IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 9, Issue 3, Aug 2021

Sentiment Classification System of Twitter Data for US Airline Service Analysis

Mrs. Lakshmi Lavanya Tumu, Mrs. Kshetravati N Sangami, Mrs. Gangula Pavani

Abstract—

The airline enterprise is a completely aggressive market which has grown hastily within side the beyond 2 decades. Airline organizations motel to conventional purchaser remarks bureaucracy which in flip are very tedious and time consuming. This is where Twitter information serves as an excellent supply to collect purchaser remarks tweets and carry out a sentiment evaluation. In this paper, we labored on a dataset comprising of tweets for six important US Airlines and achieved a multi-elegance sentiment evaluation. This method begins off evolved off with pre-processing strategies used to clean the tweets after which representing those tweets as vectors the use of a deep gaining knowledge of concept (Doc2vec) to do a phrase-degree evaluation. The evaluation changed into finished the use of 7 exceptional class strategies: Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bays and Gadabouts. The classifiers have been skilled the use of 80% of the information and examined the use of the closing. The final results of the check set are the tweet sentiment (positive/negative/neutral). Based at the effects obtained, the accuracies have been calculated to draw an assessment among every class method and the average sentiment rely changed into visualized combining all six airlines.

I. INTRODUCTION

Customer comments could be very critical to Airline businesses as this facilitates them in enhancing the fine of offerings and centers furnished to the clients. Sentiment Analysis in Airline enterprise is methodically finished the usage of traditional comments techniques that contain consumer satisfaction questionnaires and forms. These tactics would possibly appear quite easy on a top level view however are very time eating and require plenty of manpower that includes a price in analyzing them. Moreover, the statistics accumulated from the questionnaires is regularly faulty and inconsistent. This may be due to the fact now no longer all clients take those feedbacks seriously and can fill in inappropriate info which brings about noisy information for sentiment analysis. Whereas on the opposite hand, Twitter is a gold mine of information with over 1/sixtieth of the world's population the Usage of it which almost quantities to a hundred million people, more than 1/2 of one billion tweets are tweeted every day and the number maintains developing with each passing day. With the rising call for and improvements of Big Data technology in the beyond decade, it has emerge as simpler to gather tweets and apply statistics evaluation strategies on them [4]. Twitter is a miles greater dependable supply of statistics because the customers tweet their genuine emotions and feedbacks hence making it greater appropriate for investigation [6].

1,2,3 Assistant Professor
1,2,3 Department of CSE
1,2,3 Global Institute of Engineering and Technology Moinabad, Ranga Reddy District, Telangana State.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 3, Aug 2021

For example, with the phone X marketplace release, the organization can carry out a sentiment evaluation on the tweets associated with the product as part of their marketplace research to improvise their product. Once the airline tweets are collected, they go through pre-processing to remove useless information in them. Sentiment classification strategies are then carried out to the wiped clean tweets. This gives statistics scientists and Airline organizations a broader perspective approximately the emotions and critiques in their customers. The main motive of this paper is to provide the airline industry a more comprehensive view about the sentiments of their customers and provide to their needs in all good ways possible. In this paper, we go through several tweet pre-processing techniques followed by the application of seven different machine learning classification algorithms that are used to determine the sentiment within the tweets. The classifiers are then compared against each other for their accuracies.

II. DATA EXTRACTION

In this work, the dataset includes diverse tweets that had been taken from the same old Cagle Dataset: Twitter US Airline Sentiment launched through Crowd Flower. A general of 14640 tweets had been extracted which fashioned the experimental dataset. The tweets accrued had been for 6 main US Airlines that are: United, US Airways, Southwest, Delta and Virgin America. The tweets had been a combination of positive, terrible and neutral sentiment. The tweets are pre-labeled with the type of sentiment which led us to follow the approach of supervised machine learning [1]. The implementation of the code was entirely done using Spider which a powerful development environment for Python language with is advanced editing, testing and numerical computing environment. The following table gives the tweets sentiment distribution.

TABLE I. SENTIMENT DISTRIBUION OF TWEETS

Sentiment	Tweet Count
Positive	2363
Negative	9178
Neutral	3099

III. DATA PREPROCESSING

Data preprocessing is a statistics mining approach that transforms actual global statistics into comprehensible format. Twitter statistics is regularly inconsistent and lacks sure features (lacking values) which want to be treated earlier than acting any type of analysis. The tweets go through numerous degrees of preprocessing to get the wiped clean tweets which can be used for similarly analysis. The tweets are tokenized which transforms the tweets right into a listing in which every phrase within side the tweet is a detail of the listing. A lot of words in tweets are irrelevant and do not add any additional meaning to the sentence, such words are known as stop words. Example of stop words are: and, I, the, for, should, is etc. These words are eliminated using notch's stop word list. Words such as 'not', 'wasn't', 'isn't' have not been removed from the tweets as they add a meaning to the sentence. After stop word removal the tweets are then lemmatized. Lemmatization is the process where a word is reduced to its base form with the use of vocabulary.

For example, the word 'advised' and 'advising' will be reduced to 'advice'. This avoids confusion by reducing the number of words fed to the classifier. Since the tweets are a form of human expression it may contain symbols and punctuations which are eliminated. The sentiment analysis is done for words that belong to English vocabulary, so any occurrence of non-English words is eliminated.

IV. WORD EMBEDDINGS AND DOCUMENT VECTORS

Word Embeddings is a way in which every phrase is given a completely unique vector illustration with its semantic which means taken into consideration. The various illustrations of textual content facts is a leap forward for the overall performance of deep getting to know strategies on NLP problems. Each phrase is mapped to a vector in a predefined vector space. These vectors are discovered the usage of neural networks. The learning process can be done with a neural network model or by using unsupervised process involving document statistics. In this sentiment analysis we will be making use of a neural network model which incorporates a few aspects of deep learning. A. Doc2Vec Model Numeric illustration of phrases is a difficult and tough task. There are opportunity strategies such as Bag of Words (BOW) version which offers mediocre results and does now no longer take phrase ordering into

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 3, Aug 2021

consideration. To triumph over this drawback, we're using Genesis's deep getting to know library for phrase embeddings- Doc2vec. Doc2vec is a shape of sentence embedding in which every sentence is mapped to a vector in space. Doc2vec is Genesis's prolonged library of word2vec that's a library to locate vector representations for every phrase. The key distinction between doc2vec and word2vec is the algorithms used. Word2vec makes use Continuous Bag of Words (CBOW) and skip-gram version while doc2vec makes use of disbursed reminiscence version (DM) and disbursed bag of phrases version (DBOW) [5]. B. Working of Doc2Vec: Distributed Memory Model Doc2vec technique of getting to know paragraph vectors is stimulated with the aid of using Word2vec technique. It contains how the phrase vectors can expect the subsequent phrase in a given context or tweet. In doc2vec framework each paragraph is mapped to a particular vector that's represented with the aid of using a column in matrix D and each phrase is mapped to particular vector mapped in matrix W. The phrase and paragraph vectors are then concatenated to expect the subsequent phrase.

The paragraph token acts as a memory and remembers the missing word in the tweet which is why it is called as the distributed memory model of paragraph vector. The reason for using Doc2vec is that it overcomes the weaknesses of bag-of-words model by considering the semantics of the words. The other advantage of using this model is that it takes the word ordering into consideration.

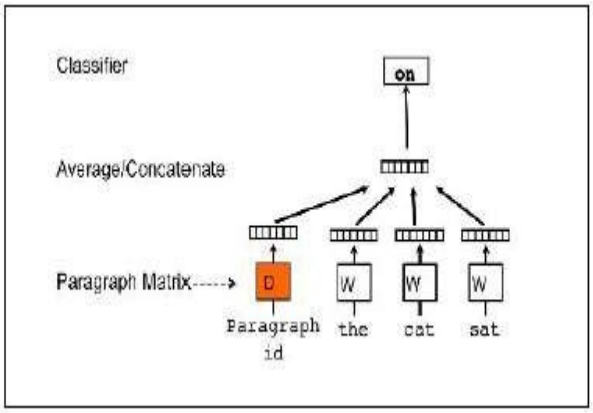

Fig. 1. Framework of Doc2vec Distributed Memory Model (PV-DM). The average of the vectors of the three words is calculated to predict the fourth word in the sentence. The paragraph id holds the information about the missing word and thus acts as a memory.

V. CLASSIFICATION TECHNIQUES

Here we describe seven extraordinary classifiers using extraordinary category strategies. These category strategies are commonly used for textual content category can be extensively utilized for twitter sentiment analysis.

***A.* Decision Tree Classifier**

Decision tree classifier is a simple and popularly used algorithm to classify data. Decision Tree represents a tree like structure with internal nodes representing the test conditions and leaf nodes as the class labels. This classification approach poses carefully crafted questions about the attributes of the test data set. Each time an answer is received another follow up question is asked until we can correctly classify the class of the test data. This classifier handles over-fitting by using post pruning approaches.

B. Random Forest Classifier

Random forest classifier is an ensemble learning classification algorithm. It is very similar to decision tree but contains a multitude of decision trees and the class label is the mode value of the classes predicted by individual decision trees. This algorithm is efficient in handling large datasets and thousands of input variables without their deletion. This model can deal with over fitting of data points. For a dataset, D, with N instances and an attributes, the general procedure to build a Random Forest ensemble classifier is as follows. For each time of building a candidate Decision Tree, a subset of the dataset D, d, is sampled with replacement as the training dataset. In each decision tree, for each node a random subset of the attributes A, a, is selected as the candidate attributes to split the node. By building K Decision Trees in this way, a Random Forest classifier is built. Random forest uses majority vote and returns the class label that is has maximum votes by the individual decision trees. Headings, or heads, are Organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

C. Logistic Regression Classifier

This algorithm was named after the core function used in it that is the logistic function. The logistic function is also

Known as the sigmoid function. It is an S-shaped curve that takes real values as input and converts it into a range between 0 and 1. The sigmoid function is defined as follows:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1} \qquad (1)$$

D. Support Vector Machine Classifier

This algorithm works on a simple strategy of separating hyper planes. Given training data, the

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 3, Aug 2021

algorithm categorizes the test data into an optimal hyper plane. The data points are plotted in an n-dimension vector space (n depends upon the features of the data points). SVM algorithm is used for binary classification and regression tasks but in our case, we have a 3-class sentiment analysis making it multiclass SVM classification. We adopt the pair wise classification technique where each pair of classes will have one SVM classifier trained to separate the classes. The overall accuracy of this classifier will be accuracies of every SVM classification included [2]. Then on performing classification we find a hyper plane that differentiates the 3 classes very well.

E. Gaussian Naïve Bays Classifier

Naïve Bays is a popular text classifier. This classifier is highly scalable. This algorithm makes use if the Bays Theorem of conditional probability [7]. Since we are dealing with continuous values we make use of the Gaussian distribution. Gaussian NB is easier to work with as we only need to compute mean and standard deviation from the training data. This classifier passes each tweet and calculates the product of the probabilities of every feature present in the tweet for each class label i.e. positive, negative and neutral. The class label is assigned to the tweet based on the sentiment label that has biggest sentiment product. The equation for normal distribution is described as

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \qquad (2)$$

F. Gadabouts Classifier

Adaptive Boosting or Gadabouts is a meta-algorithm formulated by You Freund and Robert Schapiro. It is used with other learning algorithms to get an improved performance. The output of the weak learners (other classifiers) is combined into a weighted sum which gives us the output of the Gadabouts Classifier. One drawback of this classification is that it is very sensitive to noise points and outliers. The training data fed to the classifier must be of high quality.

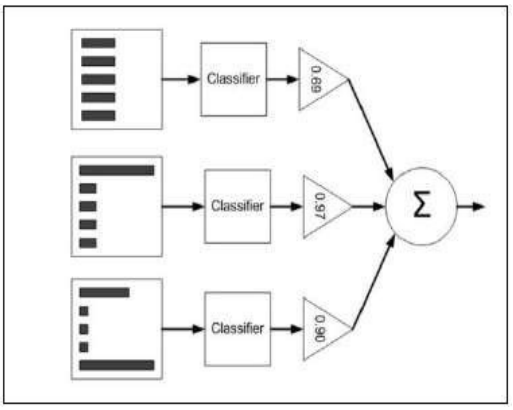

Fig. 2. Framework of AdaBoost Classifier (Ensemble Classifier)

G. K- Nearest Neighbor Classifier

KNN Classifier is an instance-based learner used for both classification and regression tasks. This algorithm does not use the training data to make any generalizations. It is based on feature similarity. A test sample is classified based on a majority vote of its neighbors; the class assigned to the test sample is the most common class among k nearest neighbors [3]. When used for regression the output value is the average of the outputs of its k nearest neighbors. This classifier is a lazy learner because nothing is done with the training data until the model tries to classify the test data. We have taken the k value to be 3 which gave us the most accurate result. The k value must not be too large that it includes the noise points or points that belong to the neighboring class.

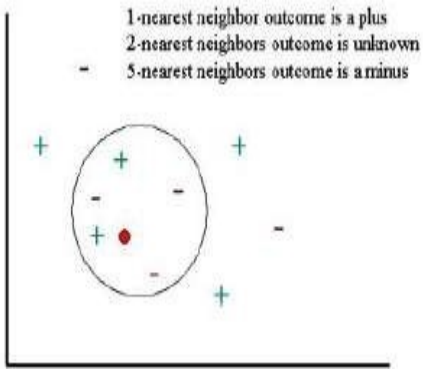

Fig. 3. Representation of Classification by KNN

VI. EXPERIMENT AND EVALUATION

The dataset includes 14640 tweets on which we perform a train-take a look at break up the use of the

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 3, Aug 2021

80-20 rule wherein 80% of the records are used for education and the ultimate 20% is used for testing. The universal sentiment is counted number which debts for the total wide variety of tweets in every sentiment class i.e. positive, poor or impartial for all 6 Airlines become visualized in Fig.4 the use of Matplotlib library that is a Pythons' version of Mat lab. On gazing the graph, majority of the tweets expressed poor sentiment, this perhaps due to the fact people normally use the social media platform to deliver their dissatisfactory remarks. The sentiment distribution for United and Virgin America airline is likewise plotted in Fig.5 & Fig.6 respectively. The classifiers listed in the previous section were trained using the training data and tested on the test set for their accuracies. In accuracy evaluation, we consider precision, recall and F- Measure to evaluate the overall accuracy of the classifier. Here, precision is the fraction of correctly classified instances for one class of the overall instances which are classified to this class and recall is the fraction of correctly classified instances for one class of the overall instances in the dataset [8]. F- Measure is a comprehensive evaluation which integrates both precision and recall. The Table 2 shows the accuracies of each classifier. The reasons for the negative feedback from the customers as mentioned in the dataset were also plotted and presented in the form of a graph in Fig.7.

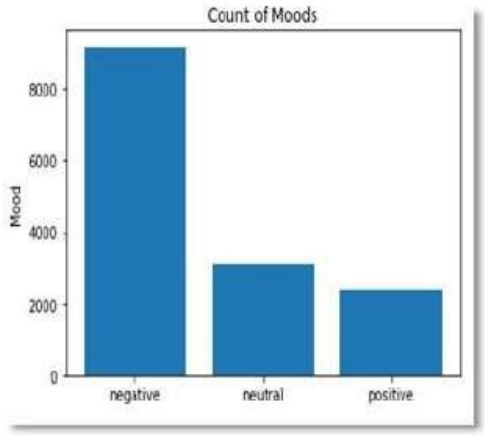

Fig. 4. Overall Sentiment Count

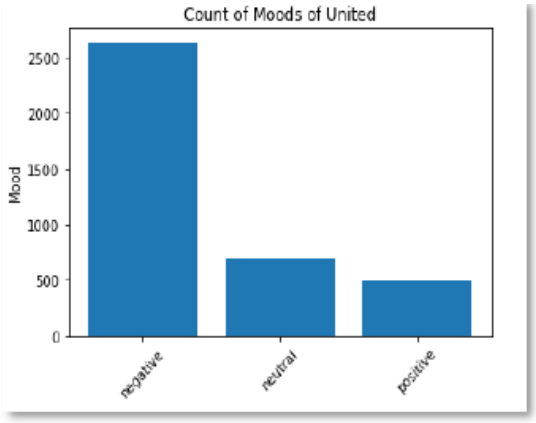

Fig. 5. Sentiment Count for United Airline

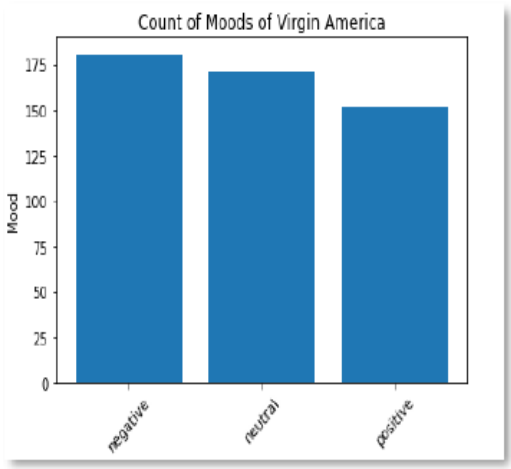

Fig. 6. Sentiment Count for Virgin America Airline

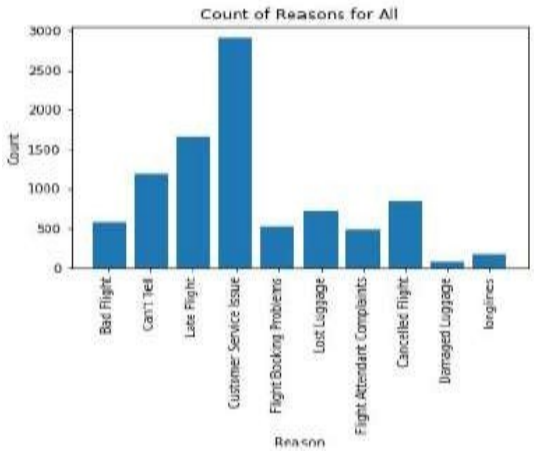

Fig. 7. Reasons for Negative Feedback

TABLE II. ACCURACY OF CLASSIFIER FOR 3- CLASS DATASET

ISSN 2347–3657
Volume 9, Issue 3, Aug 2021

Classifier	Precision	Recall	F- Measure
Decision Tree	63%	64.6%	64.5%
Random Forest	85.6%	86.5%	86.5%
SVM	81.2%	84.4%	84.8%
Gaussian Naïve Bayes	64.2%	64.7%	64.6%
AdaBoost	84.5%	83.5%	83.5%
Logistic Regression	81%	81.6%	81.9%
KNN	59%	59.2%	59.3%

VII. CONCLUSION

This paper makes empirical contribution to the sphere of data technology and sentiment evaluation. In this paper, we compare diverse conventional category strategies and compare their accuracies. In the area of sentiment evaluation for airline offerings little or no studies has been executed. The past painting that has been executed does a phrase stage evaluation of tweets without maintaining the phrase order. However, on this studies we've got executed a phrase-stage evaluation of tweets using report vectors (Doc2vec) which considers the phrase ordering as well. The category strategies used include ensemble methods along with Gadabouts which combine numerous different classifiers to shape one robust classifier and give an accuracy of 84.5%.

The accuracies attained through the classifiers are excessive sufficient to be utilized by the airline industry to put into effect patron exceptional investigation. There is nonetheless scope for development on this evaluation because the major setback is the restricted variety of tweets utilized in education the version. By growing the variety of tweets, we will construct a more potent version therefore ensuing in higher classification accuracy. The method defined on this paper may be used through airline businesses to investigate the twitter data.

REFERENCES

[1] Pang, Bo, and Lillian Lee, "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.J. Clerk Maxwell, a Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Xia, Rue, Chengqing Zong, and Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences 181.6 (2011): 1138-1152.

[3] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[4] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," Knowledge-Based Systems, vol. 69, 2014, pp. 1–2.

[5] Quoc Le, Tomas Mikolov. "Distributed Representations of Sentences and Documents" , Cornell University, 2014.

[6] S Kamal, N. Dey, A.S Ashour, s. Ripon, V.E. Balas and M. Kaysar, "FbMapping: An automated system for monitoring facebook data",Neural Network World, 2017.

[7] Pak, Alexander, and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[8] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.