**Twitter Sentiment Analysis**

**By:**

Ruksha Dahal

Pushpalatha

Pawan Sai

KambakkamYeswanth

Contents

[1. Abstract 1](#_Toc120965398)

[2. Introduction 1](#_Toc120965399)

[3. Motivation 1](#_Toc120965400)

[4. Literature Review 2](#_Toc120965401)

[5. Proposed Methodology 3](#_Toc120965402)

[5.1 Data Collection 3](#_Toc120965403)

[5.2 Tools Used 3](#_Toc120965404)

[5.3 Implementation: 4](#_Toc120965405)

[5.3.1 Data Preprocessing: 4](#_Toc120965406)

[5.3.1.2 Steps involved in Data Preprocessing: 4](#_Toc120965407)

[5.3.2 Feature extraction: 4](#_Toc120965408)

[5.3.2.1 TF-IDF: 5](#_Toc120965409)

[5.3.2.2 PCA: 5](#_Toc120965410)

[5.3.2.3 Classification: 5](#_Toc120965411)

[5.3.2.3.1 Logistic Regression: 5](#_Toc120965412)

[5.3.2.3.2 Bernoulli Naive Bayes 5](#_Toc120965413)

[5.3.2.3.3 Random Forest Classifier: 5](#_Toc120965414)

[5.3.2.4. Model Evaluation: 5](#_Toc120965415)

[5.3.2.4.1 Evaluation metrics: 5](#_Toc120965416)

[6. Result and comparison: 6](#_Toc120965417)

[6.1 Model and Accuracy: 6](#_Toc120965418)

[6.2 Classification Report 7](#_Toc120965419)

[6.3 Confusion matrix: 8](#_Toc120965420)

[7. Conclusion 9](#_Toc120965421)

[8. References: 10](#_Toc120965422)

# 1. Abstract

Sentiment analysis is a natural language processing task that analyses human sentiments as positive, negative, or neutral. Twitter sentiment analysis is the application of sentiment analysis that allows users to express their sentiments and opinions in the form of tweets and contains a vast number of tweets. With sentiment analysis, we can keep track of customer trends and interests, and it would be highly beneficial for the organization to improve customer service, grow its influence and understand the brand more deeply. Moreover, with respect to large companies, it is necessary to analyze hundreds of thousands or even millions of comments made about various items and is very crucial to automatically predict whether the customers are satisfied or not. The challenge of identifying the emotional tone of a text has been resolved with the development of machine learning and sentiment analysis. The main aim of this project is to perform a sentiment analysis on 1.6 million Twitter data and classify the tweet as positive or negative with different feature extraction techniques and classifiers.

Sentiment analysis is a natural language processing task that involves analyzing human emotions expressed in text as positive, negative, or neutral. This analysis can help organizations to monitor customer trends, improve customer service, gain a deeper understanding of their brand, and expand their influence. For large companies, it is necessary to analyze large volumes of comments to predict customer satisfaction. With machine learning and sentiment analysis, identifying the emotional tone of text has become easier. The objective of this project is to conduct sentiment analysis on 1.6 million Twitter data and use different techniques to extract features and classify the tweets as positive or negative.

# 2. Introduction

With the help of the Internet, people can convey their ideas globally through online discussion forums, blogs and other channels, and tweets. For instance, before purchasing a product, the customer will first read its reviews and comments. However, it would be impossible to read every single review manually and waste time on it. This process can be automated using Machine Learning, and sentiment analysis methods fall under machine learning, which helps the systems to understand the sentiment behind an utterance. The sentiment analysis will help the business to gather data about its products and get an insight into the product's performance, which would help them scale up its business. Sentiment analysis is important as any organization must understand its customers' needs, opinions, and level of satisfaction with its products.

The proposed solution is to perform sentiment analysis on Twitter data as Twitter is one of the well-known sources that contain a significant amount of data in the form of tweets, as people can share their feelings and thoughts quickly by sending tweets, making it the ideal source to discover people’s opinions on different matters. We have performed sentiment analysis on 1.6 million Twitter data and have classified the tweet as positive or negative. The various preprocessing steps are applied to the Twitter data to generate a cleaner tweet. The features from the text are extracted using the TF-IDF feature extraction method. Moreover, we applied a PCA over the TF-IDF to extract the valuables features from the TF-IDF feature vector. Using PCA was beneficial as the computational time was reduced significantly. We have used three classifiers and build the model using TF-IDF and PCA. Out of all the classifiers, the logistic regression with the TF-IDF feature extraction method outperforms with an accuracy of 77%.

