**TWITTER SENTIMENT ANALYSIS**

Harshitha, HK, and Kodali

George Mason University, hkodali@gmu.edu

Sri Vidya, SVG, and Garimella

George Mason University, gsrividy@gmu.edu

Vinay Kumar, VKV, and Vishwakarma

George Mason University, vvishwak@gmu.edu

The objective of our project is to analyze sentiment of twitter. Twitter is one of the popular social media services, where people share their thoughts and connect with other people. In our day-to-day life, we have been observing huge abusive comments in many social media platforms. In-order to report these comments, we are trying to build a classifier model to distinguish between “Abusive” and “Non-Abusive” comments. Now-a-days we are encountering many hateful comments in social media websites, so our primary goal of the project is to implement different classification models which can analyze abusive comments.

**Additional Keywords and Phrases**: Logistic Regression – LR, Random Forest Classifier – RFC, 'negative':0, 'neutral':1, 'positive':2

ACM Reference Format:

Harshitha, HK, Kodali, Sri Vidya, SVG, Garimella and Vinay Kumar, VKV, Vishwakarma. 2022.

Twitter Sentiment Analysis: ACM Conference Proceedings Manuscript Submission Template: An analysis performed on the Twitter’s tweet dataset for predicting sentiment analysis of the tweet.

1. Introduction

Internet and social media are powerful instruments for mobilization of people. They are gaining popularity as they allow people to share and express their thoughts and post messages across the world. There is a huge data present and generated in the social networking sites like Twitter, Facebook, Google etc. Twitter is one of the social media that is gaining popularity. Twitter offers organizations a fast and effective way to analyze customers perspective towards the success in the marketplace.

Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. Our main motivation is on sentiment analysis of twitter data which is helpful to analyze the information in the tweets their opinions are highly unstructured, heterogeneous and are either “Positive”, “Negative” or “Neutral”. It can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets, and database sources through Natural Language Processing (NLP). Sentiment analysis also includes many tasks such as sentiment extraction, sentiment classification, summarization of opinions or opinion spam detection among others. It aims to analyze people’s sentiments, attitudes, emotions towards elements such as products, individuals, topics, organizations, and services.

Our dataset contains four Columns/Labels which are ID, Comments on the twitter, Selected Abusive comments, and Sentiment of the comments. Our Specific question is “Can we build a strong classifier to predict Abusive, Non-Abusive and Neutral comments using Logistic Regression and Random Forests classifier?”. Our approach is to pre-process the data and fit the data into our model for training and predict our results, based upon our results we try to change the data and the pre-processing techniques to improve our accuracy. We are evaluating our project using confusion matrix which is used to describe the performance of the classification model. With the help of confusion matrix, we were able to calculate the accuracy, recall, precision, and f1-score.

1. mETHOD

We primarily got our dataset from Kaggle, and Columns/Labels are part of the dataset. Our dataset contains four Columns/Labels which are listed below:

ID – Unique ID

Comments on the twitter – The text of the tweet

Selected Abusive comments – Particular comments from the tweet

Sentiment of the comments – The respective sentiment of a tweet

The dataset contains all the comments which are registered on Twitter and predicts & lists them out as “Positive”, “Negative”, and “Neutral”. We are splitting entire dataset as 75% Training data and 25% as Test Data, and we will evaluate our model using 25% of Test Data.

* 1. Python Modules

A python module is a file containing python definitions and statements.

**NumPy –** It is used to calculate mathematical functions like linear algebra, matrices etc.

**Pandas –** It is used to analyze data and working with data sets.

**Nltk (Natural Language Toolkit) –** It is used to work with human language.

**Re (Regular Expression) –** It is a sequence of characters that forms a search pattern and checks if a string contains a specified search pattern.

**Seaborn –** It is library that uses a Matplotlib underneath to plot graphs. It is used to visualize random distribution.

**Standard Scaler –** It is used to standardize the data values into a standard format.

**Matplotlib –** It is a low-level graph plotting library that serves as a visualization utility.

* 1. Data Pre-processing

Data mining is the process of converting raw data into useful information that can be further analyzed. Raw data can be sometimes in a cluttered condition that is completely unusable. So, we need to pre-process the data before further analysis.

**Stop-words –** They are Common words in English language, their removal can help to focus more on important keywords in the text.

**Removal of punctuation –** It is important to remove punctuation because they do not contribute much to the meaning of the text

**Stemming –** It is a process of producing more morphological variants of a root words. It is a technique used to extract the base form of the words by removing affixes from them. Stemming is desirable as it may reduce redundancy as most of the time the word stem and their derived words mean the same.

**Lemmatization –** This process contrasts with stemming. It looks beyond word reduction and considers a language’s full vocabulary to apply morphological analysis to words, aiming to remove inflectional endings only and to return the base form of the word.

* 1. Vectorization Methods[1]

**TFIDF (Term Frequency Inverse Document Frequency) –** It is usually one of the best metrics to determine if a term is significant to a text. It is a product of Term Frequency (TF) and Inverse Document Frequency (IDF). The term Frequency refers to the number of occurrences of a word in the selected bodies of text. Inverse Document Frequency is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

**Bag of Words (BOW) –** It is a method often used for document classification. This method turns text into fixed length vectors by simply counting the number of times a word appears in a document.

**N-gram –** They are a set of co-occurring words within a given window and when computing the n-grams we typically move one word forward.

* 1. Classification Models

We are using Logistic Regression and Random Forests classifiers to implement this project.

**Logistic Regression [2] -** It is a simple and more efficient method for binary and linear classification problems. Logistic Regression is a process of modelling the probability of a discrete outcome given an input variable. It essentially uses a logistic function / sigmoid function (1 / 1 + e-x) to model a binary output variable. The primary difference between linear regression and logistic regression is that logistic regression’s range is bounded between 0 and 1 and it does not require a linear relationship between input and output variables. This is due to applying the non-linear log transformation to the odd ratio. Logistic Regression uses a loss function referred as “Maximum Likelihood Estimation” (MLE) which is a conditional probability. If the probability is greater than 0.5, then the prediction will be classified as class 0, otherwise class 1 will be assigned .

Chart, histogram

Description automatically generated

**Random Forests –** A Random Forest [3] is a supervised machine learning algorithm that is constructed from decision tree algorithms. It is mainly used to solve regression and classification problems. A random forest algorithm consists of many decision trees. The “forest” generated by random forest algorithm is trained through bagging or bootstrap aggregation. Bagging creates a training subset from sample training data with replacement and the final output is based on majority voting. In Random Forest n number of random records are taken from the dataset having k number of records Individual decision trees are constructed for each sample. Each decision tree will generate an output. Final output is considered based on majority voting.

1. RESULTS   
     
   We performed our testing using the 3 different vectorization techniques on 2 classification models. Since our data is not balanced, we use a confusion matrix to evaluate our models. A confusion metrics describes the performance of the classification model generated using the set of data. It not only gives insight into the errors being made by our classifier but also the types of errors that are being made. We found some interesting results. Table 1 shows the evaluation of datasets against each vectorization technique. In Figure 1, we present a confusion metrics of RFC using TF-IDF vectorization technique.

Table 1: Dataset Performance against Classification Models

| S. No. | Classification Model | Accuracy | F1-score |
| --- | --- | --- | --- |
| 1. | LR (Bag of Words) | 68.5% | 68.5% |
| 2. | RFC (Bag of Words) | 68% | 68% |
| 3. | LR (TF-IDF) | 68.4% | 68.4 |
| 4. | RFC (TF-IDF) | 70.1% | 70% |
| 5. | LR (N-grams) | 69.6% | 69.6% |
| 6. | RFC (N-grams) | 67.4% | 67.3% |

RFC model performed best when we used TF-IDF vectorization technique. However, performance of the models was nominal when we used other vectorization technique.

A picture containing text

Description automatically generated Chart

Description automatically generated

Figure 1: Confusion Metric and RUC curve of RFC model for TF-IDF vectorization

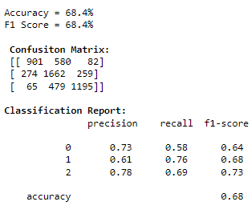
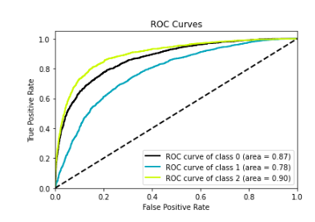
 

Figure 2: Confusion Metric and RUC curve of LR model for TF-IDF vectorization

From the above classification report, we can see that the model has slightly higher precision for positive (2) sentiment as compared to the others. It almost averages out when we calculate the f1-score. And we get almost the same f1-score for both ‘neutral’ and ‘Negative’ sentiment.

In Figure 1, we also see the visualization of ROC curve. ROC curve depicted the rate of ‘True Positive’ with respect to ‘False Positive’ and helps us to predict the performance of the classification algorithm.

1. CONCLUSION

In this project we have learnt different pre-processing techniques like Stemming, Lemmatization, Stop Words

which can improve our accuracy. As the data is huge, by using these pre-processing techniques we were able to

clean the data. Testing with different classification models gave us a full insight of how the models performs on

huge data.

Initially, we were not aware of many vectorization methods. But after research we came to know about various methods like Word2Vec (By google), CBOW, GloVe FastText, BOW , TFIDF and N-gram which converts text data into vectors . Out of all we used few techniques which helped us improve our accuracy.

1. REFERENCES

[1]. <https://neptune.ai/blog/vectorization-techniques-in-nlp-guide>  
[2]. [https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html](https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/)  
[3]. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>  
[4]. [https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification](https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/)

1. Video Link

<https://web.microsoftstream.com/video/fe93430e-0458-46e1-a58e-1a407d38c3c5>