OPINION MINING / SENTIMENT ANALYSIS FOR USER REVIEWS

Liya Mathew
Scholar,Amal Jyothi
College of Engineering,
Kanjirapally,Kerala

Ann Dona James
Scholar,Amal Jyothi
College of Engineering,
Kanjirapally,Kerala

Anjima Shaji
Scholar,Amal Jyothi
College of Engineering,
Kanjirapally,Kerala

Abstract: Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is an active research area in natural language processing and is also studied in data mining, Web mining, and text mining. In fact, it has grown widely due to its importance to business and society as a whole. The growing importance of sentiment analysis matched with the growth of social media such as reviews, Twitter, and other social networks. In this paper,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. These online sites are being widely used by people to express their emotions, beliefs as well as opinions towards any entity ranging from product, person, and event and so on. Apart from business enterprises, sentiment analysis of user comments is of immense use for buyers too. For instance, if someone wants to either buy a product or access any service, generally the initial step would be to go through online reviews and generate a discussion regarding it on the social media before taking any decision. However, it is not possible for a user to analyze all the reviews considering the massive amount of user reviews and comments available on online platforms. Hence, several sentiment analysis techniques have been proposed in order to automate this analysis process.

What is a Review:

A review is an evaluation of a publication, service, or company such as a movie review, video game review, book review etc.,. In addition to critical evaluation, the review's author assign the work a rating to indicate its relative merit. More loosely, an author may review current events or trends in the news. A compilation of reviews itself may be called a review.

A user review refers to a review written by a user for a product or a service based on his/her experience as a user of the reviewed product. Popular sources for consumer reviews are e-commerce & social media sites such as Amazon.com, TripAdvisor etc., E-commerce sites often have 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 help the user to select a respective product. A consumer review of a product usually comments on how well the product measures up to expectations based on the specifications provided by the manufacturer or seller. It talks about performance, reliability, quality defects, if any, and value for money. Consumer review, also called 'word of mouth' and 'user generated content' differs from 'marketer generated content' in its evaluation from consumer or user point of view. Often it includes comparative evaluations against competing products. Observations are factual as well as subjective in nature. Consumer review of sellers usually comment on service experienced, and dependability or trustworthiness of the seller. Usually, it comments on factors such as timeliness of delivery, packaging, and correctness of delivered items, shipping charges, return services against promises made, and so on.

What is Sentiment Analysis:

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

- *Polarity:* If the speaker expresses a positive or negative opinion.
- *Subject :* The thing that is being talked about,
- *Opinion holder:* The person, or entity that expresses the opinion.

Currently, sentiment analysis is a topic of great interest and development since it has many practical 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 commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service.

Sentiment Analysis Algorithms

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

- **Rule-based** systems that perform sentiment analysis based on a set of manually crafted rules.
- **Automatic** systems that rely on machine learning techniques to learn from data.
- **Hybrid** systems that combine both rule based and automatic approaches.

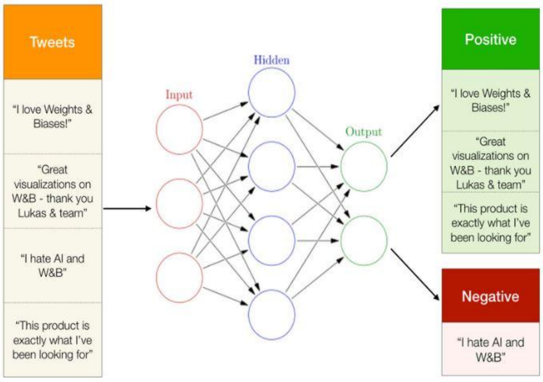
Tweets
"I love Weights & Biases!"
"Great visualizations on W&B - thank you Lukas & team"
"I hate AI and W&B"
"This product is exactly what I've been looking for"
Input
Hidden
Output
Positive
"I love Weights & Biases!"
"Great visualizations on W&B - thank you Lukas & team"
"This product is exactly what I've been looking for"
Negative
"I hate AI and W&B"

Fig 1 : Processing in Sentiment analysis

Sentiment Analysis Scope:

Sentiment analysis can be applied at different levels of scope:

- **Document level** sentiment analysis obtains the sentiment of a complete document or paragraph.
- **Sentence level** sentiment analysis obtains the sentiment of a single sentence.
- **Sub-sentence level** sentiment analysis obtains the sentiment of sub-expressions within a sentence.

Types of sentiment analysis

1. **Fine-grained sentiment analysis** provides a more precise level of polarity by breaking it down into further categories, usually very positive to very negative. This can be considered the opinion equivalent of ratings on a 5-star scale. e.g.: Very Positive = 5 stars and Very Negative = 1 star.
2. **Emotion detection** identifies specific emotions rather than positivity and negativity. Examples could include happiness, frustration, shock, anger and sadness.
3. **Intent-based analysis** recognizes actions behind a text in addition to opinion. For example, an online comment expressing frustration about changing a battery could prompt customer service to reach out to resolve that specific issue.eg:"Your customer support is a disaster. I've been on hold for 20 minutes".
4. **Aspect-based analysis** gathers the specific component being positively or negatively mentioned. For example, a customer might leave a review on a product saying the battery life was too short. Then, the system will return that the negative sentiment is not about the product as a whole, but about the battery life.

II. LITERATURE REVIEW

Sentiment analysis played a great role in the area of researches done by many researchers, there are many methods to carry out sentiment analysis. Still many researches are going on to find out better alternatives due to its importance in this scenario.

