Sentiment analysis for Twitter data

Sampada Mohan Naik
sampadanaik698@gmail.com
Sai Vidya Institute of Technology,
Rajanukunte, Bangalore, Karnataka

Mythri B.
mythrib.17cs@saividya.ac.in
Sai Vidya Institute of Technology,
Rajanukunte, Bangalore, Karnataka

Neha T. S.
nehats.17cs@saividya.ac.in
Sai Vidya Institute of Technology,
Rajanukunte, Bangalore, Karnataka

Prathiksha M.
prathiksham.17cs@saividya.ac.in
Sai Vidya Institute of Technology,
Rajanukunte, Bangalore, Karnataka

Sunil G. L.
sunil.gl@saividya.ac.in
Sai Vidya Institute of Technology,
Rajanukunte, Bangalore, Karnataka

ABSTRACT

Sentiment Analysis or Analysing emotion can be the process of understanding various textual opinions. Also known as opinion mining or emotional AI. People are interested in posting comments on social media that refer to their event experience to understand if most people had a positive or negative experience with same incident. This classification is being achieved using Sentiment Analysis. Sentiment analysis takes a few unstructured text comments , events, etc. In all comments posted by multiple users and classifies the comments into different categories as positive or negative opinions. This is also called Polarity classification. Sentimental Analysis is performed by text analysis and linguistics. This work aims to compare the performance of various machine learning algorithms when performing sentiment analysis for Twitter data.

Keywords: Machine learning, Naïve Bayes, Logistic regression, Sentiments, XGB classifier.

1. INTRODUCTION

Analysing sentiment is process of searching for textual opinion. People post social media reviews that mention their own experience in an incident and they are also have positive or negative experience with same incident or product. Analysing sentiment can be process of knowing user feeling about some specific item which can individual do on topic, incident or recent trends.

Analysing sentiment takes place in three stages and they are sentence, aspect, and document level. Twitter is likely to be very diverse data source that determines the standard for all high products. Twitter platform uses textual tweets to express opinion. Therefore, sentiment analysis at sentence level is used to identify emotions. Analysis of sentiment through twitter provides good ratings for businesses that require personal opinion on products launched. The goal is to calculate the sentiment/emotional accuracy of the text extracted from tweet.

Sentiment analysis is often performed on twitter data to classify tweets as positive or negative. This analysis helps relevant organizations find views on products, events, and tweets.

Pointing out opinions is the most difficult part of sentiment analysis. Opinions can always be positive or negative, depending on the situation. The meaning of the content is not changed by processing system even if there is slight change in word. However, analysis can change meaning of the content that changes the two words. Processing is done statement by statement. Informal twitter statements are not understood by the system, but by the users. We proposed a system to try and do sentimental analysis for twitter data using machine learning techniques and algorithms.
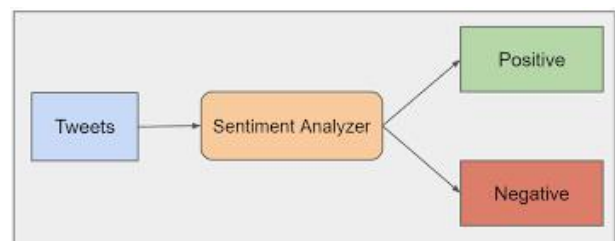

Fig. 1: Tweet Classification

2. LITERATURE SURVEY

- **A. Esuli and F. Sebastiani:** Sentiment classification can be new subfield of text classification with the opinions expressed by the document, rather than the topic of the document. The ability to extract some of the opinion from a text we will be able to determine the direction of term "subjective" contained in text.
- **Sonsoy, O.aA., Rady, S., Badr, N. L., & Hashem.M. :** Understanding how business trends behave requires collecting vast amounts of datas from miscellaneous sources. Location based social medias can provide vast amounts of data that is used for analysing business

behaviours. Collected data is used to predict business meetings.
- **Ferguson, Rebecca (2012):** In traditional ways, data (inputs) and programs are provided to computer (machine) which computer processes and provides results (output). However, in the case of the machine learning method, we provide data (input) and result (output), and computer (machine) itself deduces program according to input-output relation.
- **B. Pang, L. Lee, and S. Vaithyanathan:** Performance of Naive Bayes algorithm, Maximum Entropy and Support Vector Machines algorithm in analysing sentiment are compared with various functions, such as searching only for basic adjectives of unigram and bigram two combination parts of speech and positional information.
- **Singh B, Kushwaha N., & Vyas O. P. :** This paper suggests that analysing sentiment or emotions plays important role in business intelligence, it collects consumer feedback and behavior which helps in better prediction and higher cognitive process. So they have combined two techniques NLP (Natural Language Processing) and the Machine learning approach.
- **Ghiassi, M., Zimbra, D., & Lee, S.:** Here they discussed the company's valuable opinions about it's brand and products from social media statements on twitter. The work here focused on task of distributing brand-related tweet sentiment classes from twitter. They used SVM supervised learning model to classify emotions in two characteristics related to twitter.

3. PROPOSED SYSTEM AND METHODOLOGY
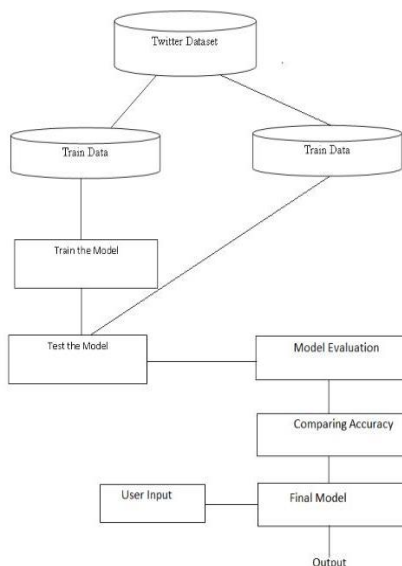Based on the tweets given by the user in the Front-end, our system classifies tweets as Positive or Negative.

