Exploration of Opinion from Twitter Data

Nilesh Korde,Gaurav Kawade,Sunita Rawat,Kavita Kalambe,Abhijeet Thakare

*Abstract***:** *To share information nowadays, people use social media sites from all around the world. For example, Twitter is a social media site enable users with facilities like reading, sending post recognized as 'tweets' and interrelate with diverse peoples. People often post their sentiments about their day-to-day lives, the whole thing for example places and brands. Businesses makes profit from this vast social media site by gathering data related to sentiments of people. Presenting a model that can accomplish opinion analysis on collected data from Twitter is the aim of this paper. To analyze highly unstructured and unorganized data in Twitter makes it difficult to manage and use. In our proposed model we are combining the work of unsupervised and supervised algorithms. Extracting and classifying each tweet depending on its opinion considered to be a neutral, positive or negative. Zomato and Swiggy are the two subjects about which data were collected to show which online food delivery business has more popularity. We have used diverse machine learning algorithms for testing. Testing metrics like f-score and cross validation were used for testing the result from these models. Our model has shown performance which is considered to be robust on directly mining Twitter texts.*

Keywords: Opinion analysis, Machine learning, Twitter, social media

I. INTRODUCTION

Instagram, Facebook and Twitter are the online social media which allow users to interconnect with the whole world. Online social media allow users share their views and sentiments about any particular product, company and even political issues. For example, Twitter, almost every single giant company has a Twitter account to know about their services or products and their customer's feedback. Opinion analysis, known for mining sentiments form text data. [5].
This paper, focuses on Opinion scrutiny for categorizing English word on two online food delivery business, Zomato and Swiggy. our research was determining which among to two has more popularity.

Nilesh Korde*, Assistant Professor, Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management Nagpur (An Autonomous Institute) , Nagpur, India.
Gaurav Kawade, Assistant Professor, Shri Ramdeobaba College of Engineering and Management, Nagpur, India.
Sunita Rawat, Assistant professor, Computer Science and Engineering Department, Nagpur, India.
Kavita Kalambe, Assistant Professor, Department of CSE Shri Ramdeobaba College Of Engineering And Management - RCOEM, Nagpur, india
Abhijeet Ramesh Thakare, Ph.D. from Visvesvarya National Institute of Technology, Nagpur, India

II. RELATED WORK

In [1] for performing opinion mining on English language written tweets, belonging to different telecommunication companies of Saudi Arabia.
Overall an emphasis on analyzing Term frequency (TF) which is inverse document frequency (IDF) to quantity how words are important to an exact tweet. In [2] focus is on analysis of sentiments embedded in public Facebook and tweets comments. They used algorithms such as Naïve Bayes, binary model (BM), Support Vector Machine
 (SVM) and TF-IDF which are supervised machine learning to understand the accuracy of sentiment analysis based on several weighting functions. In [3] researchers applied opinion mining on twitter data of Arabic dialect using natural language scrutiny for text in Arabic language. Researchers in [4] give an emphasis on analysis of opinion on corpus such as review data of products and tweets which is in Egyptian dialect.

III. METHODOLOGY

Our goal is to make sure the popularity of Zomato or Swiggy among people in terms of how positive/negative/neutral reviews are. For any business evaluating people's feelings and their views about a product could always be a valuable thing. API of Python is used for research. Python used for artificial intelligence, data analysis, machine learning algorithms and scientific computing. With the aim of extracting tweets, we have used Twitter API for creating application for Twitter and for getting authorization. In PyQt which is a GUI and environment for Python.Some of the packages are (Tweepy, Numpy, Scipy). By using tweepy package you can extract up to 3200 tweets only (see Table 1).

TABLE 1. NUMBER OF EXTRACTED TWEETS

Zomato	Swiggy
3000	3000

Formerly preparing, filtering and cleaning the tweeter data, we need to discover the data and get insight into it. The occurrence of tweets during the day for both Zomato and Swiggy is shown in Figure I and Figure II.
From Figure I we come to know that people send post about Zomato when they have free time during the day. people send the posts between 9AM to 8 PM. The interesting fact about it is that it varies from day to day.
Similarly, from Figure II we come to know that people send post about Swiggy when they have free time during the day, also we come to know people send the posts between 10AM to 7 PM. The interesting fact about it is that it varies from day to day.

Figure I. Zomato tweets frequency.

Figure II. Swiggy tweets frequency

Moreover, for getting further means of sense about the words we developed word cloud through package "wordcloud" in Python language, creating word cloud from post text provides better logic (see Figure. III). (a)

Figure. III(a). Text word cloud (Zomato)

Figure. III(b). Text word cloud (Swiggy)

A. Preprocessing

Before analyzing text, it should be cleaned and prepared first. Many tasks are involved in preprocessing text as it depends on analysis requirement. We are looking for text only as our research work is based on it [6- 7].

We converted text to data frame from extracted tweets, detached text URLs,needless spaces, removed numbers, removed punctuations also renovating encrypting Emojis to ASCII.

The analysis as well as mining are performed on text subsequently removing and cleaning the text (see Table 2).

TABLE 2. TWEET BEFORE CLEANING AND AFTER CLEANING

Tweet before Cleaning
"@WSJIndia #1934897 At #WSJDLive, @Zomato's co-founder Gaurav Gupta says he thinks food-delivery services could be more profitable in India than in China because the costs are lower\� "
Tweet after Cleaning
" co-founder Gaurav Gupta says he thinks food-delivery services could be more profitable in India than in China because the costs are lower "

B. Model Building

At this stage, after removing unnecessary symbols from each tweet, it was labelled as 0,1, -1. Tweets were classified using a lexicon-based model. By means of two text files comprising a list of negative and positive words, together with more words associated to our domain. Words present in the tweets are compared on the basis of negative or positive words used so that it can categorized as positive or negative tweet. As a result, the findings are indicated in Table 3 from this model.

TABLE 3. CLASSIFICATION BASED ON LEXICON

Topic	# of tweets	Positive	Negative	Neutral
Zomato	3000	2113	511	376
Swiggy	3000	1901	507	592

Next, for the purpose of training multiple supervised learning algorithms applied which are as follows

- Decision tree: It is flexible algorithm allocating label grounded on the uppermost score.
- Random forest: It is utilized for developing numerous choice tree and it belongs to supervised algorithm.
- Support vector machine: It is used for regression analysis and classification. It is defined as supervised model.
- Naïve Bayes: It is a classifier accustomed to decide the most plausible class name for each object.
- Bagging: It is a classifier which uses every single sample independently to build a forecast model enrapturing different irregular samples.
- Maximum entropy: large variety of text classification is classified by this algorithm.

IV. DISCUSSIONS AND RESULTS

In this paper, to prepare and test the models we removed information straightforwardly from the Twitter Programming interface. To discover the feeling of each tweet a dictionary-based classifier utilized a physically made vocabulary. Utilizing both supervised and unsupervised demonstrating in our proposed system is a novel methodology in our undertaking. Thus, the forecast indicated upgrades in contrast with present work any place a piece of label information is available.

A. Measurements

- Recall: is demarcated as number of true positives (TP) divided by the number of true positives (TP) plus the number of false negatives (FN) as indicated in (1).

$$TP| (TP+FN) \qquad (1)$$

- Precision: is demarcated as the number of true positives (TP) divided by the number of true positives (TP) plus the number of false positives (FN) as indicated in (2).

$$TP| (TP+FN) \qquad (2)$$

- Fscore: is a degree of how accurate a model is by using recall and precision following the formula in (3):

$$FScore = 2 * ((Precision * Recall) / (Precision + Recall)) \qquad (3)$$

B. Cross validation

For cross validation, the unique training data set is split into four groups, four-fold cross validation for training and testing.

Therefore, we come to know how accurate the model's predictions are when equating the actual labels of the data points and the model's predictions on the validation set. The prediction accuracy after applying validation techniques on the models, is found as indicated in Table 4 and 5.

TABLE 4. ACCURACY RESULT (ZOMATO)

Zomato				
Algorithm	Accuracy			
	Precision	Recall	Fscore	Cross Validate
Decision Tree	75%	49%	60%	49%
Random Forest	38%	16%	23%	53%
SVM	55%	38%	45%	51%
Naïve Bayes	58%	53%	56%	47%
Bagging	55%	38%	45%	48%
Maxtent	55%	27%	37%	69%

TABLE 5. ACCURACY RESULT (SWIGGY)

Swiggy				
Algorithm	Accuracy			
	Precision	Recall	Fscore	Cross Validate
Decision Tree	61%	63%	62%	59%
Random Forest	67%	84%	75%	52%
SVM	69%	66%	67%	61%
Naïve Bayes	43%	39%	41%	47%
Bagging	71%	79%	75%	65%
Maxtent	61%	75%	67%	79%

Furthermore, the model which has given best result for both Zomato and Swiggy data as compared to other models which is Maxent.

