Sentiment Analysis Using Machine Learning

Pratyush Anand[1], Sunidhi Gupta[2], Khushi Sahu[3]

[1]Student, [2]Student, [3]Student
[1,2,3] MIT World Peace University, Pune, Maharashtra, India

[1] p.anand00136@gmail.com, [2] sunidhig1203@gmail.com, [3] khushisahu4326@gmail.com

Abstract

Population storming at social media sites is leading to giant data that are fascinating analysts, researchers and business companies to follow dots of user's conviction and thinking in abundant sections. Twitter being the most widespread one where users can express and write messages with the constraint of number of characters to 280 which are termed as tweets. Sentiment analysis turns up to be the most useful for analysing these tweets and helping the business companies to get assessment and improve upon the areas. This research paper sights on the reviews of airline which were twitted. The dataset is being has been taken from kaggle. The accuracy and f1 score of various machine learning classifiers employed in this study to categorise imbalanced and balanced datasets are compared. The outcomes tell that, Random Forest Classifier, ID3 and Support Vector Classifier (SVC) gave a stronger accuracy and f1 score level than other machine learning classifier.

Keywords: *Machine Learning, Classification, Pre-Processing, Balanced and Imbalanced Datasets. Sentiment Analysis, Reviews*

1. Introduction

Social media makes it easier for people in different virtual circles to exchange suggestion, beliefs, and information. Social media as one of the strongest networking aids over the Internet has integrated with social and commercial aspects in the real-life [1]. The most effective way to build loyalty is to get to know the customers who further can tailor the marketing to suit their needs and so, many enterprises by examining the content of evaluations, can analyse their commercial capabilities on social media. The count of twitter's users is reaching around 400 million in the entire world and this is generating almost 6000 tweets per second. The tweet can be in the form of text, video, photo related to any content such as airlines and news [2]. Teens are not only the ones that are using twitter, many other business companies have also extended their circle to twitter. Due to convenient usage of twitter, the count of users is increasing speedily [3][4]. Airlines are companies that provide transportation by airplane for people and things. They have a fleet of airplanes that provide passenger flights. It is possible to improve their services by analysing and collecting customer's reviews which is the information gathered from customers about their experiences with services. This research is intended to discuss aspects of the approach in sentiment analysis of airline datasets which also involves a NLP (Natural Language Processing) approach. NLP is used in text pre-processing to get meaning from textual data. The work in this paper has covered both balanced and imbalanced data. For imbalanced data, hybrid sampling techniques are used to balance it before applying any machine learning classifier because it is anticipated to increase the accuracy and F1 score with this way.

2. Literature Survey

Nowadays, the majority of the public uses social media to express their thoughts or engage in discussions about various topics. Due to this, massive amounts of data are gathered at each social media site, and researchers are delving far into the analysis of the data. This section discusses previously surveyed text classification methods using data from Twitter in various fields.

Researchers of [5] investigated Indian Railway reviews from twitter. The data was collected over the period of 02.08.2018 to 12.08.2018. A total of 554,499 tweets were included in the data, of which 15,777 were unique tweets with labels designating positive, negative, or neutral sentiment. SVM, RF, and BPANN are three machine learning techniques that the authors utilised. Precision, Recall, F-Score, and AUC were used as the classification measures for these approaches' evaluation. Although BPANN with 1000 training iterations produced great results, RF and SVM also produced nice results. The accuracy has been found to increase as the training set size has grown, according to the authors. F1-Scores for RF and SVM were 79.5% and 82.87%, respectively.

In [6] Authors have discussed on movie review sentiment analysis, the dataset consisted of 21,000 tweets from which authors utilised 1200 of these tweets (600 positive, 600 negative, and 600 neutral) to train the classifiers. The text pre-processing involved following steps: tweets were changed to lower case in the initial stage. The next step was to remove every URL and replace it with plain text. The generic word AT_USER had then been used in place of "@username." The punctuations, hashtags and stop words were eliminated with a single white space. Then, the authors trimmed the character that was repeated more than once. In comparison to SVM, Naive Bayes had higher precision but somewhat worse accuracy and recall. SVM outperformed Naive Bayes in terms of accuracy, precision, and recall. SVM accuracy is 75%, whereas Naive Bayes accuracy is 65%. Researchers demonstrated the excellence of the feature vector chosen for their project's domain and the feature vector improved sentiment analysis regardless of the classifier chosen.

Authors of [7] gathered and classified more than a thousand reviews of Bangla food from numerous web sources, including Foodpanda, Hungrynaki, Shohoz food, Pathao food, etc. There were 1040 data in all out of which favourable reviews and 520 negative ones. Researchers divided the data into training and testing groups in an 80:20. Count Vectorizer, Glove Vector, Word2Sequence, and TF-IDF were used to pre-process the texts. For the conventional machine learning classifier, authors employed the Decision Tree, Linear SVM, Logistic Regression, Random Forest classifier and Multinomial NB and trained these models with unigram and bigram features. For deep learning, they employed the LSTM, GRU, and RNN. The best testing accuracy for all the models was produced by the LSTM with word sequence model. Random forest, linear SVM, Naïve Bayes, decision tree and logistic regression with count vector gave 74.52%, 68.75%, 69.23%, 69.71% and 71.15% respectively.

In [8], a system has been put out by authors that analyses the sentiment of product reviews gathered from Amazon. Data about data and product scrutiny from Amazon are also included, totalling millions from 1996 MAY to 2014 JULY. Music instruments, office supplies, home goods, graphic games, sports, cosmetics and recreation, cell phones and accessories, toys and games, wellness program are among the categories for which we have chosen datasets with total records of 920395 and 9 attributes. Logistic showed accuracy of 89% on the bag of words and TF-IDF score of 88%, while Naive Bayes produced 88.9% and 88.8%.

