SENTIMENT ANALYSIS ON ONLINE PRODUCT REVIEW

Raheesa Safrin[1], K.R.Sharmila[2], T.S.Shri Subangi[3], E.A.Vimal[4]

[1, 2, 3](B.Tech, Final year student Department of Information Technology, Kumaraguru College of Technology, Tamil Nadu, India)
[4] Assistant Professor-III, Department of Information Technology, Kumaraguru College of Technology, Tamil Nadu, India

---***---

Abstract - *Sentiment analysis is a rapidly emerging domain in the area of research in the field of Natural Language Processing (NLP). It has gained much attention in recent years. Sentiment classification is used to verify or analyze the comments given by the user to extract the opinion from it. Sentiment analysis is a machine learning approach in which machines classify and analyze the human's sentiments, emotions, opinions etc. about the products which are expressed in the form of text, star rating, thumbs up and thumbs down. The data used in this study is online product reviews collected from the sample website that we have created. Words such as adjectives and adverbs are able to convey opposite sentiment with the help of negative prefixes. Negation phrase identification algorithm is used to find such words. The performance is evaluated through evaluation measures. At last, we also give insight into our future work on sentiment analysis.*

Key Words: **Sentiment analysis, negation phrase identification, product reviews**.

1. INTRODUCTION

Sentiment is an emotion or attitude prompted by the feelings of the customer. Sentiment analysis is also called as opinion mining which studies people's opinion towards the product. The dataset is collected from the website.

1.1 SENTIMENT ANALYSIS

Sentiment analysis is often referred to as opinion mining, because the opinion collected from the customer will be mined to reveal the rating of the product. It comes under machine learning. Since the online data's are tremendously growing data-by-day, it is considered to be very important in the current situation because, lots of user opinionated texts are available in the web now. Sentiment analysis is considered to be the study of user's thought and feeling towards a product. Both SA and OM are interchangeable.

The importance of the sentiment analysis or opinion mining is increasing day by day, as data grows day by day. Machines must be reliable and efficient to interpret and understand human emotions and feelings.

1.2 Challenges of sentiment analysis

Some of the major challenges in sentiment analysis,

> ➢ The comments given by user for a product is considered positive at one situation and negative at other situation.
> ➢ Some people don't express opinions in the same way. Most reviews will have both positive and negative comments, which somewhat manageable by analyzing sentences one at a time.
> ➢ Sometimes people may give fake comments about the product, which gives the bad review about the product.

> The sentiment analysis problem can be sometimes managed by manual methods.

2. RELATED WORKS

Xing Fang and Justin Zhan [1] describes that the subjective contents are extracted, it consist of sentiment sentences which contain at least one positive or negative word. These sentences are tokenized into separated English words. Depending on parts of speech in the words, corresponding tags are used. Feature vector formation - The sentiment tokens and scores are information extracted from the original dataset. These are known as features. In order to classify them these features are to be transformed to vectors called feature vector.

Huge amount of content produced by amateur authors on various topics are considered in **[2]**. Sentiment analysis (SA) aggregates users' sentiments. Machine learning (ML) techniques for natural language. In lexicon-based techniques prevent over fitting. Corpus-based statistical techniques for stabilization. This paper highlights natural language processing (NLP) specific open challenges.

Two typical approaches to sentiment analysis – lexicon look up and machine learning are used by Ji Fang and Bi Chen in [3]. Lexicon look up starts with a lexicon of positive and negative words. Current sentiment lexicons do not capture such domain and context sensitivities of sentiment expressions. The proposed system present an alternative method that incorporates sentiment lexicons as prior knowledge with machine learning approaches such as SVM to improve the accuracy of sentiment analysis.

Unsupervised learning algorithm for classifying reviews as thumbs up or thumbs down by the average semantic orientation is carried out by Peter D. Turney as referenced in [4]. The semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor".

Feature driven opinion summarization method is considered in [5]. For each product class, general features are extracted and for each product, specific features and feature attributes are extracted. Then polarity is assigned to each the feature using Support Vector Machines and Sequential Minimal Optimization.

