Sentiment Classification on Twitter and Zomato Dataset Using Supervised Learning Algorithms

Rajkumar S. Jagdale
Department of CS and IT,
Dr. B. A. M., University, Aurangabad
Aurangabad, Maharashtra, India
rajkumarjagdale@gmail.com

Sonal S. Deshmukh
Department of MCA,
JNEC, MGM University,
Aurangabad, Maharashtra, India
sonaldeshmukh@jnec.ac.in

***Abstract*— Natural Language Processing (NLP) is a special type of Machine Learning that takes care of unstructured data from the real world. Sentiment Analysis is the 'computational' method that decides whether the given sentence or paragraph is positive, negative or neutral. It derives a people's feelings, opinion and attitude etc. Using sentiment analysis tools in Twitter data to evaluate views and attitude which may help companies understand how people think about various events or products. In this paper, we have collected tweets dataset from twitter of hashtag COVID-19 and CORONA VIRUS. For Twitter dataset, we performed sentiment analysis and on Zomato dataset, Supervised Machine Learning Classification algorithms. Random Forest has got highest accuracy as 94.90 %.**

Keywords—Machine Learning, Natural Language Processing, SVM, NB, Sentiment Analysis

I. INTRODUCTION

Sentiment analysis is the automated process by which the subjective information underlying a text is identified and extracted. This can be an opinion, a judgment or a feeling about a specific topic or subject. The most common form of sentiment analysis is called 'polarity detection' and is graded as 'positive,' 'negative,' or 'neutral'. For example, let's take this sentence: "The situation is not good in COVID-19". A sentiment analysis model would automatically tag this as Negative. Sentiment analysis is a sub-field of Natural Language Processing, with its many exciting business applications, it has been gaining a lot of attention in recent years. Twitter is a best social media platform where people can expresses their opinion on different campaign, events etc. Analysis of feelings is especially useful for tracking social media as it goes beyond the amount of likes or retweets, by offering qualitative insights. About 80 % of digital data in the world are unstructured, and a significant proportion of this includes social media data. Sentiment analysis systems automatically organize unstructured text data using machine learning and natural language processing. Sentiment analysis algorithms can learn from data samples in order to detect the Tweets polarity in real time. All you need to do is train sentiment analysis tools to understand feeling in tweets, and they're going to do the rest. In this COVID-19 situation more People have reacted on social media like Twitter, Facebook etc. which results huge amount of data has generated.

This data need to analyze to find out some insights. Sentiment Analysis is technique where we can calculate polarity of tweets and visualize people's reaction on COVID-19. In this work, we have extracted tweets from Twitter associated to COVID-19 and performed different sentiment analysis techniques. After calculating sentiment score, Machine learning algorithms have been applied to calculate accuracy of each classifier and completed comparative analysis.

II. LITERATURE SURVEY

Sentiment Analysis is a concept you must have learned if you've been long enough in the Tech industry. It is the method of determining whether a piece of information (i.e., most usually text) suggests a positive, negative, or neutral feeling about the subject matter. There are three different levels at which sentiment analysis can be carried out depending on the level of complexity required.

A. Document Level

This is the easiest form of sentiment classification in which whole document of opinionated text is considered as basic unit of information. It is supposed that document is having opinion about single object only (film, book or hotel). This approach is not appropriate if document contains opinions about different objects. Full document is considered for classification and decides whether that document is positive or negative. Inappropriate sentences need to be excluded before processing. A lot of work [1],[2],[3] has been done on document based sentiment analysis. There are two approaches to do classification I.e. Supervised and Unsupervised machine learning approach.

Training and text dataset is available for supervised machine learning approach with finite classes for classification. It classifies the documents by using one of the common classification algorithms such as Naïve Bayes, K Nearest Neighbours, Maximum Entropy and Support Vector Machine, etc. Document-based classification for news comments was achieved by [2] using various supervised approaches to

978-1-7281-5970-6/20/$31.00 ©2020 IEEE

machine learning for effective results by combining different approaches. The classification of the Naive Bayes and Neural Network is combined for classification of film reviews by[4]. They also proven that by integrating these two approaches, the precision of the sentiment analysis is improved to 80.65 percent.

Specific investigations conducted by [1], [3], [5] examine data using an unsupervised approach to machine learning. Sentiment Orientation (SO) of opinion words in document is calculated in unsupervised approach. If the SO of these terms is positive then the otherwise negative document is marked as positive. Two words, Bad and Excellent were used in the most influential research performed by [1]. The semantic orientation defines whether the sense of opinion is similar to the positive word "Excellent" or "bad" negative. To calculate the semantic orientation Point Wise Mutual information method is used. The lexicon-based approach was used to characterize sentiments by [6]. The unsupervised dictionary-based technique (WordNet) is used by [3] to assess the document-level polarity of the film reviews. Seed list in this paper includes the terms of opinion along with their polarity.

