Sentiment Classification System of Twitter Data for US Airline Service Analysis

Ankita Rane
Computer Science and Engineering
BITS Pilani, Dubai Campus
Dubai, UAE
ankitarane24@gmail.com

Dr. Anand Kumar
Electrical and Electronics Engineering
BITS Pilani, Dubai Campus
Dubai, UAE
akumar@dubai.bits-pilani.ac.in

***Abstract*—The airline industry is a very competitive market which has grown rapidly in the past 2 decades. Airline companies resort to traditional customer feedback forms which in turn are very tedious and time consuming. This is where Twitter data serves as a good source to gather customer feedback tweets and perform a sentiment analysis. In this paper, we worked on a dataset comprising of tweets for 6 major US Airlines and performed a multi-class sentiment analysis. This approach starts off with pre-processing techniques used to clean the tweets and then representing these tweets as vectors using a deep learning concept (Doc2vec) to do a phrase-level analysis. The analysis was carried out using 7 different classification strategies: Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bayes and AdaBoost. The classifiers were trained using 80% of the data and tested using the remaining 20% data. The outcome of the test set is the tweet sentiment (positive/negative/neutral). Based on the results obtained, the accuracies were calculated to draw a comparison between each classification approach and the overall sentiment count was visualized combining all six airlines.**

Keywords- Machine Learning; Classification techniques; Deep Learning; Distributed Memory Model

I. Introduction

Customer feedback is very crucial to Airline companies as this helps them in improving the quality of services and facilities provided to the customers. Sentiment Analysis in Airline industry is methodically done using traditional feedback methods that involve customer satisfaction questionnaires and forms. These procedures might seem quite simple on an overview but are very time consuming and require a lot of manpower that comes with a cost in analyzing them. Moreover, the information collected from the questionnaires is often inaccurate and inconsistent. This may be because not all customers take these feedbacks seriously and may fill in irrelevant details which result in noisy data for sentiment analysis. Whereas on the other hand, Twitter is a gold mine of data with over 1/60th of the world's population using it which nearly amounts to 100 million people, more than half a billion tweets are tweeted daily and the number keeps growing with every passing day. With the rising demand and advancements of Big Data technologies in the past decade, it has become easier to collect tweets and apply data analysis techniques on them [4]. Twitter is a much more reliable source of data as the users tweet their genuine feelings and feedbacks thus making it more suitable for investigation [6]. For example, with the iPhone X market release, the company can perform a sentiment analysis on the tweets related to the product as a part of their market research to improvise their product. Once the airline tweets are collected, they undergo pre-processing to remove unnecessary details in them. Sentiment classification techniques are then applied to the cleaned tweets. This gives data scientists and Airline companies a broader perspective about the feelings and opinions of 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 contains various tweets that were taken from the standard Kaggle Dataset: Twitter US Airline Sentiment released by CrowdFlower. A total of 14640 tweets were extracted which formed the experimental dataset. The tweets collected were for six major US Airlines that are: United, US Airways, Southwest, Delta and Virgin America. The tweets were a mix of positive, negative and neutral sentiment. The tweets are pre-labelled 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 Spyder which is a powerful development environment for Python language with advanced editing,

0730-3157/18/$31.00 ©2018 IEEE
DOI 10.1109/COMPSAC.2018.00114

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 data mining technique that transforms real world data into understandable format. Twitter data is often inconsistent and lacks certain features (missing values) which need to be dealt with before performing any kind of analysis. The tweets undergo various stages of preprocessing to get the cleaned tweets which can be used for further analysis. The tweets are tokenized which transforms the tweets into a list where each word in the tweet is an element of the list. 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 nltk'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 technique where each word is given a unique vector representation with its semantic meaning taken into consideration. The diverse representation of text data is a breakthrough for the performance of deep learning techniques on NLP problems. Each word is mapped to a vector in a predefined vector space. These vectors are learned using 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 representation of words is a tough and challenging task. There are alternative techniques such as Bag of Words (BOW) model which gives mediocre results and does not take word ordering into consideration. To overcome this drawback, we are making use of Gensim's deep learning library for word embeddings- Doc2vec. Doc2vec is a form of sentence embedding where each sentence is mapped to a vector in space. Doc2vec is Gensim's extended library of word2vec which is a library to find vector representations for each word. The key difference between doc2vec and word2vec is the algorithms used. Word2vec makes use Continuous Bag of Words (CBOW) and skip-gram model whereas doc2vec uses distributed memory model (DM) and distributed bag of words model (DBOW) [5].

