Twitter Sentiment Analysis

Kamma Sai Pujitha
Computer Science Engineering
Lovely professional university
Jalandhar, Phagwara
kammasaipujitha@gmail.com

Mutyala Kesava Sivanand
Computer Science Engineering
Lovely Professional university
Jalandhar,Phagwara
Nandu130505@gmail.com

ABSTRACT
With the advancement and development of Internet technolog y,The Internet provides more information to Internet users an d also creates more information. The Internet has become a pl atform for online learning, exchange of ideas and sharing of i deas. Social networking sites such as Twitter, Facebook,Goo gle+ are rapidly gaining popularity as they allow people to sh are and express their opinions on topics, interact with differe nt communities or broadcast messages worldwide. Many stud ies have been conducted from the perspective of data analysis . This research focuses on the sentiment analysis of Twitter d ata, which helps to determine whether the opinions in tweets are negative,heterogeneous, positive or negative or neutral. In this paper, we perform a benchmarking analysis and compari son of existing mining theories (e.g., machine learning and di ctionarybased methods). We examine Twitter data streams us ing various machine learning algorithms such as Naive Bayes , Maximum Entropy, and Support Vector Machines.
and the application form.

Keywords
Twitter, Sentiment analysis (SA), Machine learning, Naive Bayes (NB), Logistic regression, Accuracy, Random forest classifier, Xg boost classifier, Voting classifier.

1. INTRODUCTION

Today, the internet era has changed the way people express th eir thoughts and feelings. Now this is done through blog posts , online forums, product review websites, social media etc. To day, millions of people use social networking sites like Faceb ook, Twitter, Google Plus etc. in their daily lives to share their thoughts, feelings and opinions. From online communities, w e get social media where customers can share information wit h others and influence others through discussion. Social media creates a wealth of rich emotions in the form of tweets, status updates, blog posts, comments, reviews, etc. People rely on u sergenerated content on the internet to make decisions. For ex ample, if someone wants to buy a product or use a service, the y will first check online reviews of the product and social med ia before making a decision. The amount of usergenerated con tent is too large for ordinary users to analyze. Therefore, it nee ds to be automated and various analytical methods are widely used.

Sentiment Analysis (SA) informs users whether information a bout a product is satisfactory before purchase. is available upo n request.

Information retrieval technology generally focuses on process ing, searching or analyzing existing factual information. Thes e elements are mainly thoughts, feelings, evaluations, attitude s and emotions, which form the basis of emotional intelligenc e (EI). Due to the increase in information available in online r esources such as blogs and social networks**,** it provides many challenging ways to create new applications. EI can be used to predict the process by calculating the content such as positive or negative thinking of that activity.

2. SENTIMENT ANALYSIS

Sentiment analysis, often referred to as sentiment mining, is a subfield of natural language processing (NLP) and machine le arning (ML) that focuses on identifying and classifying senti ment in text. Its goal is to determine whether the opinion expr essed in a document is positive, negative, or neutral. This abili ty is more useful in many areas such as business, customer ser vice, finance, and social care, where understanding the public' s sentiment is important. The desire to make good decisions.

In today's digital age, a lot of unnecessary information is gene rated every day through social media platforms, online comm ents, blogs, and forums. Analyzing this data can provide insig hts into customer preferences, brand insights, and trends. For example, businesses can use feedback to measure customer sa tisfaction and respond quickly to negative feedback. In the co ntext of politics, opinion polls can also help track public opini /e can refer fig1 for better understanding of sentiment analysis

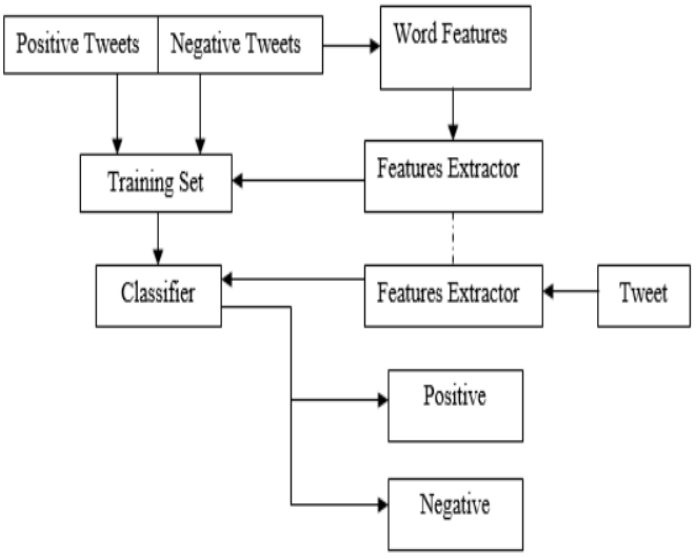

Fig. 1 Flow chart for sentiment Analysis.

