



# Sentiment Analysis Of Movie Review On Twitter Data Using Machine Learning

Sapna Manjhi , M.A. Rizvi

<sup>1</sup> M.tech student, <sup>2</sup> professor

<sup>1</sup> computer technology and science

National Institute of Technical Teachers Training and Research, Bhopal

Abstract :Under the current innovations, the whole world is rapidly changing. Through the Web being used in each and every industry, Internet has become a fundamental need for every person. People are increasingly turning to social networking sites to air their views on current events and other topics of interest, thanks to the fast growth of these apps. Analysis of people's responses to a product, a public service and so on is crucial. A common conversation preparation activity is opinion mining or sentiment analysis, its objective is to determine sentiment behind viewpoints expressed in diverse texts. Sentiment analysis is essentially the same as analysing people's emotions. Sentiment analysis researchers have lately been looking into how people feel about anything from movies to commercial products to daily societal issues. Customers may post their opinions on Twitter, a popular microblog. When it comes to "tweets" (comments), opinion research on Twitter has been an increasingly popular area over the last decade. In this paper, we focus on analysis of public reviews and data on twitter's tweets on #Squidgame by machine learning model. Performance measurement and evaluation is done by using confusion matrix and by comparing with existing algorithm.

Keywords- Sentiment Analysis, Machine Learning, Opinion Mining.

## 1. INTRODUCTION

Text mining, computational linguistics (CL), and biometric measures may all be used to analyse emotional states and subjective information via the usage of "Natural language processing" (NLP) and sentiment analysis (which is identified as opinion mining or "emotion Artificial Intelligence" as well). For most part, voice of the customer is taken into account while doing a sentiment analysis, which may be accomplished via the use of online surveys, reviews, and social networking sites. "Sentiment analysis" (SA) handles the data to identify and extract subjective information from the input data and polarity detection deals with the automatic detection of the polarity of the opinions and classifies them based on their contrariety.

Sometimes, a sentence itself contains multiple opinions that lead to aspect-level SA. In general, SA has two steps,

- a) Document processing: convert the documents into text and pre-process them using linguistic tools. The SA uses linguistic resources called lexicons.
- b) Document analysis: process the pre-processed documents and annotate them with sentiment labels. The results present to the user using variety of presentation and visualization tools.

Internet services, E-commerce, online transactions and social media are increasing digital data day by day with increase in mobile and smart devices. All these different communications generate huge unstructured data, which is a challenge for data analyst and scientists. Data mining is a powerful way for data driven administrations. Various companies frequently prefer to monitor social Medias like Facebook, Snapchat, Twitter, Instagram and WhatsApp to analyses data by applying statistical, predictive methods and fact-finding techniques. Decision makers and followers both seem influenced by social media for open communications. Lots of data in various forms, with number of dimensions (variables) is available for analysis through these interactive social media communications.

"Positive," "negative," or "impartial" are the three types of emotional polarity (neutral). As a reminder, the three stages of sentiment mining may be summarised as follows: [1]:

- A document may be classed as "positive, negative, or neutral" in this stage of sentiment categorization
- Every statement is rated as "positive,negative," or impartial at sentence-level sentiment categorization level.
- A characteristic and feature-level sentiment classification may be used to classify sentences and documents as "positive,negative, or non-partisan" based on few features of the phrases and archives, and this is widely known as "perspective-level evaluation grouping."

## 2.1 Issue and Challenges

- Data categorization :
  - i. Semi-structured-Many sources of Social Media Data
  - ii. Unstructured- Tweets, Comments, Video data, audio data, animated data
- Major Challenge for researcher:
  - i. Extracting and analyzing real time or online communication data.
  - ii. Non Textual contents graphics, Audio, video, images etc.

Imprecision, Incompleteness, variety, privacy and Security, Scale and Timeliness are confronted while managing social media data.

### 3. Text Processing

Authors have investigated Object Oriented analysis in NLP Tools to tackle NLP using OOA [2]. Class identification and elaboration using Automatic OOA in NLP System is implemented in authors approach. By establishing association relationship, static and dynamic behaviours of external and internal entities extracted. Normalization is applied to identify attributes of NLP to study dependency. In this paper, authors have projected a rational midway elucidation concerning progress of discussion process that will result in ampler scheme model in field.

