**CAPSTONE PROJECT REPORT**

**SENTIMENTAL ANALYSIS ON TOP TEN AIRLINES**

**Data Analytics for Business**

**Submitted by:**

**Priya -------------- 0779602**

**Submitted to:**

**Mrs. Savita Sherawat**

**St. Clair college Mississauga**

**Submission Date:**

**July 23, 2022**

**Contents**

[**1.** **Sample chart** 2](#_Toc109460429)

[**2.** **Table of attributes** 2](#_Toc109460430)

[**3.** **Abstract and Keywords** 3](#_Toc109460431)

[**3.** **GitHub** 3](#_Toc109460431)

[**7.** **Purpose** 4](#_Toc109460432)

[**8.** **Research questions** 4](#_Toc109460433)

[**9.** **Theme** 5](#_Toc109460434)

[**10.** **Technique and tools** 5](#_Toc109460435)

[**11.** **Literature** 5](#_Toc109460435)

[**12. Methodology** 5](#_Toc109460435)

[**12.1. Data Collection** 9](#_Toc109460438)

[**12.3 Text- Preprocessing** 10](#_Toc109460439)

[**12.4 Sentiments Distribution**](#_Toc109460440) 10

**12.5** [**Inverse Document Frequency - Term Frequency** 13](#_Toc109460443)

[**13. Split data** 14](#_Toc109460447)

[**14. Train set and test set** 14](#_Toc109460448)

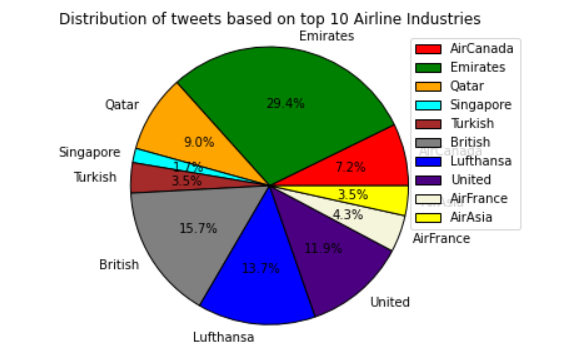
[**15.** **Machine Learning Models** 15](#_Toc109460449)

[**16.** **Confusion Matrix** 15](#_Toc109460449)

[**17.** **Conclusion** 15](#_Toc109460449)

[**18.** **References** 15](#_Toc109460449)

# **Sample chart**



# **Table of attributes**

Table 1: Integer Attributes

| Attributes | Data Type | Description |
| --- | --- | --- |
| id | integer | Unique ID of each user on twitter |
| likeCount | Integer | Count of likes on tweets |
| retweetCount | Integer | Count of retweet |
| replyCount | Integer | Count of reply on tweets |

Table 2: Object Attributes

| Attributes | Data Type | Description |
| --- | --- | --- |
| User | object | Username in tweets |
| place | object | Place of user |
| date | object | Date of posted tweets |
| content | object | Reviews given by users |
| Language | object | Language of tweets |
| sourceLabel | object | Source for tweets |
| hashtags | object | Hashtags used for tweets |

# **Abstract**

Social media is a platform where many users must connect with one another to communicate. **Twitter** is one of the most popular social networking platforms. Tweets, or quick messages, are posted on this Twitter. To learn how its customers feel about the desired service, the invited business also requires customer feedback. To classify the company's sentiment, sentiment analysis is required. The dataset **Airlines** is taken from **Twitter Developer Account**, illustrates the sentimental analysis on tweets by multiple users on twitter. These tweets can be used to get sentiments as it could be **Positive, Negative or Neutral** and can be combined with other methods of analysing the data using Machine Learning models.

A sentiment analysis project on ten major airlines (such as **Air Canada, Emirates, Qatar Airways, Singapore Airlines, Turkish Airlines, British Airways, Lufthansa, United Airlines, Air France, Air Asia**). This dataset contains **83187** **records** and **11** **features** that fetched from twitter.

**Keywords**

NLP (Natural Language Processing) and models such as Logistic Regression, K-nearest neighbor (KNN), Support Vector Machine (SVM), Random Forest.

