DAYANANDA SAGAR UNIVERSITY

**KUDLU GATE, BANGALORE – 560068**



**Bachelor of Technology in**

**COMPUTER SCIENCE AND ENGINEERING**

**Mini Project Report on**

**Twitter US Airline Sentiment**

**Natural Language Processing (NLP)**

By

**Mohammed Irfan - ENG20CS0203**

Under the supervision of

## Prof. Sasikala N

## Assistant professor

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,**

## SCHOOL OF ENGINEERING DAYANANDA SAGAR UNIVERSITY

**(2023-2024)**



School of Engineering

Department of Computer Science& Engineering

Kudlu Gate, Bangalore –560068 Karnataka, India

**CERTIFICATE**

This is to certify that the Mini Project Report of Natural Language Processing titled **“Twitter US Airline Sentiment”** is carried out by **Mohammed Irfan(ENG20CS0203)** bonafide student of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2023-2024**.

|  |  |  |
| --- | --- | --- |
| **Prof. Sasikala** | **Dr. Girisha G S** | **Dr. Uday Kumar Reddy**  **K R** |
| Assistant Professor Dept. of CSE,  School of Engineering Dayananda Sagar University  Date: | Chairman, CSE School of Engineering Dayananda Sagar University  Date: | Dean  School of Engineering Dayananda Sagar University  Date: |

**DECLARATION**

I, **Mohammed Irfan (ENG20CS0203),** is student of the seventh semester B.Tech in **Computer Science and Engineering**, at School of Engineering, **Dayananda Sagar University**, hereby declare that the Natural Language Processing(NLP) Project titled **“Twitter US Airline Sentiment”** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2023-2024**.

|  |  |
| --- | --- |
| **Student** | **Signature** |
| **Name1: Mohammed Irfan**  **USN : ENG20CS0203** |  |
| **Place : Bangalore**  **Date :** |  |

**ACKNOWLEDGEMENT**

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this Project.

First, we take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor’s degree in this institution.

We would like to thank **Dr. Uday Kumar Reddy K R, Dean**, **School of Engineering & Technology**, **Dayananda Sagar University** for his constant encouragement and expert advice. It is a matter of immense pleasure to express our sincere thanks to **Dr. Girisha G S, Chairman**, **Department of Computer Science, and Engineering**, **Dayananda Sagar University,** for providing the right academic guidance that made our task possible.

We would like to thank our guide **Prof. Sasikala, Assistant Professor**, **Dept. of Computer Science and Engineering**, **Dayananda Sagar University**, for sparing his valuable time to extend help in every step of our Project, which paved the way for smooth progress and the fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course. We would like to thank one and all who directly or indirectly helped us in the Project.

**TABLE OF CONTENTS**

Page

LIST OF ABBREVIATIONS .......................................................................................... 6

[ABSTRACT .....................................................................................................................](#_TOC_250008) 7

[CHAPTER 1 INTRODUCTION………........................................................................ 8](#_TOC_250007)

[CHAPTER 2 PROBLEM DEFINITION………............................................................ 9](#_TOC_250006)

CHAPTER 3 LITERATURE SURVEY..................................................................…... 10

[CHAPTER 4 PROJECT DESCRIPTION.......................................................................](#_TOC_250005) 12

[CHAPTER 5 SCOPE OF PROJECT…………………………………………………..](#_TOC_250004) 13

[CHAPTER 6 REQUIREMENTS ..................................................................................](#_TOC_250003) 14

[CHAPTER 7 METHODOLOGY..................................................................................](#_TOC_250002).15

[CHAPTER 8 TESTING AND RESULTS](#_TOC_250001) 16-30

[REFERENCES... .............................................................................................................](#_TOC_250000) 31

**ABSTRACT**

Social media is a powerful source of communication. Information extracted and analysed from social media offers valuable insight for companies regarding their products and services. A customer’s experience is an important concern for the air-travel industry in particular, and Twitter has become a popular platform for travellers to share feedback. Airline companies invest time and resources to enhance customer loyalty. By exploring customer feedback, airlines can allocate resources to the weak areas of customer satisfaction, increasing economic and social development of the company. Without the use of machine learning and artificial intelligence, orthodox businesses spend hours of human effort completing manual annotations of tweets to group them into positive/negative sentiments, thus increasing the Time to insights.

**CHAPTER 1 INTRODUCTION**

We certainly consider online shopping, e-banking, etc. Correspondingly, The Car Rental System is the online office to book vehicles online inside hardly any snaps in a manner of speaking. A couple of individuals can't remain to have a vehicle, for those people this system ends up being outstandingly helpful. This framework incorporates different vehicles, according to the client request and solace it put in the request and got according to the pickup-sloping edge area inside the zone. Booking should be possible by means of network access as it were. This undertaking navigates plenty of territories going from business idea to registering field and required to play out a few explore to have the option to accomplish the venture goals The prior investigations demonstrated that management data system could be used to oversee vehicle rental, expected to quicken just as chronicling administrations to clients better and more secure, making it simpler when required whenever. The online execution of the executives data framework gave and bolstered the clients to reservations, help the executives in realizing rental vehicle stock at a predetermined time, to process exchanges between branches vehicle rental, transportation exchange handling, which underpins good assistance to clients and bolster the organization's operational procedures. Online vehicle rental data framework expands the clients and helps advancement. The point of this examination is to take care of the issues that happen in Avis Indonesia; propose advancement of the electronic vehicle rental administration data framework.

**CHAPTER 2 PROBLEM DEFINITION**

Interpreting and classifying opinions and emotions in subjective data (like tweets) is referred to as sentiment analysis. Sentiment analysis of customer experience is a hot topic and is applied in various industries, such as finance, business, politics, and risk assessment. Some researchers (Kandasamy et al., 2020) (Rezaeinia et al., 2019) (Yadav et al., 2020) have discussed various machine learning, artificial intelligence, and deep learning, and other techniques for sentiment analysis such as Word embedding, Deep Belief Network, Recursive Neural Network, Recurrent Neural Network, and approaches involving Refined Neutrosophic Sets.