Opinions are given by the user through various sources about the product and their services. Sentiment analysis system using modified k means and naïve Bayes algorithm that saves running time and reduces computational complexity is the analysis done by Ashish Shukla and Rahul Misra is directed in [6]. The same system can be extended to other product review domains easily.

Data mining techniques used to discover common features across products and relationship among those features are studied in [7]. Novel incremental diffusive algorithm is being used to extract features from online product descriptions, and then employ association rule mining and the k-nearest neighbor. Machine learning method to make feature recommendations during the domain analysis process.

The textual data constitute resources that it is worth exploiting. So S. Cherfi, A.Nepoli, Y.Toussaint [8] proposed the use of Knowledge discovery from textual

databases, or for short, text mining (TM), is an important and difficult challenge, because of the richness and ambiguity of natural language (used in most of the available textual documents).

The challenges raised by sentiment aware applications, as compared to traditional fact-based analysis are considered in [9]. Summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to are included under it.

In order to overcome the drawbacks of individual algorithm, different types of features and classification algorithms are combined and used by G.Vinodhini and R.M.Chandrasekaran which is stated in [10] in their research and also benefit from each other's merits, and finally enhance the sentiment classification performance in an efficient way.

ANN is used to predict the customer comments in the social media about the restaurant. ANN provides more accuracy than the support vector algorithm is being analyzed in [11], where SVM gives less accuracy compared to ANN.

Pranali Borele, Dilipkumar, A.Boriker have used some algorithms which are used in sentiment analysis to give their best results but none can resolve all the challenges. Some researchers said that SVM has high accuracy than other algorithms, but it also has some limitations. They have used ANN, this results that ANN with fuzzy logic is an improved one which is mentioned in [12].

The different combinations of functions and its effects while using ANN is analyzed by Saravanan.K and Sasidhra.S which is referenced in [13]. They are using ANN, they are trained by back-propagation algorithm. And then BPNN is used. It's used in classifying images in remote sensing area. This method proves to be more effective than other classification algorithm.

3. PROPOSED SYSTEM

The objective is to review the product based on the comments given by the customer. The comments show the opinion of the user towards the product. These comments may be positive or negative. These comments may be in the form of sentences. In order to gain the sentiment of the user, these sentences have to be segregated into words in which the adjectives, verb and adverbs are processed using Parts Of Speech Tagging (POST). Some sentences may contain negative expression which are identified and processed using Negation Phrase Identification algorithm. Due to rapid growth of data in E-commerce, it is used to reveal the quality of product. As these websites have become the major source for the customers to get rating of a product but due to huge amount of data available it becomes difficult for them to make decisions.

Our system that could summarize the feedbacks, extracting the opinions from all this information, giving an overall view of the product, that could save time and ease the decision process of the customer.

4. IMPLEMENTATION:

To review the online product, the comments are collected from the customer in the form of textual,

thumbs-up and thumbs-down, etc. Using the feedback, they are processed and finally reviewed. Figure 1 represents the implementation process. The implementation steps may include,

(I) Creation of website and retrieving the feedback (i.e.) data Collection
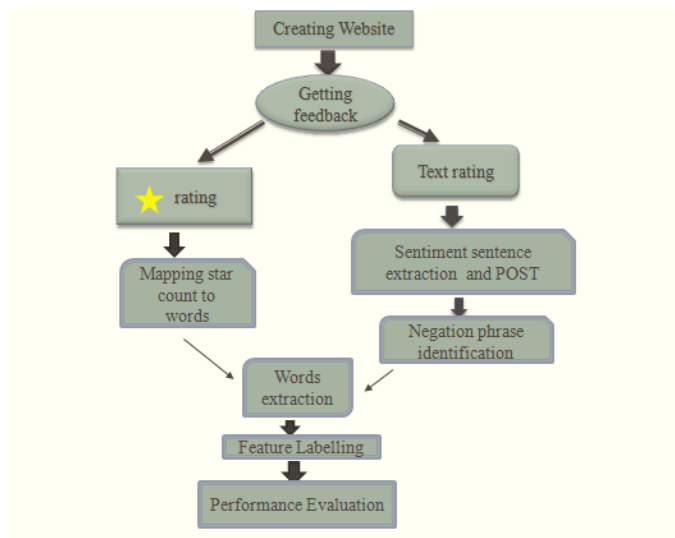
(ii) Pre-processing and NLP

(iii) Feature Labeling.

