Diagram

Description automatically generated with low confidence

**Twitter Sentiment Analysis**

SUBMITTED TO: **Prof. Savita Sehrawat** SUBMITTED BY: Prabhjot Singh STUDENT ID: 0774364

[INTRODUCTION 3-5](#_TOC_250005)

* [Abstract 3](#_TOC_250004)
* [Research Questions 4](#_TOC_250003)
* [Objective 4](#_Objectives)
* [GitHub](#_GitHub:_https://github.com/Prabhjoy) 5
* [Scope of Project](#_Scope__of) 5
* [System Features 5](#_System_Features)

[LITERATURE SURVEY](#_LITERATURE_REVIEW) 6

* [System Design 7](#_SYSTEM_DESIGN)

[DATA COLLECTION 5-14](#_DATA_COLLECTION:)

* EDA (Exploratory Data Analysis) 7
* [Data Visualization 8-14](#_Data_Visualization)
  + Visualization Of Graphs 8-13
  + Word Clouds 13-14

[**MATERIALS AND METHODOLIGIES 15-19**](#_MATERIALS_&_METHODOLIGIES)

* Complete pre-processing
* Partial pre-processing
* TF-IDF approach
* [Feature-Extraction Methods](#_Feature_Extraction_Methods:)
* [Support Vector Machine](#_SVM(Support_Vector_Machine):)
* [kNN (K Nearest Neighbour):](#_kNN_(K_Nearest)
* [Random Forest Classification](#_Random_Forest:)
* [Naive Bayes](#_Naive_Bayes)
* Evaluation of parameters

[**CONCLUSION 20**](#_CONCLUSION:)

**[REFERENCES 21](#_References:)**

# INTRODUCTION

In this study, natural language processing to do sentiment analysis on the Twitter dataset. The text field's polarity was broken down into three categories: positive, negative, and neutral. A sentiment is an attitude, belief, or conclusion brought on by a sensation. Sentiment analysis, commonly referred to as opinion mining, examines how individuals feel about particular things. The internet is a great resource for sentimental knowledge. Through different social media, including forums, microblogs, and online social networking sites, users may publish their own material. Many social media platforms disclose their application programming interfaces (APIs), which motivates academics and developers to gather and analyse data. For instance, Twitter currently has three different versions of APIs available namely the REST API, the Search API, and the Streaming API. With the REST API, developers are able to gather status data and user information; the Search API allows developers to query specific Twitter content, whereas the Streaming API is able to collect Twitter content in real time. Moreover, developers can mix those APIs to create their own applications. Hence, sentiment analysis seems having a strong fundament with the support of massive online data. The input data to the classification model must be closed using classification procedures in order to train the data. For the newly learned data, these models forecast the class label categories.

# 

# Abstract:

Today, the importance of social media cannot be denied. Most businesses collect feedback from their Twitter followers to gain insight and to gain a better understanding of their customers. Social media feedback changes rapidly, and the ability to analyze that feedback in real time is essential for the success of any business. There are several ways to discover how people react to a new product, brand, or event. For example, the sentiment expressed through tweets about a particular topic, product, brand, or event can provide an indication of the level of willingness or trust in a product. Hence, I use this premise in my project and extract trending #tags/tweets related to our desired topic every few minutes since hash-tagged tweets are more engaging. Developing a program for sentiment analysis is an approach to be used to computationally measure customers’ perceptions. This report on the design of a sentiment analysis, extracting and training a vast number of tweets.

# Research Questions:

The problem at hand consists of two subtasks:

* + - Phrase Level Sentiment Analysis in Twitter :

Given a message containing a marked instance of a word or a phrase, whether that instance is positive, negative, or neutral in that context.

* + - Sentence Level Sentiment Analysis in Twitter:

Given a message, decide whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

## **Objectives**

The objectives of this project are:

* + - To implement an algorithm for automatic classification of text into positive and negative
    - Sentiment Analysis to determine the attitude of the mass is positive, negative, or neutral towards the subject of interest
    - Graphical representation of the sentiment in form of Pie-Chart, Bar Diagram and Scatter Plot.