In [9], Authors provide a system to use machine learning techniques to assess and visualise the sentiments of Arabic tweets connected to the Saudi stock market. Twitter API was utilised to gather offline data for training and prediction, and Apache Kafka was used to stream tweets in real time. They used five machine learning algorithm with various feature extraction methods such as word embedding (word2vec) and BOW methods. SVM classifier in conjunction with the TF-IDF technique produced the best result which was 79%. From January 1 to December 13, 2019, a total of 5209 tweets in Arabic about the Saudi stock market were retrieved using the Twitter API. The classifiers which were used are Linear SVM, Logistic Regression, Decision Tree, Random Forest Stochastic Gradient Descent (SGD). The SVM classifier with the TF-IDF training model was chosen to be loaded on the website in order to predict the sentiments of the offline and real-time streaming tweets since it had the highest accuracy (79.08%).

A data collection that Andrew Maas [10] built using 50k IMDb movie reviews is used by authors in [11]. The data is divided into 25,000 reviews for testing the classifier and 25,000 reviews for training. Additionally, each set has 12500 reviews, both negative and positive and the viewers can scale from 1 to 10. Anything with fewer than 4 stars is indicated as negative, and anything with more than 7 stars is marked as good. Several data pre-processing methods are used for the IMDb dataset like all punctuation marks like "?" and "!" are eliminated, all letters are converted into lower case, links and stopwords are removed and stemming is applied on the text. Vectorization or Word embedding is used as feature extracting technique Doc2Vec model demonstrated to offer good accuracy results with lower processing cost when compared to other strategies. Best accuracy is given by LSTM algorithm with adam optimizer.

In [12] researchers have attempted to create and put into action a predictive system that would direct stock market investing. The uniqueness of their strategy is the integration of sensex points and Really Simple Syndication feeds for accurate forecasting. Data from RSS news feeds and stock market investment data are gathered over time. This module cleans the data by finding and eliminating outliers, filling in missing values, and smoothing noisy data. A noun and an adverb are combined when using the POS tagger. The final forecast is positive if the findings of the sentiment analysis and the Sensex Moving Average are both positive. If both are bad, the outcome is likewise bad. Combining the two will provide neutral effects. Whether it comes down to it, the stock market analysts can effectively determine when to buy or sell their stocks by combining sentiment polarity news and sensex points. With the use of Moving Average, ID3, and C4.5 methods, their experimental findings demonstrated a considerable increase in the precision and correctness metrics on bench marking.

The writers of [13] paper talked about the revenue of performance and online reviews for smart phones. After the collection of data from online sites the data is pre-processed with various techniques like POS Tagging, stopword removal, text transformation, clustering etc. In this data authors applied SVM machine learning algorithm to classify the negative and positive reviews. Results for different products are with Galaxy s5 the accuracy is 89.35%, Microcanvas 89.96%, Micronitro 88.03%.
In [14] authors have done a sentiment analysis on movie reviews the dataset is taken from the acl Internet Movie Database(IMDB) which is consist of 12500 each positively and negatively labelled data. They convert the text format into numerical format by Count Vectorizer and TF-IDF. Many Supervised learning techniques applied such as SVM, Naive Bayes, Stochastic Gradient Descent and n-gram method. The Best performance among various models is given by SVM with unigram 86.97%, Bigram 83.22%, Trigram 70.16%, Unigram+Bigram 88.88%, Unigram+Trigram 83.63%, Unigram+Bigram+Trigram 88.94%.

This [15] paper focuses on enhancing a robot receptionist's conversational skills. By creating a crude sentiment analysis module, they emphasise the Chabot’s responsiveness to a user's comment and the conversational atmosphere. Finally, a toxic-comment classifier that enhances a module for detecting profanity based on a lexicon has been used to guarantee that the output of the Chabot is positive and pleasant. The approach classified feelings simply into three categories: negative, positive, and neutral. The accuracy of 95.1% was the best outcome single ULMfit model.

Author	Aim	Source Data	Feature Extraction technique	Algorithm	Evaluation metric
[5]	Indian Railway sentiment analysis	Twitter	Doc2Vec	RF, SVM	P = 82.81% R = 82.94% F = 82.87% AUC = 0.845
[6]	Movie Review sentiment analysis	Twitter	Unigram	SVM, Naïve Bayes	A = 90%
[7]	Bangla food sentiment analysis	Foodpanda, Hungrynaki, Shohoz food, Pathao food	Count Vectorizer, Glove Vector, Word2Sequence, and TF-IDF	Decision Tree, Linear SVM, Multinomial NB, Random Forest classifier, Logistic Regression, LSTM, GRU, RNN	A = 74.52%
[8]	Amazon product reviews sentiment analysis	UCSD Design Lab	TF-IDF	Logistics, Naïve Bayes	A = 89%
[9]	Saudi stock market sentiment analysis	Twitter	BoW, Word2Vec	Linear SVM, Random Forest, Logistic Regression, Decision Tree, Stochastic Gradient Descent	A = 79.08%
[11]	IMDB movie review sentiment analysis	50k IMDb movie reviews	Word2vec, word embedding	LSTM	A = 89.9%
[12]	Stock investing sentiment analysis	RSS news feeds and stock market investment data	POS tagger	ID3, C4.5	A = 78.75%
[13]	Smart phone sentiment analysis	online sites	POS Tagger	SVM	A = 90.99%
[14]	Movie sentiment analysis	acl Internet Movie Database	Count Vectorizer, TF-IDF	SVM, Naïve Bayes, Stochastic Gradient Descent	A = 88.88%
[15]	Chat bot sentiment analysis	Chabot’s responsiveness to a user's comment	n-gram	ULMfit model	A = 95.1%

3. PROPOSED APPROACH

In this section the approach is discussed in brief. The authors in this paper have classified the sentiments of the review tweets by pre-processing them and then implementing machine learning classifiers, first on the imbalanced dataset then on the balanced dataset. The whole airline dataset is separated on the basis of 6 distinct airlines datasets and are individually analysed. At the end comparing accuracy and F1 score of all the classifiers – ID3, Random Forest, Naive Bayes, Decision Tree, KNN and SVC.