B. Working of Doc2Vec: Distributed Memory Model

Doc2vec approach of learning paragraph vectors is inspired by Word2vec approach. It incorporates how the word vectors can predict the next word in a given context or tweet. In doc2vec framework every paragraph is mapped to a unique vector which is represented by a column in matrix D and every word is mapped to unique vector mapped in matrix W. The word and paragraph vectors are then concatenated to predict the next word. 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.

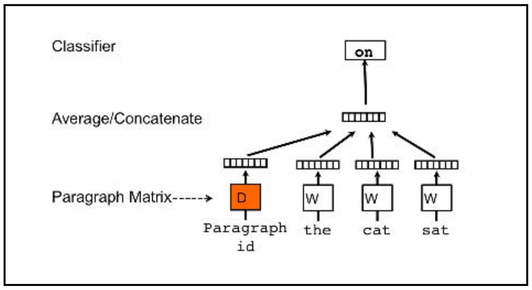

Fig. 1. Framework of Doc2vec Distributed Memory Model (PV-DM). The average of the vectors of the three words is calculated to predict the fourth word in the sentence. The paragraph id holds the information about the missing word and thus acts as a memory.

V. CLASSIFICATION TECHNIQUES

Here we describe seven different classifiers using different classification techniques. These classification techniques are generally used for text classification can be also used for twitter sentiment analysis.

A. Decision Tree Classifier

Decision tree classifier is a simple and popularly used algorithm to classify data. Decision Tree represent 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 overfitting of data points. For a dataset, D, with N instances and A 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 a 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 hyperplanes. Given training data, the algorithm categorizes the test data into an optimal hyperplane. The data points are plotted in a 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 pairwise 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 hyperplane that differentiates the 3 classes very well.

E. Gaussian Naïve Bayes Classifier

Naïve Bayes is a popular text classifier. This classifier is highly scalable. This algorithm makes use if the Bayes 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

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

F. AdaBoost Classifier

Adaptive Boosting or AdaBoost is a meta-algorithm formulated by Yoav Freund and Robert Schapire. 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 AdaBoost 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.

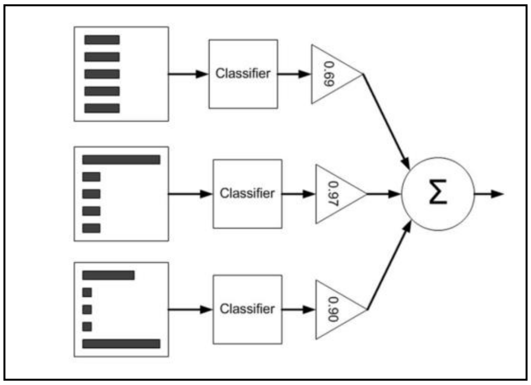

Fig. 2. Framework of AdaBoost Classifier (Ensemble Classifier)

G. K- Nearest Neighbour 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.

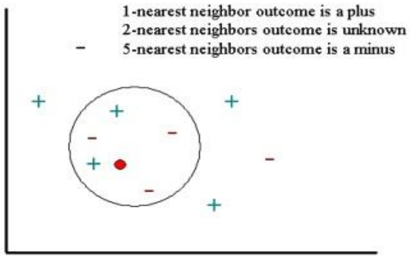
1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus
+
−

Fig. 3. Representation of Classification by KNN

VI. Experiment and Evaluation

The dataset consists of 14640 tweets on which we perform a train-test split using the 80-20 rule where 80% of the data is used for training and the remaining 20% is used for testing. The overall sentiment count which accounts for the total number of tweets in each sentiment category i.e. positive, negative or neutral for all 6 Airlines was visualized in Fig.4 using Matplotlib library which is a Pythons's adaptation of Matlab. On observing the graph, majority of the tweets expressed negative sentiment, this maybe because people generally use the social media platform to convey their dissatisfactory remarks. The sentiment distribution for United and Virgin America airline is also 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.

