International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Brand Review Prediction using User Sentiments

Arati Manjaramkar [1], Rupesh Sonkamble[2]

[1, 2]Department of Information Technology, SGGSIE&T, Nanded Department & College

---

Abstract - As the speed of innovation increases at an accelerating rate, Individuals' ways of sharing their views on different websites is also expanding. On social media sites like Facebook, Twitter, and Yelp, there are various reviews and opinions. Scores are normally provided on a range of 1 to 5 stars. Assessment of opinions through textual data analysis has played a critical role in analytics research because it offers useful options to sentiment mining. A review is an evaluation of a product or service by someone who has used the product or service or has experience with it. The ranking of any ecommerce site is heavily influenced by the opinions of its users. The goal of this paper is to describe how a Machine Learning (ML) algorithm works on Yelp's database to evaluate, anticipate, and recommend brands. With the sentimental analysis algorithm, we implemented Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, K-Nearest Neighbour, and Multilayer Perceptron classifiers. The best result of the multilayer perceptron classifier is 93.40 percent.

Key Words: Machine Learning (ML) algorithm, Brand Review Prediction, Multilayer Perceptron classifiers.

1. INTRODUCTION

In this age of modernization, we are constantly in search for some idea that will allow us to save time; reduce the task's complexity by substituting manual production. Analysis of sentiment is a unique technique that helps us to save time while still accomplishing our task. Rather than detecting emotions, data analysts use the phrase "sentiment analysis." Opinion mining, also known as sentiment mining, is a form of natural language processing that identifies the audience's feelings, opinions, and sentiments about a particular object, film, or condition. Customer reviews are an important part of web-based applications like Zomato, Yelp, Swiggy, Amazon, Foodpanda, and others, where customers can share their opinions on companies, products, and services through free-form textual reviews and numerical star ratings, typically out of five. These online reviews serve as "digital word-of-mouth" and as a consumer search factor for similar products available on the market. According to study, they have a huge influence on consumer purchase decisions, as well as brand sales and revenue. Only extracting features will not solve the problem; to achieve the best recognition rate, one must also identify the suitable classifier. We classify the Review Rating Prediction Problem as a multi-class classification problem in Machine Learning, where the class labels are the star ratings. On the raw data, we used Data Filtration, Text Preprocessing, and Feature Extraction, as well as a variety of machine learning algorithms to predict user reviews based on star ratings. The following is a breakdown of the paper's structure. The related work of predicting rating based on user sentiments is described in Section II. The Yelp dataset is discussed in section III. The process flow is explained in section IV. In Section V, the results of the experiments are discussed. The future works and conclusion are discussed in Section VI.

2. RELATED WORK

At Micro blogging is an increasingly popular broadcasting platform among the Internet population these days. Individuals share their opinions and feelings about a wide range of topics on micro blogging websites every day, including products, media, organizations, and so on. In estimation systems, perception mining, and other applications, sentiment evaluation is important. Users on Facebook, one of the micro blogging websites, are limited to 140 characters. For data recovery purposes, Twitter often offers a programmer-friendly streaming API, allowing the researcher to check several users for real-time tweets. This far has been the case. Machine learning techniques have proved to be extremely successful, delivering accurate outcomes. Any method's efficacy is essentially determined by the order. The lexical technique is a ready-to-use technique that requires no advanced experience or training. Even though machine learning involves a well-designed classifier, it needs a large number of training data sets and regression testing before it can be deployed. On twitter datasets, sentimental analysis using class-two naive bayes has an accuracy of 84 percent.

In social networking sites, the way people share their views, assessments, or feelings about individuals or brands has changed drastically. Among other social media sites, one of the free social service websites is twitter, which allows users to share their day-to-day activities. As a positive or negative tweet emotion, the authors illustrate and provide machine learning system functionality. Data collection and storage of tweets using the Twitter API and the best of the approved IPL hash tags (# IPL 2016 and #IPL 9) that can be accomplished. The result was calculated using the relevant sample dataset and Random Forest techniques. As an output parameter for each tweet, a binary-class value of '0' and '1' reflecting both positive and negative views is used. The evaluation accuracy of the proposed technique is 81.69 percent. The method of analysing natural language to determine if a piece of text contains subjective data and the type of subjective information it expresses is known as sentiment analysis. The authors used big data to examine vast volumes of tweets and calculated the polarity of letters, sentences, and whole articles. Author uses linear regression to investigate the relationship between a scalar dependent variable Y and one or more independent variables denoted X. This method of data analytics is more efficient than using vector machines and naive bayes. The method's precision is 85.23 percent, which is a major improvement over SVM. The authors use 10-fold cross validation to increase the system's accuracy.