While general frameworks have been explored, we will be utilizing these in conjunction with machine learning methods on a twitter sentiment analysis for this project. Keywords and key phrases can be a useful tool for analysing large amounts of textual material quickly by serving as a concise summary. International Encyclopaedia of Information and Library Science (Bolger et al., 1989) defines a “keyword” as a “word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document.” (Feather et al., 1996). Research suggests that people tend to be more willing to complain when they are stressed (Bolger et al., 1989) (Meier et al., 2013), and moods are affected by time (Ryan et al., 2010). Therefore, this project will provide airlines with a tool to calibrate and judge the positivity/negativity of the tweet based on the weekday (Sunday, Monday, …), which is an issue that has yet to be studied. We aim to perform text & sentiment analysis on extracted airline travel tweets while also taking into consideration when the tweet occurred, and whether it has resulted in a positive or negative effect.

**CHAPTER 3 LITERATURE REVIEW**

There have been several studies in the past which analysed the airline travel tweets across the globe. They had employed the usage of various methods and pipelines. They achieved varying reports of accuracy ranging between 50% to 90%. Several papers available in literature have analysed social media complaints such as (Tsakalidis et al., 2018). They proposed an unsupervised method, utilizing information from the network and time sensitive text. In comparison to simply text-based models, their examination in a real-time simulation environment revealed the efficacy and resilience of their technique versus competing baselines, attaining a substantial 20 percent boost in Fscore. The research in (Suryotrisongko et al., 2018) used a tweet dataset of 3000 rows which was of the feedback for Surabaya’s City Government. It was classified as a complaint or non-complaint tweet, and it resulted in systems that can categorise tweets automatically with an accuracy of 82.5 percent. The fuzzy system which was proposed in (Vashishtha et al., 2019) integrated NLP techniques with Disambiguation of Word Sense leveraging a new unsupervised 9 fuzzy rule-based system which classifies the post into either of positive/ neutral or negative sentiments. Their results helped in choosing which lexicons are best for the use case of social media. Their method of fusion of fuzzy logic with lexicons for sentiment classes can be adjusted to any lexicon. They can also be adjusted for any dataset. Recently, some papers have also analysed tweets of airline travel such as (Kumar et al., 2019), which used Python to retrieve tweets about airlines using twitter API and then used support vector machines and artificial intelligence networks methods to analyse those tweets. It demonstrated how to use ML to analyse tweets in order to improve the user experience. Word embedding with Glove dictionary technique and n-gram approach were used to extract features from tweets. CNN outperformed ANN and SVM models in this study, and it used association mining to map the link between tweets and sentiment categories. 8 In contrast, (Tiwari et al., 2019) proposed using machine learning to identify passenger tweets about aircraft services in order to better understand the emotional trend. On an actual dataset, they were able to reach an accuracy of around 80% using Random Forest (RF) and Logistic Regression (LR) to categorise each tweet into positive, negative, and neutral sentiment. In (Khan et al., 2018), the paper's major goal was to analyse tweets about airlines from four different regions: Europe, India, Australia, and America, with the goal of predicting consumer loyalty. The TextBlob analyzer was used to do the sentiment analysis.

The positive and negative mean sentiment ratings were then calculated and visually shown using the tweets. Random Forest, Decision Tree, and Logistic Regression were the three classifiers used. Random Forest has a maximum classification accuracy of 99.05 percent after 10-fold cross validation. There were other models based on Transformers. Like, for example, for Twitter sentiment analysis, (Naseem et al., 2020) employed Transformer-based Deep Intelligent Contextual Embedding. In this, Deep intelligent contextual embedding was used to improve the quality of tweets by reducing noise and taking into consideration word sentiments, polysemy, syntax, and semantic knowledge. (Jain et al., 2020) compared Random tree and Decision tree ML techniques to provide recommendations to customers using airline reviews. People are more likely to complain when they are anxious, according to research in (Meier et al., 2013). An examination of the temporal effects of weekend and weekdays was required. We found (Ryan et al., 2010) which does exactly that in addition to exploring work versus nonwork experiences on mood and other well-being indicators too. (Utama et al., 2019) did Sentiment classification on Airline Tweets. They used Mutual Information for Feature Selection. Despite recent advances in text analysis, many approaches have limitations for tweets (Wang et al., 2017). They proposed using convolutional neural networks framework. It combined implicit as well as explicit representations of short text for classification. They obtained the short text embedding by 9 joining the words and relevant concepts on top of pre-trained word vectors and further incorporating character level features into the model.

Text from twitter (tweets) fall under the category of short-texts since Twitter allows a maximum character length of 280 characters only and using supervised and unsupervised natural language processing methods, they performed analysis on approximately 350,000 tweet replies to U.S. politicians in (Jaidka et al., 2019). Therefore, unlike large paragraphs or documents, tweets might not observe the syntax of natural language thus including heavy usage of abbreviations

**CHAPTER 4 PROJECT DESCRIPTION**

Keywords and key phrases can be a useful tool for analysing large amounts of textual material quickly by serving as a concise summary. International Encyclopaedia of Information and Library Science (Bolger et al., 1989) defines a “keyword” as a “word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document.” (Feather et al., 1996). Research suggests that people tend to be more willing to complain when they are stressed (Bolger et al., 1989) (Meier et al., 2013), and moods are affected by time (Ryan et al., 2010). Therefore, this project will provide airlines with a tool to calibrate and judge the positivity/negativity of the tweet based on the weekday (Sunday, Monday, …), which is an issue that has yet to be studied. We aim to perform text & sentiment analysis on extracted airline travel tweets while also taking into consideration when the tweet occurred, and whether it has resulted in a positive or negative effect.

NEED FOR NEW SYSTEM

* To scrape the tweets mentioning the top airlines of interest.
* To identify the sentiment of the above scraped tweets.
* To analyse the relationship, if any, between time of the tweet and the sentiment of the tweet.

**CHAPTER 5 SCOPE OF PROJECT**

We extracted the day of the week from the date\_stamp of each tweet using Python’s datetime module. Once the sentiment was predicted, we then gave a score of +1 to positive, -1 to negative and 0 for the neutral sentiment. After some missing value imputations and handling outliers, we then calculated the mean sentiment score for each day of the week in our crawled twitter dataset which we crawled using the mentioned tools. After some further analysis, we got a correlation of -0.44 between day of the week and sentiment which means people were more negative towards the weekends than the weekdays.

**Keywords**: social media, customer experience, supervised learning, airline tweets, text analysis, classification.

**CHAPTER 6 REQUIREMENTS**

## 5.1 Hardware requirements: RAM: 2 GB

## Processor: Intel Core i3

**Harddisk:** 20 GB

## Software requirements:

**Operating System:** Windows or any Equivalent OS **Language:**  Python Language

**Browser:** Chrome

**Editor:** Google Colab

**CHAPTER 7 METHODOLOGY**

List of innovations and why the method and visualization are better than state of the art:

● Comparing the accuracy of the results for both logistic regression and SVM.

● Making it possible to create a good classification model with greater than 65% f1 score after manually labelling as few as 10 samples per class

● Creating a new list of better stop words and processing the text based on these custom stop words to avoid changing the meaning when filtering the tweets while ensuring that the words that do not contribute much to the meaning of the text are removed.

● Studying the behaviour and correlation of sentiment with day of week. Steps and algorithm:

1. Gathering data:

To gather data quickly and efficiently, we planned to use the web scraping software, Octoparse. Octoparse takes in a given URL and scrapes selected data. We spent some time testing this method and found that even if every team member commits to scraping data for a week, Octoparse would still fail to provide a large enough dataset. So, we explored other methods, and decided to use ScrapeHero. To determine the airlines we want to analyse, we used those that have the largest revenues and most passengers carried, thus selecting Delta, American, Lufthansa, United, Air France, Southwest, China Southern, and Ryanair. Since COVID-19 disrupted the airline industry, we took tweets that mentioned these airlines prior to the pandemic, from August 2019 to December 2019. Using the advanced search provided by Twitter, we got a url for all the tweets meeting these criteria. ScrapeHero used this url and scraped the webpage, capturing the user’s handle, the user’s name, the tweet, the number of 4 retweets, replies, and likes, and the time and date of the tweet (unix timestamp). We exported this data as a csv file. While cleaning this data, we found that ScrapeHero included tweets from outside the date range we specified, and some tweets that did not mention any airline. After excluding these, our data reduced by about 1,000 data points.

1. Collecting external resources/data:

While the above data was being labelled, we looked for a representative dataset and found the airline tweets dataset of “Crowdflower's Data for Everyone library” to train our sentiment classification model and build our model for few shots classification. We then applied these models in the relevant sections on our scraped dataset. This dataset (Crowdflower, 2019) was scraped in February 2015 and the tweets were classified into positive, negative, and neutral.

3) Data Visualization.

4) Exploratory Data Analysis(EDA).

5) Vectorization Process.

6)Train Test Split.

7) Data Scaling.

8)Defining X and Y.

9) Tuning.

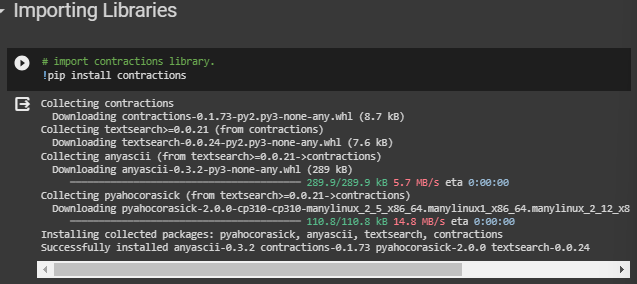
10) Building a machine learning model:

Several machine learning methods can be used to classify tweets.

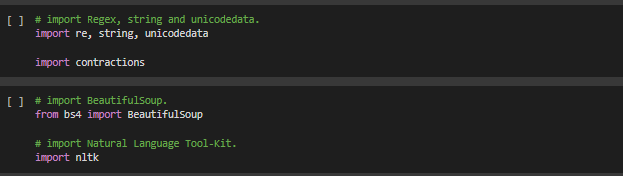
We explored 5 methods:

* + 1. SVM
    2. Decision Tree
    3. Random Forest
    4. KNN
    5. Naïve Bayes

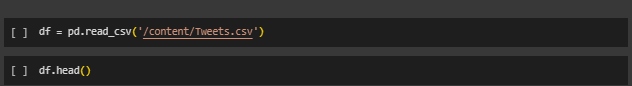
**1.Importing Libraries:**

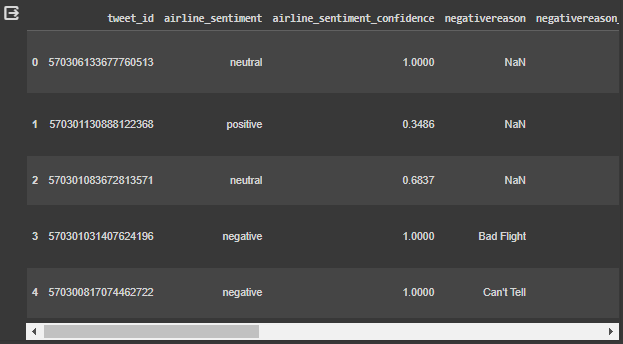




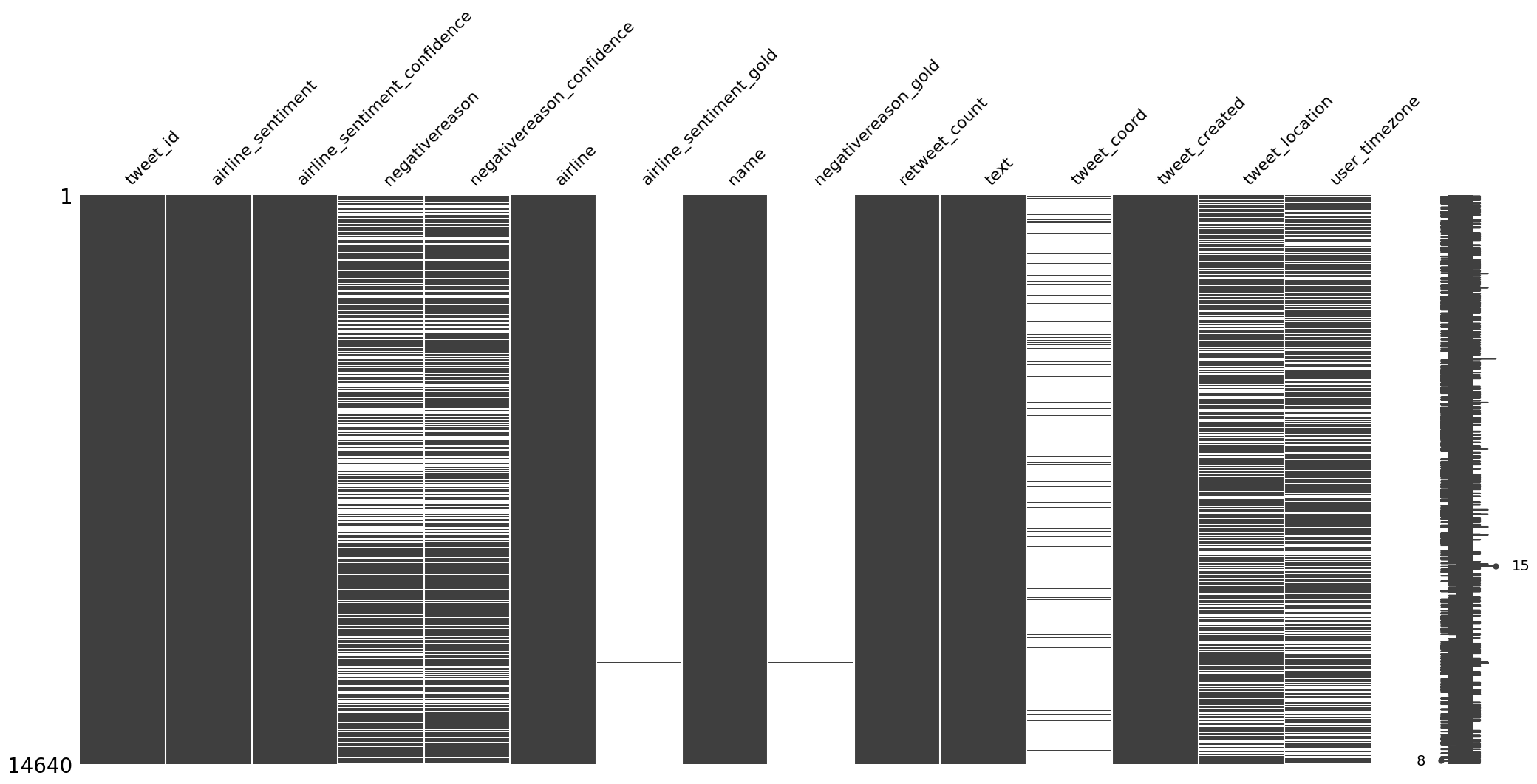


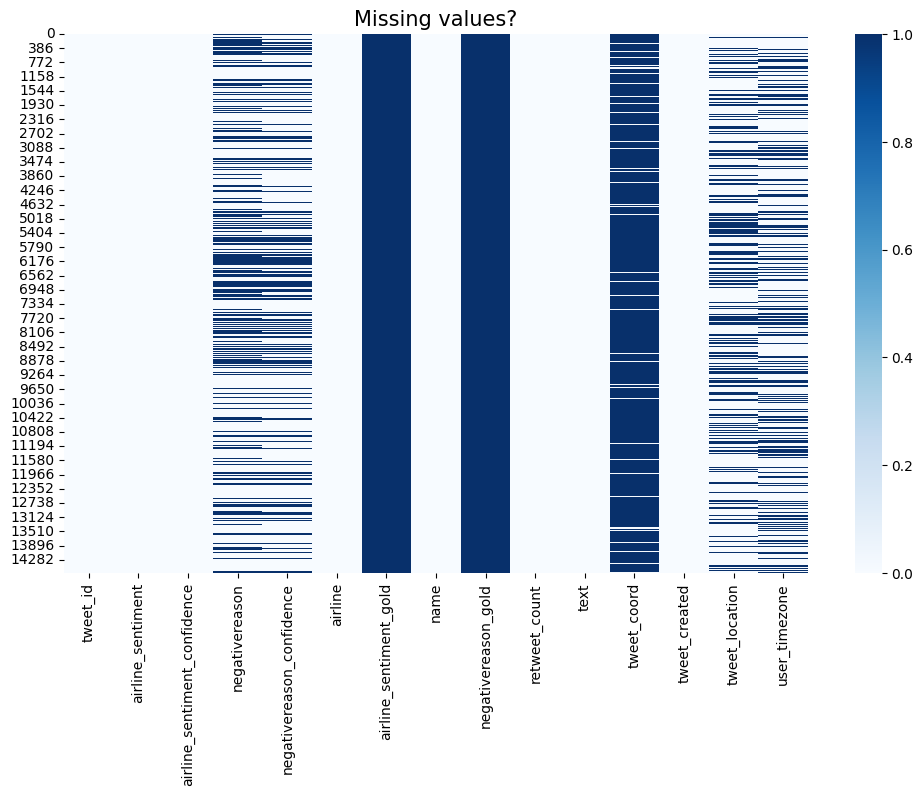
**2. Loading the Dataset.**





**3. Data Visualization**



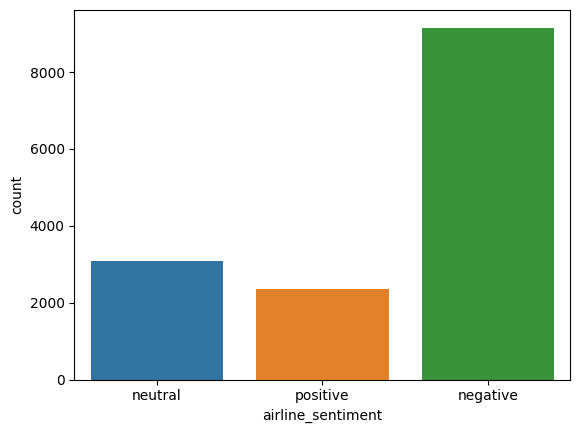


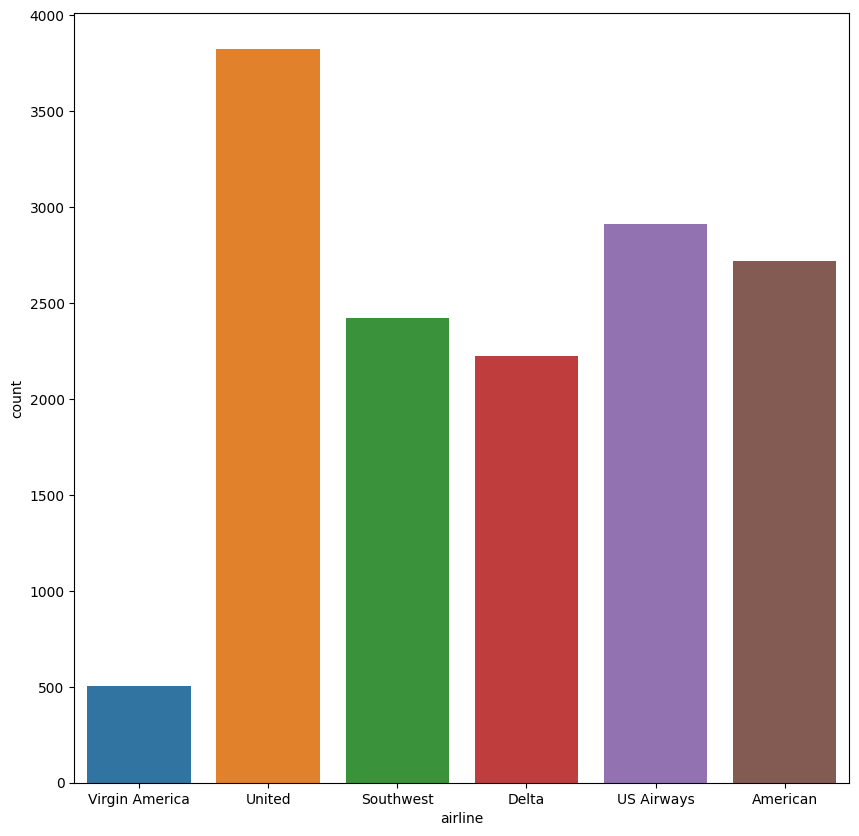
Interestingly, the only non-null values of the \_gold columns seems to be the same entries for the most part. Meanwhile, there is some but not total overlap between location and timezone in terms of missing values.

airline\_sentiment\_gold, negativereason\_gold have more than 99% missing data And tweet\_coord have nearly 93% missing data. It will be better to delete these columns as they will not provide any constructive information.

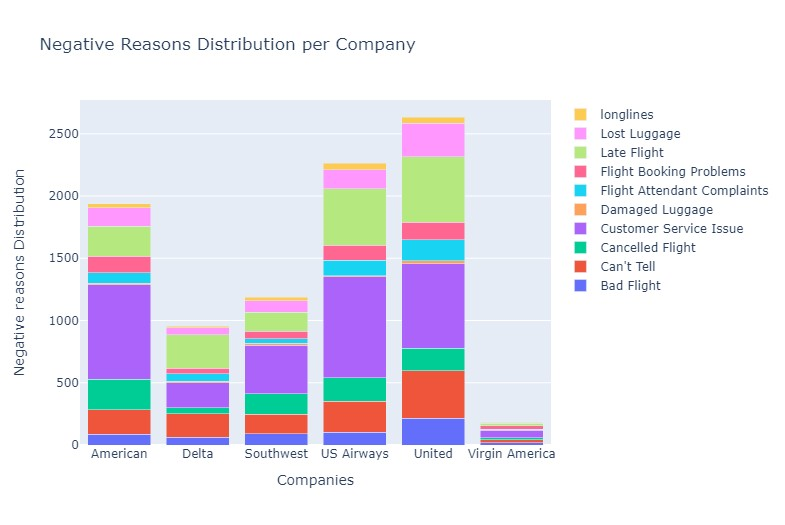
we can't fill it will affect in bad way for example we have positive reviwe and we fill the values with mode that means with Customer Service Issue it is missmatch and can be affect on train model so we keep the data as it is.

**4. Exploratory Data Analysis(EDA).**





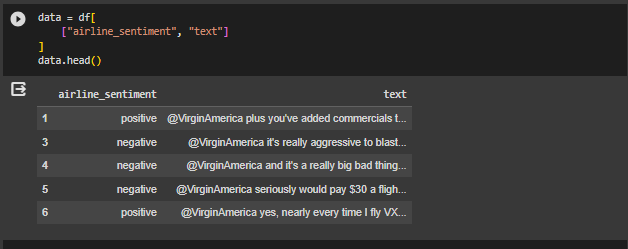
Stacked bar chart to show the distribution of reviews per company

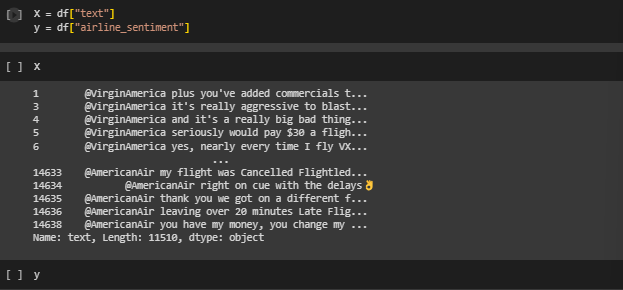


**Observations:**

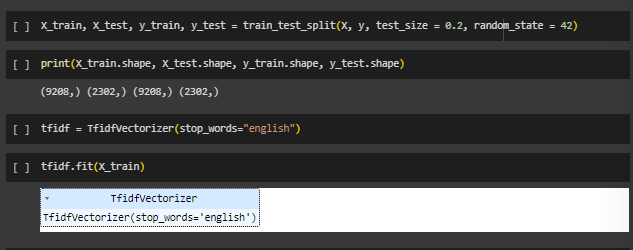
There are 15 columns in the dataset. Half of the columns have null values. Considering both dependent and independent variables not having any null values, we will not do any null value processing. Most columns in the dataset are of object type. airline\_sentiment is our dependent / target variable. text column is our independent variable that we will use for analysis. All other columns will be dropped at a later stage.

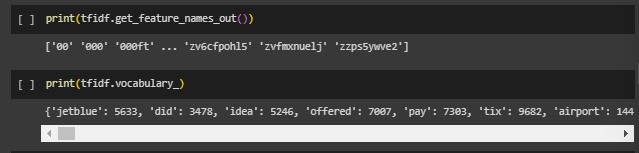
**5.Vectorization.**



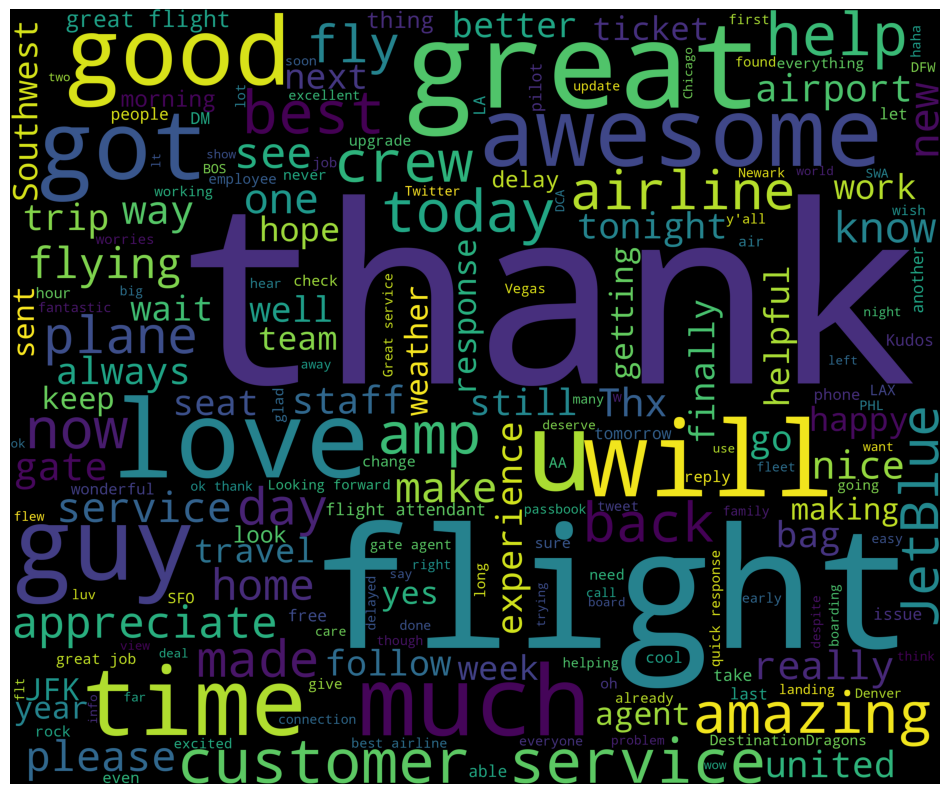


**6.Train Test Split.**

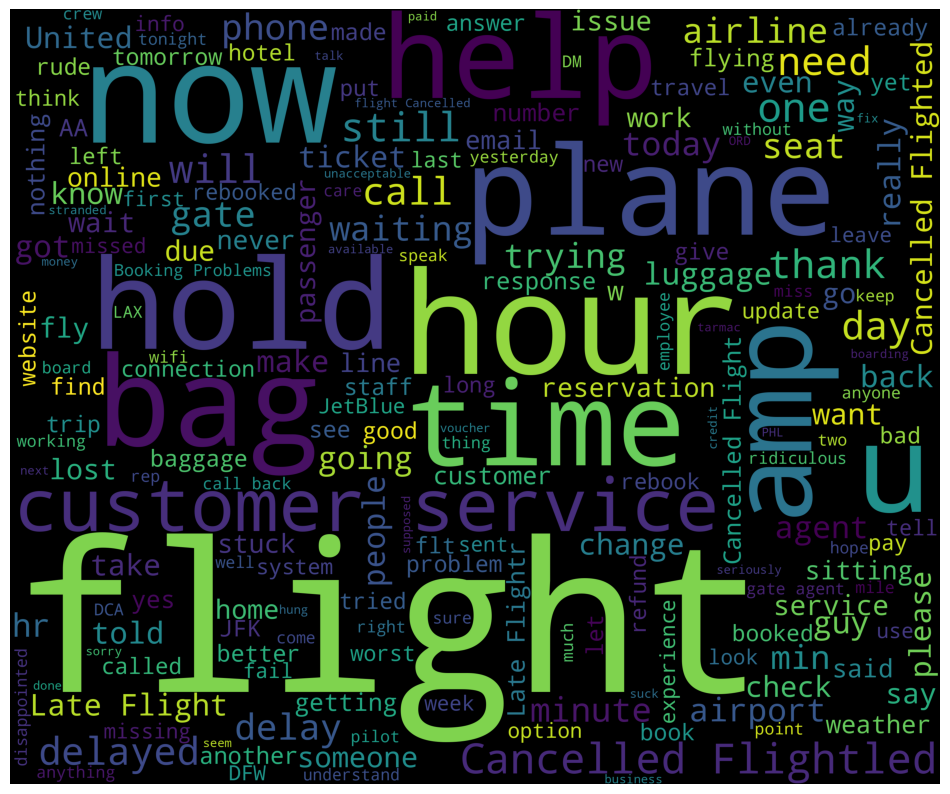




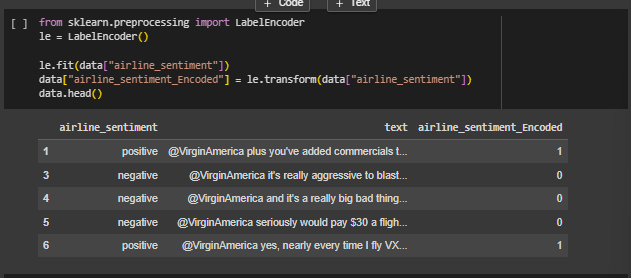
**Word Cloud For Positive Reasons**

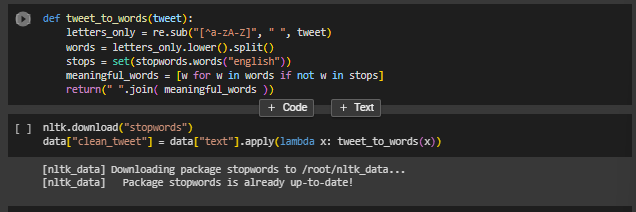


**Word Cloud For Positive Reasons**

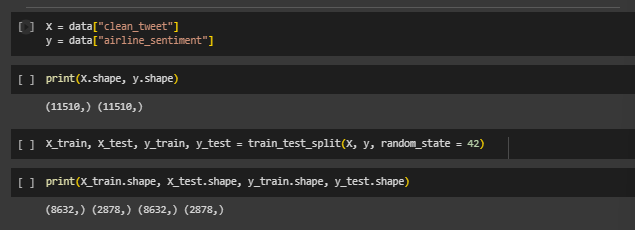


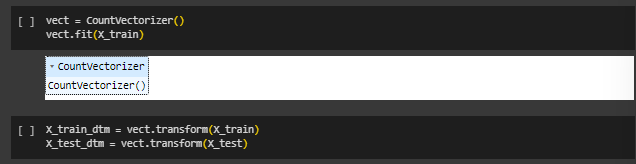
**8. Data Scaling**



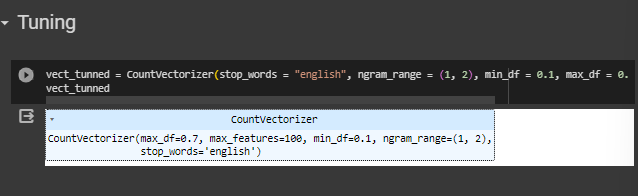


**9. Defining X and Y**

****

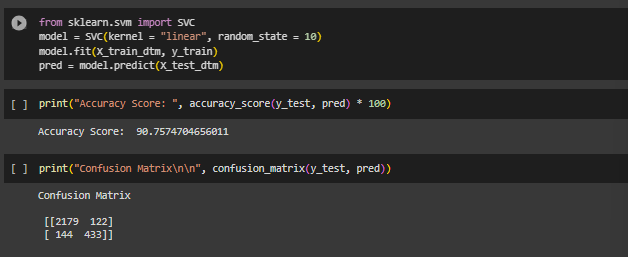


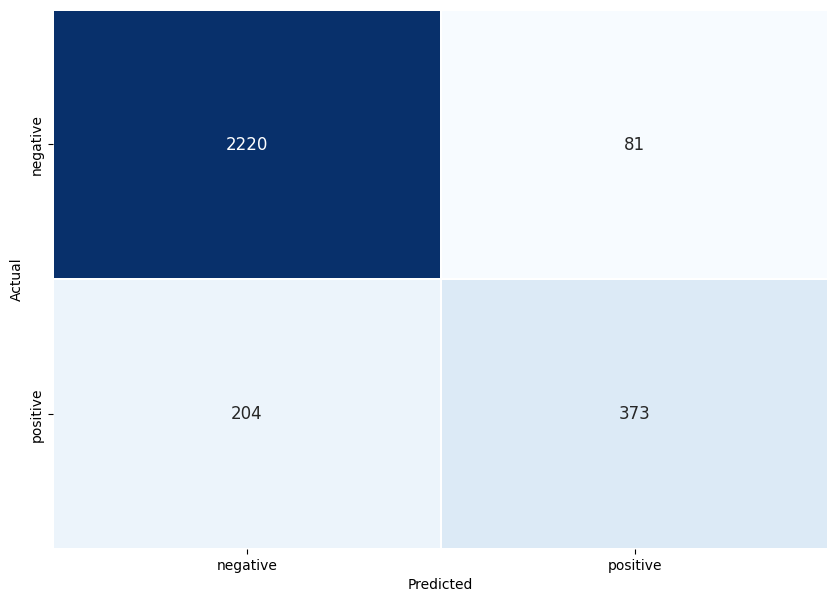