Application of right algorithm at right case of data is more important to get more accurate results in sentiment analysis which also helped to track improved research in design and development of novel concepts and framework. Authors [3] have introduced major steps as:

- Text Extraction  
Data is extracted from different sources to particular from. Original data may be of variety of form.
- Text Refinement  
Extracted data may contain untuned as well as some missing and unstructured data. This data is refined to tune up with main content and context of data.
- Text Classification  
Different categories and classifications are possible but in this classification data is classified into two categories positive data and negative data.
- Score Aggregation  
Sentiment classification is generally based on frequency or intensity weights estimated here aggregated to get overall sentiment in form score.

Sentiment analysis may also be conveyed visually. Each pair of photos taken at a certain time “t and t+1” is integrated into one picture using "Temporal Convolutional Neural Network (CNN)". In CNN research, activation of a weighted matrix at each hidden layer is convolved with that of the matrix at layer below which is trained. In contrast to two language-specific approaches, 21 sentence-level sentiment analysis approaches are available for the English language. Additionally, a multilingual sentence-level method is shown to be supported by nine specific-language datasets. [5].

### 4. Methodology

The objectives of present research were to pre-process social media text and apply classification-based system to for sentiment analysis for appropriate classification, prediction proposed analysis. For the saidresearch work the researcher has set following objectives:

- i. To extract the tweet data set from kaggle on “Squid Game” Netflix series.
- ii. Pre-process Tweets (need to remove noise and pre-process tweets like convert the tweets to lower case)
- iii. To study machine learning based sentiment analysis for opinion categorization, pattern detection and prediction.
- iv. Analysing Tweets for Sentiments.
- v. To apply grouping and sub grouping fuzzy classifications of social media data.
- vi. Compare our result with existing work.

#### 4.1 Support Vector Machine (SVM)

The dataset is represented as a collection of the points in space on graph dimensions of which depend on the amount of classes. This algorithm then predicts class to which each input belongs. SVM uses a training dataset and a margin determined by the datasets to identify a hyperplane. The Support Vector Machine (SVM) is a technique that has a great degree of versatility due to the fact that it can be utilised to distinguish among datasets that are linearly and nonlinearly separable (where there is no clarity in separation with a hyperplane between classes). Dataset may be classified into two or more categories using SVM. The midpoint of margin or gap which appropriately classifies data is known as a hyperplane. Support vectors are the datasets on the left and right margins. It is essential that the margin distance be as great as it possibly can be. The SVM algorithm supports a variety of kernels, including “dot, radial, polynomial, neural, gaussian combination, multiquadric, ANOVA, and epachnenikov,” among others. There is a specific function connected with each of these kernels' workings.

#### 4.2 Logistic Regression

- One of the most well-known and frequently used machine learning algorithms is called logistic regression. This technique, which belongs to the larger category of supervised learning, is also one of the most widely employed. Using a specified set of independent variables, it may be used to predict the category dependent variable.
- It is possible to predict the result of a categorical dependent variable using a logistic regression model. Because of this, outcome must be discrete or categorical in nature. Rather of presenting a specific value of 0 or 1, it presents the probability values that lie between 0 and 1. It is conceivable that it is either Yes or No, 0 or 1, true or false, etc.

- Logistic Regression and Linear Regression are quite similar in many ways. In the case of regression difficulties, linear regression is utilised, while logistic regression is used in the case of classification problems.
- To predict two maximum values in logistic regression, an "S" is employed instead of a straight line (0 or 1).
- This graph illustrates the probabilities of a number of different outcomes, such as whether or not the cells are malignant, whether or not a mouse is overweight based on its weight, and so on.
- Probabilities and classifications may be generated using both continuous and discrete data using logistic regression. This is an important machine learning technique.
- In Logistic Regression, the most efficient parameters for categorising observations may be readily determined using different data formats. The logistic function is shown in the following image:

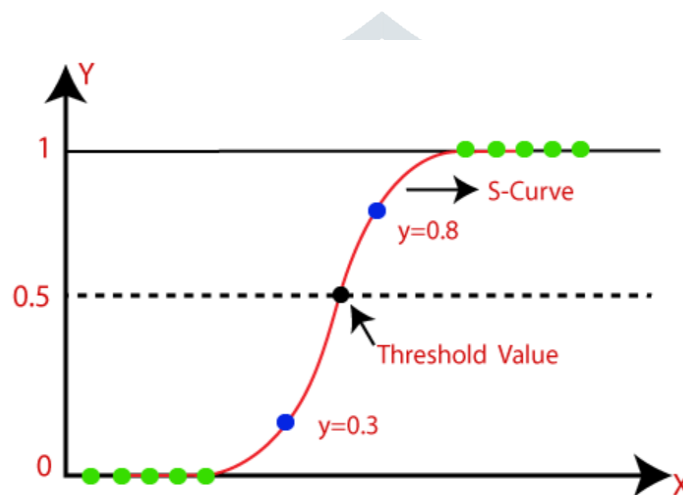


Figure: 1 Logistic Regression Representation

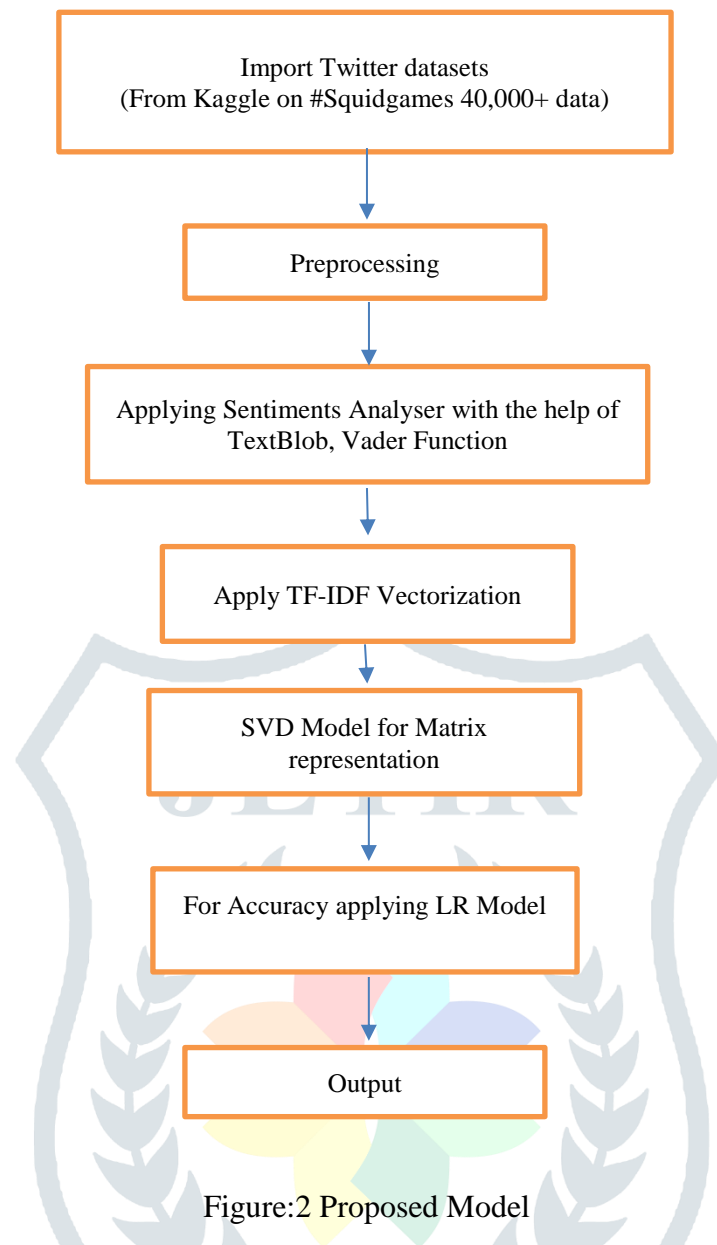


Figure:2 Proposed Model

In this projected work, the sentiment analysis is processed using the tweets that are collected. Particular text feature is extracted from the dataset and processed the five steps discussed above and analyzed. Finally, the extracted data is used for classifying sentiment using the proposed model that is based on LR.

## 5. Analysis

Various aspects of the Sentiment Analysis have been considered and analyzed in this research work. The main focus was on the tweet polarity detection process. The works considered the processes from pre-processing to classification.

Other social media platforms, such as Facebook and LinkedIn, may be taken into account to improve the presented methodologies. In addition, the hybrid mathematical model may be reinforced even more by the addition of a large number of web series and movies for the purpose of opinion prediction.