# 3. Motivation

Millions of people use social network sites to express their emotions and opinions and update their daily lives. We were more inclined to the Twitter analysis as it provides a more accurate representation of public sentiment than traditional online articles and web blogs. Public sentiment analysis is crucial when predicting a certain company's stock market rate and detecting angry customs or negative mentions before they escalate. Since Twitter allows us to download streams of geo-tagged tweets for specific regions, businesses may also estimate how well their product is performing in the market and which sections of the market are it having a favorable reaction and in which an inadequate response. If businesses can gather this data, they may evaluate the causes of regionally diverse responses and sell their products more effectively by coming up with relevant solutions like forming appropriate market groups. Another emerging use for sentiment analysis is making predictions about the outcomes of popular political elections and surveys. In one such study, which was carried out in Germany to forecast the results of federal elections, Andranik Tumasjan et al. concluded that Twitter is a decent representation of offline sentiment [[1].](#a1)

# 4. Literature Review

This section explains the related work done in this sentiment analysis and gives a quick overview of various approaches implemented for Twitter sentiment analysis.

Kaur, H.J. et al. [[2]](#a2) described the workflow for performing sentiment analysis in a new domain of natural disasters to detect public emotions in a crisis. The data is collected from Twitter using the hashtags #kashmirfloods and #jammuflods. The pre-processed data is annotated with the emotional categories as negative, positive, and neutral. For the feature extraction, a baseline model is created using a combination of unigrams and Part of speech features and the PCA technique is used for the feature reduction. They used the Naive Bayes Classifier to classify information and achieved an overall accuracy of 66.88%.

In this paper, Sai Ramesh et al. [[3],](#a3) performed sentiment analysis in Twitter data to analyze the users’ opinion of the election status between Hillary and Trump. The Twitter data was extracted using the Twitter streaming API, and a predictive model is developed for the analysis, consisting of several predictors that influence future behaviour or results. Linear regression is used for the predictive modelling approach, where the linear equation and neural network are implemented in python. With the 10-fold execution pattern, 85.23% accuracy is achieved, and this data analytic approached performed better than the support vector machine and the naïve Bayes approach.

The author Amrita Shelar et al. [[4]](#a4) aim to perform the Twitter sentiment analysis to explore the Twitter data with the tweets related to donations, fundraising or charities. The sentiment analysis was done using NLTK 2.0.4 powered text classification process. The text was categorized and identified using three positive, negative, or neutral sentiments by NLTK’s VADER. The VADER sentiment analysis handles the data in its lexicon by including all sorts of letters, and symbols, and discovers the sentiments of people in the form of polarity.

In this paper [[5],](#a5) the author Dr K. Maheswari. et al., aims to classify the tweets with respect to the sentiment value and sentiment using the KNN algorithm. The feature selection algorithm was applied to the dataset to identify the important attribute. Weights were assigned to each of the k points according to their distance from the test point for improving the KNN algorithm. To select the optimal model using the largest value, accuracy is used, and the accuracy of the sentiment is 36%.

R. Darsini et al[. [6],](#a6) performed the Twitter sentiment analysis using a machine learning method and python. The main agenda of this analysis is to determine the positive, negative, and neutral tweets by performing text analysis. Further, using more methods, they improved the existing sentiment analysis models of SVM. The proposed machine learning classifiers are efficient and have achieved 71% accuracy.

Shilpi Sharma et al. [[7]](#a7) worked on Twitter sentiment analysis using the Hadoop system that efficiently analyses a large amount of data in the Hadoop cluster. They aim to establish the actual line for automating the cluster setting. They used Hadoop with Map reduce to produce the excel graph, and in the future, this can be expanded into more analysis, such as parsing, and theme modelling.

Jaspreet Singh et al[. [8]](#a8) aims to find positive or negative polarities on social media reflecting people’s viewpoints. They have explained the text posted on social media platforms of sentiment analysis via NLP and have used four machine-learning models Naïve Bayes, J48, Tree and OneR. The Naïve Bayes was found to be quick, and OneR produced 91.3 % precision accuracy. They analyzed that deep neural networks with word embeddings can be used for enhancing the pre-preprocessing step.

# 5. Proposed Methodology

## 5.1 Data Collection

The dataset used in this project is taken from Kaggle, which contains 1.6 Million tweets (238.8 MB in size) in which the positive values are “800k”, and negative values are “800k”. The tweets have been annotated (0 = negative, 4 = positive) and can be used to detect the sentiment.

Chart, pie chart

Description automatically generated

Figure: equal number of positive and negative tweets

According to the creators of the dataset:

"Our approach was unique because our training data was automatically created instead of having humans manually annotate tweets. In our approach, we assume that any tweet with positive emoticons, like :), were positive, and tweets with negative emoticons, like :(, were negative. We used the Twitter Search API to collect these tweets by using keyword search".[9]

## 5.2 Tools Used

* Google Collab
* Python
* SKlearn

## 5.3 Implementation:

This project presents the idea of performing Twitter sentiment analysis using the machine learning method, and the following steps were performed for the implementation of this project.