Figure 1 is a flowchart that depicts our proposed process for sentiment analysis as well as the outline of this paper.

I Creation of website and retrieving the feedback (i.e.) data Collection

The website is created with different products from which the data is collected in the form of feedback. The feedback is collected in the form of two different ways like,

1. Star rating
2. Textual format.

The count of the star is mapped to certain adjectives like good, bad, excellent. The count may vary from 1 to 3. The count 1 maps to bad. The count 2 maps to good and 3 maps to excellent which is shown in figure 2.

Figure 2: Rating System for our website.

The other type is text format in which customer express their feedback in the form of sentences. The feedback may consist of sentiment sentences that show the opinion of the customer towards the product.

TYPES OF RATING

- Star-rating
- Thumbs-up and thumbs-down
- Textual
- Emoji

Feedback:

The type of feedback that we have used in our website is textual and star-rating which is shown in the figure 3.

Figure 3: Feedback format

II Data pre-processing

In this section, the input data i.e. the customer reviews dataset is preprocessed to improve the classification results. The process of data preprocessing includes two main steps. One is Parts Of Speech Tagging (POST) which is mainly for positive phrases. To find the phrases with negative prefixes we use Negation Phrase Identification algorithm. Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data an stop words. POS tagging is the process of marking a word in a text as corresponding to a particular part of speech as its context i.e. relationship with adjacent and related words in a phrase, sentence or a paragraph. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction, prepositions and determiners. It has been found that certain parts of speech such as adjectives and adverbs express polarity more often. Words such as adjectives and verbs are able to convey opposite sentiment with the help of negative prefixes. For instance, the phrase "don't like", here though like is a positive word this phrase is considered as a negative due to the presence of prefix (don't).

- The sentences that are tagged are checked for negation with file adjective or verb.
- We compare those words from file, if it is a single word it is considered as 'i' .If it is a phrase, we considered it as an 'i+1'.
- If it is a negative word, we check the next word is adjective or verb, we will return 'i'th and 'i+2'th word.
- If we didn't find the adjective or verb in that phrase then we will go for the next phrases of the sentence.
- In that case, we will return the 'i'th and 'i+4' th.

III Feature Extraction

The input data can be transformed into a reduced set of features (feature vectors). This process is called feature extraction. For feature labeling, we use two files containing positive (count include 2005) and negative (count include 4781) words collected from the dictionary. The resultant set of pre-processing is compared with these files. The positive words are labeled as '0' whereas negative are labeled under '1'.

K-means cluster

K-means clustering is used to classify the retrieved dataset through a certain number of clusters. Let the number of clusters be 2 (0 and 1). The labelled words are now taken for clustering. As the result we get the clusters of positive and negative words.

Dataset

The dataset contains online product reviews along with their associated binary sentiment polarity labels. The dataset is obtained by creating an own online shopping website. The user can view their interested product

and they can buy it. And also the user may comment about the product through various means like star rating, textual and thumbs-up and thumbs-down. Those comments are taken as review and it is considered as a dataset for our project. The number of entries in the dataset is 3100 is shown in figure 4.

Figure 4: Collected feedback

IV Performance Evaluation

Text classification rules are typically evaluated using

Performance measures from information retrieved. Common metrics for text categorization evaluation include recall, precision and accuracy. For the collected clustered dataset, a two-by-two contingency table with four cells is constructed for each classification problem.

The cells contain counts for true positive (tp), false

Positive (fp), true negative (tn) and false negative (fn).

Total data count = tp + fp + tn + fn.

The number of true positive, true negative, false positive and false negative are calculated. Using this recall, precision and accuracy are evaluated.

RECALL

Recall is the proportion of real positive cases that are correctly predicted positive.

Recall = tp/ (tp+fn)

It is the number of correct results divided by the number of results that should have been returned. The percentage obtained during recall is 90%.
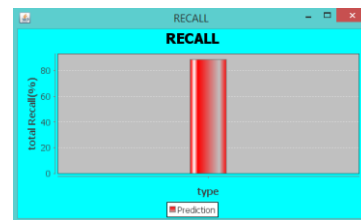
Figure 4: Recall

PRECISION

It is a proportion of predicted positive cases that are correctly real positive.

Precision = tp / (tp+fp)

It is the number of correct results divided by the number of all returned results. The percentage obtained during precision is 87%.
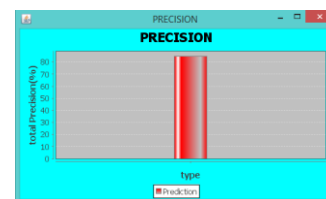

Figure 5: Precision

ACCURACY

Accuracy represents what percent of prediction were correct. The percentage obtained during precision is 90.47%.

Accuracy = (tp+tn)/ (tp+tn+fp+fn)

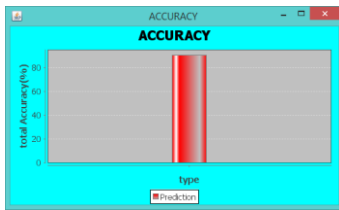**Figure 6**: Accuracy

Table 1 depicts the percentage obtained during precision, recall and accuracy.

	PERCENTAGE
RECALL	90%
PRECISION	87%
ACCURACY	90.47%

TABLE 1: Performance level

4. CONCLUSIONS

Sentiment analysis or opinion mining is the study that is used to analyze people emotions, sentiments towards the product. This paper is used to perform evaluation measure on comments obtained from the customer. Online product reviews from our website are selected as data used for this study. The POS tagging is used to extract the most relevant features to get better results in classifying the sentence as positive or negative. This positive and negative separation of comments is used to analyze the quality of the online products.

REFERENCES

[1] Xing Fang* and Justin Zhan "Sentiment analysis using product review data" Fang and Zhan Journal of Big Data (2015) 2:5

[2] Muhammad Taimoor Khan, Mehr Durrani2, Armughan Ali, Irum Inayat, Shehzad Khalid and Kamran Habib Khan "Sentiment analysis and the complex natural language" Khan et al. Complex Adapt Syst Model (2016) 4:2

[3] Ji fang, Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pp. 94–100.

[4] Turney, Peter D., "Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of Association for Computational Linguistics, Philadelphia, PA. July 2002, pp. 417-424.

[5] Turney, Peter D., "Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of Association for Computational Linguistics, Philadelphia, PA. July 2002, pp. 417-424.

[6] Ashish Shukla* Rahul Misra M.tech Scholar, CSE Department Assistant Professor, CSE Department, Pranveer Singh Institute of Technology, Kanpur Pranveer Singh Institute of Technology, Kanpur U.P.T.U., Luck now, Uttar Pradesh, India U.P.T.U., Luck now, Uttar Pradesh, India "Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm"

[7] Negar Hariri, Carlos Castro-Herrera, Member, IEEE, Mehdi Mirakhorli, Student Member, IEEE, Jane Cleland-Huang, Member, IEEE, and Bamshad Mobasher, Member, IEEE "Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings" IEEE

TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 12, DECEMBER 2013.

[8]   H. Cherfi · A. Napoli · Y. Toussaint "Towards a text mining methodology using association rule extraction".

[9]   Opinion mining and sentiment analysis, Bo Pang1 and Lillian Lee2. 1 Yahoo! Research, 701 First Ave. Sunnyvale, CA 94089, U.S.A., bopang@yahoo-inc.com, 2 Computer Science Department, Cornell University, Ithaca, NY 14853, U.S.A., llee@cs.cornell.edu

[10] Sentiment Analysis and Opinion Mining: A Survey G.Vinodhini* Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002. RM.Chandrasekaran, Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002. India.

[11] Customers behaviour prediction using artificial neural network BichenZheng, Keith Thompson, Sarah S.Lam, Sang Won Yoon, Nathan Gnanasambandam.

[12] An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques Pranali Borele, Dilipkumar A. Borikar

[13] Review on classification based on artificial neural netwoks, Saravanan K1 and S. Sasithra2.