2.1 Data Collection:

Data collection is the process of collecting raw data or analysis for use in machine learning. This is an important step because the quality and quantity of data directly affect the performance and reliability of the final model.

2.1.1 Importing Libraries:

We have imported the libraries that we have to use

1. Pandas (import pandas as pd): Pandas is a powerful data and analytics library. It provides a data structure suitable for creating data structures such as Data Frame.

2. NumPy (import numpy as np): NumPy is a simple package for computing with Python. Provides support for multi

dimensional arrays and various arithmetic operations.

3. Scikit-learn:
It is a popular machine learning library in Python. It provides simple and effective tools for data exploration and data analysis. The train_test_split function is used to split the dataset into random training and testing subsets.

4. Matplot:
Matplotlib is a powerful Python library for creating static plots, visualizations, and dialogs in Python. It provides a variety of drawing tools commonly used in fields such as data science, machine learning, engineering, and scientific research.

2.1.2 Data Description
Preliminary data is an important step in any machine learning process, including Twitter sentiment analysis. Refer fig 2

Data:

	id	flag	target	text
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...

Fig. 2 First 5 rows of dataset.

2.1.3 Work Flow

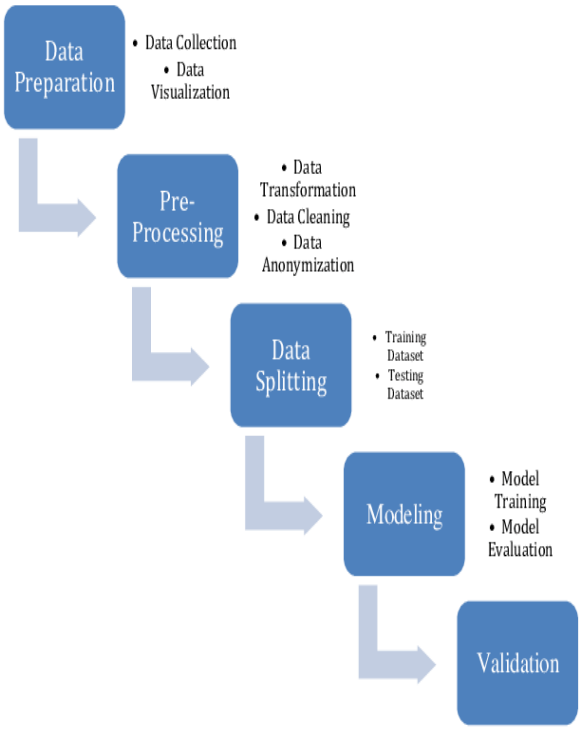

Fig. 3 Work flow of machine learning models.

2.2 Data Preprocessing
Reading data is an important step in data analysis or machine learning. In Python, the Pandas library is often used to read and manipulate data, especially when creating data such as CSV files, Excel files, and SQL databases.

2.2.1 Removing the stop-words
Sentimental analysis is an important part of natural language processing (NLP) and involves identifying the emotional to ne behind the text. An important first step in emotional asse ssment is to eliminate stop words. Abandoned words are usu ally words that do not have any significant meaning and do not reflect the main meaning of the text.
Abandoned words are words that occur frequently in a langu age but contribute little to the meaning of a sentence. Exam ples include objects (the, a, one), prepositions (in, of, of), co njunctions (and, but, or), and pronouns (he, her, it). While th ese words are useful grammatically, they often do not provi de useful information when analyzing the thought or content of a text. Important points for improving data quality Word s that help define ideas or topics.

2.2.2 Tokenization
Tokenization is an important step in natural language proces sing (NLP), especially when analyzing sentiment expressed in tweets. Given the unique characteristics of Twitter profile s, effective tokenization is essential for the dissemination of meaningful comments. This article explores the tokenization process in Twitter's understanding of identity
Tokenization is the process of breaking text into smaller pie ces called tokens. These tokens can be words, phrases, or sy mbols. In the context of Twitter sentiment analysis, tokeniza tion typically focuses on a wordlevel tag that is important fo r understanding the sentiment expressed in tweets.

