OPINION MINING / SENTIMENT ANALYSIS FOR USER REVIEWS

Liya Mathew[1], Ann Dona James[2], Anjima Shaji[3], Ms. Sona Maria Sebastian[4]

[123]Student, Amal Jyothi College of Engineering, Kanjirapally, Kerala
[4]Assistant Professor, Amal Jyothi College of Engineering, Kanjirapally, Kerala

Abstract: Sentiment analysis, also called the Opinion Mining is a type of Natural Language Processing (NLP) in which the people's opinions, sentiments, emotions, attitudes etc., are extracted from the text based on the polarity. i. e, the positivity or negativity in the text. It is an active research area in the field of study of NLP and the data mining. It has grown widely due to its importance to business and society. The Opinion Mining / Sentiment Analysis is used widely in the field of social media and other reviewing systems. In this paper, by using the K-Nearest Neighbour algorithm, the sentiment analysis is used for reviewing a system.

Keywords: Opinion Mining, Polarity, Sentiments, Natural Language Processing.

I. INTRODUCTION

Nowadays, all day-to-day applications are going online. Due to the growth of technologies, people use their smart phones, tablets and laptops for all applications. Everyone can view/buy/use a product/service online and can review a product or service through online. These online sites are being widely used by people to express their emotions, beliefs as well as opinions towards any product or services. The sentiment analysis of user reviews can be useful for the other customers who need to use that product or service too apart from other business enterprises. For example, if someone wants to buy a product or use a service, the first step is to go online and check the user reviews about the particular product. However, it is not possible for the users to check the massive amount of user reviews online. Therefore, several sentiment analysis techniques have been proposed to automate this analysis process. In this paper, a classification method called the K-Nearest Neighbour algorithm used to implement a system in which the user reviews are categorized as positive, negative and neutral reviews and generate a score based on the polarity.

What is a Review?

A review is an evaluation procedure of a service, a product, a publication or any other company. The user can give a review about anything such as movie, songs, books, etc. A compilation of reviews itself may be called a review.

A user review refers to a review, which is written by a user for a particular product, or a service based on his/her experience as a user of the reviewed product. Some popular sites where we can see the consumer reviews are e-commerce sites such as Amazon.com, Trip Advisor etc. E-commerce sites have the consumer reviews for products and sellers separately. Usually, consumer reviews are in the form that is of several lines of texts along with a numerical rating. This helps the other user to select a best product or service from a group of products. A customer review of a product or service usually comment on how well the product or service is useful for them and how well they meet the measures up to expectations based on the specification given by the seller. It talks about anything such as the performance, quality defects, if any, and value for money. Customer review, also known as the user-generated content, is entirely different from that of the marketer-generated content. Because, in the marketer-generated content, the review answers are generated by the marketer itself and the user need to choose any of the answer given by the marketer. Perceptions are true as well as emotional in nature.

What is Sentiment Analysis?

Sentiment Analysis, also known as the Opinion Mining is a type Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text based on the polarity. Besides evaluating the opinions, this system also evaluates the emotions of the users also. e.g.:

- *Polarity:* Positive or negative opinion expressed by the speaker.
- *Subject:* The topic that is being talked about.
- *Opinion holder:* The person or the entity that gives the opinion

Currently, sentiment analysis is a topic of great interest and development since it is used in various applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media.

With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for different commercial applications.

Sentiment Analysis Algorithms

There are many methods and algorithms to implement sentiment analysis systems, which can be classified as:

- **Rule-based systems**:
  - Perform sentiment analysis based on a set of manually created rules.
- **Automatic systems**:
  - Depends on machine learning techniques to learn from data.
- **Hybrid systems**:
  - Combination of both rule based and automatic approaches.

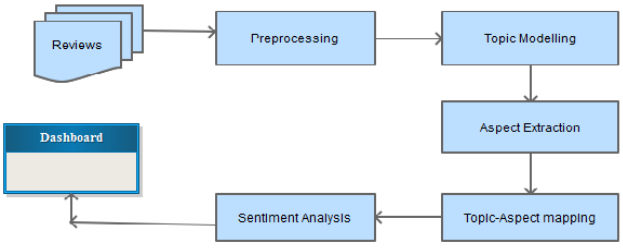

Fig 1: Processing in Sentiment analysis

Levels of Sentiment Analysis:

Sentiment analysis can be performed at different levels of scope:

- **Document level sentiment analysis**:
  - Evaluates the sentiment of an entire document or paragraph.
- **Sentence level sentiment analysis**:
  - Evaluates the sentiment of a single sentence.
- **Sub-sentence level sentiment analysis:**
  - Evaluates the sentiment of sub-expressions among a sentence.

Types of sentiment analysis

1. **Fine-grained sentiment analysis**:
   - Provides a more precise level of polarity by breaking down the text into further categories, usually very positive to very negative. e.g.: Very Positive = five stars and Very Negative = one star.
2. **Emotion detection:**
   - Identifies emotions rather than positivity and negativity. Examples could include happiness, sorrow, frustration, shock, anger, sadness, etc.
3. **Intent-based analysis:**
   - Identifies actions behind a text in addition to opinion.eg:"Your customer support is very poor. I holded for 15 minutes", which shows the fault in customer support.
4. **Aspect-based analysis**:
   - Identifies a specific component, which is positively or negatively mentioned. For example, if the battery life of a product is too short, then, the system will return that the negative sentiment is not about the whole product, but about the battery life only.

II. LITERATURE REVIEW

Sentiment analysis played a dominant role in the area of researches done by many researchers; there are many methods to carry out sentiment analysis. Many researches are going on to find out better alternatives due to its importance.

In [1], Soudamini Hota & Sudhir Pathak compares the K-Nearest neighbour [KNN] algorithm & Support Vector Machine[SVM] . It shows that the analysis is improved further by using KNN algorithm to train the classifier than the SVM technology. It is improved further by employing distance weighted KNN algorithm that involves associating weights with the nearest neighbors based on their proximity to the data point.

In [2], Lopamudra Dey compares the KNN algorithm with the Naive Bayes Classification. Their experimental results show that the classifiers yielded better results for the movie reviews with the Naïve Bayes'approach. It giving above 80% accuracies and outperforming than the K-NN approach.

In [3], Surya Prakash Sharma says that, first extracts the feature, modifier and opinion from the dataset and then using clustering mechanism divide them into discrete clusters by user's opinion. A feature wise opinion mining system to determine the polarity of the opinions in reviews documents using Senti-WordNet.

In [4], Devika M D had made a comparative study of different approaches in sentiment analysis. She conclude that, in the internet world majority of people depends on social networking sites to get their valued information, analyse he reviews from these blogs will yield a better understanding and help in their decision-making.

In [5], Wararat Songpan uses two methods to calculate the sentiment analysis called the Naive Bayes classification & Decision tree algorithm. The classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques, which gives correct classifier to 94.37° ~ compared with decision tree Techniques.

In [6], Vidisha M. Pradhan had done a Survey on Sentiment Analysis Algorithms for Opinion Mining. Dictionary based technique takes less processing time even though the accuracy is not up to the mark. However, the supervised learning approach provides better accuracy. From the

survey, it summarizes the supervised techniques provide better accuracy compared to dictionary based approach.

III. IMPLEMENTATION

K-Nearest Neighbour [KNN] Algorithm:

In this sentiment analysis implementation, the K-Nearest Neighbour method is used, which is a non-parametric supervised learning technique in which we try to classify the data point to a given category with the help of training set. In other words, it captures data of all previously stored cases and classifies new cases on the basis of a similarity. It is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbours. The case being assigned to the class is the most common among its K nearest neighbours measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski, levenshtein and Hamming distance. If K = 1, then the case is assigned to the class of its nearest neighbour.

The algorithm looks at the different centroids and compares distance using some of functions, and then analyzes those results and assigns each point to the group so that it is optimized to be placed with all the closest points to it.

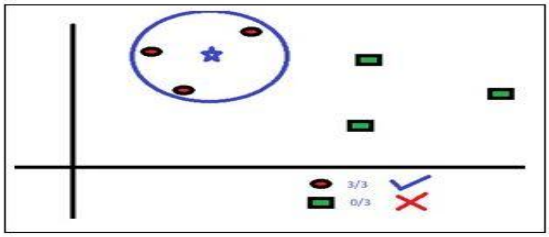

Fig 2: Example for KNN algorithm

The score can be calculated using:

$$Positivity\ Score = \left(\sum_1^j score(pos) + \sum_1^k score(neg)\right) \Big/ \sum_1^s maximum\ score \quad - (1)$$

Here s=j+k, ie. Count of both positive and negative together. In weighted k-NN method, they first tokenize the sentences and removed the stop words from the comments they have fetched. The algorithm proposed by the authors of is carried out in two parses.

- After the first phrase, a positive score is assigned to each review.
- This is passed for second parsing & a neutral review input is given.

Using this, the score is modified if required. It is done for better positivity determination and an output file consisting of review ID and its positive score is determined.

Strings are broken into tokenized arrays of single words. These words are analyzed against txt files that contain different words with the ratings. A score is then calculated based on this analyses and this forms the "Sentiment analysis score".

Pseudo-code for KNN:

The implementation of a KNN model is as follows:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
    1. Calculate the distance between test data and each row of training data.
       Here we will use levenshtein distance as our distance metric.
       The other metrics that can be used are Euclidean, Chebyshev, cosine, etc.
    2. Sort the calculated distances in ascending order based on distance values
    3. Get top k rows from the sorted array
    4. Get the most frequent class from these rows.
    5. Return the predicted class.

Levenshtein Distance:

The Levenshtein distance is a string metric to calculate the difference between two sequences of words. In other words, the Levenshtein distance between two words is the minimum number of edits required to change one word to the other**.** (i.e. insertions, deletions, or substitutions etc.)

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & if min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases} \quad - (2)$$

Here, $1_{(ai \neq bj)}$ is the indicator function equal to zero, when $a_i \neq b_j$. In addition, equal to one otherwise, and $lev_{a,b}(i,j)$ is the distance between first i characters of 'a' & first j characters of 'b'.

A. *Phrase Analysis*

This function is to identify whether the phrase in questions can be compared to phrases that we have analyzed and stored before. It uses Levenshtein distance to calculate distance between word length phrases against the dataset we already have. We also make use of PHP's similar_text to double verify proximity. The phrases can be more accurately scored against historical data and more phrases that had analyzed previously improve the entire dataset.

1. The phrase is broken into n-gram lengths texts.
2. Then, the array is sorted in reverse.
3. Phrases are matched against positive, negative and neutral phrases with the appropriate TXT files.

4. Only matches that meet the minimum levenshtein min distance & similiarity_min_distance are kept

IV. METHOD OF IMPLEMENTATION

Technologies used to implement proposed system:

- PHP: a server-side scripting language designed specifically for web development.
- MySQL: a free, open-source, database management system.

V. RESULT

The precision of the proposed system is approx. 82 percent. The recall of the proposed system is 81.5 percent. The accuracy achieved by the proposed system is up to 86 percent.

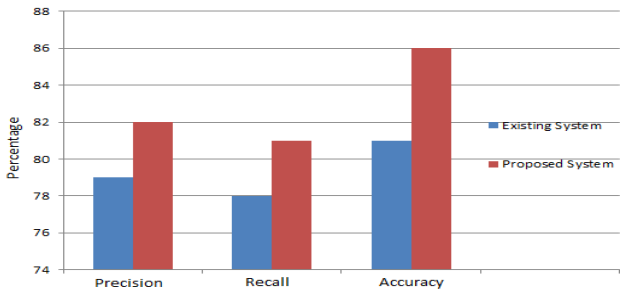

Fig 3: Performance Analysis

When processing time is considered it is shown that the processing time is totally depend upon the size of test set as the size increases the processing time increases and remain same for these classifiers and if different number of documents (and of different test size) are used then we can observed the processing time differences.

VI. FUTURE WORK

When classification methods are applied on same data sets to find the optimal result shows that K-NN classification method gives more accuracy (~83.65%) as compare to naïve Bayesian classification method that gives the accuracy result (~75.77%).The classification can be further be improved by using various other attributes and increasing the number of cases for training and testing. The efficiency of result can be further increased by using better feature selection methods like CHI Square, Information Gain, etc.

VII. CONCLUSION

Various sentiment analysis methods and their level of analysing have been seen in this paper. Our ultimate aim is to come up with Sentiment Analysis, which effectively calculate and categorize the user reviews. Research work is carried out for better analysis methods in this area, Here, the K-Nearest Neighbour algorithm is used to effectively calculate the polarity of the reviews, by grouping the reviews as positive, negative and neutral values. In the world of Internet majority of people depend on social networking sites to get their valued information, analysing the reviews from these blogs will yield a better understanding and help in their decision-making.

REFERENCES

[1] Soudamini Hota , Sudhir Pathak, "KNN classifier based approach for multi-class sentiment analysis of twitter data"- *International Journal of Engineering & Technology, 7 (3) (2018) 1372-1375*

[2] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *I.J. Information Engineering and Electronic Business,* 2016.

[3] Surya Prakash Sharma, Dr Rajdev Tiwari, Dr Rajesh Prasad ,"Opinion Mining and Sentiment Analysis on Customer Review Documents"- A Survey, *International Conference on Advances in Computational Techniques and Research Practices-Vol. 6, Special Issue 2, February 2017*

[4] Devika M D, Sunitha C, Amal Ganesha, "Sentiment Analysis:A Comparative Study On Different Approaches"- *Procedia Computer Science 87 (2016).*

[5] Wararat Songpan ,"The Analysis and Prediction of Customer Review Rating Using Opinion Mining"-, *2017 IEEE SERA 2017, June 7-9,2017, London, UK.*

[6] Vidisha M. Pradhan, Jay Vala,Prem Balani ,"A Survey on Sentiment Analysis Algorithms for Opinion Mining"-, *International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016*