Besides, there was minor difference between the number of negative and positive for both Zomato and Swiggy (see Table III). Still, more people liking and disliking Zomato as compared to Swiggy.

V. CONCLUSION

Opinion scrutiny is an area of interest for studying sentiments conveyed in manuscript in numerous societal media sites. Several algorithms were used by our proposed model to improve the accurateness. In our proposed model, there is a combination of unsupervised machine learning algorithm along with lexicon-based algorithm on previously labeled data. Basically its existence was not there first. Afterward data were fed into numerous supervised models.

Several metrics used for testing it have revealed that maximum entropy has the maximum accurateness which is based on cross validation. Consequently, Zomato is popular than Swiggy. Detecting rumors on Twitter similar procedure can be used in several area. Algorithms which are automatically classifying tweets will be a fascinating area of research in future.

Retrieval Number: B3854078219/2019©BEIESP
DOI:10.35940/ijrte.B3854.118419
Journal Website: www.ijrte.org

11

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. Ahmed, S. Qamar, A. and Alsuhibany, S. (2017). (IJACSA) International Journal of Advanced Computer Science and Applications,[online]8.Availablehttps://thesai.org/Downloads/Volume8No1/Paper_50-Sentiment_Classification_of_Twitter_Data_Belonging.pdf [Accessed 1 Feb. 2018].
2. I.Qarqaz and R. M. Duwairi , "A framework for Arabic sentiment analysis using supervised classification" , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016.
3. Sherif M. Abdou, Hossam S. Ibrahim, Mervat Gheith, "Sentiment Analysis For Modern Standard Arabic And Colloquial", International Journal on Natural Language Computing (IJNLC), Vol. 4, No.2, pp. 95-109, April 2015.
4. Emam, A. Assiri, A. and Al-Dossari, H.(2016). Saudi Twitter Corpus for Sentiment Analysis. International Journal of Computer and Information Engineering, [online]Availableat:http://waset.org/publications/10003483/saudi-twitter-corpus-for- sentiment-analysis [Accessed 1 Mar. 2018].
5. Levy, M. (2016). Playing with Twitter Data. [Blog] R-bloggers.Availableat:https://www.r-bloggers.com/playing-with-twitter-data/ [Accessed 7 Feb. 2018].
6. https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
7. https://stackoverflow.com/questions/37860411/extracting-tweet-from-twitter-api-using-python

AUTHORS PROFILE

Nilesh Korde has received Master of Technology degree in Information Technology from YCCE an Autonomous Institute at Nagpur University India in 2014. Presently he is working as an Assistant Professor in Computer Science and Engineering Department at Shri Ramdeobaba College of Engineering and Management Nagpur (An Autonomous Institute). He has published research papers on Parallel Computing, IOT, Cache Management and Clustering. Recently he is working on Natural language Processing, Data Retrieval and Machine Learning.

Gaurav Kawade has received Master of Technology degree in computer Science and Engineering from Visvesvaraya National Institute of Technologyisv.Nagpur in 2015. Presently he is working as an Assistant Professor at Shri Ramdeobaba College of Engineering and Management, Nagpur. His research areas are Wireless Sensor Network, Network Security, parallel computing. He has published research papers in National and Internal Journals and has 2 conferences proceeding in Scopus Indexed journal.

Sunita Rawat presently working as a Assistant professor in Computer Science and Engineering department and pursuing my PhD in the field of NLP. I have some publications in National and International Conferences. and have 2 publications in Scopus Indexed journal. I have ISTE membership.

Kavita Kalambe currently working as Assistant Professor in the Department of CSE at SRCOEM. She received B. E. degree from KDK college of engineering, Nagpur in 2010 and M. Tech degree in CSE specialization from G. H. R. I. E. T. W. Nagpur, Nagpur University in 2015. Her current research interests include Algorithm designing, High Performance Computing Architecture, Wireless Sensor Network, Microprocessor and Microsystems.

Abhijeet Ramesh Thakare has received his Ph.D. from Visvesvarya National Institute of Technology, Nagpur, India in Computer Science and Engineering and his M.E. from Walchand College of Engineering, Sangli, India. He is currently working as Assistant Professor in the Department of Computer Application at Shri Ramdeobaba College of Engineering and Management, Nagpur, India. He has total of 16 years of experience, R&D experience of 2 years in Dassault Systems, France, 4 years of experience in various IT companies and 10 years of Academic experience.