## **GitHub:** [**https://github.com/Prabhjoy25/Twitter-Sentiment-Analysis**](https://github.com/Prabhjoy25/Twitter-Sentiment-Analysis)

# Scope of project

This project will be helpful to the companies, political parties as well as to the common people. It will be helpful to political party for reviewing about the program that they are going to do or the program that they have performed. Similarly, companies also can get review about their new product on newly released hardware or software. Also, the movie maker can take review on the currently running movie. By analyzing the tweets analyzer can get result on how positive or negative or neutral are peoples about it.

**System Overview**

This proposal is used to analyze the tweets. We will be performing sentiment analysis in tweets and determine where it is positive, negative, or neutral. This web application can be used by any organization office to review their works or by political leaders or by any others company to review about their products or brands.

# System Features

The main feature of our web application is that it helps to determine the opinion about the peoples on products, government work, politics or any other by analyzing the tweets. Our system can train the new tweets taking reference to previously trained tweets. The computed or analyzed data will be represent in various diagram such as Pie- chart, Bar graph and Scatter Plot

# LITERATURE REVIEW

Sentiment analysis has been managed as a Natural Language Processing task at many levels of granularity. Starting from being a document level classiﬁcation task , it has been handled at the sentence level and more recently at the phrase level. Microblog data like Twitter, on which users post real time reactions to and opinions about “everything”, poses newer and different challenges. Some of the early and recent results on sentiment analysis of Twitter data use distant learning to acquire sentiment data. They use tweet sending in positive, negative, neutral. They build models using Naive Bayes, Support Vector Machines (SVM), and they report SVM outperforms other classiﬁers. They perform a different classiﬁcation task though: subjective versus objective.

In addition, explore a different method of data representation and report signiﬁcant improvement over the unigram models. Another contribution of this paper is that we report results on manually annotated data that does not suffer from any known biases. Our data will be a random sample of streaming tweets unlike data collected by using speciﬁc queries.

They use polarity predictions from three websites as noisy labels to train a model and use tweets for tuning and another tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. We extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classiﬁers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contribute to the classiﬁer accuracy.

One fundamental problem in sentiment analysis is categorization of sentiment polarity. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative(or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level. The document level concerns whether a document expresses negative or positive sentiment, while the sentence level deals with each sentence’s sentiment categorization. The entity and aspect level then targets on what exactly people like or dislike from their opinions.

# SYSTEM DESIGN

## Use Case Diagram

Diagram

Description automatically generated

Figure 1: Case Diagram

# DATA COLLECTION:

The dataset used in this study, dubbed "Twitter Sentiment," has 12 columns and a total of 30,124 tweets. The tweets included a mix of positive, negative, and neutral feelings, and they each included a confidence rating for the assigned label as well as the justification for a negative classification. The features that are included are tweet id, sentiment, confidence score in sentiment, negative reason, confidence in negative reason, airline, confidence in sentiment gold, name, retweet count, tweet text, tweet coordinates, time, date, and location of the tweet as well as the time zone of the user who posted it.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute’s Name | Count | Data Type | Null Values |
| User\_Name | 30124 | Object | 0 |
| User\_loaction | 18423 | object | 11701 |
| User\_description | 28968 | object | 1156 |
| User\_created | 30124 | object | 5462 |
| User\_followers | 30124 | int64 | 0 |
| User\_friends | 30124 | object | 0 |
| User\_favourites | 30124 | Int64 | 0 |
| User\_verified | 30124 | bool | 0 |
| Date | 30124 | object | 0 |
| Text | 30124 | object | 0 |
| Hashtags | 21687 | object | 8437 |
| Source | 30124 | object | 0 |
| Is\_retweet | 30124 | bool | 0 |

Table 1 :Data Information

# Data Visualization

Histogram

Description automatically generated

Figure 2: Top 20 Source

This Bar plot gives the information about of each source count. It Shows information of top 20 source used while entering the tweet on twitter.