## 5.3.1 Data Preprocessing:

The Twitter data taken from Kaggle needs to be preprocessed, as data preprocessing is an important step in any pattern recognition and machine learning problem as no data are perfectly fine and clean to use directly. For the Twitter sentiment analysis project, as we are using text data and using the below pre-processing steps for the sentiment analysis.

### 5.3.1.2 Steps involved in Data Preprocessing:

* + Removed the null values as no models handle the NULL values on their own.
  + Performed the tokenization process on the text data, as this process will help us better understand the data to develop the model. Tokenization is the process of splitting raw data into a list of tokens.
  + Lemmatization process was performed on the tokenized data as it allows to map of the multiple words to the common root words; as a result, the words are treated similarly.
  + The stop words were removed as these words repeatedly appear in the text. However, they do not add much value to it. Removing them will shift the focus to more unique data that holds significant information.
  + All the special characters/punctuation were removed in the preprocessing step, which is an important step the data preprocessing, and a cleaner tweet was generated with lowercase characters.

Diagram

Description automatically generated

Figure: Proposed Methodology

## 5.3.2 Feature extraction:

After completing the data preprocessing step, we apply the feature extraction technique to extract the features from the textual data. Feature extraction helps to identify the useful features reducing the dimension of features to be used in training machine learning model. Thus, this also reduces computational complexity at the end. With the feature extraction techniques, we converted the text into a matrix of features. Some of the popular feature extraction techniques are Bag-of-words, TF-IDF, Word2vec, and PCA. In our project, we have used the below feature extraction method.

### 5.3.2.1 TF-IDF:

We have used TF-IDF as the feature extraction method where the TF-IDF stands for term frequency-inverse document frequency. The idea behind the approach is to determine how relevant those words are to a given document by using the frequency of words. They can be calculated as:

-TF = (Number of times term t appears in a document)/ (Number of terms in the document)

-IDF = log(N/n), where N is the number of documents and n is the number of documents a term t has appeared in. TF-IDF = TF\*IDF

### 5.3.2.2 PCA:

In our project, we have applied PCA over TF-IDF to extract the meaningful features from the TF-IDF feature matrix. PCA is the method of obtaining important variables in the form of components from the large set of variables available in a dataset.

## 5.3.2.3 Classification:

In our model-building phase, we have split the train-test data as 70% for training and 30 % for testing. We have used three models named Logistic regression, Bernoulli Naive Bayes and Random Forest Classifier and build these models with PCA and TF-IDF feature extraction methods, separately.

5.3.2.3.1 Logistic Regression: Logistic regression is a classification algorithm used to predict the binary outcome based on the number of independent variables [[9]](#a9). The independent variables can be continuous, discrete ordinal and discrete nominal.

5.3.2.3.2 Bernoulli Naive Bayes: Bernoulli Naive Bayes performs well with binary classification problems, a variant of Naïve Bayes. Here, the features are in binary form. With respect to this algorithm, the features are accepted as binary values like true or false,0 or 1 etc.

5.3.2.3.3 Random Forest Classifier: For classification problems, a random forest classifier can be used, and it makes use of many decision trees to make predictions. The missing value is handled by this classifier by using median values to replace the continuous variables, and the proximity-weighted average of the missing value is computed.

## 5.3.2.4. Model Evaluation:

Once the model is build, we evaluated all these models with PCA and TF-IDF on the test data, respectively, in terms of accuracy, confusion matrix, F1 score, precision and recall. Out of all the models used, Logistic regression with the TF-IDF feature extraction method outperformed with an overall accuracy of 77%.

### 5.3.2.4.1 Evaluation metrics:

Accuracy: The ratio of the number of accurate predictions to the total predictions is known as accuracy.

Precision: For predicting the specific category, precision tells how well the model is good at it.

Recall: For detecting a specific category, Recall specifies how many times the model was able to detect the category.

F1 Score: F1 score performs well in the case of imbalanced data set by considering both false positive and false negative and taking the mean of precision and recall.

Confusion Matrix: A confusion matrix is a simple matrix that defines the model performance on the test data. It is the combination of predicted and the actual values. The basic terminologies in the confusion matrix are:

* True positives (TP)
* True negatives (TN)
* False positives (FP)
* False negatives (FN)

Fig: Confusion Matrix

# 6. Result and comparison:

After the model evaluation phase, we compared the accuracy of all the models by applying TF-IDF alone and PCA, and the below results are observed:

## 6.1 Model and Accuracy:

Chart, bar chart

Description automatically generated

Fig: Models and accuracy

From the above bar diagram, we can observe that the **Logistic regression model with the TF-IDF feature extraction method is the best performing model** as compared to the other models with an overall accuracy of 77%. We can also observe that while applying PCA, the accuracy has slightly decreased as using PCA, we can lose some spatial information that can be important for the classification. However, PCA reduces the computation time.