Fig 2: System Architecture Training and Testing of Data

Data set is split into Train Data and Test Data in the ratio 80:20

3.1 Data set
For our model we have used Twitter data set of an online website such as Kaggle. Train the model using another Twitter data set. The data set used in this project has 99989 rows of data present.

3.2 Data Pre-processing
When considering data, it is always a very large dataset with a large number of rows and columns are displayed. However, this is not always the case when data can be in different formats such as images, audio and video file structured table. Machines do not understand images, video and text data as they are. Machines understand only 1s and 0s.

3.3 Steps in Data Pre-processing:
- Data cleaning: In this step, input "missing values", "Smooth noisy data", "finding or deleting outliers ", and "Resolution of inconsistencies is done."
- Data Integration: Here, addition of some databases, information files or information sets is executed..
- Data transformation: Aggregation and normalization are executed to adjust to a specific value.
- Data reduction: This section gives you an overview of dataset that is very small in size but produces the same analysis results as before.
(a) Stop words: "Stop words are the nation words that don't add much aspiring to a sentence." It is safely ignored without forgoing the sense of the sentence.
(b) Tokenization: "Tokenization is that the process of splitting a stream of the manuscript into phrase, symbols, words, or any expressive elements named as tokens."
(c) Lemmatization: Lemmatization is that the process of converting a word to its base form.

Software Specification
Operating system used is Windows XP / 7, Coding Language is Python, HTML and Version of python used is Python 3.6.8, IDE is Python 3.6.8 IDLE, ML Packages are Numpy, Pandas,Sklearn, Matplotlib, Flask, PymySql, NLTK. ML Algorithms are Logistic Regression, XGB Classifier and Naïve Bayes and the NLP Techniques are Count Vectorizer, TFIDF

Transformer and, RE (regular expression), word_tokenizer, WordNetLemmatizer and Other Requirements required are Notepad and XAMPP Control Panel.

4. ALGORITHMS
1. Logistic Regression: Input: Twitter Dataset
Steps:
- Take user's input and convert it into arrays.
- Pre-process data used as input
- Train the model
- Finally, use the learned model to produce a prediction

2. Naive Bayes
Input: Twitter Dataset
Steps:
- Pre-process the data used as input
- Use the Count Vectorizer and TFIDF Vectorizer to build our feature vector
- Train the model
- Use the learned model to produce a prediction

3. XGB Classifier
Input: Twitter Dataset
Steps:
- Pre-process the data used as input
- Split the data into Train and Test data in the ratio 80:20
- Train the model using training data
- Use the learned model to produce a prediction

5. RESULTS
Our model is trained using multiple algorithms to identify and compare the results for greater accuracy. After knowing the accuracy of all the algorithms, we select Logistic Regression as the finalized algorithm.

Table 1:Accuracy Table

Sl.no	Algorithm	Accuracy
1	Naïve Bayes	73.70
2	Logistic Regression	75.87
3	XGB Classifier	73.20

In the Front end, the user will enter the tweet. The tweet is taken as input to our finalized algorithm to predict the output. The Output will be either the tweet is Positive or Negative.

6. CONCLUSION

Sentiment analysis is process of identifying opinion in statement. People post reviews on social media that mention their experience in an incident or event and also want to know other people experience on same event whether it is positive or negative. Here goal is to calculate sentiment accuracy of statements that has been extracted from text of tweets. Analysing sentiment of tweet helps to determine whether tweet has sentiment positive or negative on particular products, events and so on. Sentiment analysis was developed to investigate public opinion on tweet/hashtag. Analysis of sentiment can be determining factor in business process decision.

So, we proposed a system with the help of machine learning techniques, NLP techniques like count vectorizer, TF-IDF Transformer and Machine learning algorithms like Logistic Regression, XGB Classifier and Naïve Bayes to predict the Tweet is positive or negative based on tweet data entered by the user in the front end.

We implemented our proposed system and logistic regression gave the more accuracy of 76% in the prediction of sentiment of the tweet.

7. ACKNOWLEDGEMENT

8. REFERENCES
[1] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proc. 14th ACM Int. Conf. Inf. Knowl.Manage., 2005, pp.617-624

[2] Sonosy, O. A., Rady, S., Badr, N. L., & Hashem, M. A study of spatial machine learning for business behaviour prediction in location based social networks. 2016 11th International Conference on Computer Engineering & Systems(ICCES).

[3] Ferguson, Rebecca. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6) pp. 304–317.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," inProc.ACL02Conf.Empirical Methods Natural Lang. Process., 2002, pp. 79–86

BIOGRAPHIES

Sampada Mohan Naik
Student,
Sai Vidya Institute of Technology, Rajanukunte, Bangalore, Karnataka, India

Neha T. S.
Student,
Sai Vidya Institute of Technology, Rajanukunte, Bangalore, Karnataka, India

Mythri B.
Student,
Sai Vidya Institute of Technology, Rajanukunte, Bangalore, Karnataka, India

Prathiksha M.
Student,
Sai Vidya Institute of Technology, Rajanukunte, Bangalore, Karnataka, India

Sunil G. L.
Asst Professor, Dept. of CSE
Student,
Sai Vidya Institute of Technology, Rajanukunte, Bangalore, Karnataka, India