



Social Media Analytics on Swiggy using Sentiment analysis : The Good, The Bad & The omg..

Dr.S.Ajitha¹
Janani Prabu²
Sai saranesh B³
Abhinayah B⁴

Abstract

In recent times, social media is the main resource for collecting the information about an individual. All businesses are turning online, so the data generation rate is gradually increasing day by day. In this technical paper, we have applied the sentiment analysis on the reviews of Swiggy. Swiggy is an online food delivery start-up. It solves the crucial logistics problem in the local environment by having own delivery fleet. It is a high risk for them because the other food delivery start-ups like UberEats, food Panda rely on their partner restaurants for delivery. So if any problem occurs, the partner restaurants will accept the blame. So if sentiment analysis is made in Swiggy, it will be easier to elevate the quality of their service to the customers. Sentiment analysis is a tool used in social media analytics which identifies and categorizes the opinions of the customer based on the text. By that we can figure the attitude of the customer towards the product whether it is positive or negative or neutral.

Keywords: Sentiment analysis, Tokenizing, Mining, Random Forest Classifier.

1. Introduction

In the previous era, the data is not accurate as it was collected manually. Those types of data are called traditional data. With those data we can't analyze and decide a conclusion in any business. It was very hard at those times to handle a massive amount of data using the file system. But that situation has changed now. In recent days, everything is data. There is a enormous amount of data around us. We should analyze that data to find the hidden insights and improve the business requirement. Data is extracted from different resources and it is categorized to analyze the behavior patterns now. Data analytics is a method of collecting both quantitative and qualitative information about a product. Therefore it allows the marketers to make a informed decision about their customers. The main purpose is to find the patterns and allow marketers to find some hidden insights from the data. There are different types of data generated: Transactional data, Social data, scientific data, and Sensor data. Here the social data can be analyzed by using the social media .Social media platform is huge. To track the traffic in the social media marketing and to measure the engagement between the product and the customer, the social media analytics is done. Social media data is providing a large amount of information for government to analyze and predict the future events for some vital decisions. Social media analytics is tracking, measuring and analyzing the insights from the social media platform. Social media analytics is mainly used in the online sales and digital marketing sales. Sentiment analysis or opinion mining is a technique in the social media analytics that build systems that extract the words separately and assign the score according to the subject of the word. Opinions can be categorized into direct opinions and comparative opinion. Direct opinions give the opinion about the brand directly. Comparative

¹ Assistant Professor, ITM Business School, Chennai, ajitha22@gmail.com

² Student, ITM Business School, Chennai, janani2706@gmail.com

³ Student, ITM Business School, Chennai, bsaisaranesh@gmail.com

⁴ Student, ITM Business School, Chennai, abhinayah20@gmail.com

opinions give the opinion about the brand comparing with other brands. Sentimental analysis is implemented in the documental level, sentence level and the word level.

2. Literature Review

2.1 Opinion Mining and Sentiment action Classification

Social media provides a lavish data that reflects current trends on a collection of topics. Information visualization is vital in drawing out the options of interest in social media. A system can be provided which can apace adapt to user desires. The information in which the user is interested is featured prominently in the application (D.M., Bruce, J., Dowson, S., Love, O., and McGrath, L.). Parallels to know the level of data provided by twitter, The Australian government described an ongoing work to detect, assess, summarize, and report messages of interest for crisis coordination published by Twitter. The advanced medium and applicant tools, collectively describes the Emergency Situation Awareness-Automated web text mining system was used to display how relevant Twitter messages can be identified and used to inform the situation awareness of an emergency incident as it unfolds (Cameron, M. A., Power, R., Robinson, B. and Yin, J.). Sentiment Analysis is a Natural Language Processing task that identifies the user's opinions expressed in a positive, negative or neutral comments based on the text. This survey gives an analysis of the possible efficient techniques, recent advancements and the future research directions within the field of Opinion Mining and Sentiment action Classification (Sindhu, Chandrakala). Opinion summarising is process of generating summary from multiple reviews. Opinion mining is a topic in Text mining, Natural Language Processing, and Web mining discipline. Though user generated contents has proven helpfull in many applications, challenges still exist in process of opinion mining due to unstructured data and the data which provides many meaningless information on websites.

