US AIRLINE TWITTER DATA ON SENTIMENT ANALYSIS USING DEEP NEURAL NETWORK

[1]Ajai Rai, [2]Dr. Daya shankar pandey
[1]Department of Information Technology, [2]Department of Computer science
RKDF-Institute of Science &Technology
SRK University
Bhopal, India

Abstract*—Many people like expressing their opinions on the internet on anything and everything, including social events, specific goods, or services from any business. Sentiment analysis may be used to assess people's attitudes or sentiments by analyzing data from sites that include a lot of opinions, such as Twitter. There are a variety of airline services available across the globe, each of which offers a unique set of amenities to its consumers. Customers may be satisfied or dissatisfied with the services provided by the airlines. Customers are unable to share their opinions instantly, thus airline services offer a Twitter blog where they may submit feedback on their services and products. The number of people using Twitter has risen to improve the quality of services. This research introduces a novel deep classification approach to increase the accuracy of sentiment analysis by using a large amount of data. The tweets of services are divided into two polarities, which are positive and negative, respectively. Specifically, the purpose of this study is to give an analysis of the content tweets text of users' emotions in the services provided by United States airline companies, as well as to examine the application of Machine Learning (ML) approaches to predict sentiment from United States airline tweets. The Machine Learning (ML) methods classification method is a Deep neural network measuring the accuracy achieved by the sigmoid and relu activation function was compassed in the training and validation phase. The precision, recall, f1-score, and accuracy metrics for sentiment analysis have been identified for all of the classification techniques described above in this study. In addition, the average predictions of classifiers, as well as the accuracy of the average predictions of classifiers, were calculated to achieve high accuracy. The results show that Deep Neural Networks (DNNs) outperform Logistic Regression (LR) classifiers when it comes to classification accuracy.

Keywords*— Sentiment Analysis, Opinion Mining, Tweets, Machine Learning, Logistic Regression, Neural Network.

I. INTRODUCTION

People's thoughts and opinions are now expressed differently because of the rise of the Internet. The majority of it is now done via blogs, internet forums, product review sites, social media, and so on. In today's world, millions of individuals use social networking sites like Facebook, Google+, Twitter, and so on to communicate their thoughts, feelings, & opinions. As a result of online communities, we now have a medium that allows users to share information & exert influence on one another. There is a massive amount of sentiment-rich information being generated by social media in the form of posts, tweets, and other changes. Additionally, social media offers companies a platform to communicate with their consumers to advertise their products and services. When it comes to making decisions, most people rely heavily on user-generated material. If a person is considering purchasing a product or using a service, they will first research the product or service's online reviews and debate the matter on social media. The enormous volume of data created by consumers is beyond the capabilities of the average consumer. Automated sentiment analysis approaches are frequently employed since there is a pressing need to do so (A. and Sonawane, 2016). One of the most talked-about subjects on Twitter right now is air travel. Twitter is a common medium for airline customers to discuss their experiences. If this data is processed utilizing machine learning methods, it may yield insights that assist determine a passenger's degree of comfort throughout the journey. In the area of sentiment analysis, there is an enormous amount of research available. In terms of long-distance travel, air travel is one of the most convenient options, both domestically and globally. Globally, there is a big no. of ASPs (airline service providers). The airline industry is driven to compete for consumers by the competitive environment. Travelers, on the other hand, think about numerous factors before settling on a certain airline. These factors may include costs like fares, journey duration & luggage allowance, as well as reviews from previous customers and so on. As a result, all ASPs are attempting to enhance their facilities and in-flight comforts to attract more consumers (Alqahtani and Al-qahtani, 2021).

Opinion mining (OM), also known as sentiment analysis (SA), is a computational study of people's sentiments, emotions, assessments, & attitudes toward various types of entities, including products, services, organizations, persons, concerns, events, & topics, and also the aspects associated with each. Review sites, forums, blogs, micro-blogs, Twitter, and other Web-based social media have all sprung up at the same time that this new industry is exploding, thanks to the unprecedented amount of user-generated content that is now available in digital form. One of the most active study topics in natural language processing has been sentiment analysis since the early 2000s. Data mining, Web mining, text mining, & info retrieval are all areas where it is used extensively. Since its value to society and business has grown beyond

computer science, it has been incorporated into a wide range of disciplines from marketing to politics to health care and beyond. The expansion of views is because they play a fundamental role in almost all human activity and have a significant impact on our actions. What we believe and how we make decisions are heavily influenced by what others believe and how they view and interpret the world. For this purpose, we frequently seek the advice of others while making a choice. As well as people, companies may benefit from this as well (Zhang, Wang and Liu, 2018).

Text sentiment analysis is an ML technique that detects polarity in text, whether it is a document, paragraph, phrase, or clause. SA may be used to any kind of text. Emotion detection and fine-grained sentiment analysis are only a few of the forms of SA available. Other types include aspect-based analysis with intent analysis. If a person is sad, pleased, or furious, emotion detection inspects their language to determine if they are in that condition. This information may be used to influence any public choice. Faintly nuanced sentiment analysis is concerned with the public's perceptions of a certain product or a particular candidate in elections to determine if they are favorable, negative, or neutral in their feelings toward that candidate. Following doing an aspect-based analysis, it is possible to extract opinions of consumers about a single portion of a product, such as an automobile engine, and utilize the information to improve that part after the study (Saad, 2020) (Akhmetgaliev, Gafarov and Sitdikova, 2020)(Saad, 2020).