2.2.3 Stemminng or Lemmatization
Stemming and lemmatization are two simple methods in nat ural language processing (NLP) that aim to reduce a word to its base or root form. Both methods are important for activit ies such as text analysis, data retrieval, and sentiment analys is because they help generate informative data by identifyin g morphological changes. However, the methods differ in te rms of accuracy and computational requirements.
Stemming and lemmatization play an important role in natur al language processing by reducing words to their simplest f orms. Stemming provides speed and simplicity through heur istics, while lemmatization provides clarity and expressiven ess through complex word analysis. Understanding these dif ferences allows practitioners to choose appropriate methods based on specific applications and needs in word processing tasks.

2.2.4 Feature Extraction
Feature extraction is the process of identifying and extractin g relevant features from raw data to create more informative data. The goal is to reduce the complexity (often referred to as "dimensionality") of the data while preserving as much r elevant information as possible. This simplification helps in crease the efficiency and effectiveness of machine learning algorithms and simplifies the analysis process.
It is an important technique in machine learning and natural language processing that involves transforming raw data int o a format suitable for analysis. Specifically in the context o f text, feature extraction technology converts unstructured te xt into a digital representation that can be processed by mac hine learning algorithms.It will provide an indepth look at v arious extraction methods, their applications.

Feature Extraction Techniques:

1. Bag of Words(BOW)

The bag of words model represents a data file based on the collection of single words, regardless of grammar or word order . Each unique word in context becomes unique and its frequency in the data is counted.

Advantages: Easy to use, works well for simple text classificati on.

Disadvantages: Ignore word order and semantics which can lea d to loss of information content.

2. Term Frequency-Inverse Document Frequency (TF-IDF)

TFIDF is an extension of BoW and assigns a weight to each wor d based on its frequency in the document based on its occurrence in each document. This tool helps to separate important words fr om common words. Refer fig 4

Advantages: Provides better representation than BoW by taking into account keywords.

Disadvantages: Still can't capture orders or content

$$\mathrm{TF\,IDF}(t,d) = \mathrm{TF}(t,d) \times \log\left(\frac{N}{\mathrm{DF}(t)}\right)$$

where N is the total number of documents, $DF(t)$ is the number of documents containing term t, and $TF(t,d)$ is the term frequency of t in document d.

Fig. 4 Formula for TF-IDF

3. Word Embeddings

Word embeddings like Word2Vec or GloVe transform words into dense vector representations that capture semantic relationship between words based on their context in large dataset.

Advantages: Captures meaning and relationships between words, similar words have similar vector representations.

Disadvantages: Requires substantial amounts of data for training; may not perform well on small datasets.

Feature extraction is an important step in preparing data for mac hine learning in NLP. By converting raw text into digital represe ntations using methods such as Bag of Words, TFIDF, Ngram, a nd word embedding, practitioners can improve performance and translation while reducing noise and length.

2.3 Splitting Data into Training and Testing Sets

To evaluate model performance effectively, split the standardized dataset into training and testing subsets. A common practice is to allocate 80% of the data for training and 20% for testing.

2.4 Model Training:

After splitting the data into training and testing, the next step is to train the model using the training data. Here are the basics of training machine learning models in Python using scikitlearn.Model selection is an important step that affects th e efficiency and accuracy of the sentiment classification proce ss. Model selection depends on many factors as the nature of t he product, the complexity of the analysis task, and the perfor mance evaluation to be performed

Models for Twitter Sentiment Analysis

1. Logistic Regression
2. Naïve Bayes

Logistic Regression:

Logistic regression is a statistical method used for binary classification; The goal here is to estimate the probability that a sample belongs to a particular class. Despite its name, logistic regression is not a regression algorithm, but a classification algorithm.

Here's a simplified explanation of how the logistic regression works in a model:

Input variables: Logistic regression takes input variables (also known as features or independent variables) and assigns weights to them. Input variables can be continuous, categorical or binary.

Logistic function: Logistic regression uses the logistic function (also known as the sigmoid function) for a linear combination of input variables and their weights. The logistic function transforms the output into a value between 0 and 1 that represents probability that the sample belongs to the positive class.