2.2 Text processing

Everyday creation of user-generated content in a large scale and because it involve user attitude there is need of opinion mining.(Vijay. B. Roth et al). There are various challenges in Sentiment analysis. The first is a opinion word that is represented to be positive in one situation may be shown up as negative in another situation. A second challenge is that people will not always express opinions in a same manner. Most traditional way of text processing relies that small differences between two pieces of text will not change the meaning very much. In Sentiment analysis, however, "the picture was great" is not the same from "the picture was not great". People can be contrary in their statements. Mostly the reviews will have both positive and negative comments. In more informal social mediums like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for the people to understand, but more difficult for a computer to define. Sometimes even people will have difficulty in understanding what someone thought based on a short piece of text because it lacks context. (G. Vinodhini).

3. Research Methodology

3.1 Method

The research method used here is described in the flow chart (Fig 3.1.1)

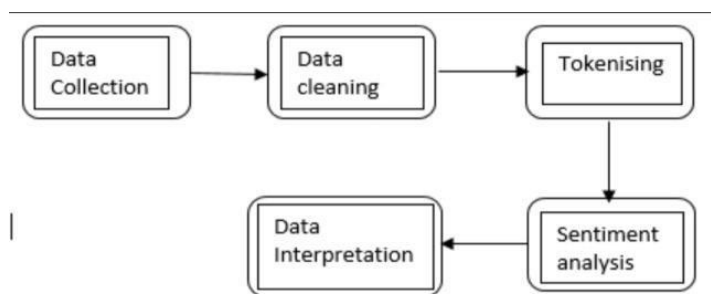


Fig 3.1.1 Steps followed in the sentiment analysis

The dataset is imported in the specific platform (Jupyter notebook) to perform the sentiment analysis. Then the stop words from each individual data should be removed. Then, each and every word from a individual data is divided into tokens. And the stemmer is performed. Stemmer is a process which converts all the words into a specific tense. And each word is compared to the database of positive and negative words. If it is a positive word, the score is assigned as 1. And if it is a negative word; the score is assigned to as -1. If a specific word is neither is in positive and negative database, then it is considered as a neutral word and assigned as zero.

3.2 Data Collection and Data Cleaning

The reviews are scraped from the Google Play store.

3.2.1 Tokenization

The special characters, the punctuations, the numbers and the short words are removed. Then the words are tokenized (Fig 3.1.2). Word tokenization is that the process of splitting an outsized sample of text into words. There is a requirement in the natural language processing tasks where each word has to be captured and subjected to further analysis like classifying and counting them for a specific sentiment etc. The Natural Language Tool kit (NLTK) is a library used to achieve this.

```
In [9]: ## Tokenization

token=rev_df['Reviews'].apply(lambda x: x.split())
token.head()

Out[9]: 0    [Inaccurate, delayed, Customer, care, service,...
1    [This, sucks, when, comes, apply, coupon, paym...
2    [days, missing, items, there, refunds, Custome...
3    [overall, quite, good, have, been, facing, iss...
4    [Horrible, experience, Showing, false, rates, ...
Name: Reviews, dtype: object
```

Fig 3.1.2 Tokenization process

3.2.2 Stemming

Now the stemming process is done. We come across situation where two or more words have a common root (Fig 3.1.3). For example, the three words - love, loving and loveable have the same root word love. A search involving any of these words should treat them as the same word which is the root word. This process is called stemming. We can visualize the most commonly used words by using the 'WordCloud'. Also the unique words can be listed by means of 'Bag of Words' feature.

```
In [13]: ## Stemming process

from nltk.stem.porter import *
stemmer=PorterStemmer()

#Apply stemmer to the tokens

token=token.apply(lambda x: [stemmer.stem(i) for i in x])
token.head()

Out[13]: 0    [inaccur, delay, custom, care, servic, functio...
1    [thi, suck, when, come, appli, coupon, payment...
2    [day, miss, item, there, refund, custom, execu...
3    [over, quit, good, have, been, face, issu, wit...
4    [horribl, experi, show, fal, rate, differ, tha...
Name: Reviews, dtype: object
```