The remaining work is summarized as follows: Section II addresses past research that has been conducted on the issue of this study and is relevant to it. It examines prior airline Twitter apps using sentiment categorization to determine their effectiveness; Section III describes the steps involved in putting the suggested technique into action; A full analysis of each outcome is presented in Section IV, which presents the results and assessments of the experiment; & As the study comes to an end in Section VI, we review the experiment's results and recommend further research.

II. LITERATURE REVIEW

Many types of research have been done in the Sentiment analysis field. They analyze the behaviors of users' live data to extract the feelings of ordinary people towards any subject, trend, product, etc. several studies focus mainly on extracting useful information from the users' natural language and processing it to get the real feelings. It has generated interest with the ever-rising use of the Internet by people to share their opinions. Previous research has used Machine Learning algorithms to classify tweets for different airline companies. The majority of researchers have looked at the technique of sentiment analysis by looking for emotions in the content of tweets.

(Yimga, 2021) According to this paper, the COVID-19 pandemic has had an adverse influence on flight delays in the United States airline sector. Our findings, which are based on daily data on COVID-19 cases and aircraft on-time performance, but which are adjusted to account for product, airline, & market variables, show that rises in reported COVID-19 cases are related to decreases in both arrivals and departures delays. Specific to COVID-19 situations, a standard deviation increase lowers arrival time by 1 min 42 seconds & departure time by 2 minutes on average. While the epidemic has had an adverse economic impact, our findings imply that one silver lining has emerged: planes are leaving and arriving with less delay as a consequence of the pandemic.

(Hrazi et al., 2021) Designing and implementing a SA that will categorize genuine tweets gathered via the Twitter API into one of the following categories above is their primary aim. With the use of supervised learning techniques including such Logistics Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), and Decision Trees (DT), researchers were able to conduct a sentiment analysis using machine learning. To do this, researchers preprocess the data set that they have selected to train our classification algorithm. Researchers extracted features from the preprocessed tweets using bag-of-words & TF-IDF approaches, respectively. In addition, researchers utilized unigrams, bigrams, and trigrams to rank our features to determine which characteristics had the highest prediction accuracy for the classifier. The classification strategies that we utilized performed the accuracy for both validation as well as test data when employing tri-gram features with stop words & TF-IDF feature extraction methodology, which is the most accurate classification method.

(Saad, 2020) According to the findings of this study, researchers suggested an ML model classify Twitter postings into three categories: good, negative, or neutral. Researchers tested their model on a dataset that included tweets from six different airlines in the United States. They used six different machine learning approaches. Twitter classification algorithms include SVM, LR, RF, XgBoost (XGB), NB, and DT. Finally, during the validation phase, researchers divided the data into two groups: 70 percent for training and 30 percent for testing. The K-Fold Cross-Validation approach was used to test and validate the data throughout the testing and validation phases. Lastly, they computed Accuracy, Precision, Recall, & F1-score for each classifier using the aforementioned metrics. After evaluating the outcomes of each classifier, we discovered that SVM had the best accuracy, with an accuracy of 83.31 percent.

(Flores, Andrew, 2020) In this study, researchers use the NB Classifier to extract sentiment from social media postings, namely tweets, to determine consumer satisfaction with the use of five commercial airlines' services. When preparing text data for classification, researchers use natural language processing methods to ensure that the data is ready for categorization. An NB Classifier is utilized, which accepts as input tokenized lists of text that have been cleaned up. The model is then trained using samples of positive & negative tweets that have been categorized. After that, researchers input the model unlabeled tweets, which is what it

categorizes as either negative or positive based on its classification. By examining performance indicators like accuracy, precision, recall, and a confusion matrix (and the amount of time it takes to train and validate the data), researchers can show the correctness and effectiveness of the model. Every time they conducted an investigation, we used a different amount of data. Their NB model can obtain an average accuracy rate of 90.8 percent for the top tests, which is much higher than the industry standard. According to our findings, both personal and corporate users may benefit from a tool that can extract sentiment from unstructured data, such as social media postings.

(Sharma, Rahamatkar and Sharma, 2019) The purpose of this effort is to determine whether or not people are pleased as a consequence of their descriptive Tweets by analyzing their words and their facial expressions in a quantitative fashion. The innovative aspect of this study is the examination of ambiguous Tweets as well as the neutralization of such tweets using the suggested algorithm. The whole approach is founded on a Twitter dataset; in this case, we are utilizing a US airline dataset, & performing several levels of mining and processing to get the most accurate results. It has been presented here to enhance the sentiment analysis model based on the NB classification method to categorize tweeted messages according to their emotions and transform tweets from ambiguous to positive or negative.