Chart, bar chart

Description automatically generated

Figure 3: Sentiment Analysis

This plot explains that positive tweets were often as compared to negative tweets. However, apart from this Neutral sentiment almost equally contributed, due to which percentage of these elements can be more helpful.

|  |  |  |
| --- | --- | --- |
|  | **Total** | **Percentage (%)** |
| **Positive** | 13676 | 45.4% |
| **Neutral** | 12979 | 43.09% |
| **Negative** | 3469 | 11.52% |

Table 2 Total Number of Sentiments of Tweets with Percentage

The given table delineates the info about sentiments of the tweet and the percentage of tweets as from this graph, we can make conclusion that most of tweets are negative with 11.52%.

Chart, pie chart

Description automatically generated

Figure 4: Pie Chart for sentiment analysis

This pie plot shows the Percentage of each sentiment for better understanding of dataset. From here it can be observed that positive and neutral sentiment have just the difference of 2.31% .while, negative tweets with only 11.5% is at least frequent.

Chart, bar chart

Description automatically generated

Figure 5: Total Number of tokens in tweets for positive analysis

Total number of tokens used in sentiment tweets for positive: This figure, depicts about the positive token of all positive tweets

Chart, bar chart, histogram

Description automatically generated

Figure 6:Total Number of tokens in tweets for negative analysis

Total number of tokens used in sentiment tweets for negative tokens: This figure, tells us about the negative token of all negative tweets

Chart, histogram

Description automatically generated

Figure 7: Total Number of Neutral token

Total Number of Neutral tokens in all tweets for Neutral analysis

# Word Clouds:

Figure 8: Word cloud about the negative tweets.

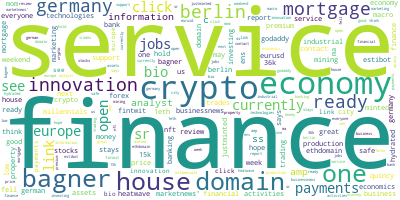


Figure 9: Word cloud about the positive tweets.



Figure 10: Word cloud about the neutral tweets.

# MATERIALS & METHODOLIGIES

The dataset used for sentiment analysis is described in this part, along with how it was visualised and the methodology that was suggested for applying sentiment analysis to the chosen dataset.

The dataset underwent pre-processing as part of the research's methodology procedures. This stage made use of a variety of programmes and libraries, including the Natural Language Toolkit. At the pre-processing level, this study took into consideration two strategies:

**Complete pre-processing:** Data cleaning was done during thorough pre-processing to increase the learning effectiveness of machine learning models. If the data are pre-processed, machine learning models exhibit increased classification accuracy. The Python's natural language toolkit was used for the pre-processing. Punctuation, stop-words, and the use of both lower- and capital letters in tweets can all have an impact on a model's ability to learn.

Punctuation was taken out of the data because it wasn't necessary for the study's text analysis. Although punctuation makes sentences more readable, it makes it harder for models to distinguish between punctuation and other letters [30]. The tweets' numerical values were eliminated in the following stage because they had no bearing on text analysis. deleting numbers from values decreases the complexity of training the models.

**Before:** @VirginAmerica I didn’t today... Must mean i need to take another trip for 2 months!

**After:** I did not today Must mean i need to take another trip for months

**Partial Pre-processing:** To examine the effects of pre-processing steps on classifier accuracy, this study also took into account the usage of partial pre-processing in addition to complete pre-processing. The partial pre-processing excludes "stop-words elimination" and stemming.

# Feature Extraction Methods:

The corpus was split into a "training subset" and a "testing subset" following the pre-processing stage. For training and testing, it was split in a 70:30 ratio, respectively. The training subset was then subjected to feature extraction techniques, as depicted in Figure 6, which illustrates the employed methodology. Both the training data used to develop the selected models and the testing data used for classification both underwent feature extraction techniques.

**TF-IDF:** A common scoring method in information retrieval (IR) and summarization is TF- IDF. The goal of TF-IDF is to demonstrate how pertinent a phrase is in a particular document. TF-IDF feature extraction takes both TF and IDF into account. IDF awards tokens that are uncommon across the board in a dataset. (term frequency (TF), inverse document frequency (IDF)).

Next, in this section compared bagging classifiers and non-bagging classification techniques.