B. Sentence Level

In this, for each sentence, polarity is determined as each sentence is regarded as a separate entity, and each sentence may have different opinions.

A statement may be either subjective or factual. Objective Phrase includes the truth. It will therefore take no part in determining the review's polarity and should be filtered out. The benefit of the sentence level review lies in the classification of subjectivity / objectivity [7]. Under supervised machine-learning methodology, several different approaches are explored and compared [8]. Sentence can be [9], [10] classified as positive, negative or neutral depending upon the opinion words present in the sentence. It mainly focuses on finding how to classify the text effectively.

C. Feature Level

Feature level sentiment analysis is capable of providing more fine-grained SA on some opinion targets and has a broader variety of E-business applications. This study proposes an approach for feature-level SA, based on comparative subject corpora. They [11] obtained a corpus from Twitter, which has been manually labelled at aspect level as positive, negative, or neutral. It achieved best results through the N-gram around method with a precision of 81.93%, a recall of 81.13%, and □- measure of 81.24%

These are some of the main advantages of Twitter sentiment analysis:

1. *Real-Time Analysis:* Analysis of Twitter sentiment is important to track rapid changes in consumer moods, to identify whether concerns are on the increase and to take action before issues escalate.
2. *Business:* In the field of marketing, marketers use it to improve their strategies, to understand the feelings of customers towards goods or brands, how people react to their ads or product launches and why consumers don't buy any Items. Items.
3. *Politics:* This is used in the political field to keep track of political opinions, to identify ambiguity and inconsistency between claims and behavior at government level. This can also be used for forecasting election outcomes.
4. *Scalability:* It would take hours of manual processing, and as your data grows it would be impossible to scale in case of huge amount of data. We can convert this manual task to automated and gain valuable insights in a very short time.
5. *Consistent Criteria:* Two members of the same team will interpret the same tweet differently. You can use one set of parameters to evaluate all of your data by training a machine learning model to perform sentiment analysis on Twitter data, so that results are consistent.
6. *Public Actions:* Analysis of opinion is often used to track and interpret social trends, to identify potentially harmful circumstances and to assess the blogosphere's general mood.

III. PROPOSED METHODOLOGY

Following flowchart shows the proposed methodology for this research work.

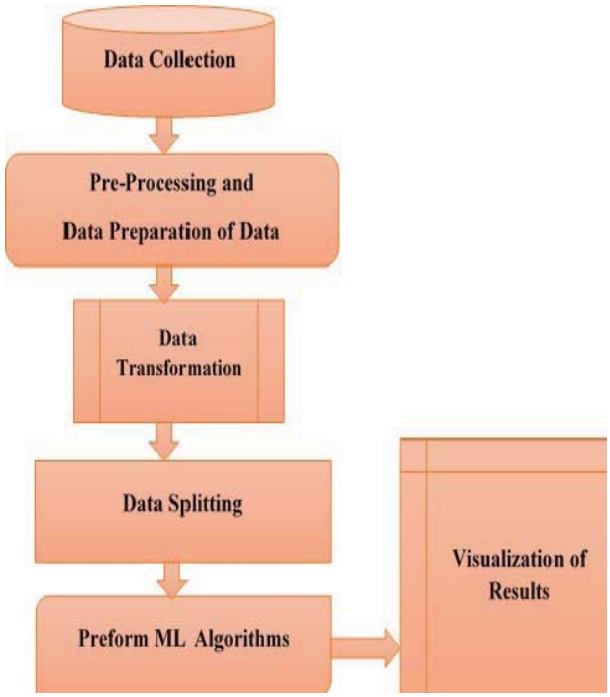

Fig 1. Proposed methodology

A. Data Collection:

In this paper, Two dataset has been used for sentiment classification.

There are different data sources where we get different types of datasets like Text Data, Video Data, Audio data and many more. In this paper, we have been downloaded dataset from kaggle [12] which is publically available. This Zomato dataset has 51717 rows (records) and 17 Columns (Attributes).

Following Table I. Shows the different attributes in Zomato Dataset:

TABLE I. LIST OF ATTRIBUTES OF ZOMATO RATING DATASET

url	address	name	online_order	book_table
rate	votes	phone	location	rest_type
dish_liked	cuisines	approx_cost (for two people)	reviews_list	menu_item
listed_in (type)	listed_in (city)			

For the Twitter Dataset, We extracted total 133226 tweets related to two hashtags i.e. COVID-19 and CORONA VIRUS.