(Rane and Kumar, 2018) The research included working with a dataset that included tweets from six major US airlines as doing a multi-class SA on the data. The study was carried out by using seven different classification techniques: the DT, the RF, the SVM, the K-Nearest Neighbors (KNN), the LR, the Gaussian Nave Bayes, as well as the AdaBoost. The DT was the most often used classification strategy. In this study, 80 percent of data was applied to train classifiers, while the remaining 20 percent was applied to test them. The result of the test set is a sentiment of tweets (good, negative, or neutral). Data that has been collected, the accuracy of each categorization technique was computed to make a comparison between them, and the total sentiment count, which included all six airlines, was shown.

(Wan and Gao, 2016) This paper presents an ensemble sentiment classification strategy focused on the Majority Vote principle, which was employed in various classification techniques, which included the NB, SVM, Bayesian Network, C4.5 DT, as well as Random Forest methodologies, all of which were trained on the same dataset. The classifiers in these studies were validated using a 10-fold assessment on the same dataset of 12864 tweets, which was utilized to train as well as evaluate six individual classification techniques as well as the suggested ensemble technique. The findings demonstrate that the suggested ensemble strategy outperforms the individual classifiers in this airline service Twitter dataset. According to our findings, the ensemble strategy may be able to increase the overall accuracy of Twitter sentiment categorization for additional services in addition to Twitter.

III. RESEARCH METHODOLOGY

This section contains the description of the problem identification as well as the proposed methodology to perform the sentiment analysis on the selected dataset.

A. Problem Identification

Twitter Sentiment Analysis is a technique for identifying attitudes and conjectures in tweets, which is becoming more popular as people express their ideas and opinions on social media sites such as Twitter and Facebook. The major classification stage in this approach is to determine the contradiction or mood of the tweets and then categorize them into three categories: positive, negative, & neutral tweets. The problem with sentiment analysis is identifying the proper congruous sentiment classification algorithm to use to list the tweets logically. Currently, the airline industry is a significant player on the global stage. It is necessary to examine opinion mining to maintain that area alive and current. Tweets are one of the most important sources of opinion mining data since they include a big no. of tweets that must be processed & evaluated to create choice or improve a certain service. Furthermore, sentiment mining has been expanded to solve issues including discriminating between objective and subjective concepts and predicting future behavior.

B. Proposed Methodology

In this study, we investigated consumer comments regarding airline services using opinion mining, which is one of the text sentiment applications. Tweets are one of the most important sources of opinion mining data since they include a big no. of tweets that must be processed & evaluated to make a choice or improve a certain service. This study suggested an ML model for categorizing Twitter tweets into positive and negative categories, which was tested and validated. We tested our model on a dataset that included tweets from six different airlines in the United States. For this, the data is taken from Github and labeled as a positive and negative category before preprocessing it. Pre-processing of the dataset was carried out as part of the methodological phases of this study. We started our model by preprocessing steps where Using a set of criteria, we cleaned tweets as well as used them as training data for multivariate statistical models. This is similar to how knowledge-based expert systems are commonly used as training data. To transform the text data into numerical values or vectors to represent them as feature vectors, we used the Bag of Words (BoW) model, which we developed later. Afterward, we divided our data into two groups: 90 percent for training, and 10 percent for testing. The last phase is fitting the data into several machine learning models, which are described below. In the latter category, there are many alternatives, the most popular model in sentiment analyzing is neural networks (NN). This model uses several deep dense layers so the model is called a Deep neural network. Finally, it tests and validates the model on the test dataset for analyzing the feedback about US airlines in two categories.

1. Preprocessing of Twitter Data

Twitter data may be in an unstructured format that is not good for extracting features. Tweets may consist of empty spaces, stop words, slang, special characters, hashtag, emoticons, time stamps, abbreviations, URLs, etc. for mining these data we should have to pre-process the data. During this phase, we completed five cleaning procedures. As a first step, we eliminated duplication and complexity by eliminating phrases such as "to, for, and how," which were previously used to impede the flow of information. Using the @ sign as well as the @airline company, we eliminated all punctuation from across all tweets in the second stage. Because each tweet was begun by the @airline company, we eliminated the "@" symbol as well as the airline company from all tweets in the third phase. The final stage included changing all of the letters to lowercase to unify all of the terms since the machine is case-sensitive. The next stage is stemming, which involves determining the word's origin. For illustration, the word "flew" becomes "fly" once it has been stemmed. Last but not least, we are prepared to begin constructing our corpus, which is a collection of text that represents our cleaned tweets. Following that, we utilized BoW to encode tweets as a feature vector, which can then be used in ML models to make predictions.

Punctuation was eliminated from the data since it does not add to the text analysis in this research. Punctuation makes texts more understandable but also degrades the models' ability to distinguish between punctuation and other characters. The numeric values from the tweets were eliminated in the next stage since they did not affect the text analysis. The removal of numeric information reduces the complexity of the model training process.

Stemming is an essential approach in pre-processing since it helps to improve the effectiveness of the system by eliminating affixes from words & transforming them into their base form. For example, words in a book may appear in a variety of forms, all of which have fundamentally the same meaning. Porter stemmer is used to stemming in this position.

Tokenization is the process of breaking down the text into tokens before translating them into vectors. It is also simpler to filter out tokens that aren't needed. For example, you may break down a text into paragraphs or phrases into individual words. Specifically, we are tokenizing the reviews into words in this instance.

2. Bag of Words (BoW)

The boW is used in this research text convert to numeric data or vectors. BoW model is a decrease & summarised depiction of a text document that is constructed from selected portions of text that are related to specific criteria, including such frequency of occurrence of words. Several disciplines, including computer vision, natural language processing (NLP), Bayesian spam filters, document categorization as well as information retrieval via Machine Learning, make use of the BoW approach. BoW refers to a body of text, including a document or a phrase, as a "bag of words" in which words are collected. The BoW procedure results in the creation of lists of words. These words in a matrix are not sentences that structure sentences and grammar, as well as the semantic link between these words, is not taken into consideration in their gathering & development. The words in a sentence are often indicative of the substance of the phrase. While grammar as well as order of occurrence is disregarded, the number of times a word appears is tallied, and this number may be used to define the emphasis areas of the documents later on.

Feature vector: An arbitrary list of integers extracted from the output of a neural network layer is known as a vector. Using this vector, which is a dense representation of the input picture, a user may do a range of operations like as ranking, classification, including clustering on the image data. The approach of starting with a model that has previously been trained on a big dataset is popular in the field of machine learning.

3. Machine Learning Classifiers

It is possible to discover Sentiment Analysis from a tweet using two fundamental techniques: one is the use of lexicon-based methods, the other was the use of machine learning techniques. For this work, the neural network classification technique has been used as an ML technique of numerous forms for text categorizing & SA of Twitter data. These strategies are used to practice algorithms, with the task of engaging the algorithm with the train data and the test data being assigned to it by the programmer. For determining the correctness of a given dataset, a variety of techniques, including neural networks and logistic regression, are used.

Logistic Regression (LR):

The dataset is analyzed using the LR statistical approach, which yields a binary conclusion as a result. It is possible that the dataset included one or more autonomous variables in addition to the dependent variables. It is these factors, which are inherently dichotomous, that decide the outcome. As a consequence, there are only two conceivable outcomes. It is a subcategory of regression that is used to predict binary and categorical outcomes most efficiently.

A technique known as the LR approach is used to control the influence of a large number of autonomous variables that are presented at the same time. It also estimates either one of the two independent groups of factors, which is another advantage of this strategy. The maximum likelihood approach is used to create the best-fitting function, and this method is used to optimize the probability of categorizing the recognized data into the appropriate division (Celine, Dominic and Devi, 2020).

There are just two broad categories in an LR, hence the dependent variable has only 2 possibilities. According to most conventions, the presence of an event is classified as 1, while its absence is coded as zero. Recognizing, however, that codification affects the signal of the coefficients as well as, as a result, changes the substantive meaning of the coefficients It is vital to grasp the logic of regression analysis as a whole to comprehend how logistic regression works. Look at the conventional notation for the linear model:

$Y=\propto\beta X+\varepsilon$…….(1)

Y is the dependent variable, which is the thing we are

attempting to analyze, explain, or forecast. The independent variable is denoted by the letter X.

Deep Neural Networks (DNN):
In the field of classification, neural networks have emerged as an effective tool. Because of the current flurry of research activity in neural classification, it has become clear that neural networks offer an attractive alternative to a variety of traditional classification approaches. It is the similar mathematical elements that distinguish neural networks from other types of networks: The first thing to keep in mind regarding neural networks is that they are truly data-driven self-adaptive approaches, meaning that they may modify themselves in response to the data without any explicit characterization of the underlying model's functional or distributional structure being provided. Second, neural networks are universal functional approximators, which implies that they can estimate any function with arbitrary accuracy at any time and in any environment. Because every classification approach aims to establish a functional link between the members of a group as well as the properties of an item, the accurate detection of the underlying function is unquestionably critical to the process. Third, neural networks are nonlinear models, which allows them to be more flexible When modeling complicated interactions in the actual world. Furthermore, neural networks are capable of estimating an posterior probability, which serves as the foundation for developing classification rules and doing statistical analysis on the data collected. In contrast, several computer-based empirical assessments of neural networks for classification issues have been carried out under a range of different scenarios (Zhang, 2000).
It is a common network design that contains a new training technique, which is called Deep Neural Networks. Essentially, the DNN is a multilayer network (usually deep as well as containing many hidden layers), that each pair of linked layers represents a dense pair of connections. When described in this manner, a DNN is equivalent to an accumulation of dense layers.

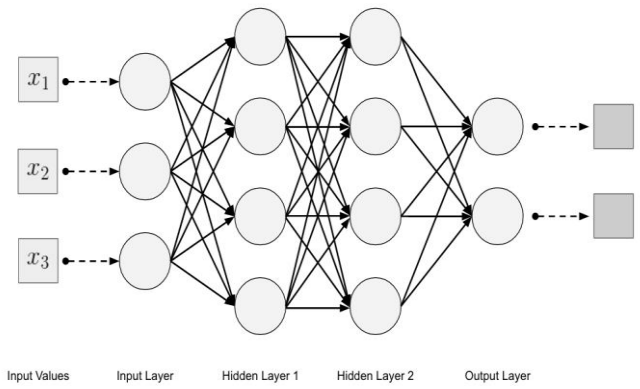

Figure 1: Modular Deep Neural Network Architecture

This DNN is made up of many layers that are interconnected. While the input layer gets data from different sources, like the attribute values of the relevant data entry, as well as the output layer creating the output of the network, the hidden layers are responsible for connecting the input as well as output layers. Node input values are calculated at every layer by multiplying the total number of incoming nodes by the appropriate weight associated with the connectivity between the nodes, which is calculated for each layer separately. This dense layer is used to train the network.

Sigmoid function: Every node's output value is determined by combining all of its input values with a preset function that is used by all other nodes in the network. Oj (sigmoid) functions are frequently utilized, & are described in the following manner (Erb, 1993):

$$o_j = \frac{1}{1+e^{-i}_j} \dots\dots (2)$$

here i_j is a sum of input nodes of j.

With its normalization to values between 0 to 1 as well as nonlinear nature, Erb considers that this function has two major benefits: it facilitates network learning as well as avoids overloading & dominance impacts. Domination happens when a single or few qualities have a very large impact on the projected target attribute, rendering another attributes useless and so dominating them (Erb, 1993).

Relu: Rectified Linear Unit (ReLU) is the most often utilized activation function in deep learning models. For just about any negative input, the function returns 0; for any positive input, it returns that value. In comparison to the sigmoid function, ReLu is much quicker to calculate, as well as its derivative is also much faster to calculate For neural networks, it has a major impact on the training & inference time required. As a result, it may be written as:

f(x)=max(0,x) (3)

A. PROPOSED ALGORITHM

Step 1. Start
Step 2. Collect the twitter-airline-sentiment dataset.
Step 3. Perform EDA on the dataset to analyze the dataset
Step 4. Then fine the dataset by applying preprocessing phase in which included punctuation and stop word removal, stemming, tokenizing, etc.
Step 5. Use BoW denotes data of tweets as a feature vector
Step 6. Split data into training and test set
Step 7. Apply a Neural network classifier for classification.
Step 8. Perform training and learning of the model
Step 9. Test and estimate the performance of the model by applying it to the test set.
Step 10. Achieved predicated outcomes and polarize the sentiments.
Step 11. End

B. PROPOSED FLOWCHART

As mentioned earlier, our goal is to do SA for Twitter data. By using various kinds of machine learning classifiers, we will build a neural network classifier. once it gets trained then we have followed different steps to sentiment analysis as mentioned below diagram:

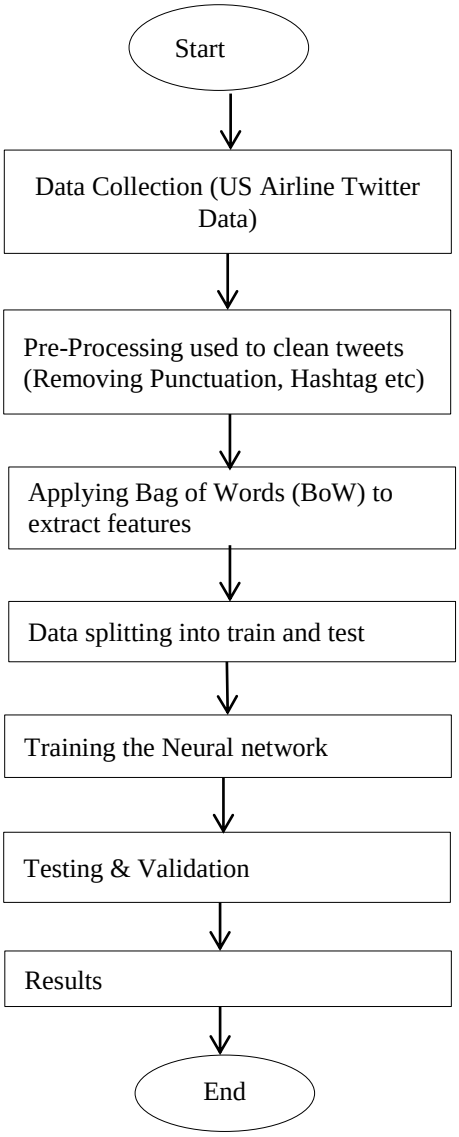

Figure 3. Flow Chart of proposed Methodology

IV. RESULTS AND DISCUSSION

Description of the dataset used for sentiment analysis, as well as its visualization, are included in this section, as well as discusses the analysis of experiments performed in the current research work. The proposed methodology in this research has been implemented in python 3.0 using the Tweets API dataset. The result is evaluated using various performance parameters that are accuracy, precision, recall, & F-measure.

A. Dataset Description

In this study, sentiment analysis using natural language processing (NLP) is performed on a Twitter US airline dataset. We utilized the "twitter-airline-sentiment" dataset, which was retrieved from the Kaggle platform. (Kaggle, 2020). The opinions of passengers are represented in the dataset via the use of Twitter messages. In all, 14640 tweets signify 6 airlines in the US, for a total of 14640 records. ach record is subdivided into three categories: positive and negative.

B. Explorational data analysis (EDA) Plots

EDA, to recognize the structure behind data include there, may be designated as how one or more datasets are examined

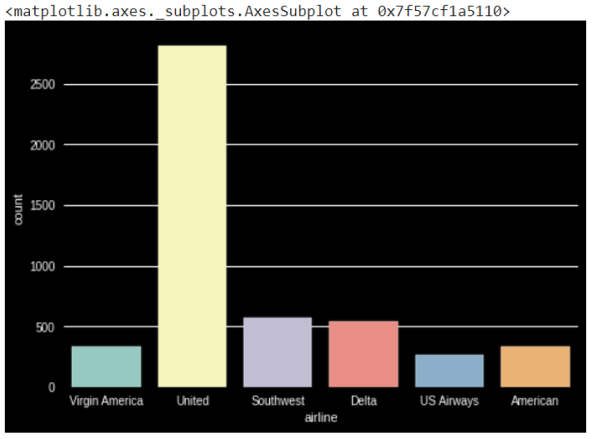

Figure 4: Airline count plot

Fig 4 represents the airline count plot for each airline. The six United States airline corporations, as well as the total amount of tweets gathered on each airline, are included in US Airline Twitter Trends.

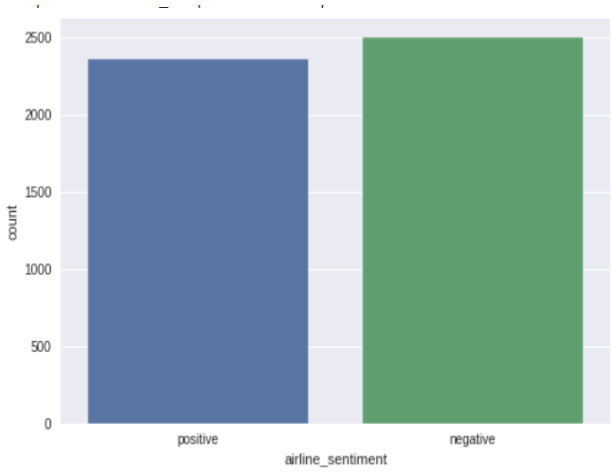

Figure 5: Sentiment count plot

Figure 5 depicts how customer sentiment varies across various airline services. United has the most amount of tweets with negative emotion, as well as the greatest number of tweets overall.

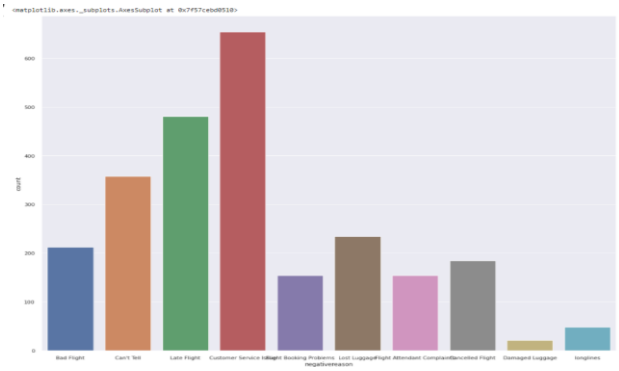

Figure 6: Negative reason count plot

Fig. 6 represents the word count for the negative tweets. These are the frequently repeated words for the negative comments of the airline services.

A. Performance Parameters

The study has been evaluated using the precision, recall, as well as f-score values that were acquired from the data. Various ML algorithms were being carried out to discover the algorithms that were the most appropriate for the system.

True Negatives (TN) - The values estimated as negative (meaning the real class value is 'no' and forecasted class value is 'not') are perfectly accurate. E.g., when a real class shows that this passenger hasn't survived & the forecasted class tells you the same. The real class occurs when the predicted class, a false positive as well as a false negative occurs in opposition.

True Positives (TP) - The real class value and the predicted class value are both correct. Consider, for instance, the difference between actual class value, which is "has survived this passenger," and predictive class value, which tells you, "This passenger is likely to be the same one next time."

False Negatives (FN) – Real class is yes, whereas forecasted class is not. In other words, for example, when we see how valuable each passenger class is, it may tell us that passengers have survived or that passengers are likely to die.

False Positives (FP) – In the situation when the class is “No” and the forecasted class is “Yes,” In other words, if the class real says that this passenger has not lived, but the class forecast predicts that he will, this passenger has died.

Accuracy - It is a fundamental accuracy indicator that is proportional to the total measurements. Symmetrical datasets provide better statistical accuracy since the proportion of false negatives and false positives are almost equal.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Positive observations accurately predicted as a percentage of total positive observations expected is known as the predictive accuracy ratio.

Precision = TP/TP+FP

Recall (Sensitivity)- It is a ratio of positive comments that were accurately predicted against all actual class observations - yes.

Recall = TP/TP+FN

F1 score - It is recall & precision weighted average. This score, therefore, takes into account both false negatives & false positives.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Training Phase: When doing machine learning, it is necessary to have a labeled dataset that can be utilized to train the classifier. The training dataset comprises examples that have an input item as well as a label or a class, which are grouped as a training set. First, an algorithm analyses the labeled data, and then it extracts features from the input data that contain the necessary information, allowing the sentiment analysis to be completed using simply a representation of the original data, as opposed to utilizing the original data itself. Finally, the algorithm generates a function that may be employed in the categorization of previously unknown data or the testing of the method.

Model Training and Validation: The bag of words from all tweets becomes the representation of every tweet after the features are produced. The dataset is then separated into 2 types: training set, which comprises 90% of the total, as well as the testing set, which contains the residual 10%. To prevent overfitting, the dataset was divided into two subgroups. When a classifier is trained as well as evaluated on the same set of data, overfitting may occur. The classification of previously seen data is a success, whereas the classification of yet unobserved data is a failure.

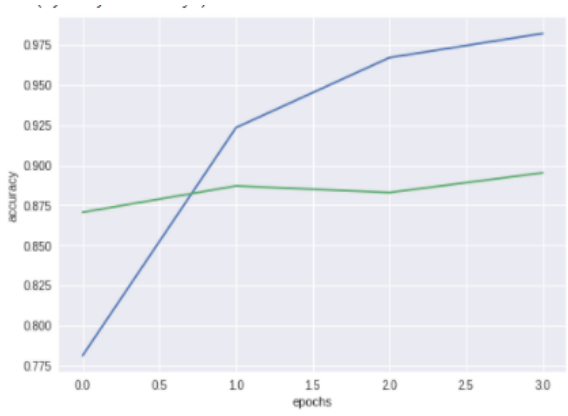

Figure 7: Training and Validation accuracy

Above figure 7 shows training & validation accuracy. In this graph, X-axis represents the no. of epoch & the Y-axis represents accuracy. This graph shows the highest accuracy at the 30th epoch which is approx. 97.5 % for training represented by blue color and for 89 %validation represented by green color.

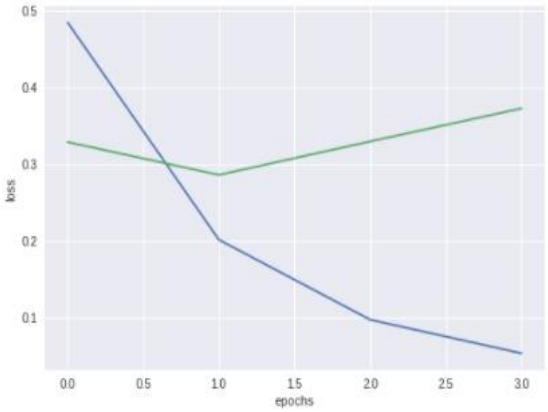

Figure 8: Training and Validation Loss

The training & validation losses are shown in the picture above figure 8. The number of epochs is shown on the X-axis, while the loss values are represented on the Y-axis. While the training loss is a measure of how well the model is fitting training data, the validation loss is a measure of how well the model fits fresh data. This graph shows the lowest loss value at the 30th epoch which is approx. 0.04 for training and 0.39 for validation.

A. Confusion Matrix

There are several ways to measure the efficacy of machine learning classification. TP, FP, TN, as well as FN values, may all be directly compared using confusion matrices. The model's final assessment confusion matrix. The confusion matrices generated over the test data are shown in figures number 9 proposed method, respectively.

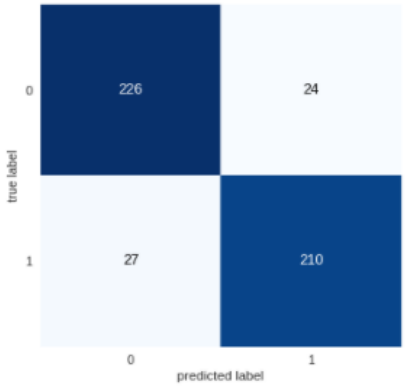

Figure 9: Confusion matrix of the proposed (deep neural network) Method

B. Classification Outcomes

This section discussed classification outcomes obtain by applying a deep neural network. It shows the performance of classification outcomes in terms of accuracy, recall, and f1-score for the US Airline Twitter Dataset.

	precision	recall	f1-score	support
0	0.89	0.90	0.90	250
1	0.90	0.89	0.89	237
accuracy			0.90	487
macro avg	0.90	0.90	0.90	487
weighted avg	0.90	0.90	0.90	487

Figure 10: Classification report of deep neural network

Figure 10 depicts the classification report of the deep neural network for the proposed work. The negative tweets show the precision is 89%, recall is 90%, and f1-score has 90% value while positive tweets show the precision is 90%, recall is 89%, and f1-score has 89% that is the same for all three metrics to US Airline Twitter Dataset.

Table 1: Comparison table of base and propose results

Metric	Base (LR)	Propose (Deep NN)
Accuracy	86%	89%

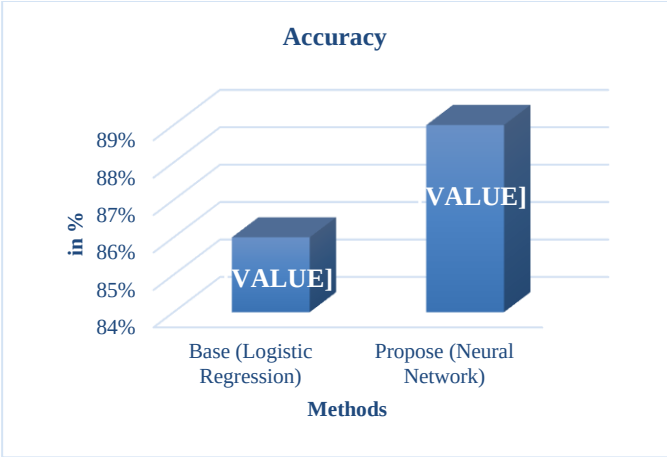

Figure 11: Accuracy graph of base and proposed methods

Based and recommended approaches can be seen in the graphs shown in table 1 (figure 11). During our research, we tested two distinct machine learning algorithms. ' The main goal was to categorize tweets into positive and negative categories. We discovered that applying the BoW model as well as extracting the most representative features resulted in greater accuracy, precision, recall, or f1-score

when compared to previous studies. Deep neural networks were shown to have an accuracy of 86% and 89% respectively, based on the trials. Deep NN surpassed other LR classifiers with an accuracy of 89%, according to the study's results.

V. CONCLUSION

A very huge amount of data is generated every second for microblogs, content sharing via Social media sites and social networking. Twitter is an important popular microblog where people voice their opinions about daily issues. Recently, analyzing these opinions is the main concern of Sentiment analysis (or opinion mining). It's been a challenge for academics to efficiently capture, aggregate, and analyze the feelings of people. This research proposes a very accurate model for sentiment analysis of tweets to overcome these difficulties. In this article, we employed a neural network classifier, a machine learning approach, to improve the accuracy & performance of sentiment classification. Research also focuses on the importance of pre-processing or feature vector representation in the sentiment classification method. The research was based on tweets from passengers who had expressed their opinions regarding their experiences flying with US airlines. The tweets in the text were divided into two categories based on the classification models that were most often utilized. For this comparison, we used both LR and DNNs. In comparison, the suggested deep NN classifiers have an accuracy rate of 89%, while the LR classifiers have an accuracy rate of 86%.

FUTURE RESEARCH

Research in the future might focus on finding ways to address the existing dataset's imbalance. This, we believe, may enhance the outcome. A novel transform model with the use of other languages, like Arabic reviews or English, might be useful in doing further tests, as could utilizing 10-fold cross-validation to examine alternative hyper parameters for deep neural approaches. Our investigation may have had an impact on the accuracy of the deep learning classifiers since we had a limited quantity of data to work with.

REFERENCES

A., V. and Sonawane, S. S. (2016) ‘Sentiment Analysis of Twitter Data: A Survey of Techniques’, *International Journal of Computer Applications*, 139(11), pp. 5–15. doi: 10.5120/ijca2016908625.

Akhmetgaliev, A. I., Gafarov, F. M. and Sitdikova, F. B. (2020) ‘Solving the problem of sentiment analysis using neural network models’, *International Journal of Pharmaceutical Research*. doi: 10.31838/ijpr/2020.12.01.162.

Alqahtani, R. and Al-qahtani, R. (2021) ‘Models Predict Sentiment of Airline Tweets Using ML Models’.

Celine, S., Dominic, M. M. and Devi, M. S. (2020) ‘Logistic Regression for Employability Prediction’, *International Journal of Innovative Technology and Exploring Engineering*, 9(3), pp. 2471–2478. doi: 10.35940/ijitee.c8170.019320.

Erb, R. J. (1993) ‘Introduction to Backpropagation Neural Network Computation’, *Pharmaceutical Research: An Official Journal of the American Association of Pharmaceutical Scientists*. doi: 10.1023/A:1018966222807.

Flores, Andrew, and H. F. (2020) ‘Sentiment Analysis on Airline Tweets Using Naïve Bayes Classifier’, *athena.ecs.csus.edu*.

Hrazi, M. M. *et al.* (2021) ‘Sentiment analysis of tweets from airlines in the gulf region using machine learning’, in *2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021*. doi: 10.1109/WIDSTAIF52235.2021.9430231.

Rane, A. and Kumar, A. (2018) ‘Sentiment Classification System of Twitter Data for US Airline Service Analysis’, in *Proceedings - International Computer Software and Applications Conference*. doi: 10.1109/COMPSAC.2018.00114.

Saad, A. I. (2020) ‘Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques’, in *16th International Computer Engineering Conference, ICENCO 2020*. doi: 10.1109/ICENCO49778.2020.9357390.

Sharma, N. K., Rahamatkar, S. and Sharma, S. (2019) ‘Classification of airline tweet using naïve-bayes classifier for sentiment analysis’, in *Proceedings - 2019 International Conference on Information Technology, ICIT 2019*. doi: 10.1109/ICIT48102.2019.00019.

Wan, Y. and Gao, Q. (2016) ‘An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis’, in *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*. doi: 10.1109/ICDMW.2015.7.

Yimga, J. (2021) ‘The airline on-time performance impacts of the COVID-19 pandemic’, *Transportation Research Interdisciplinary Perspectives*. doi: 10.1016/j.trip.2021.100386.

Zhang, G. P. (2000) ‘Neural networks for classification: A survey’, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. doi: 10.1109/5326.897072.

Zhang, L., Wang, S. and Liu, B. (2018) ‘Deep learning for sentiment analysis: A survey’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi: 10.1002/widm.1253.