# The classification techniques are

* **Support Vector Machine**
* **kNN (K Nearest Neighbour):**
* **Random Forest Classification**
* **Naive Bayes**

**Experimental Design:**

The dataset was split into train (70%) and test (30%) sets based on data for subjects, e.g. 21 subjects for train and nine for test.

* **training set**—a subset to train a model.
* **test set**—a subset to test the trained model.

**Classification Algorithms:**

Data is divided into several unique classes using the classification technique, and labels are given to each class. The primary goal of classification is to establish a class from which to launch fresh data by analysis of the training set and correct boundary identification. Predicting the target class and the procedure are collectively referred to as classification. The different types of classification methods used are discussed below:

# SVM(Support Vector Machine):

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example, linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. Non-linear data set are difficult to be separated using a linear hyperplane. SVM algorithm is related to finding the hyperplane which separates the data based on maximum margin.

# kNN (K Nearest Neighbour):

KNN Is a algorithm which used to predict categorical values. K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.

# Random Forest:

A "forest" of decision trees is created by the supervised machine learning technique known as random forest. For a more precise forecast, Random Forest produces numerous decision trees that are then combined.

# Naive Bayes

Naive Bayes classification is based on the Bayes theorem and takes the confidence between each pair of features into consideration. A limited amount of training data is required for Nave Bayes to measure the required parameters. Comparing this approach to more complex classifications, it is quick.

|  |  |
| --- | --- |
| Model | Score |
| SVM | 0.851381 |
| kNN | 0.753920 |
| Random Forest | 0.689320 |
| Naive Bayes | 0.781926 |

Table 3: Comparison of All models

**Chart, line chart

Description automatically generated**

Figure 11: Comparison of All models

# CONCLUSION:

To conclude the whole survey or project, as my dataset is “Twitter Sentiments”, as the name described that its about the sentiments of the people regarding their experience while using twitter. The dataset contains 30124records and 12 attributes from which have dropped some attributes and changed the datatype according to the need as well as replaced the null values. Post this, by removing stop-words and punctuations from the ‘text’ attribute, all in all text pre-processing or text-analysis is done based on the requirement. Then, to normalize the data TF-IDF vectorizer is used and for handling imbalance SMOTE is used.

Post that, data visualization is also done based on the research papers studied. Word cloud about the negative, neutral, and positive is described as well.

# References:

1. Jacobson, R. 2.5 *Quintillion Bytes of Data Created Every Day. How Does CPG & Retail Manage It*; IBM: Indianapolis, IN, USA, 2013*.*
2. Wang, Q.; Kealy, A.; Zhai, S*. Introduction for the Special Issue on Beyond the Hypes of Geospatial Big Data: Theories, Methods, Analytics, and Applications. Comput. Model. Eng. Sci. 2019, 119, 245. [CrossRef]*
3. Pang, B.; Lee, L. *Opinion mining and sentiment analysis.* Found. Trends Inf. Retr. 2008, 2, 1–135.
4. Rainie, L.; Horrigan, J. Election 2006 online. *In Pew Internet & American Life Project Report; Pew Research Center: Washington, DC, USA, 2007.*
5. Qaisi, L.M.; Aljarah, *I. A twitter sentiment analysis for cloud providers: A case study of Azure vs. AWS.* In Proceedings of the IEEE 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016.
6. Liza Wikarsa, SherlyNoviantiThahir, *“A Text Mining Application of Emotion Classifications of Twitter’s Users Using Naïve Bayes Method” .*
7. E.Prabhakar and K. Sugashini, *“New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data”,*
8. The SIJ Transactions on Industrial, Financial & Business Management (IFBM), Vol. 6, No. 1, 2018. Adeborna, Esi, and KengSiau. 2014. *"An approach to sentiment analysis-The case of airline quality rating."* Pacific Asia Conference on Information Systmem.
9. Nadia F.F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka, *"Tweet sentiment analysis with classifier ensembles”* Science direct, 2014.
10. D. D. Das, S. Sharma, S. Natani, N. Khare, and B. Singh*, “Sentimental Analysis for Twitter data,” in IOP Conf. Series:* Materials Science and Engineering, 2017*.*