1. **GitHub Link**

https://github.com/itspriya666/Airlines-Project

1. **Introduction**

The **sentiment analysis** is taken in the form of movie review, product review, social discussion, trip review, etc. It made a difference between computer and human. Airline industry is the growing billion-dollar industry and it also have millions of passengers. Sentiment analysis on Airlines is also taken. Regarding the airline industry, sentiment analysis has the potential to help customers decide which airline is best by analyzing other customer opinions. It is based on the classification of twitter data into positive, negative, and neutral command.

Every piece of information shared on social media carries an emotion, sentiment or feeling. These emotions can be positive, negative, and neutral. All these emotions may come from a travel trip, restaurant trip, exhibitions, movies, elections, hospital visits etc. These emotions carry some hidden information related to comfort/discomfort in related areas. Hence, there is a good scope of analyzing this information to detect the patterns of the emotions. This analysis can help us to understand the emotions of the people in respective domain and the reasons behind it.

# **Purpose**

The objective of this study is to conduct topic **modelling and sentiment** analysis on the posts of **tweets (Twitter),** where there are many interests and participation of the people who have used or are willing to use it for airlines. The purpose of people gathering from tweets is to make better choices using the actual experiences of other customers who have experienced airlines. Online tweets written by **passengers with experience** using airlines in all over the world were collected. The data collected were online reviews from top ten airlines, with more than 83 thousand reviews. Topic modelling and sentiment analysis were used with the collected data to figure out what kinds of **important words** are in the tweets.

# **Research questions**

1. **Which Airline Industry has highest tweets from 2021 to 2022?**

This question helps us to find the most preferable Airline which can be targeted by multiple passengers. They will have many queries regarding this industry.

1. **For one year period, what will be the public response about top ten airlines?**

This question will help to determine the response of passengers that will give an idea about the top ten airlines. Either response can be positive, negative, or neutral.

1. **Predict and compare the accuracy of sentiments using different machine leaning models.**

This question will help to predict the accuracy of sentiments and comparing the models.

# **Theme**

* **Sentimental Analysis**: Based on Tweets.
* **Prediction and Classifications:** Using ML Models.

# **Technique and tools**

Use different techniques and tools in python to solve our problems with some libraries (such as pandas, NumPy). Along with it, Natural Language Toolkit is being used for pre-processing the text that can remove the noise from tweets.

1. **Literature Review**

Air travel is one of the most convenient modes for long distance travel at both national and international level. There are many airline service providers (ASPs) around the world. The competitive world motivates the airlines company to attract the customers. However, a traveler considers several points before selecting any airline. These points can be airfare, travel time, number of stoppages, number of baggage allowed, and existing customer feedback etc. Therefore, all ASPs are working in all these customer service areas to improve their facility and in-flight comfort to attract the customers. **Adarsh & Ravikumar (2018)** used tweets relating to Indigo Airlines, Emirates Airlines, and Qatar Airlines from their Twitter and customers who tweeted about these airlines. The approach of detecting sentiments on Twitter was proposed by considering the tweets from three popular Airlines. The definition of positive, negative, neutral sentiments was based on the score computation, which was the difference between the positive and negative words for each tweet. They found that Emirates Airlines had more positive sentiments compared to the other two; Indigo Airlines had more negative sentiments, and Qatar Airlines had more neutral sentiment tweets. The problem of this approach is that the presence of positive and negative words may not give relevant results in the case of sarcastic tweets as the placing of positive and negative words in a sentence gives different conclusions.

It is very important to understand the needs and comfort level of customers i.e., customer satisfaction during the flight. Therefore, customer feedback is very important for any airline industry. There could be several possible ways to collect the customer feedback. The easiest and most traditional way is the customer feedback form available during the journey. However, most of the passengers do not show any interest in filling feedback forms. Another shortcoming of this approach is that it may or may not have appropriate questionnaire and may be biased on certain parameters i.e., the feedback form may only have certain specific questions. Other approaches for customer feedback collection could be through online website or online mobile applications of the airlines. After the journey, an email with a link can be sent to the passenger to request for feedback. However, there is no guarantee of its success. Another approach is to send a message on passenger’s mobile phone and ask them to rate your service (1 for poor and 5 for excellent) on certain parameters. All these traditional methods opted by the industry are restricted to certain parameters only. The more convenient way for a passenger is to express their feedback, as they want. **Rossi F (2013),** the most convenient way for the passengers to share their opinions is the social media instead of feedback form. Social media provides a platform where a user can freely express his feedbacks on any issues they observed during flight. Twitter is one of the popular platforms worldwide. The information from Twitter can be utilized to develop a recommender system). In addition, travelers are more comfortable in sharing their views about travel experiences on Twitter.

**Classification Models:** Once the preprocessed features are extracted, machine learning methods can be used to develop the classification model for sentiment analysis. After the crucial investigation of literature, I decided to use Support Vector Machines (SVM), Artificial Neural Networks (ANNs), and Convolutional Neural Networks (CNN) in our study. **Han J, Kamber M. (2006)**, Machine learning and big data technologies made it possible to analyze huge database and to develop highly accurate prediction or classification models. In this article, machine learning techniques are used to develop a binary class sentiment classification model for twitter data for some of the popular airlines worldwide. The study opted to develop a classification model for two categories of sentiments i.e., positive, and negative. The reason behind selecting only positive and negative sentiments is that neutral sentiment does not provide any information whether the service was nice or not. It may also depend on a neutral and kind passenger who takes things normally. **Zomaya AY, Sakr S. Handbook** and. **Berlin: Springer (2017)**, Neutral sentiment just pretends that everything was normal, but airline industries are more interested in more specifically “what was good or bad”. Therefore, only positive, and negative sentiments classes have been opted for analysis in this study. Support Vector Machine (SVM) and Artificial Neural Networks (ANN) were trained on the preprocessed tweets. Further, convolutional neural network (CNN) is trained on the data and its performance were compared with the best model among SVM model. Results shows that CNN outperformed all other models in terms of accuracy and performance. Further, association rule mining is used to map the relationship between several issues related to passenger’s comfort during flight with the nature of emotions (positive or negative). **Serrano-Guerrero J (2015),** To get more insights from the text, the entity level analysis can be performed which is a finest grained level analysis. At this level of analysis, the base for sentiment analysis is the main context. Therefore, this level of analysis is slightly more complex as high precision is required on output. **Ojokoh BA, Kayode O (2012),** Feature based sentiment analysis and summarization comes into this category of analysis. On the other part, machine learning has played an excellent role in NLP. Machine learning is broadly categorized into supervised and unsupervised learning. Supervised techniques such as Naïve Bayes (NB), support vector machines (SVM), Maximum Entropy are widely used for sentiment analysis The data set with unlabeled set of documents can be analyzed using unsupervised ML techniques. **Kim K, Lee (2014),** In addition, hybrid version of ML techniques that combined both supervised and unsupervised can also be used for sentiment analysis.

1. **Methodology**

**Data collection Data Analysis**

Evolution

Accuracy,

Confusion Matrix

Data in form of Tweets

Exploratory Data Analysis

ML Classification Models

* Logistic Regression
* k- nearest Neighbor
* Support Vector Machine

Split Data

Test data

Train Model

Text Preprocessing

Remove

* Stop Words, Punctuations, hyperlinks.
* Tokenization the words
* Lemmatization the words
  1. **Data Collection**

Social network sites like Facebook and Twitter. Opinions and feelings are expressed in different way, with different vocabulary, context of writing, usage of short forms and slang, making the data huge and disorganized. Manual analysis of sentiment data is virtually impossible. Therefore, special programming languages like ‘R’, ‘Machine Learning’ are used to process and analyze the data.

* 1. **Exploratory Data Analysis**

# **Check the information of attributes**

Cleaning up the given dataset is made easier with an understanding of it. Checking the shape, description of the dataset as well as addressing missing values or human error.

# **Removing and changing attribute name**

|  |  |
| --- | --- |
| Attribute Names | Rename |
| id | ID |
| user | User\_Name |
| date | Date |
| content | Tweets |
| place | Location |
| likeCount | Like\_count |
| retweetCount | Retweet\_Count |
| replyCount | Reply\_Count |
| lang | Language |
| sourceLabel | Source |
| media | Media |

# **12.2.3 Checking missing values**

Incomplete data attributes or the records. This needs to check the NAN,s and handle them by replacing or removing specific value or attribute. After dropping only those features which are important. Then, dataset contains only 9 features except Location and Media.

|  |  |
| --- | --- |
| Attribute Name | Null values |
| Location | 77953 |
| Media | 46139 |

# **12.3 Text- Preprocessing**

**12.3.1 Change the datatypes**

|  |  |
| --- | --- |
| Attribute Name | Changed Data Type |
| Date | object to datetime |
| Tweets | object to string |

**12.3.2 Cleaning the text**

Remove all the **URLs, emojis, special characters** and **white spaces**.

Graphical user interface, text, application

Description automatically generated

**12.3.3 Word Tokenize**

Tokenization in NLP is the process by which a large quantity of text is divided into smaller parts called tokens. We use the word tokenize () method to split a sentence into tokens or words.

**12.3.4 Stop words**

When working with text mining applications, we often hear of the term “stop words” or “stop word list” or even “stop list”. Stop words are basically a set of commonly used words in any language, not just English. In my dataset I fetch all tweets of English language. So, the stop words like me, my, myself, we, our, ours, ourselves, you, you’re. In my dataset stop words are like a, is, this, icao, alt, min, yyz etc.

**12.3.5 Lemmatization**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. In contrast to stemming Lemmatization is more powerful because it removes inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

For example:

Delayed ---- Delay

Waiting ----- Wait

**12.3.6 Text Blob**

With the help of Text Blob, we can get the sentiments of the sentences, we use the polarity with the help of this we get the polarity score it varies from -1 to +1, if the score is less than 0 then sentence is negative , if it is greater than 0 then sentence will be positive , if score is equals to 0 , then it will be neutral, we will get the polarity score on the basis of emotions and feelings what the user has been expressed. The words like good, graceful shows positive sentiments and apathy, dreadful shows negative sentiments.

# **Sentiments Distribution**

* + 1. **Pie chart distribution**

Chart, pie chart

Description automatically generated

Above pie chart shows that 47.7% are positive tweets, 36.8.2% are neutral tweets and very less percent of negative tweets i.e., 15.5. All in all, the highest of tweets have positive response.

* + 1. **World Cloud**

World Cloud on Positive Tweets

Text

Description automatically generated

# 

World Cloud on Negative Tweets

# 

**Text

Description automatically generated**World Cloud on Neutral Tweets

* 1. **Feature Extraction using TF-IDF**

# Usually take a selection before performing TF-IDF memory is limited because of the size of the dataset. Term Frequency & Inverse Document Frequency, sometimes known as TF-IDF, is a potent feature engineering technique used to locate significant or, more precisely, uncommon words in text data. In this data has more than 60% words that are featured.

# **Inverse Document Frequency - Term Frequency**

# TF - how often a term appears in a document.

# IDF - Weight of rare words

# The word weight, which is used sparingly in the sentence, has a high IDF score.

# **13 Split data**

The dataset into a training set and a test set, the training dataset does not have enough data for the model to learn an effective mapping from input to output. Also, there is not enough data in the test set to effectively evaluate the performance of the model.

In this dataset 70% of the actual data is training set and 30% of the actual data is test set. The target attribute is Score that contains three categories Positive, Negative and Neutral.

# **14 Train set and test set**

Train/Test is a technique for determining the accuracy of your model. It is called Train/Test because the data set is divided into two parts: a training set and a testing set. The important task in machine learning algorithms is splitting the dataset into training set and testing set. Here, most of the known data is separated into train set and rest of the test set is compared with the train set to check its similarity. As a result, one can avoid the incompatibility of the dataset and can better understand the features of the data.

**14.1 Data Balancing**

**SMOTE (Synthetic Minority Oversampling Technique):** Smote is one of the most widely used oversampling methods to solve the imbalance problem. It always goes to highest value.

Neutral 30687

Positive 24984

Negative 10159

After Smote all the three categories get balanced.

Negative 30687

Positive 30687

Neutral 30687

# **Machine Learning Models**

Machine learning models are algorithms that have been trained to recognise patterns in new data and predict outcomes. These models are presented as a mathematical function that receives requests in the form of input data, processes that data to create predictions, and then returns an output. These models are first trained using a collection of data, and then an algorithm is given to them so they can analyse data, find patterns, and learn from the data. These models can be used to forecast the unknown dataset once they have been trained.

* 1. **Logistic Regression**

It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

* 1. **KNN**

KNN stands for "K-Nearest Neighbour" meaning the closest neighbour. It is an algorithm for supervised machine learning. Problem statements involving classification and regression can both be solved using the technique. The sign "K" stands for the number of closest neighbours to a new unknown variable that needs to be predicted or categorised.

* 1. **SVM**

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outliers’ detection. It is still effective in cases where number of dimensions is greater than the number of samples. By using a subset of training points in the decision function (called support vectors), so it is also memory efficient. SVMs are used in applications like **handwriting recognition, intrusion detection, face detection, email classification, gene classification**, and in **web pages**.

* 1. **Random Forest Classifier**

In a random forest algorithm, there are many different decision trees. The random forest algorithm creates a "forest" that is trained via bagging or bootstrap aggregation. The accuracy of machine learning algorithms is increased by bagging, an ensemble meta-algorithm.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

1. **Accuracy Prediction**

|  |  |  |
| --- | --- | --- |
| Series | Model | Accuracy |
| 1 | Logistic Regression | 93.21 |
| 2 | KNN (K-nearest neighbour) | 56.25 |
| 3 | SVM (Support Vector Machine) | 95.70 |
| 4 | Random Forest Classifier | 82.60 |

**16.1 Comparison of Model’s Accuracy**

**Chart, line chart

Description automatically generated**

1. **Confusion Matrix**

**16.1 Logistic Regression Confusion Matrix**

This Confusion Matrix depicts that the actual and predicted value for positive tweets is 5628, neutral tweets is 7588 and negative tweets is 2126. The highest value is predicted in favour of neutral tweets.

Chart, waterfall chart

Description automatically generatedChart

Description automatically generated

**16.2 KNN Confusion Matrix**

This Confusion Matrix shows that the actual and predicted value for positive tweets is 3557, neutral tweets is 3738 and negative tweets is 2264. The highest value is predicted in favour of neutral tweets.

**16.3 SVM Confusion Matrix**

This Confusion Matrix depicts that the actual and predicted value for positive tweets is 5864, neutral tweets is 7666 and negative tweets is 2221. The highest value is predicted in favour of neutral tweets.

Chart, waterfall chart

Description automatically generatedChart, waterfall chart

Description automatically generated

**16.4 Random Forest Classifier Confusion Matrix**

This Confusion Matrix depicts that the actual and predicted value for positive tweets is 5039, neutral tweets is 6841 and negative tweets is 1715. The highest value is predicted in favour of neutral tweets.

**17. Conclusion**

This study conducted by using twitter sentiments on the bases of top ten Airlines that has overall positive response. As Emirates airline industry gained more popularity by getting highest tweets over the period between 2021 and 2022. As well as learning algorithms, Logistic Regression, KNN, SVC and Random Forest Classifier on Sentimental analysis. Overall, SVM reached at 95% accuracy, it means Support Vector Machine is the best fit model for this project.

**18. References**

1] S. Kumar and M. Zymbler, “A machine learning approach to analyze customer satisfaction from airline tweets,” Journal of Big Data, vol. 6, no. 1, 2019.

2] M. Al-Ghalibi, A. Al-Azzawi, and K. Lawonn, “NLP based sentiment analysis for Twitter's opinion mining and visualization.

3] D. Tiwari and N. Singh, “Ensemble Approach for Twitter Sentiment Analysis,” I.J. Information Technology and Computer Science, Aug. 2019.

4] R. Khan and S. Urolagin, “Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data,” IJACSA) International Journal of Advanced Computer Science and Applications,vol. 09, no. 6, 2018.

5] Y. Zhang and B. Wallace, “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification, 2015.

6] Liu, Y.; Bi, J.W.; Fan, Z.P. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms, 2017.

7] Catal, C.; Nangir, M. A sentiment classification model based on multiple classifiers. Appl. Soft Comput. 2017.

8] Yun Wan, Dr.Qigang Gao, An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis, IEEE, 2015.

9] Liu, Y.; Bi, J.W.; Fan, Z.P. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. Expert Syst. Appl. 2017.

10] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, “Tweets Classification on the Base of Sentiments for US Airline Companies,” Entropy, Nov. 2019.

11] N. Smith, “A disciplined approach to neural network hyperparameters: Part 1 -- learning rate, batch size, momentum, and weight decay.” 2018.

12] Y. Zhang and B. Wallace, “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification, 2015.

13] Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 2013 April.

14] A. Rane and A. Kumar, “Sentiment Classification System of Twitter Data for US Airline Service Analysis,” 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 2018.

15] Ashi MM, Siddiqui MA, Nadeem F. Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets. Adv Intell Syst Comput.2019.