B. Data Pre-processing and Preparation

In this step, we performed different techniques for pre-processing the dataset and made clean for further analysis.

In the data formatting we performed operations like hanging the Columns Names to proper names, Remove the NaN values from the dataset etc. we formatted data properly and did further process of cleaning.

Data cleaning step is very important in Data analytics. It helps to delete unwanted data and made useful data only for further analysis. We have deleted unnecessary Columns like 'url','dish_liked' and 'phone' etc which are not important Classification. Removed the Duplicates records which increases redundancy of the data which affects on EDA process. We also removed '/5' from Rates and replace proper rating and converted into Numerical form and also converted cost object type into numerical type.

C. Data Transformation

Data transformation is important in feature extraction. Features should be in numerical form in Machine Learning techniques. Scaling helps to convert all numerical numbers into one scale.

- *Scaling:*

In this step, categorical data has been converted into numerical. Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. In this all attributes like online_order, book_table, rate, votes, location, rest_type, cuisines, cost, menu_item are encoded using LabelEcoder.

- *Feature Extraction:*

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. In this paper, we have selected features like online_order, book_table, rate, votes, location, rest_type, cuisines, cost and menu_item as input features and rate is target feature. In the Twitter dataset, positive and negative words have been extracted for further analysis.

D. Data Splitting

Splitting the dataset into Train set and Test set in very important for train the model and calculate accuracy of model. Training data is used to train an algorithm, typically making up a certain percentage of an overall dataset along with a testing set. Test data is used to see how well the machine can predict new answers based on its training. In this paper, from Zomato dataset, while applying Machine Learning Classifier 80 % data is used for training and 20 % data is used for testing purpose.

E. Supervised Machine Learning Algorithms

In this Learning, data is well labelled and need to train machine which helps to predict the target variables for unknown feature. Two types of supervised machine learning techniques are regression and classification.

- *Logistic Regression*

If the dependent variable is dichotomous, logistic regression is the correct regression analysis to perform. Logistic regression is used to characterize data and to illustrate the relationship between one dependent binary variable and one or more ordinal type independent variables.

- *Decision Tree*

Decision tree builds a tree structure in the context of classification or regression models. This breaks down a collection of data into smaller and smaller subsets while at the same time incrementally creating an related decision tree. There are two or more branches of a decision node. Leaf node reflects a ranking or judgment.

- *K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) is one of the simplest algorithms used for regression and classification problems in Machine Learning. This takes the data and classifies new data points based on measures of similarity (e.g., distance function). Classification to its neighbour's is achieved by majority vote.

- *Random Forest*

The random forest is a classification algorithm that consists of several trees for decisions. While constructing each individual tree, it uses bagging and features variability to try to construct an uncorrelated forest of trees whose prediction by committee is more reliable than that of any individual tree.

- *Support Vector Machine*

A Support Vector Machine (SVM) is a supervised model of machine learning which uses classification algorithms for classification problems of two groups After giving an SVM model sets of labelled training data for each category, they are able to categorize new text. So you're working on a text classification problem.

- *Gradient Boosting*

Gradient boosting is a kind of boost to machine learning. It relies on the assumption that when paired with previous models, the best possible next model minimizes the total error in prediction. The main concept for this next model is to set target outcomes to minimize the error.

- *Naïve Bayes*

It is a Bayes Theorem-based classification technique with an assumption of independence among predictors. The classifier Naive Bayes believes that the inclusion of a specific feature in a class is irrelevant to any other function.

IV. Results and Discussion

After pre-processing on Twitter dataset sentiment score is calculated and divided tweets into positive, negative and neutral tweets. Following Table II. Shows distribution of tweets.

TABLE II. SENTIMENT ANALYSIS ON TWITTER DATASET

Sr. No	Polarity	Tweet Count
1	Positive Tweets	88797
2	Negative Tweets	22411
2	Neutral Tweets	21989
	Total	133206

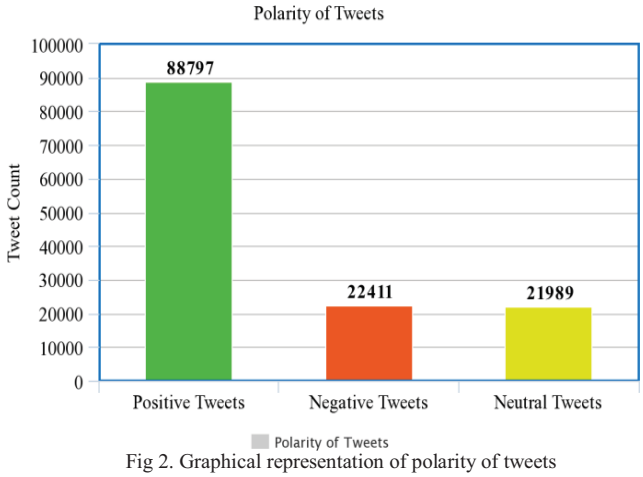

Fig 2. Graphical representation of polarity of tweets

After analysis of Twitter dataset, more people react on Twitter positively about Corona virus. It may be because people might have been helping to increasing confidence of the people and make people positive.