In [1],Soudamini Hota & Sudhir Pathak compares the K-Nearest neighbour [KNN] algorithm & Support Vector Machine[SVM] .And 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 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 divides them into discrete clusters by user's opinion.A feature wise opinion mining system to determines 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 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.

In [5], Wararat Songpan uses two methods to calculate the sentiment analysis called the Naive Bayes clsiification & Decision tree algorithm. this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques, which gives correctly 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 approach takes less processing time than supervised learning approach but accuracy is not up to the mark.Supervised learning approach provides better accuracy. From the survey, it is concluded that supervised techniques provide better accuracy compared to dictionary based approach.

In [7], Divyashree N explains the Opinion Mining and Sentimental Analysis of TripAdvisor.in for Hotel Reviews. The J48 algorithm is to classify Positive words and Negative words in reach review of all the hotels on the TripAdvisor website.

In [8], Smita Bhanap explains Sentiment analysis of mobile datasets using naïve bayes algorithm**.**Naïve Bayes algorithm is applied to train and test data to find sentiment polarity of overall sentence.

III. IMPLEMENTATION

K-Nearest Neighbour [KNN] Algorithm:

In this sentiment analysis project,we use the K-Nearest Neighbour method,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 simple words, it captures information of all training cases and classifies new cases based on a similarity. It is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is the most common among its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski,levenshtein and Hamming distance. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing KNN modeling.

The algorithm looks at different centroids and compares distance using some sort of function (usually Euclidean), 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.

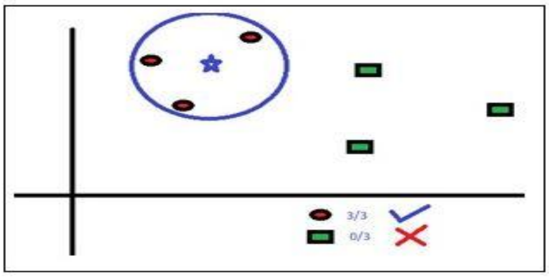

Fig 2: Example for KNN algorithm

The score can be calculated using:

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

-(1)

Here s=j+k, ie. Count of both positive and negative together. In weighted k-NN method they first of all tokenise 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. A positive score is assigned to each reviews after the first parse. This is passed for second parsing and an input of neutral review 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 tokenised arrays of single words. These words are analysed against TXT files that contain emotion words with ratings, emoticons with ratings, booster words with ratings and possible polarity changers. A score is then calculated based on this analyse and this forms the "Sentiment analysis score".

Pseudo-code for KNN:

We can implement a KNN model by following the below steps:

```
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 Euclidean distance as our distance metric since it's the most popular method.
       The other metrics that can be used are Chebyshev,Levenshtein, 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 of these rows
    5. Return the predicted class
```

Levenshtein Distance:

The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions, or substitutions) required to change one word into the other**.**

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

-(2)

Here, $1(a_i \neq b_i)$ is the indicator function equal to 0 when $a_i \neq b_i$ and equal to 1 otherwise, and $\boldsymbol{lev_{a,b}(i,j)}$ is the distance between the first i characters of a and the first j characters of b.

A. Phrase Analysis

This function is key to identifying whether the phrase in questions can be compared to phrases that we have analysed 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.This means that the more phrases we have analysed previously improves the entire dataset and allows phrases to be more accurately scored against historical data.

1. The phrase is broken up into n-gram lengths.

2. The array is reverse sorted so we compare 10 word length phrases first, then 9, and so on

3. Phrases are matched against positive, negative and neutral phrases in the relevant 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 : PHP is a server-side scripting language designed specifically for web development.
- MySQL: MySQL is 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.

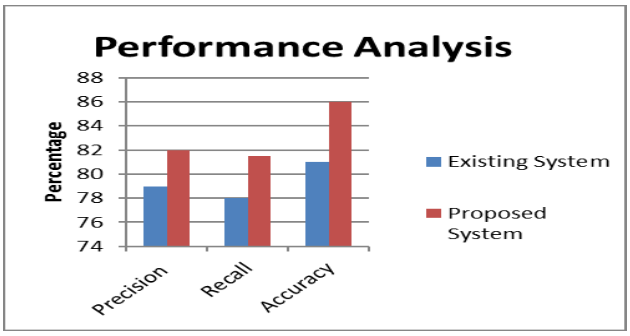

Fig 3:Performance Analysis

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 (approx 83.65%) as compare to naïve bayesian classification method that gives the accuracy result (approx 75.77%) .The classification can be further be improved by incorporating 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, Relevance Factor, Information Gain and other weighted features.

VII. CONCLUSION

Various sentiment analysis methods and its different levels of analysing sentiments have been studied in this paper. Our ultimate aim is to come up with Sentiment Analysis which will efficiently categorize various reviews. Research work is carried out for better analysis methods in this area, We use the K-Nearest Neighbour algorithm to effectively calculate the polarity of the reviews. 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*

[7] Amit G. Shirbhate, Sachin N. Deshmukh ,“Feature Extraction for Sentiment Classification on Twitter Data”-, *International Journal of Science and Research (IJSR), Volume 5 Issue 2, February 2016*

[8] Smita Bhanap, Dr. Seema Kawthekar ,“Sentiment analysis of mobile datasets using naïve bayes algorithm”-, *Volume 9, No. 2, March-April 2018 International Journal of Advanced Research in Computer Science.*