Kao and Huang [4] suggest a data mining method for evaluating the association between the emotion of fan page users and client buying acts to predict past client buying activity. Businesses create their own fan pages to advertise their products and broadcast advertisements. Subscribers to the fan page often post their opinions on the website, which helps to spread customer opinion. Because all these viewpoints will be disseminated across the social network to every nook and cranny, some business analysts wonder whether they will help or hinder the sale of goods. The first phase of the proposed method is to determine a single fan's sentiment rating based on his or her shared messages. The term "public sentiment" refers to a summary of all of the fans' emotions. To find the sequence like "if modern message of positive sentiment tends to increase, the quantity of products purchased will improve over the next two days," the inter-transaction association rule mining approach is then used to discover the connection between public sentiment and the quantity of purchases.

Ren and Wu [5] focus on the difficult task of estimating user sentiments on topics that they haven't directly addressed before, which the authors refer to as user topic opinion estimation. On a real-world Twitter data set, suggested techniques outperform new collaborative filtering methods, as evidenced by research findings. The author conducts tests to assess the proposed ScTcMF technique. Social context and topical context are helpful in improving the quality of user-topic sentiment analysis.

Consumer satisfaction polls, in which questions are explicitly designed to gather customer input on particular goods or services, have traditionally been used to test public opinion. Zhu at el. [7] and colleagues. Propose aspect-based opinion polling versus unlabelled free-form user conceptual feedback with no requests for answers to questions. Then a multi-aspect bootstrapping device should be purchased. A new methodology called aspect-based segmentation method suggests segmenting multi-aspect sentences into multiple single-aspect bases as general units for opinion polling. The proposed method of opinion polling focused on aspects achieves. In the experimental findings, there was 75.5 percent accuracy on real Chinese restaurant reviews.

On sites like Amazon, Netflix, and Yelp, there are several reviews and ratings. Scores are typically given on a scale of one to five stars. Reviews are a type of free-form text that consists of many words. Textual sentiment analysis has been described as a critical component of analytics research because it offers useful options for opinion mining. We may recommend new products, films, and restaurants to a person based on their feedback and ratings. Through observing similar customers and developing feedback, recommenders can fit consumer actions. Developing suggestions. From the point of view of a user's responsibility of the film and interests, a scheme with sentiment tags coupled with regular recommendations appears to be a creative and rational solution. Support Vector Machines, Naive Bayes, Unigram, Bigram, Trigram, and Random Forests are among the methods used by many. In comparison to traditional techniques, the researchers have used the Random Forest approach to significantly boost sentiment mining. Better data stemming and pre-processing processes led to a decrease in root-mean-squared-error (RMSE). The recommendation system's calculation time is greatly decreased when Spark is used.

Wang at el. [9] performed a sentiment analysis based on Yelp user feedback. Author uses user input to build predictive models for scores in each dataset. The author finds a 5- or 4-star Yelp rating to be positive and a 3, 2 or 1 rating to be negative. They used supervised learning algorithms such as perceptron learning algorithm, Naive Bayes, and SVM to predict sentiment as expressed in the Yelp rating system for sentiment prediction. The features from the Yelp user reviews material was extracted using several language systems. The key language model uses stop words and complex symbols to halt and delete them, and the perceptron learning algorithm typically offers the best predictive outcomes.

3. DATASET

Yelp [11] is the biggest online comment forum in the United States. It includes businesses in the restaurant, grocery, hotel, and hospitality sectors. Yelp allowed people to review and rate local businesses in the same way they would online for products. Yelp's website and mobile applications have a diverse range of suggestions. "Yelp's Best: City" is one of Yelp's most well-known categories. Yelp ranks businesses based on Yelp user feedback and ratings and choose the most popular places to visit in a town in this category. When it comes to discovering a company, most users rely solely on the average star rating. Although it offers a clear interpretation of a person's opinion of a company, it involves a lot of different types of data, such as features, customers, and the review text itself.