After implementing different Machine Learning classifier on both datasets, we achieved following classifier accuracy.

TABLE III. CLASSIFIER ACCURACY FOR ZOMATO RATING DATASET

Sr. No.	ML Classifier Name	Accuracy (%)
1	Logistic Regression	59.29
2	Decision Tree	93.67
3	KNN	89.81
4	Random Forest	94.90
5	Support Vector Machine	56.05
6	Gradient Boosting	69.18
7	Naive Bayes	45

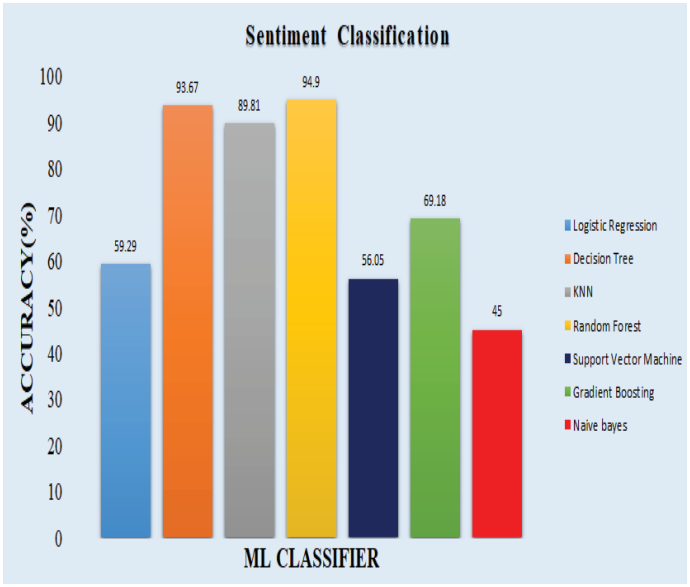

Fig. 3 Graphical representation of accuracy of classfier

V. Conclussions and Future Scope

We can conclude that we got highest accuracy for Random Forest classifier i.e. 94.90 % and lowest accuracy for Naïve bayes 45 %. In future, we can increase in dataset size and also apply Deep learning algorithms. Web based application can be made on Restaurants Rating system online which is useful for common people to find out proper and popular Restaurants in any location.

REFERENCES

[1] Turney, Peter D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." arXiv preprint cs/0212032, 2002.

[2] Yan Zhao, Suyu Dong and Leixiao Li, "Sentiment Analysis on News Comments Based on Supervised Learning Method", International Journal of Multimedia and Ubiquitous Engineering, Vol.9, No.7, 2014, pp.333-346.

[3] R. Sharma, S. Nigam and R. Jain, "Opinion Mining Of Movie Reviews At Document Level", International Journal on Information Theory (IJIT), Vol.3, No.3, 2014, pp.13-21.

[4] L. L. Dhande and G. K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", International Journal of Emerging Trends & Technology Computer Science (IJETTCS), Volume 3, Issue 4, 2014, pp.313-320.

[5] G.Kumar, P. K. Goel, S.K. Chauhan, A.K. Pandey, "Opinion mining and summarization for customer reviews, International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.08, 2012, pp.3688-3692.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll , and M. Stede, "Lexicon-based methods for sentiment analysis, Association for Computational Linguistics, vol. 37, No. 2, 2011, pp. 267-307.

[7] S. ChandraKala and C. Sindhu, "Opinion Mining And Sentiment Classification: A Survey", ICTACT Journal on Soft Computing, Vol- 03, ISSUE: 01, 2012, pp.420-425.

[8] S. Padmaja and Prof. S Sameen Fatima, "Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC), Vol.4, No.1, 2013, pp. 21- 23.

[9] V. S. Jagtap, K. Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology, Volume 2 Issue 3, 2013, pp.164-170.

[10] Raisa Varghese, Jayasree M, "A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology (IJRET), Vol. 02 Issue 11, 2013, pp. 312-317.

[11] Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., & Valencia- Garcia, R.,"Sentiment analysis on tweets about diabetes: an aspect-level approach", Computational and mathematical methods in medicine, Vol. 2017.

[12] Poddar, H., 2020. Zomato Bangalore Restaurants. [Online] Kaggle.com. Available at: https://www.kaggle.com/himanshupoddar/zomato-bangalore- restaurants [Accessed 25 July 2020.