Decision boundary: Logistic regression calculates a decision boundary (or threshold)that separates events in to two groups based on their predicted probability of occurrence. In general, if the predicted probability is greater than 0.5, the sample is classified as a good class.

Model training: Train a logistic regression model using optimization techniques to find the most weighted model that minimizes the difference between the predicted probability and the class map in the training data.

Prediction: Once the model is trained, it can predict a list of new events by applying the learned weights to the communication variables and feed the min to the logistics study.

Logistic regression is widely used in many fields, such as medicine(e.g., predicting disease incidence), finance(e.g., credit risk assessment), business(e.g., stopping people guessing), and more. It is interesting for its simplicity, interpretation and efficiency, especially considering that the relationship between different inputs and outcomes is linear.

Naïve Bayes

Naive Bayes is a popular classification technique based on Baye s' theorem and is widely used in many applications including tex t classification, spam detection, and sentiment analysis. This topi c will take an indepth look at the principles behind Naive Bayes, its types, advantages, limitations, and practical applications. Naive Bayes is a family of bestfit methods that use Bayes' theore m and the "naive" assumption of independence of each pair of fe atures. This means that the presence of a feature in a category is assumed to be independent of the presence of other features. Des pite this simplicity, Naive Bayes often performs very well in pra ctice, especially for large datasets.

Bayes Theorem

The basis of Naive Bayes is Bayes' theorem, which describes the probability of a hypothesis based on prior knowledge of events associated with the hypothesis. It can be represented mathematic ally as: Refer fig 5

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Where:

- $P(Y|X)$ is the posterior probability of class Y given feature X.
- $P(X|Y)$ is the likelihood of feature X given class Y.
- $P(Y)$ is the prior probability of class Y.
- $P(X)$ is the prior probability of feature X.

Fig. 5 Formula for Bayes Theorem.

Types of Naive Bayes Classifiers

Gaussian Naive Bayes:
The features follow a normal (Gaussian) distribution. It is suitable for continuous data.

Multinomial Naive Bayes:
Used for discrete counts (e.g., word counts for text classification). It is particularly effective for document classification tasks.

Bernoulli Naive Bayes: Similar to Multinomial Naive Bayes but assumes binary features (e.g., presence or absence of a word).

2.5 Model Evaluation

Model evaluation is an important step in the machine learning pr ocess and helps measure the effectiveness and quality of the trai ning model. This process involves using various measures and m ethods to understand how the model performs on invisible object s, making it perform better than just memorizing information. Be low is a detailed description of the measurement model, includin g its importance, common metrics, measurement methods, and b est practices.

Importance of Model Evaluation

Performance evaluation:
Allows practitioners to evaluate how well the model predicts bas ed on new data, which is important for determining the effective ness of the model in real-world use in practice.

Identify strengths and weaknesses:
By evaluating the model, data scientists can identify its strengths and areas for improvement.

Avoid overfitting:
Testing performance on a separate test helps ensure that the mod el not only remembers the training data but can also generalize t o new situations.

Decision making information:
Information obtained from evaluating the model can inform dec isions about using the model, improving it, or abandoning the m odel altogether.

2.6 Common Evaluation Metrics:

Classification Tasks:

1. Accuracy: The proportion of correct predictions made by the model out of all predictions.
Accuracy = (True Positives + True Negatives)/Total Predictions

2. Precision: The ratio of true positive predictions to the total predicted positives, indicating how many selected items are relevant.
Precision = (True Positives)/(True Positives + False Positives)

3. Recall (Sensitivity): The ratio of true positive predictions to all actual positives, indicating how many relevant items were selected.
Recall = (True Positives)/(True Positives + False Negatives)

4. F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.
F1 = 2 · (Precision · Recall / Precision + Recall)

5. Confusion Matrix: A table that provides a detailed breakdown of correct and incorrect classifications for each class, helping visualize performance.

Area Under the ROC Curve (AUC-ROC): Measures the ability of a model to distinguish between classes at various threshold settings.

Regression Tasks:

Mean Absolute Error (MAE): The average absolute difference between predicted and actual values. Refer fig 6.

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

Fig. 6 Formula for MAE

Mean Squared Error (MSE): The average squared difference between predicted and actual values. Refer fig 7

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2$$

Fig. 7 Formula for MSE

R-squared: Indicates how well the independent variables explain the variability of the dependent variable.

Model evaluation is an important part of the machine learning lif ecycle, allowing practitioners to assess the performance of their models and ensure they are suitable for use in realworld applicati ons. By using appropriate metrics and evaluation techniques, dat a scientists can understand the strengths and weaknesses of their models and ultimately make higherquality, more reliable predicti ons across a wide range of functions.

2.7 Enhance Text Classification

Text classification has evolved with the advent of natural langua ge processing (NLP) technology. This system uses deep learning and advanced algorithms to increase the accuracy and efficiency of text classification.

1. Word Embeddings

Word embedding is an important process to transform a wor d into a dense vector image that captures relationships and r elations. Popular models include:
Word2Vec: Learns word associations from large corpora.
GloVe: Focuses on global statistical information.
fastText: Extends Word2Vec by considering sub word information, improving performance on morphologically rich languages.
These embeddings allow the model to understand the simila rities and differences between words, improving classificati on accuracy.

2. Deep Learning Architectures

Deep learning has introduced several architectures that excel in text classification:
Convolutional Neural Networks (CNNs): Originally designed for image processing, CNNs have been adapted for text by capturing local patterns through convolutional layers. They effectively learn hierarchical representations, improving classification outcomes.
Recurrent Neural Networks (RNNs): Specifically, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are adept at handling sequential data. They

maintain memory of previous inputs, making them suitable for tasks where context is crucial.

Transformer Models: Models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized NLP by using self-attention mechanisms to capture contextual relationships across entire text sequences. These models achieve state-of-the-art performance in various benchmarks and can be fine-tuned for specific classification tasks

3. Random Forest Classifier

Random forest classifier is a powerful learning method wide ly used in classification and reprocessing in machine learnin g. It works by creating multiple decision trees during trainin g and generating predictive models for distribution function s or averages for regression functions. This approach impro ves the accuracy of prediction and control on the job, makin g it suitable for complex data.

4. XG Boost Classifier

XGBoost (eXtreme Gradient Boosting) is an effective and p owerful gradient boosting widely used in supervised learnin g such as classification and regression. Its popularity stems from its performance in competitive machine learning and it s ability to process large amounts of data.

5. Voting Classifier

Polling is an integrated machine learning system that combi nes predictions from several underlying models to improve overall classification accuracy. It works on the principle that aggregating different inputs from various distributions redu ces bias and variance, leading to better predictions.

Types of Voting:

- **Hard Voting**
  In forced voting, the class prediction results are det ermined by the class that receives the most votes fr om each classifier.
- **Soft Voting**
  Soft voting determines the estimated probability of each class across all distributions and selects the cl ass with the highest probability above the mean

2.7 Significance of Twitter Sentiment Analysis:

- **Instant insights:** One of the biggest benefits of Twitter sentiment analysis is the ability to provide instant feed back on public opinion. With over 400 million active us ers tweeting on a variety of topics every day, businesses can track customer reviews and feedback. This allows organizations to respond to customer concerns immedia tely, resulting in positive feedback.

- **Market research and trend analysis:** Companies can i dentify new trends and consumer preferences by analyz ing tweets about products, services, or events. This info rmation can inform marketing strategies, product devel opment, and competitive analysis. For example, brands can measure the effectiveness of their marketing campa igns by measuring perceptions of support.

- **Crisis Management:** Emotional assessment as an early warning system for public relations. By monitoring neg ative sentiment around their brand or product, organizat ions can take important steps to resolve issues before th ey escalate. This capability is especially important in to day's fastpaced digital environment, where misundersta ndings can spread quickly.

- **Opinion polls:** In politics, opinion polls can provide pu blic insight into policies, candidates,events. Researcher s and analysts can track changes in sentiment over time helping to inform strategic planning and policy decision s.

2.8 Challenges in Twitter Sentiment Analysis:

Noise information: The informal nature of tweets often includes slang, abbreviations, typos, and emojis, which can hinder analys is. This noise can make sense if it doesn't work well with the first time.

Short text: Tweets were limited to 280 characters (previously 1 40), limiting the amount of content available for commentary. T he brevity of tweets can obscure meaning and make it difficult f or algorithms to correctly classify them.

Neutral sentiment: A large portion of tweets may express neutr al sentiment or no clear sentiment. This may lead to bias if the m odel is not trained to handle such situations.

Understanding the situation: Thinking is often contextdepende nt; words can have different meanings depending on how they ar e used in a tweet. For example, sarcasm or mockery can lead to i ncorrect classification models if not included in different context

Feature Engineering: Eliminating positive influences is critical to feature design, but can be challenging due to the multitude of ways users express their emotions on Twitter. The model should include multiple languages for example n grams, partofspeech labels and also consider specific languages.

2.9 Accuracy score:

Training Models	Accuracy Score
MN Naïve Bayes	36%
Logistic Regression	54%
Voting classifier	83%
Xg boosting	89%
Random Forest	97%

Fig. 8 Accuracy Scores of different training models

Conclusion:

Sentiment analysis on Twitter has become important for quickly understanding public opinion by utilizing the large amount of re dundant information generated on the platform. As social media continues to play a significant role in communication, businesses , researchers, and policymakers are increasingly turning to senti ment analysis to gain insight into consumer behavior, public opi nion, and the economy. This paper presents the main results and implications of Twitter analytics theory, highlighting its impact, challenges, and future directions.

Sentiment analysis on Twitter is a powerful tool that allows orga nizations to effectively engage with the public. By using advanc ed machine learning techniques and solving problems such as no ise data and content understanding, participants can gain better i nsights that drive the right decisions on everything from busines s to crisis management to political analysis. As technology conti nues to evolve and new methods emerge, the applications of Twi tter analytics will expand even further, making it an essential too l in today's social media information world.

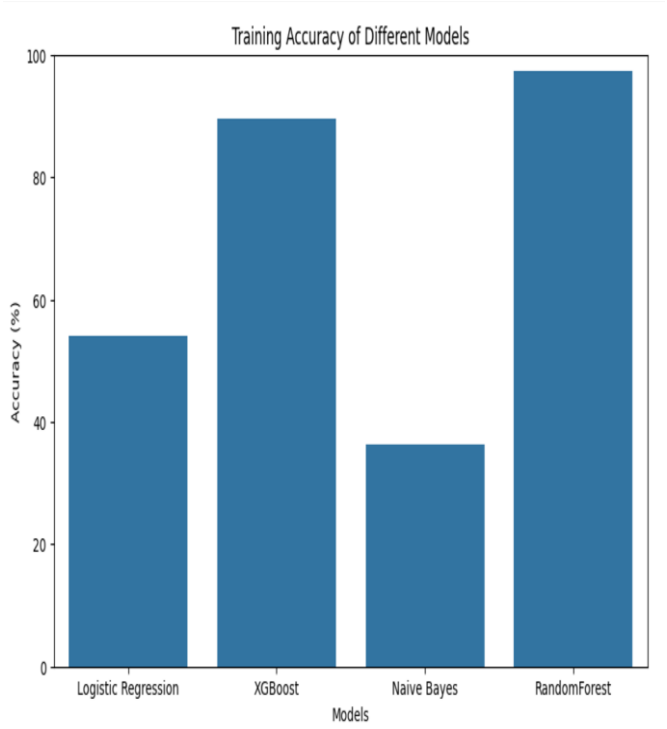

Fig. 9 Visualization of training accuracy of different models

By referencing the above fig 9 we get to know that we have implemented the four different training models successfully that is Logistic regression, Xg boost, Multinomial naïve bayes, Random forest classifier. Were Random forest classifier is having the highest prediction value than remaining training models. It is better to predict the target is positive, negative, neutral, irrelavent.

2.10 References:

1. R. Parikh and M. Movassate, "Sentiment Analysis of User Generated Twitter Updates using Various Classication Techniques",CS224N Final Report, 2009

2. A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evalua tion, 2010, pp.1320-1326

3. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

4. Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

5. Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

6. R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

7. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M.."Lexicon basedmethods for sentiment analysis". Computational linguistics, 2011:37(2), 267-307.

8. Li, S., Xue, Y., Wang, Z., & Zhou, G.."Active learning for cross-domainsentiment classification". In Proceedings of the Twenty-Thirdinternational joint conference on Artificial Intelligence (pp. 2127-2133).AAAI Press,2013

9. Bollegala, D., Weir, D., & Carroll, J.. Cross-Domain SentimentClassification using a Sentiment Sensitive Thesaurus. Knowledge andData Engineering, IEEE Transactions on, 25(8), 1719-1731,2013

10. Pang, B.and Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

11. V. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.

12. https://arxiv.org/pdf/1601.06971