4. Experiment Setup

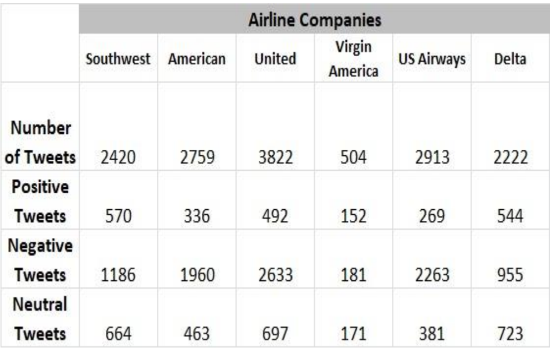
Here, the whole setup and process of sentiment analysis is discussed in detail. The dataset has been taken from kaggle. The environment chosen for implementing this is Jupyter notebook because it is a preferable option for delicate files that should not be stored in the cloud. You never have to worry about your GPU or runtimes being throttled when you install the laptops on your own hardware.

4.1 Data Description

Originally the dataset contained 15 columns and 14640 rows, from which 2 columns has been selected for the analysis: Airline Sentiment and Text.

4.2 Data Pre-processing

In NLP techniques, the dataset must go through a few pre-processing processes before any classification or prediction algorithms are applied. In table 1 the whole summary of imbalanced dataset is mention including, the number of tweets in each distinct airline datasets, number of positive, negative and neutral tweets. The steps that are followed while building this model is discussed in this section.



	Airline Companies					
	Southwest	American	United	Virgin America	US Airways	Delta
Number of Tweets	2420	2759	3822	504	2913	2222
Positive Tweets	570	336	492	152	269	544
Negative Tweets	1186	1960	2633	181	2263	955
Neutral Tweets	664	463	697	171	381	723

Table 1: Summary of Imbalanced Dataset

4.2.1 Removal of stop words

A set of words frequently used in daily discourse is called "stop words" like a, the, I, am etc. These words don’t change the meaning of the sentiment of the text hence not significant for the analysis. So the removal of stop words allows our text to focus more on the important information by eliminating low level information. The stop words list has been extended with ‘@’ and the airline names, since those are common and less important in all reviews.

4.2.2 Tokenizing

The reviews were in form of sentences, so these are tokenized first. Tokenization breaks the sentences into chunks of words so that machine can target each word and understand it in better way which would have been difficult in corpus form.

4.2.3 Lemmatization

It is a technique used for switching words from conjugated forms to their root forms, even after the breakdown it keeps the same meaning.

4.2.4 TF-IDF (Term Frequency-Inverse Document Frequency)

Term Frequency tf(t,d): Equation (1) represents tf - It is a term's total count across a document. The frequency with which phrase t appears in document d.

$$tf\ (t,d)\ =\ \frac{Count\ of\ t\ in\ d}{Number\ of\ words\ in\ d} \qquad (1)$$

Inverse Document Frequency idf(t, D): Equation (2) represents inverse document frequency - It is the included factor that reduces the weight of terms that occur frequently in the document collection and enhances the weight of terms that occur infrequently. It is a logarithmically scaled inverse proportion of the texts that contain the word, df is the document frequency – occurrence of t in documents.

$$Idf\ (t,\ D) = log\left(\frac{N}{df+1}\right) \qquad (2)$$

Equation (3) is of Term Frequency – Inverse Document Frequency: it is term frequency times the inverse document frequency of an document. To construct uni-grams and bi-grams transcript, ML algorithms employ the TF-IDF vectorizer as an input.

$$TF - IDF\ = tf\ (t,d) * idf\ (t,D) \qquad (3)$$

4.2.5 Training Data

Whole dataset is divided into 6 datasets as per the **'airlines'** which are further divided into two parts: Train-Set with 66% of the total reviews and Test-Set with remaining 34% of the total reviews. Figure 1 is the flowchart of the whole process that will be followed for building this model involving input, text pre-processing, feature extraction, classifiers and evaluation.

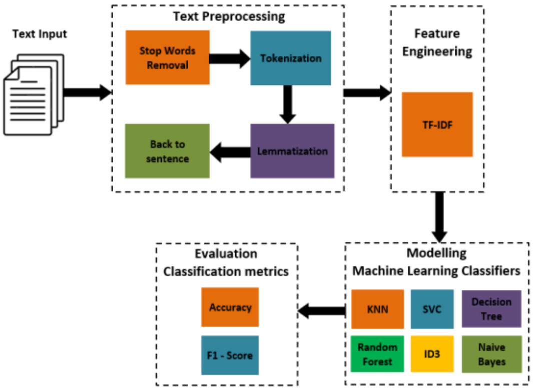


Figure 1: Flow chart of the whole process

4.3 Dataset Classification

In this section, the implementation of machine learning classifiers are discussed. The classifiers such as ID3, Random Forest, Naïve Bayes, Decision Tree, SVC and KNN are widely used in automatic text classification and the benefits of each are different; Others are more complex, but they are more resilient and adaptive, like SVC. Some are simple to implement, like KNN, while others are more complicated. For evaluation, classification performance metrics is used.

Accuracy (ACC) – Equation (4) is of accuracy where Tp is true positive, Tn is true negative and Total is the total of all. Accuracy is how often is the model correct. A model's accuracy is deemed to be at its highest if and only if we have a balanced dataset in which the values of false positive and false negative for the two-class issue are nearly equal [16].

$$ACC = \frac{Tp + Tn}{Total} \qquad (4)$$

F1 Score (F) – Equation (5) represents F1 score where, Pr is precision, R is recall. F1-Score is harmonic average of Recall and Precision. Since uneven class distribution happens in the majority of real-world text classification tasks, F1-score is a more useful statistic to test a model [16].

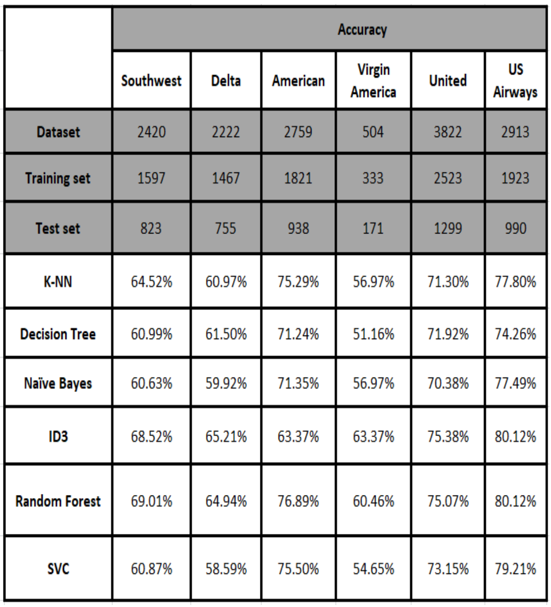
$$F = 2 * \frac{\Pr * R}{\Pr + R} \qquad (5)$$

5. Results Discussion

Here, evaluation of all the datasets after training on both imbalanced and balanced set is discussed including some tables and visuals for quick understanding.

5.1 Imbalanced Datasets

Table 2 represents accuracy of the machine learning classifiers on imbalanced dataset for all airline datasets also including the number of tweets after splitting for train and test.



	Accuracy					
	Southwest	Delta	American	Virgin America	United	US Airways
Dataset	2420	2222	2759	504	3822	2913
Training set	1597	1467	1821	333	2523	1923
Test set	823	755	938	171	1299	990
K-NN	64.52%	60.97%	75.29%	56.97%	71.30%	77.80%
Decision Tree	60.99%	61.50%	71.24%	51.16%	71.92%	74.26%
Naïve Bayes	60.63%	59.92%	71.35%	56.97%	70.38%	77.49%
ID3	68.52%	65.21%	63.37%	63.37%	75.38%	80.12%
Random Forest	69.01%	64.94%	76.89%	60.46%	75.07%	80.12%
SVC	60.87%	58.59%	75.50%	54.65%	73.15%	79.21%

Table 2: Accuracy of classifiers on imbalanced dataset

Figure 2 below shows the bar chart of the above accuracy table where blue, red, olive green, purple, sky blue and orange represents KNN, Decision Tree, Naïve Bayes, ID3, Random Forest and SVC respectively.

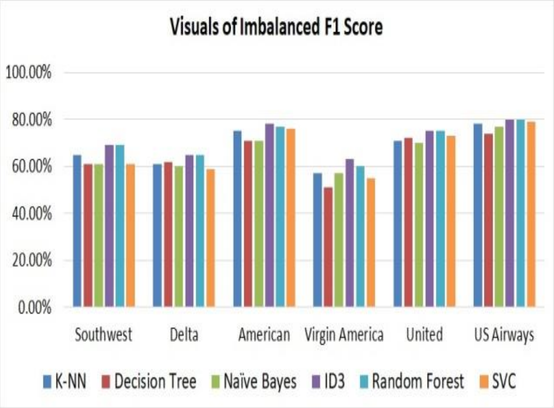
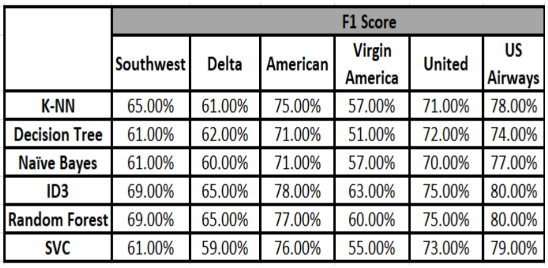


Figure 2: Bar chart of accuracy for imbalanced dataset

Table 3 shows the F1-Score of the machine learning classifiers on imbalanced datasets for all airlines in percentage for every distinct airline dataset.



	F1 Score					
	Southwest	Delta	American	Virgin America	United	US Airways
K-NN	65.00%	61.00%	75.00%	57.00%	71.00%	78.00%
Decision Tree	61.00%	62.00%	71.00%	51.00%	72.00%	74.00%
Naïve Bayes	61.00%	60.00%	71.00%	57.00%	70.00%	77.00%
ID3	69.00%	65.00%	78.00%	63.00%	75.00%	80.00%
Random Forest	69.00%	65.00%	77.00%	60.00%	75.00%	80.00%
SVC	61.00%	59.00%	76.00%	55.00%	73.00%	79.00%

Table 3: F1 Score of classifiers on imbalanced dataset

The bar chart of the aforementioned F1-Score table is shown in Figure 3 below. KNN, Decision Tree, Nave Bayes, ID3, Random Forest, and SVC are each represented by blue, red, olive green, purple, sky blue, and orange, respectively.

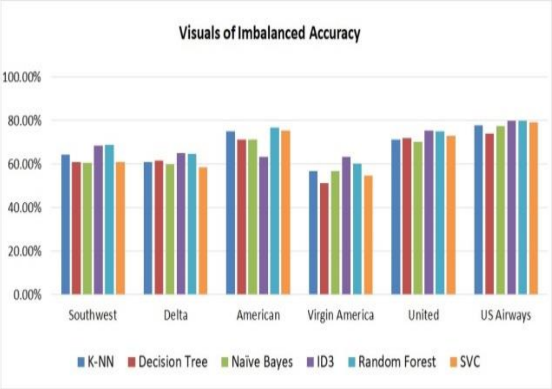
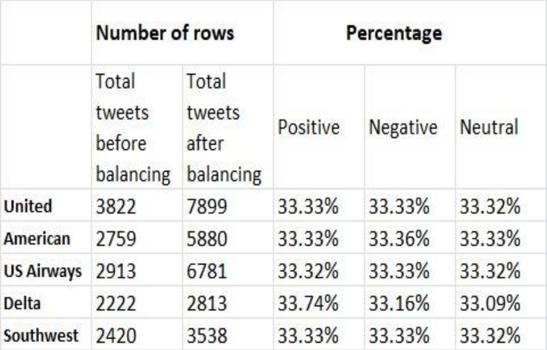


Figure 3: Bar chart of F1 Score for imbalanced dataset

All the algorithms did pretty well on US Airways and United datasets compared to others. Few algorithms gave good results on selected datasets, dataset size being the result for imbalanced and few did not deliver the expected results. Random Forest, ID3 were nearly consistent in their result for all the datasets. Whereas, other classifiers have peak and pit in their result for the same. Random Forest resulted into highest accuracy with 80.12%. The Second highest is of SVC with 79.21% .As we can see in the Table 2 the dataset is imbalanced, thus accuracy is not as expected.

5.2 Balanced Datasets

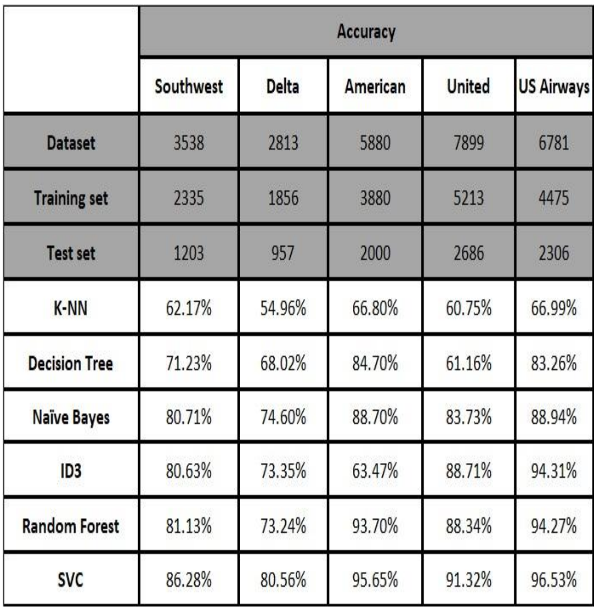
In balanced dataset, there are same number of input samples for target class. For doing the same, the technique which is used to modify the unequal data classes to create balancing, is sampling. Due to its balanced position, Virgin America is not considered. Table 4 below shows the information of total tweets before and after balancing with percentage of tweets as positive, negative and neutral.



	Number of rows		Percentage		
	Total tweets before balancing	Total tweets after balancing	Positive	Negative	Neutral
United	3822	7899	33.33%	33.33%	33.32%
American	2759	5880	33.33%	33.36%	33.33%
US Airways	2913	6781	33.32%	33.33%	33.32%
Delta	2222	2813	33.74%	33.16%	33.09%
Southwest	2420	3538	33.33%	33.33%	33.32%

Table 4: Percentage tweet summary of before and after balancing

After applying different classifier on the balanced datasets, Table 5 represents accuracy of the machine learning classifiers of the same for all airline datasets also including the number of tweets after splitting for train and test.



	Accuracy				
	Southwest	Delta	American	United	US Airways
Dataset	3538	2813	5880	7899	6781
Training set	2335	1856	3880	5213	4475
Test set	1203	957	2000	2686	2306
K-NN	62.17%	54.96%	66.80%	60.75%	66.99%
Decision Tree	71.23%	68.02%	84.70%	61.16%	83.26%
Naïve Bayes	80.71%	74.60%	88.70%	83.73%	88.94%
ID3	80.63%	73.35%	63.47%	88.71%	94.31%
Random Forest	81.13%	73.24%	93.70%	88.34%	94.27%
SVC	86.28%	80.56%	95.65%	91.32%	96.53%

Table 5: Accuracy of classifiers on balanced dataset

Figure 4 below displays the accuracy table's bar chart, where the colours blue, red, olive green, purple, sky blue, and orange, respectively, stand in for KNN, Decision Tree, Nave Bayes, ID3, Random Forest, and SVC.

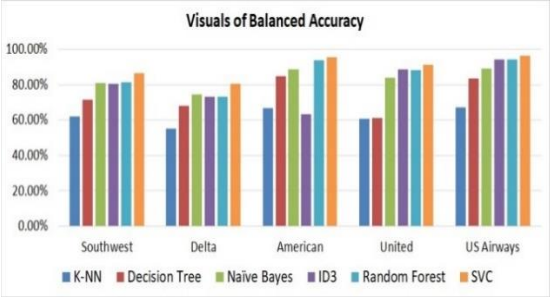
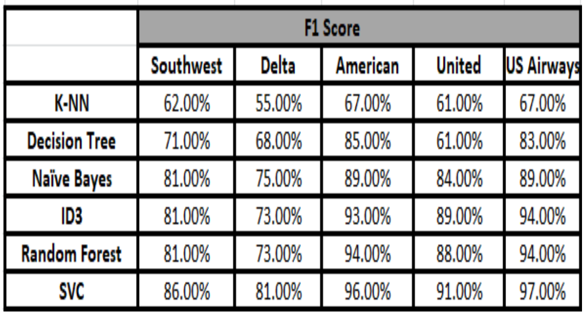


Figure 4: Bar chart of accuracy on balanced dataset

Table 6 displays the percentage F1-Score for each unique airline dataset for machine learning classifiers on balanced datasets for all airlines.



	F1 Score				
	Southwest	Delta	American	United	US Airways
K-NN	62.00%	55.00%	67.00%	61.00%	67.00%
Decision Tree	71.00%	68.00%	85.00%	61.00%	83.00%
Naïve Bayes	81.00%	75.00%	89.00%	84.00%	89.00%
ID3	81.00%	73.00%	93.00%	89.00%	94.00%
Random Forest	81.00%	73.00%	94.00%	88.00%	94.00%
SVC	86.00%	81.00%	96.00%	91.00%	97.00%

Table 6: F1 Score of classifiers on balanced dataset

Figure 5 below displays the bar chart for the aforementioned F1-Score table. The colours blue, red, olive green, purple, sky blue, and orange, respectively, stand in for KNN, Decision Tree, Nave Bayes, ID3, Random Forest, and SVC.

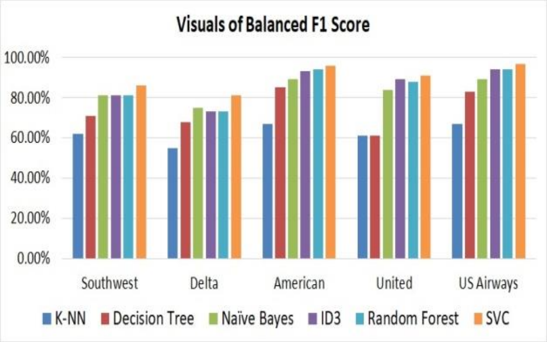


Figure 5: Bar chart of F1 Score on balance dataset

Despite its poor performance on all datasets, KNN gave greater accuracy in imbalanced than balanced positions on a handful of datasets. With the exception of the 'American' airline. Dataset, ID3 performed well on rest. SVC and Random Forest Classifier leads the league with a consistent difference of around 15-20% between imbalanced and balanced positions in all datasets. Naïve Bayes also showed promising results.

5.3 Random Forest Classifier

Leo Breiman introduced Random Forest in 2001. It has the capacity to significantly improve performance based on developing concepts from Ensemble Learners, commonly known as Decision Trees. Growing ensembles of trees and allowing them to vote for the most popular class has been proven to significantly increase categorization accuracy [17]. Bootstrapping (bagging) is used to create tree forest to avoid highly correlated trees. For each decision tree, nodes importance is calculated using entropy. Equation (6) shows the Entropy equation where pi is the probability of picking an element of random class i.

$$\text{Entropy} = \sum_{i=1}^{c} - p_i * \log_2(p_i) \qquad (6)$$

The quantity of information that was improved in the nodes before they were split to make new judgments is referred to as information acquired in the decision tree, equation for calculating Equation (7) represents Information Gain, where H(S) is the entropy for root node, Sv is number of sample of particular node and S is the total number of sample. Information Gain is the entropy of parent node minus the weighted average times entropy of child nodes.

$$\text{Gain}\ (S, f) = H(S) - \sum_{i \in N} \frac{|\, s_v \,|}{|\, S \,|} \qquad (7)$$

Figure 6 demonstrates how the random forest classifier selects the tree that can predict the test text set with the greatest accuracy by voting for the trees that is to be best trained.

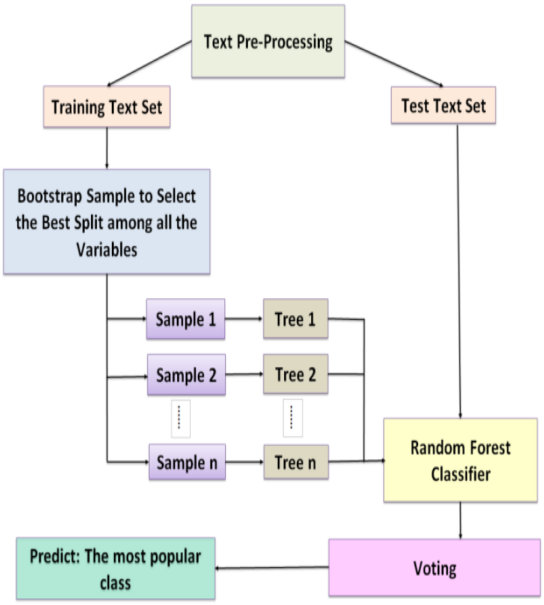
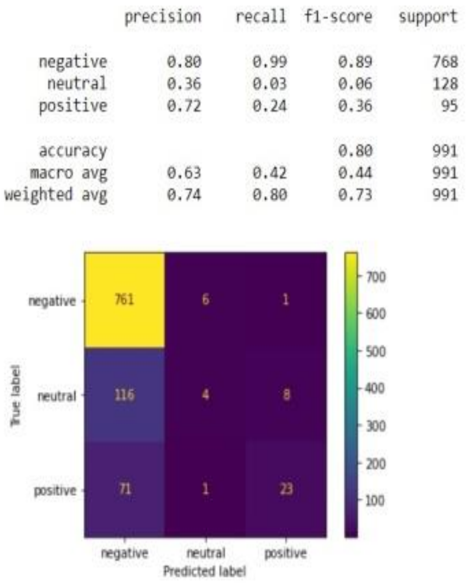


Figure 6: Random Forest on text classification

To begin building the decision tree, the feature with the greatest information gain should be taken into consideration as the root node. Information gain is used by the ID3 algorithm to build the decision tree whereas CART algorithm uses Gini index to do the same.

Pruning: The main idea behind Pruning is to reduce the size of decision trees to prevent growing to its full depth that lack the ability to classify instances. One can trim the trees and stop them from overfitting by adjusting the decision classifier's hyper parameters. The below flow chart is representing how random forest classifier works on textual data.

Figure 7(a), (b) shows the confusion matrix of the best performing airline on both imbalanced and balanced datasets. Random forest shows highest result on US Airways dataset with 80% F1 Score in imbalanced and 94% F1 Score in balanced. Confusion matrix is an N x N matrix called a confusion matrix is used to assess the effectiveness of a classification model, where N is the total number of target classes. This information is used to determine performance metrics including precision, F1-score, accuracy, and recall as well as the number of accurate and incorrect predictions generated by a classifier.

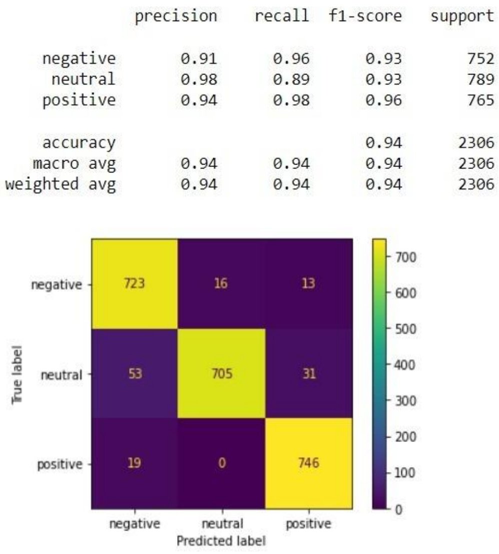


```
              precision    recall  f1-score   support

    negative       0.80      0.99      0.89       768
     neutral       0.36      0.03      0.06       128
    positive       0.72      0.24      0.36        95

    accuracy                           0.80       991
   macro avg       0.63      0.42      0.44       991
weighted avg       0.74      0.80      0.73       991
```

Figure 7(a): Confusion matrix of the best performing airline on imbalanced dataset for RFC



```
              precision    recall  f1-score   support

    negative       0.91      0.96      0.93       752
     neutral       0.98      0.89      0.93       789
    positive       0.94      0.98      0.96       765

    accuracy                           0.94      2306
   macro avg       0.94      0.94      0.94      2306
weighted avg       0.94      0.94      0.94      2306
```

Figure 7(b): confusion matrix of the best performing airline on balanced dataset for RFC

Algorithm 1 is of random forest classifier representing how it processes the text input and outputs the prediction

Algorithm Ensemble Method: Majority Voting

```
Input: US Airways Dataset
        Model = Random Forest Classifier
        Input rows size = 6781
        No. of class = 3
Output: class label predicted (y) via majority voting
Output classes: (Positive, negative or neutral)
BEGIN:
    function(text):
            text = text.lower()
            words = nltk.word_tokenize()
            Lemmatize = [wl.lemmatize(word) for word in words if word not in stopwprds('English')]
  US_Airways['text'] = US_Airways['text'].apply(function)
  Text = TfidfVectorizer()
  Train_Test_split in the ration of 66% and 34%
  Rfc = RandomForestClassifier(n_estimator=250, criterion = 'entropy')
  print (Rfc.predict())
END
```

Algorithm 1: Random forest algorithm

5.4 Support Vector Classifier

It is the classification method of Support Vector Machines (SVMs). Vladimir N. Vapnik and Alexey Y. Chervonenkis developed the first SVM algorithm in 1963. By using the kernel trick on maximum-margin hyperplanes, Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik proposed a method to build nonlinear classifiers in 1992. SVC uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. It is also versatile in that different kernel functions can be specified for the decision function. SVC is effective in high dimensional spaces and is still effective in situations where the number of dimensions is greater than the number of samples. The core notion of SVM is that classes can be separated using hyperplanes. Maximal margin classifier – divider that increases the space between the classes. There is a soft margin between the threshold and the observations – allowing bias variance trade off to perform miss classification. In SVM, the data is projected to a higher dimension using kernels, where a hyperplane is used to separate the data.

Equation (8) is of hyperplanes in a **p** dimension space for parameter **β.** A flat subspace with **p-1** dimensions is referred to as a hyperplane in a **p**-dimensional space.

$$\beta_{\circ} + \beta_{1} X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p = 0 \qquad (8)$$

If a point is defined as the hyperplane, X = (X_1, X_2, …., X_p)^T gratifies equation (8), X is then lying in the hyperplane, depicting in equation (9). On the other hand, if X does not meet equation (8) and instead,

$$\beta_{\circ} + \beta_{1} X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p < 0 \qquad (9)$$

then we know X lies on one side of the hyperplane. As an alternative,

$$\beta_{\circ} + \beta_{1} X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p > 0 \qquad (10)$$

X is then known to be located on the other side of the hyperplane depicting in equation (10). The separating hyperplane that is furthest from the training observations is referred to as the greatest margin hyperplane. The margin is the shortest (perpendicular) distance between the training observations and the hyperplane; the maximum margin hyperplane is the one with the biggest margin. The next stage is to define the largest margin classifier, which we achieve by resolving the following optimization problem as shown in equation (11) it makes sure that there is as much space as feasible between the separation plane and the closest observations.

$$\text{Maximize} = M \qquad (11)$$

$$\beta_\circ, \beta_1 \ldots \beta_p, M$$

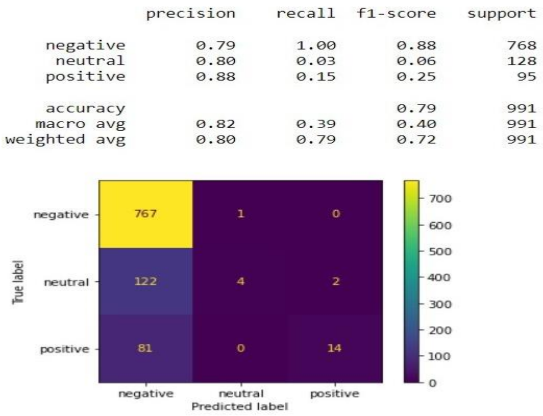
So we are essentially trying to choose the best beta coefficients that can maximize this margin, but we do have some constraints subject to shown in equation (12) it guarantees the exclusivity of the hyperplane.

$$\sum_{j=1}^{p} \beta_j^2 = 1 \qquad (12)$$

$$y_i\ (\beta_\circ + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip}) >= M(1-\epsilon_i) \qquad (13)$$

Using Equation(13), each observation is guaranteed to land on the proper side where this $\boldsymbol{\epsilon}$ is a relationship between some set of misclassifications [18]**.** Different kernel functions available are: linear, poly, rbf and sigmoid.

Figure 8(a), (b) shows the confusion matrix of the best performing airline on both imbalanced and balanced datasets. SVC shows highest result on US Airways dataset with 79% F1 Score in imbalanced and 97% F1 Score in balanced.

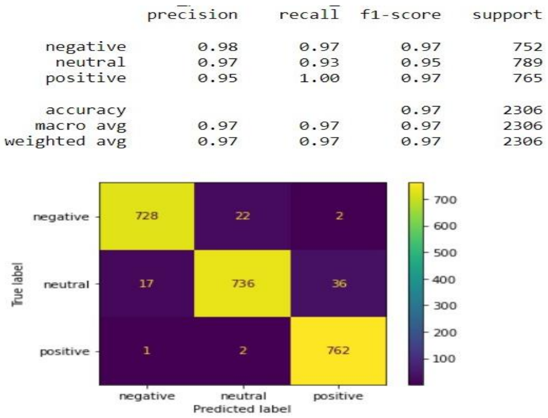


```
              precision    recall  f1-score   support

    negative       0.79      1.00      0.88       768
     neutral       0.80      0.03      0.06       128
    positive       0.88      0.15      0.25        95

    accuracy                           0.79       991
   macro avg       0.82      0.39      0.40       991
weighted avg       0.80      0.79      0.72       991
```

Figure 8(a): confusion matrix of the best performing airline on imbalanced dataset for SVC



```
              precision    recall  f1-score   support

    negative       0.98      0.97      0.97       752
     neutral       0.97      0.93      0.95       789
    positive       0.95      1.00      0.97       765

    accuracy                           0.97      2306
   macro avg       0.97      0.97      0.97      2306
weighted avg       0.97      0.97      0.97      2306
```

Figure 8(b): confusion matrix of the best performing airline on balanced dataset for SVC

6. Business Aspects

Opinion mining, also referred to as business sentiment analysis, the action of recognising and indexing texts based on the tone they convey, it can be a game-changer in completely revitalising a brand. Making quick decisions and eliminating guesswork are made possible by information that is rich in insights. The airline company can target its weak points and enhance its services by forecasting the percentage of favourable, negative, or neutral attitudes. They can make adjustments to the current state of the market and improve client satisfaction based on the reviews. There are an overwhelming number of uses for sentiment analysis in business. What tool is selected and how effectively it is employed to the advantage will determine how much more commercial value it can derive for company from sentiment analysis.

7. Conflict of Interest

KNN being a distance based machine learning algorithm should have resulted better after balancing but it not did so.

8. Conclusion

Nowadays, social media plays a crucial role in how people connect in various spheres of life and how work is accomplished. There are several social media sites out there. The majority of young people, corporate sectors, government sectors, and other sectors utilize Twitter. A Tweet is any type of content shared on Twitter. In this study, reviews submitted by airline passengers are taken into consideration. Reviews are classified as either positive, negative, or neutral. It would be preferable for airlines to target and address bad customer feedback in order to prevent future travellers from having the same experience, which is why this model was created. The technique utilised for this sort of work is sentiment analysis. It is a technique for examining text data to determine its purpose. To estimate overall sentiment, the objective is to automatically identify and classify opinions stated in the text. Natural Language process is used for sentiment analysis in this study. The dataset was obtained from kaggle. The texts, which are tweets, have undergone some pre-processing, such as tokenization, stopwords removal, and the elimination of a few additional frequently used terms associated with tweets. The dataset has been divided into six datasets depending on the different airlines US Airways, Delta, American, Virgin America, United, and Southwest, after that features are extracted using the TF-IDF approach. The datasets were imbalanced, so the classifiers were initially trained on the imbalanced data, then on the balanced data which were balanced using hybrid sampling. Random Forest, SVC, KNN, Naive Bayes, Decision Tree, and ID3 are the classifiers that are used for training the data. The accuracy and F1 Score evaluation measures of each classifier were compared. For a clearer and faster comprehension of the classifiers' results, tables, bar charts, and confusion matrices are given throughout the study. Random Forest performed well on imbalanced side. After balancing with hybrid sampling, SVC resulted with highest accuracy. In future, the business would be more efficient and the service would be better if this model could retrieve data in real-time. There are several feature selection techniques that can maximise accuracy and Deep Learning methods that are pertinent could improve the outcomes.

Acknowledgement

We are very grateful to our mentor who supported us to work on this paper. We would like to thank her for giving valuable suggestions and our team members for coming up with exceptionally great ideas.

References

[1] S.A. Salloum, C. Mhamdi, B. Al Kurdi, K. Shaalan, “Factors affecting the Adoption and Meaningful Use of Social Media: A Structural Equation Modeling Approach,” International Journal of Information Technology and Language Studies, 2(3), 2018.

[2] Hastari Utama, "Sentiment Analysis in Airline Tweets Using Mutual Information for Feature Selection", 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE).

[3] M. Alghizzawi, S.A. Salloum, M. Habes, “The N role of social media in tourism marketing in Jordan,” International Journal of Information Technology and Language Studies, 2(3), 2018.

[4] S.A. Salloum, W. Maqableh, C. Mhamdi, B. Al Kurdi, K. Shaalan, “Studying the Social Media Adoption by university students in the United Arab Emirates,” International Journal of Information Technology and Language Studies, 2(3), 2018.

[5] Sachin Kumar, Marina I.Nezhurina, "Sentiment Analysis on Tweets for Trains Using Machine Learning" (2020).

[6] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan, "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques", International Journal of Engineering and Technology, 2016.

[7] Mohd. Istiaq Hossain Junaid, Faisal Hossain, Udyan Saha Upal, Anjana Tameem, Abul Kashim, Ahmed Fahmin, "Bangla Food Review Sentimental Analysis using Machine Learning", 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022.

[8] Vineet Jain, Mayur Kambli, "Amazon Product Reviews: Sentiment Analysis" (2020).

[9] Amal Alazba, Nora Alturayeif, Nouf Alturaief, Zainab Alhathloul, "Saudi Stock Market Sentiment Analysis using Twitter Data", International Conference on Knowledge Discovery and Information Retrieval - KDIR 2020.

[10] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol.1, pp. 142-150, 2011.

[11] Saeed Mian Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory", 2011.

[12] Shri Bharathi, Angelina Geetha, "Sentiment Analysis for Effective Stock Market Prediction", International Journal of Intelligent Engineering & systems, INASS, 2017.

[13] Upma Kumari, Dr Arvind K Sharma, Dinesh Soni, "Sentiment Analysis of Smart Phone Product Review using SVM Classification Technique", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017), 2017.

[14] Saeed Mian Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory", 2011.

[15] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentiment Reviews using N-gram Machine Learning Approach", Expert System with Applications, 2016.

[16] Nanlir Sallau Mullah (Member, IEEE), Wan Mohd Nazmee Wan Zainon (Member, IEEE) ,"Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review", IEEE Access, 2021.

[17] Breiman L: Random Forests.Machine Learning.Vol.45(2001), p.5-32, Kluwer Academic Publishers, 2001.

[18] Isaiah Boone, “SVM and the Application of Prediction Rules”, Pomona College Senior Thesis in Mathematics, 2016.