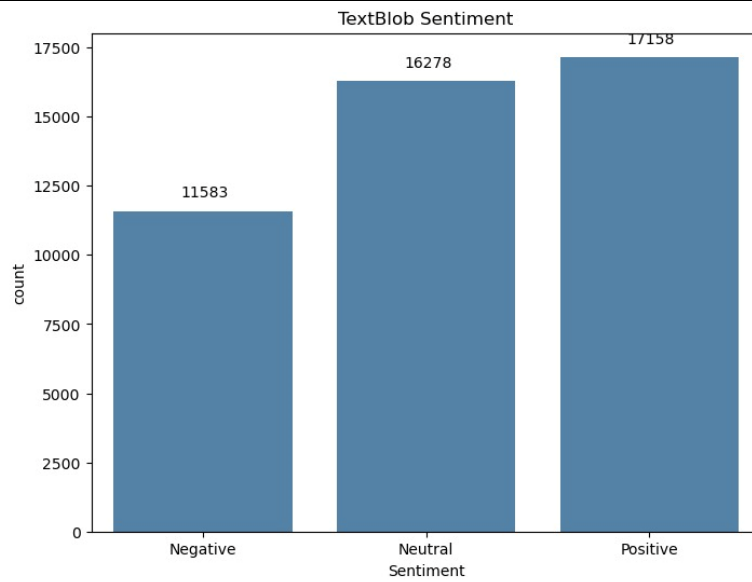


Figure:3 Sentiment analysis using Textblob

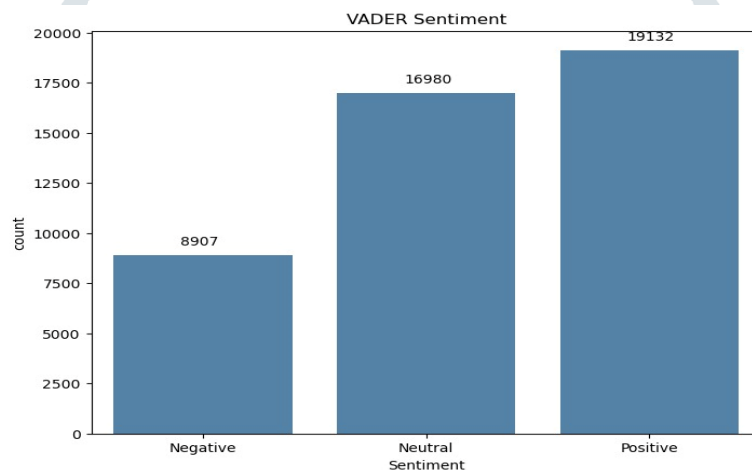


Figure:4 Sentiment analysis using VADER

	Precision	recall	f1-score	support
Negative	0.91	0.72	0.80	1781
Positive	0.88	0.97		3827
Accuracy				5608
Macro avg.	0.90	0.84	0.86	5608
Weighted avg	0.89	0.89	0.89	5608
F1 score : 0.8889087018544936				

Figure:5 Highest Accuracy of LR Algorithm

Table:1 Comparing the accuracy with Existing work.

Model	Accuracy
Existing Wok [6]	83%
Proposed Work	89%

## 6. Conclusion

Social media sites such as Facebook, Twitter, and Instagram are excellent places for academics to acquire data for further study. Opinion analysis is another term for sentiment mining, is used to conduct research on the user's feelings and evaluate the results. The piece of text that is recorded by the users is parsed in order to extract thoughts, which are then classified, and this information is then utilised to ascertain the attitude towards a certain subject or product. The contribution of research on the thesis consists of doing a sentiment analysis for the purpose of making a forecast about the Netflix web series "Squid Game." The information used in the sentiment analysis comes from social networking or social media sites like "Twitter, RSS news feeds, and news bias" associated with stock news. Primary objective of the body of research is to enhance the sentiment analysis performance and to bring more precision to prediction via the use of the Logistic Regression Algorithm. The outcome suggests an improvement prediction accuracy of LR Algorithm is 89%.

## 7. References

1. R. Sharma, S. Nigam, and R. Jain, "Polarity detection at sentence level," *International Journal of Computer Applications*, vol. 86, no. 11, 2014.
2. K. Li, R.G.Dewar, R.J.Pooley (2005), " Object-Oriented Analysis using Natural Language Processing", in *Technical Report Index* at <http://www.macs.hw.ac.uk/cs/techreps/browse.html>
3. Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam(2014), "Comparative Study of Classification Algorithms used in Sentiment Analysis", *International Journal of Computer Science and Information Technologies*, Vol. 5 (5) , 6261-6264
4. Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM) 2016 Dec 12* (pp. 439-448). IEEE.
5. Araujo M, Reis J, Pereira A, Benevenuto F. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing 2016 Apr 4* (pp. 1140-1145). ACM.
6. Nair, Anu J., G. Veena, and Aadithya Vinayak. "Comparative study of twitter sentiment on covid-19 tweets." *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2021.