Fig 3.1.3 Stemming process

3.3 Algorithm

Here the Random Forest Classification Algorithm is used. Random forest is a supervised statistical learning algorithm that has been proved having unsurpassable accuracy among classification techniques. It is simply a collection of many decision trees where the output of each individual tree is aggregated and averaged. The final ensemble model is a Random Forest classifier. A decision tree is a model with a tree-like structure where each node in the tree splits the input data points into two children nodes based on value of a particular feature. For a classification tree, each leaf node is usually a target class or a histogram. The histogram describes the probability of a data point belonging to each class, if the data point reaches this leaf. To train a decision tree model means to recursively construct a tree using training data. As for the forest, the aggregate method we adopt to combine all trees is simply an arithmetic average of all output. The key to understand why Random Forest performs so well is the randomness achieved by bootstrapping and bagging. The source of randomness lies in two steps of building an individual tree: the random bootstrapped sample and the randomly set features selected, as described by the following algorithm.

1. For $b = 1$ to B :
 - (a) Draw a bootstrap sample Z of size N from the training data.
 - (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the subsequent steps for every terminal node of the tree, until the minimum node size is reached:
 - i. Select m variables at random from the p variables.
 - ii. Pick the best variable/split-point among the m .
 - iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}$ B

where B is the number of trees. Bootstrapping helps to reduce the noise in data, and the random selection of features further reduces the risk of over-fit, since each tree only focus on a subset of features. Especially in our project, text analysis usually has large number of features and noisy data, Random Forest suits the most. The required packages are imported. After the Random Forest model is optimized, we randomly bootstrap 70% of our corpus as training data and use the rest 30% as test data to validate of our mode. After training the model using the Random Forest Classification Algorithm the accuracy of predicting the reviews whether it is positive or negative or neutral is calculated. (Fig 3.1.4)

```
In [16]: from sklearn.metrics import classification_report, accuracy_score

        ##Calculating the accuracy score

        report = classification_report(y_test, predictions)
        accuracy = accuracy_score(y_test, predictions)
        accuracy

Out[16]: 0.826272829
```

Fig 3.1.4 Accuracy score

4. Conclusion

The results from sentiment analysis are useful in social media monitoring as it permits the users to get a deep insight behind certain topics. The applications of sentiment analysis are widely used and strongly favorable. Changes in sentiment on social media have been shown to correlate with changes in the stock market. It also plays a major role in market research and customer service approach. Not only it predicts the customer's view about our own product, but also it will analyze what people think about the competitors too. The overall customer experience of the users can be collected quickly with sentiment analysis. Also it is used to maintain the brand reputation. Customer is the main source for any kind of business. To improve the customer experience and to understand the customer view, the sentiment analysis is performed.

5. References

1. Best, D.M., Bruce, J., Dowson, S., Love, O., and McGrath, L. 2012. "Web-Based Visual Analytics for Social Media," in AAAI ICWSM SocMedVis: Workshop on Social Media Visualization (AAAI Technical Report WS-12-03), Dublin, Ireland, pp. 2-5.
2. Cameron, M. A., Power, R., Robinson, B. and Yin, J. (2012). "Emergency Situation Awareness from Twitter for Crisis Management." Proceedings of the International Conference on World Wide Web.Lyon: France, 695-698.
3. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)
4. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proc. 8th Conf.on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1997)174–18
5. Miller, G.A.: Wordnet: A lexical database. Communications of the ACM 38(11) (1995) 39–41
6. Rudy Prabowo, Mike Thelwell, "Sentiment Analysis: A Combined Approach".
7. Nidhi Mishra et al, "Classification of Opinion Mining Techniques", International Journal of Computer Applications, Vol 56, No 13, Oct 2012Pg No 1-6.
8. Sindhu, Chandrakala, "A Survey on Opinion Mining and Sentiment Polarity Classification", International Journal of Emerging Technology and Advanced Engineering.Vol 3 Issue 1, Jan 2013.
9. Vijay. B. Roth et al, "Survey on Opinion Mining and Summarization of User Reviews on Web", International Journal of Computer Science and Information Technologies (IJCSIT),Vol 5(2), 2014. 1026-1030.
10. G. Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012.