**10. Tuning**

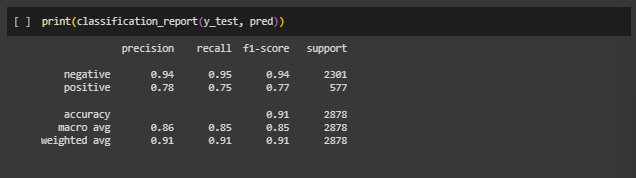
****

**11. Model Building**

1. **SVC**

****



****

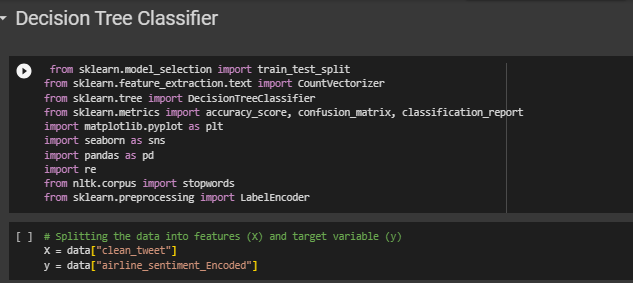
**Conclusions:**

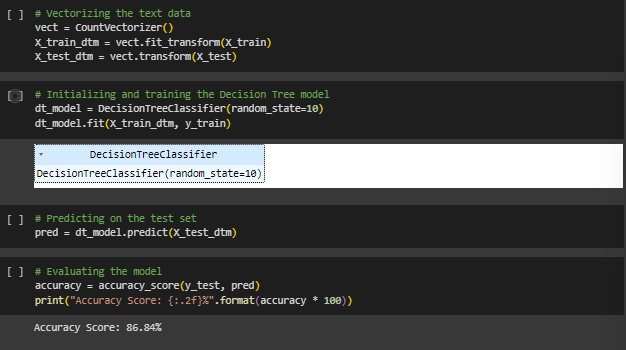
As we you can see above we have plotted the confusion matrix for predicted sentiments and actual sentiments (negative and positive)

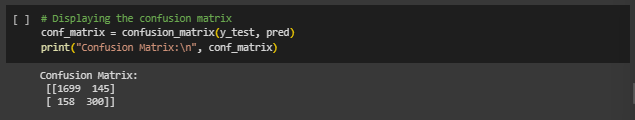
SVM Classifier gives us the best accuracy score i.e 91% precision scores according to the classification report

The confusion matrix shows the TP,TN,FP,FN for sentiments(negative, positive).

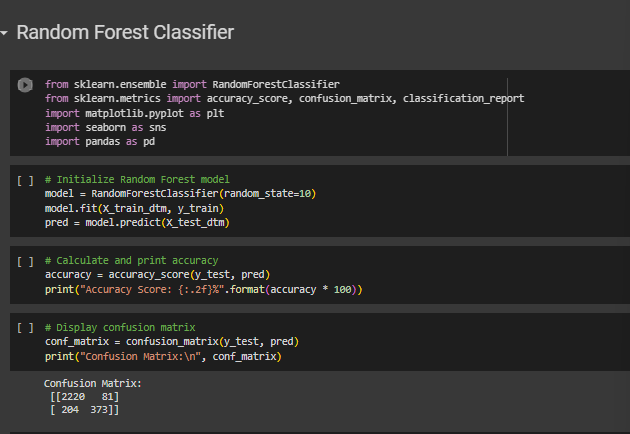
**2. Decision Tree**

****

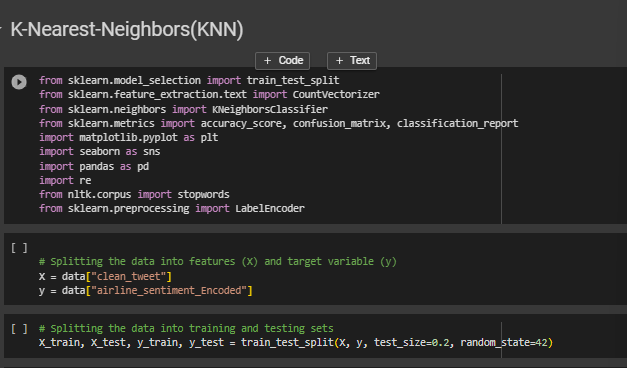
****

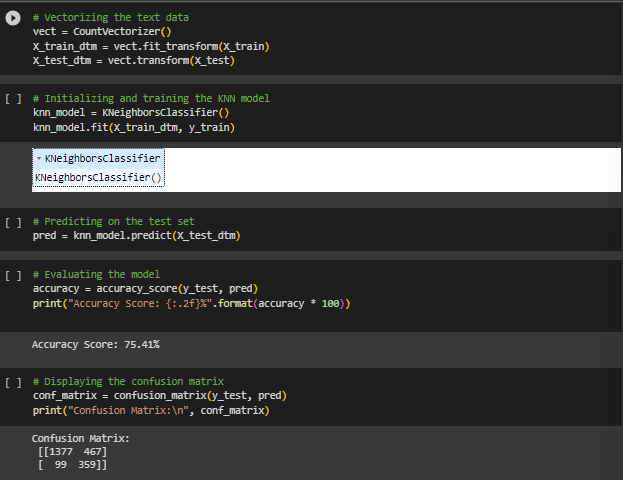
****

**3. Random Forest Classifier**

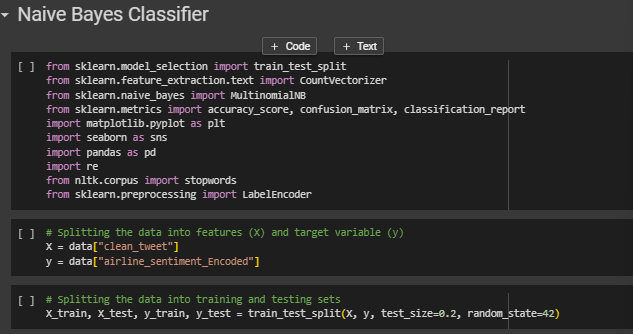
****

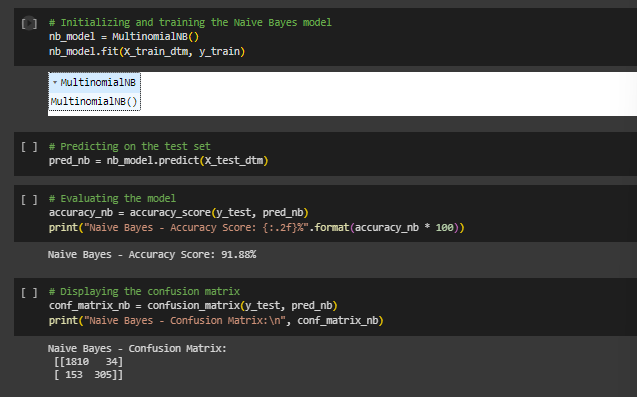
**3. KNN**

****

****

**4.Naive Bayes**

****

****

**CHAPTER 8 TESTING AND RESULTS**

**Conclusions:**

**SVM**

\*\*As we you can see above we have plotted the confusion matrix for predicted sentiments and actual sentiments (negative and positive)\*\*

\*\*SVM Classifier gives us the best accuracy score i.e 91% precision scores according to the classification report\*\*

\*\*The confusion matrix shows the TP,TN,FP,FN for sentiments(negative, positive).\*\*

**The Decision Tree Classifier**

\*\*The Decision Tree Classifier exhibits a reasonable overall performance with an accuracy of\*\* 87%.\*\* This suggests that the model effectively predicts the correct class for a majority of instances based on the given features.\*\*

**The Random Forest Classifier**

\*\*The Random Forest Classifier demonstrates strong overall accuracy, achieving a score of 90%. This indicates the model's ability to correctly classify instances into their respective classes based on the given features.\*\*

\*\*While the model performs well in identifying the "negative" class with high precision and recall (0.92 and 0.96, respectively), it faces challenges in accurately classifying the "positive" class, showing lower precision (0.82) and recall (0.65). This suggests potential areas for improvement, particularly in capturing positive instances.\*\*

**Naive Bayes**

The \*\*Naive Bayes\*\* Classifier demonstrates strong overall accuracy as compared to other ML models, achieving an accuracy score of \*\*92%\*\*. This indicates the model's ability to correctly classify instances into their respective classes based on the given features.

The classifier performs exceptionally well in identifying the "negative" class, with high precision (0.92) and recall (0.98). However, there is room for improvement in classifying the "positive" class, as reflected by a slightly lower precision (0.90) and recall (0.67).

This suggests potential areas for fine-tuning, particularly in capturing positive instances.

The macro and weighted average F1-scores are both 0.86, indicating a balanced performance in terms of precision and recall across classes. The model provides a robust trade-off between precision and recall, offering a comprehensive evaluation of its effectiveness.

**REFERENCES**

1. Bolger, N., DeLongis, A., Kessler, R. C., & Schilling, E. A. (1989). Effects of daily stress on negative mood. Journal of Personality and Social Psychology, 57(5), 808–818. <https://doi.org/10.1037//0022-3514.57.5.808>

2. Crowdflower dataset. (2019). <https://www.kaggle.com/crowdflower/twitter-airlinesentiment>

3. FastText common crawl vectors. (2020). <https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M-subword.zip>

4. Feather, J., & S. P. (1996). Implementing online archiving. International encyclopedia of information and library science, London and New York.

5. Gautam, A., Mahata, D., Gosangi, R., & Shah, R. (2020). Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in social media. Proceedings of the 3rd Workshop on E-Commerce and NLP.

6. Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of Twitter: The constraint affordance and political discussion. Journal of Communication.

7. Jain, P.K., & Pamula, R. (2020). Sentiment analysis in airline data: Customer rating -based recommendation prediction using Weka. Machine learning algorithms for industrial applications, Springer.

8. Kandasamy, I., Vasantha, W. B., Obbineni, J. M., & Smarandache, F. (2020). Sentiment analysis of tweets using refined neutrosophic sets. Computers in Industry, 115. https://doi.org/10.1016/j.compind.2019.103180 33.