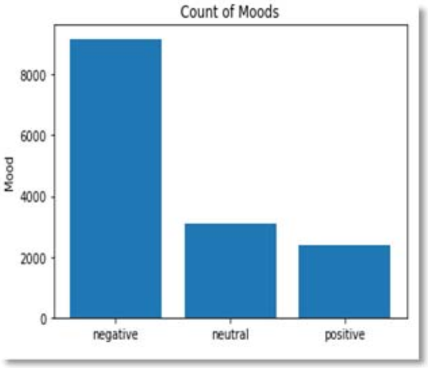

Fig. 4. Overall Sentiment Count

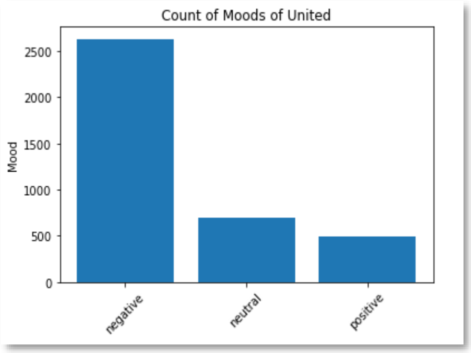

Fig. 5. Sentiment Count for United Airline

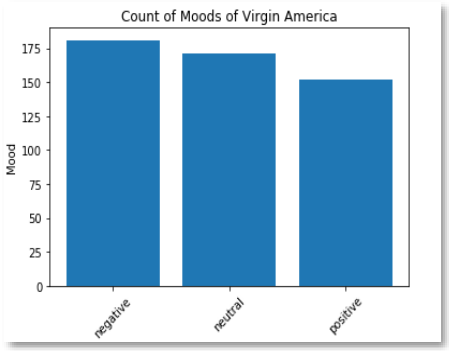

Fig. 6. Sentiment Count for Virgin America Airline

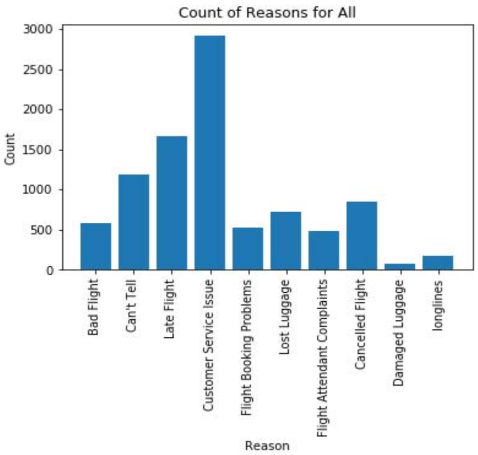

Fig. 7. Reasons for Negative Feedback

TABLE II. ACCURACY OF CLASSIFIER FOR 3- CLASS DATASET

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 field of data science and sentiment analysis. In this paper, we compare various traditional classification techniques and compare their accuracies. In the domain of sentiment analysis for airline services very little research has been done. The past work that has been done does a word level analysis of tweets without preserving the word order. However, in this research we have done a phrase-level analysis of tweets using document vectors (Doc2vec) which considers the word ordering as well. The classification techniques used include ensemble approaches such as AdaBoost which combine several other classifiers to form one strong classifier and give an accuracy of 84.5%. The accuracies attained by the classifiers are high enough to be used by the airline industry to implement customer satisfactory investigation. There is still scope for improvement in this analysis as the major setback is the limited number of tweets used in training the model. By increasing the number of tweets, we can build a stronger model thus resulting in better classification accuracy. The approach described in this paper can be used by airline companies to analyze 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, Rui, 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.