classification Algorithms used in Sentiment Analysis" at (IJCSIT) International Journal of Computer Science and Information Technologies , 2014

[20] Rodrigo Moraes, Joao Francisco Valiati, Wilson P. GaviaoNeto,"Document-lavel Sentiment classification: An empirical comparison between SVM and ANN" Article from "Expert Systems with Applications" (2012) Journal Homepage: www.elsevier.com/locate/es

★★★