## 6.2 Classification Report

Classification report is used to measure the quality of predictions from the classification algorithm. Different model classification reports in terms of precision, recall and F1-score are shown below. Out of all the models used, Logistic regression with the TF-IDF feature extraction performs the best.

In the below table, we can see that for the logistic regression with TF-IDF model, the class negative sentiment has a precision of 0.78 which means that when it predicts the tweets as negative it is correct 78% of the time and recall of 0.76 means that it correctly classifies 76% of negative tweets. Likewise, the class positive sentiment has a precision of 0.77 which means that when it predicts the tweets as positive it is correct 77% of the time and recall of 0.78 means that it correctly classifies 78% of positive tweets. While the F1-score is the mean of precision and recall.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Feature extraction method | Class | Precision | Recall | F1-Score |
| Logistic Regression | TF-IDF | Negative Sentiment (0) | 0.78 | 0.76 | 0.77 |
| Positive Sentiment (1) | 0.77 | 0.78 | 0.77 |
| PCA | Negative Sentiment (0) | 0.74 | 0.75 | 0.74 |
| Positive Sentiment (1) | 0.75 | 0.75 | 0.75 |
| Bernoulli Naïve Bayes | TF-IDF | Negative Sentiment (0) | 0.77 | 0.74 | 0.75 |
| Positive Sentiment (1) | 0.75 | 0.78 | 0.76 |
| PCA | Negative Sentiment (0) | 0.69 | 0.70 | 0.69 |
| Positive Sentiment (1) | 0.69 | 0.68 | 0.69 |
| Random Forest Classifier | TF-IDF | Negative Sentiment (0) | 0.75 | 0.77 | 0.76 |
| Positive Sentiment (1) | 0.76 | 0.74 | 0.75 |
| PCA | Negative Sentiment (0) | 0.68 | 0.71 | 0.70 |
| Positive Sentiment (1) | 0.70 | 0.67 | 0.69 |

Fig: Classification report

## 6.3 Confusion matrix:

Chart, waterfall chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Graphical user interface, chart, application

Description automatically generated

Fig: Confusion Matrix

From the above confusion matrix, we can observe that Logistic regression model with TF-IDF model is the best performing model on the test data as it has correctly predicted the positive value [True Positive]as 39.2% and the negative value [True negative] as 38.2 %. However, the model has incorrectly classified the negative value as positive [False positive] with 11.80% and the positive value as negative [False negative] with 10.94%.

# 7. Conclusion

In this project, we discussed the application of sentiment analysis in different areas and its importance. The sentiment analysis is important as Companies can utilize it in the marketing industry to create their plans, understand how consumers feel about their items or brands, how people react to their advertisements or new product releases, and determine why some products aren't purchased by consumers. Our focus was on Twitter sentiment analysis since it is a well-known micro-blogging site that includes human expressions of liking, disliking, and comments on various issues in the form of tweets. We pre-processed the Twitter data, applied PCA and the TF-IDF feature extraction method, and compared the result with the three models. With PCA, the computation time was reduced during the model training phase in comparison with TF-IDF. Out of all models, the Logistic regression model performed well with the TF-IDF feature extraction method with an overall accuracy of 77%.

# 8. References:

[1] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[2] Kaur, H. J. (2015), Mamgain, (2016). Anto, M. P. (2016). N. M. Sentiment Analysis from Social Media in Crisis Situations.

[3] A Razia Sulthana, A K Jaithunbi, L Sai Ramesh P., 2018. Sentiment analysis in twitter data using data analytic techniques for predictive modelling. National Conference on Mathematical Techniques and its Applications (NCMTA 18) IOP Publishing. IOP Conf. Series: Journal of Physics: Conf. Series 1000 (2018) 012130 doi :10.1088/1742-6596/1000/1/012130

**[4]** Amrita Shelar and Ching-yu Huang. P., 2018.Sentiment Analysis of Twitter Data.2018 International Conference on Computational Science and Computational Intelligence (CSCI).

**[5]** Dr. K. Maheswari. P., 2018.Improving Accuracy of Sentiment Classification Analysis in twitter Data Set Using knn.E ISSN 2348 –1269, PRINT ISSN 2349-5138

[6] S.Siddharth, R. Darsini, Dr. M. Sujithra. P., 2018. Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python. Research Gate.

[7] Malika Acharya, Shilpi Sharma. P., 2018.Semantic Analysis of Twitter Posts.978-1-5386-7709- 4/18/$31.00 c 2018 IEEE

[8] Jaspreet Singh, Gurvinder Singh, and Rajinder Singh. P., 2017.Optimization of sentiment analysis using machine learning classifiers. Singh et al. Hum.Cent. Comput.Inf. Sci. (2017) 7:32 DOI 10.1186/s13673- 017-0116-3.

[9] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc