NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

DATA MINING LA 2 REPORT

on

AIRLINE SENTIMENTAL ANALYSIS ON TWEETS

Submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

Darshan MR	1NT19CS063
JVSS Pavan Kumar	1NT19CS087
Vishesh K R	1NT19CS217
Saiel K Gaonkar	1NT19CS166

Under the Guidance of

Dr. Vani V

DESIGNATION, Dept. of CS&E, NMIT

NITTE EDUCATION TRUST

Department of Computer Science and Engineering

(Accredited by NBA Tier-1)

2021-2022

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM,
APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

CERTIFICATE

This is to certify that the **Video Summarization on e-sport** is an authentic work carried out by **Darshan MR (1nt19cs063)**, **JVSS Pavan Kumar (1ntcs087), Vishesh K R (1nt19cs217) and Saiel K Gaonkar (1nt19cs166)** bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belgavi during the academic year ***2021-2022.*** It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report. This project has been approved as it satisfies the academic requirement in respect of project work presented for the said degree.

Guide

Dr. Vani V

Professor, Dept CSE, NMIT

Bangalore

DECLARATION

We hereby declare that:

(i) The project work is our original work

(ii) This Project work has not been submitted for the award of any degree or examination at any other university/College/Institute.

(iii) This Project Work does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This Project Work does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a) their words have been re-written but the general information attributed to them has been referenced;
   b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

NAME	USN	Signature
Darshan MR	1NT19CS063	
JVSS Pavan Kumar	1NT19CS087	
Vishesh K R	1NT19CS217	
Saiel K Gaonkar	1NT19CS166	

Date: 17/01/2022

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Sarojadevi H.** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our ***Dr. Vani V, Professor***, Department of Computer Science & Engineering on **her** periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

NAME	USN	Signature
Darshan MR	1NT19CS063	
JVSS Pavan Kumar	1NT19CS087	
Vishesh K R	1NT19CS217	
Saiel K Gaonkar	1NT19CS166	

Date: 17/01/2022

ABSTRACT

Sentiment analysis is the computational study of people's opinions, attitudes, sentiments and emotions expressed in written language. It is one of the most active research areas in text mining and natural language processing in recent years. It has a wide range of applications since opinions are central to almost all human activities and are key influencers of our behaviors. We want to hear others' opinions, whenever we need to make a decision. Companies can use sentiment extremity and opinion point acknowledgment to pick up a more profound comprehension and the general extent of estimations. These experiences can progress focused insight, enhance client benefit, accomplish better brand picture, and upgrade competitiveness. Python sentiment analysis is a methodology for analyzing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text. Data mining has been used to produce the presents positive, negative sentiment, and their correlation about customer tweets. Here we have implemented associative rule mining and Random Forest Classifier algorithm for prediction before which data pre-processing has been done in order to obtain the training and test models.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Motivation

In the recent years, we have been witnessing the explosion of what is usually called participatory sensing. Ordinary people take a proactive role in publishing comments and complaining online, increasingly using technology to record information about events and problems in all dimensions of their political and social life. Data collection and opinion mining approaches are seen as the cornerstones of large-scale collaborative policy-making.

1.2 Problem Domain

Here the problem domain is based on the inability to perform well in different domains, inadequate accuracy and performance in sentiment analysis based on insufficient labeled data, incapability to deal with complex sentences that require more than sentiment words and simple analyzing.

1.3 Aim and Objectives

The main aim of this project is to make the sentimental analysis based on the data given from the particular dataset taken. We analyzed the data set and applied tf-idf algorithm for text conversion and Random Forest Classification model for training the dataset and the later on we predict the accuracy and f1 score.

CHAPTER 2:DATA SOURCE AND DATA QUALITY

2.1 Dataset used

A Data set is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set. Here we import the dataset repository and then display the dataset in the compiler.

We have 11 attributes in the dataset as listed below:

1.**tweet_id** : This column contains the unique id of every user.

2.**airline_sentiment** : This column says whether the tweet was negative, positive or neutral.

3.**airline_sentiment_confidence** : contains the confidence of the analyzed tweets.

4.**negativereason**: This column tells the negative word that was present in the tweet.

5.**negativereason_confidence**: This column displays the confidence of negativereason.

6.**airline**: Contains the name of the airline.

7.**name** : The column contains the name of the twitter user .

8.**retweet_count** : This column displays the count of retweets.

9.**text** : This column contains the texts of the tweets based on which the sentimental analysis was carried out.

10.**tweet_created** : This column contains the date and time of creation of tweets.

11.**user_timezone** : This column contains the timezone of the user who had tweeted.

2.2 Data Preprocessing

Preprocessing the dataset is a very important part of the analysis, it is used to remove outliers and duplicates from the dataset. Data cleaning is important as, it helps in faster convergence and give better results.

Steps involved in data cleaning are:

- Identifying the duplicate values
- Drop the duplicate values
- Remove Nan values from dataset
- Perform transformation

CHAPTER 3: INTRODUCTION

3.1 Data Mining Questions

Using the dataset, we can find:

- The different type of airline companies used for analyzing the tweets.
- Retweets count
- Total Number of tweets.
- The total number of positive ,negative and neutral tweets of each airline.

And using the above data of tweets, the user can select the best airline for travelling.

3.2 Data Mining Algorithms

Here we have used the TF-IDF(Term Frequency Inverse Document Frequency) algorithm to transform the text into meaningful representation of numbers . It's an effective way of distilling and manageable abstraction. The term tf is called as the term frequency and the document will count how many times it shows the document. The df word is called as the document frequency and many times it will show in all the documents. It transforms the count matrix to normalize or tf-idf.

This transformed dataset is used by datamining model for prediction the accuracy of model.

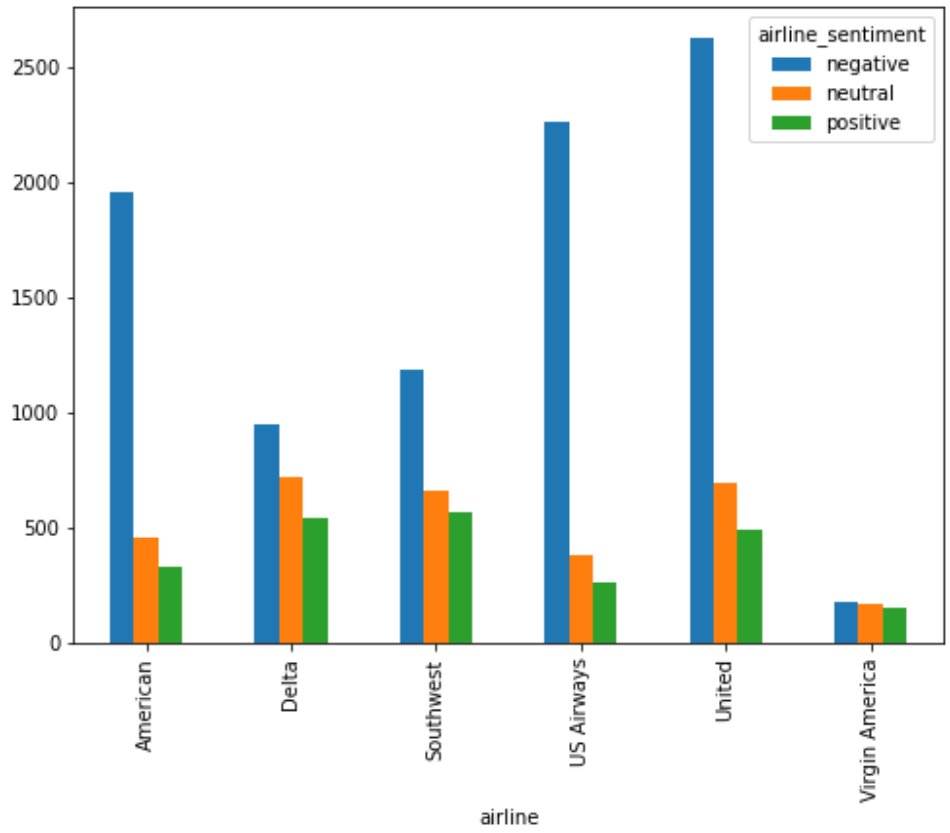
3.3 Data Mining model

We use RandomForestClassifier for training the transformed inputs. The random forest classifier is a supervised learning algorithm which you can use for regression and classification problems. It is among the most popular machine learning algorithms due to its high flexibility and ease of implementation.

It is an ensemble algorithm. The algorithm generates the individual decision trees through an attribute selection indication.

Chapter 4: Model Evaluation and Discussion

The following image represents the number of positive, negative and neutral tweets of all 6 airlines.




```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions))
print(accuracy_score(y_test, predictions))
```

```
[[1723  108   39]
 [ 326  248   40]
 [ 132   58  254]]
              precision    recall  f1-score   support

    negative       0.79      0.92      0.85      1870
     neutral       0.60      0.40      0.48       614
    positive       0.76      0.57      0.65       444

    accuracy                           0.76      2928
   macro avg       0.72      0.63      0.66      2928
weighted avg       0.75      0.76      0.74      2928

0.7599043715846995
```

The results of this analysis shows that we achieve an accuracy of 75.99% for sentimental analysis of tweets using TF-IDF algorithm and RandomForestClassifier model.

Chapter 5: Conclusion and Future Direction

We implemented Sentimental Analysis project using tf-idf algorithm and Random Forest classification model, where tf-idf algorithm is used for text conversion and Random Forest Classification for training the dataset, we then do data preprocessing for removing the duplicate values or data that is present in the dataset we can also try to implement other methods like svm(Support vector machine) and logistic regression for predicting the accuracy and f1 score for the same dataset. Future direction for this project can be done where it can analyze complex sentences and try to give more accurate solution for sentiment analysis.

Chapter 6: Reflection Portfolio

By doing this project we learnt about different algorithms and methods that are present in data mining and which can be used for sentimental analysis and how this algorithms and methods are used for the particular raw dataset to get the preprocessed dataset. Through this project we got an opportunity learn more about data mining algorithms and its methods and how sentimental analysis can be improved.

Chapter 7 : References

1. **https://pythonclass.in/sklearn-tfidfvectorizer.php**
2. **https://www.upgrad.com/blog/random-forest-classifier/**
3. **https://itechindia.co/blog/which-of-the-3-algorithms-models-should-you-choose-for-sentiment-analysis-2/**
4. **https://www.mindsmapped.com/data-pre-processing-in-python/**