Customers can leave comments, check ins, ratings, and other information about any company they like or dislike on the Yelp website. People often "Yelp" the location to which they are about to fly or use Yelp's assistance to decide which businesses are doing well and which are not. Users have accounts, and on Yelp they can also make friends. Unlike many other appraisal and suggestion systems, Yelp builds its own social network with a graph of people connected to his or her peers, rather than relying on traditional social networks like LinkedIn or Facebook. Users may assign a business a star rating from 1 to 5, as well as write a text review to back up the number ranking. Those ratings are a great guide for users searching for local businesses, helping them to find out which one is right for them and making Yelp a reliable source of feedback.

```
Shape of the dataset:
(10000, 10)
Column names:
Index(['business_id', 'date', 'review_id',
'stars','text', 'type', 'user_id',
'cool', 'useful', 'funny'],
      dtype='object')
 Datatype of each column:
 business_id     object
 date            object
 review_id       object
 stars            int64
 text            object
 type            object
 user_id         object
 cool             int64
 useful           int64
 funny            int64
 dtype: object
```

Figure 1. Shape of the yelp dataset [11]

Every business has a general ranking, which is simply the sum of the star ratings for all the reviews the company has evaluated. Users can also vote in favour of reviews written by other users, with votes ranging from helpful to cool too funny. User information, company information, user-to-business access history, user scoring, and responses are all included in the Yelp dataset. Candidates were asked to use the dataset they had access to generate research ideas that could help Yelp in Yelp's own Dataset Contest, nicknamed "Yelp Dataset Challenge." Yelp has made a portion of the data accessible to the scientific and educational communities by organizing competitions, opening many possibilities to explore and use this data in different ways for various purposes. The Yelp dataset contains information on local businesses in 11 cities across four continents. Based on the statistics, there are 86872 companies and 686556 consumers. There are 2835065 ratings and 48732456 check-in data in the set of datasets. Shape of the academic business data set for Yelp is shown in the Fig. 1.

4. PROPOSED METHOD

Fig. 2 displays the implemented system's fundamental proposed flow diagram. The steps are discussed in detail here:

A. Load Dataset:
We'll start by loading the raw data. Every text analysis is divided into two parts: a positive and a negative section. To begin with raw text data and no other details, we combine them collectively.

B. Text Preprocessing:
In its raw form, any unstructured information is not well configured. Text pre-processing involves using a range of techniques to turn raw text with typical structure and notation into pre-defined patterns. To standardize text info, we used the following pre-processing strategies.

a) Tokenization: Divide the text into words.
b) Text Cleaning: Delete useless words containing digits.
c) Remove useless stop words like 'there', 'a' etc.
d) Convert the text into small letters.

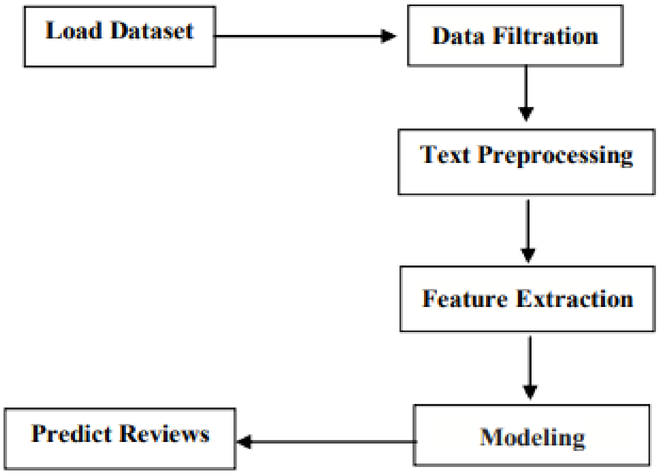

Figure 2. Proposed flow diagram

C. Data Filtration:
We will filter the dataset to only include 1- and 5-star ratings because their polarities are opposite, making it simpler to turn the issue into a binary classification.

D. Feature Extraction:
Since we can conclude that customer feedback is strongly associated with how they feel about the restaurant, we begin the process by designing features for sentiment analysis. The feature extraction technique facilitates in the translation of structured text into numerical or categorical functions. To learn, we use various supervised learning models on extracted features. This approach is also known as vectorization since each text is translated into a feature vector that can be fed into supervised classification systems. For feature extraction, we used the Bag of Words Model.
Bag-of-Words is a very analytical approach to this issue, with the following methods:

1) Implement a sequence of some sample to break the documents into tokens.
2) Each token should be assigned a weight based on how often it appears in the text and/or corpora.
3) Make a document-term matrix by placing a token in each column and a text in each row.

E. Modeling:
To train our model, we must first choose the features that will be fed into the system as input. Then we break our data into two categories:

1) First section is used to train our model.
2) Second section to determine its efficiency.

We used supervised machine learning approaches to model the method, including Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, K-Nearest Neighbor, and Multilayer Perceptron.

F. Predict Reviews:
The best score goes to the naive bayes classifier, which we use to predict a random positive review, a random mediocre review, and a random negative review.

5. RESULTS

On the raw data, we used Data Filtration, Text Preprocessing, and Feature Extraction, as well as a variety of machine learning algorithms to predict user feedback based on star ratings. Now, as seen in Fig. 3, we have printed some word clouds to see what kind of terms appear in our reviews. Birthday, Love, Concept, Place, Occasion, and other terms are all linked to the reviews. Some terms are more aligned with the customer's contact with the venue.

Figure 3. WordCloud from user review

From Table II shows different machine algorithm in terms of precision, recall, f -1 score, support, and accuracy of 5-star rating reviews has better result than 1-star rating reviews.

TABLE II COMPARISON OF VARIOUS MACHINE ALGORITHMS FOR 5 STAR RATING

Machine Learning Algorithm	5-Star Rating		
	Precision	*Recall*	*f-1 support*
Naive Bayes	0.95	0.97	0.96
Random Forest	0.88	0.98	0.93
Decision Tree	0.92	0.93	0.93
Support Vector Machine	0.95	0.94	0.95
K-Nearest Neighbors	0.83	0.99	0.81
Multilayer Perceptron	0.95	0.93	0.93

6. CONCLUSIONS

In this paper, we have identified an investigation into determining the polarisation of a user's textual analysis as positive or negative automatically. There is a strong need for such research because rankings and stars are becoming increasingly relevant in assisting prospective clients in making decisions or purchasing products. For our tests, we used the Yelp data collection. To predict the review, we used a machine learning algorithm. Proposed system achieved a recognition rate of 93 percent, which is higher than the state of the art system's recognition rate.

REFERENCES:

1. Thakkar, Harsh & Patel, Dhiren. (2015). Approaches for Sentiment Analysis on Twitter: A State-of-Art study, https://arxiv.org/abs/1512.01043
2. Arti, K. P. Dubey and S. Agrawal, "An Opinion Mining for Indian Premier League Using Machine Learning Techniques," 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, 2019, pp. 1-4.doi: 10.1109/IoT-SIU.2019.8777472
3. A. Razia Sulthana, A. K. Jaithunbi, and L. Sai Ramesh, "Sentiment analysis in twitter data using data analytic techniques for predictive modeling", J. Phys. Conf. Ser., vol. 1000, no. 1, 2018 pp. 1-4.
4. L. Kao and Y. Huang, "Predicting purchase intention according to fan page users' sentiment," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, 2017, pp. 831-835, doi: 10.1109/SMC.2017.8122712
5. F. Ren and Y. Wu, "Predicting User-Topic Opinions in Twitter with Social and Topical Context," in IEEE Transactions on Affective Computing, vol. 4, no. 4, pp. 412-424, Oct.-Dec. 2013, doi: 10.1109/T-AFFC.2013.22
6. R. Chakraborty, M. Bhavsar, S. K. Dandapat and J. Chandra, "Tweet Summarization of News Articles: An Objective Ordering-Based Perspective," in IEEE Transactions on Computational Social Systems, vol. 6, no. 4, pp. 761- 777, Aug. 2019, doi: 10.1109/TCSS.2019.2926144.

7. J. Zhu, H. Wang, M. Zhu, B. K. Tsou and M. Ma, "Aspect-Based Opinion Polling from Customer Reviews," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 37-49, Jan.-June 2011.doi: 10.1109/TAFFC.2011.2.

8. B. S. S. Govind, D. R. Tene and K. L. Saideep, "Novel Recommender Systems Using Personalized Sentiment Mining," 2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, 2018, pp. 1-5, doi: 10.1109/CONECCT.2018.8482394.

9. J. Wang, "Predicting Yelp Star Ratings Based on Text Analysis of User Reviews," 2015.

10. J. Chen et al., "Double layered recommendation algorithm based on fast density clustering: Case study on Yelp social networks dataset," 2017 International Workshop on Complex Systems and Networks (IWCSN), Doha, 2017, pp. 242-252,doi: 10.1109/IWCSN.2017.8276534.

11. Yelp dataset challenge, http://www.yelp.com/dataset_challenge