IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Advanced Classification Method of Twitter Data using Sentiment Analysis for Airline Service

T. Hemakala[1*], S. Santhoshkumar [2]

[1]Department of Computer Applications, Alagappa University, Karaikudi
[2] Department of Computer Science, Alagappa University, Karaikudi

[*]*Corresponding Author: hemakala25@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 16/Jul/2018, Published: 31/July/2018

***Abstract*—** The social media has immense and popularity among all the services today. Sentiment Analysis is new way of machine learning to extract opinion orientation (positive, negative, neutral) from a text segment written for any product, organization and all other entities. In this research, design a framework for sentiment analysis with opinion mining for the case of airlines service feedback. Most available datasets of hotel reviews are not labelled which presents a lot of works for researchers as far as text data pre-processing task is concerned. Twitter is a SNS that has a huge data with user posting, with this significant amount of data, it has the potential of research related to text mining and could be subjected to sentiment analysis. 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. In this work, worked on a dataset comprising of tweets for 6 major Indian 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 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 outcome of the test set is the tweet sentiment (positive/negative/neutral).

***Keywords*—** Sentiment Analysis, Machine Learning, Classification techniques, Deep Learning, Distributed Memory Model, Twitter Analysis

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.

© 2018, IJCSE All Rights Reserved

II. RELATED WORK

Most existing studies to Twitter sentiment analysis can be divided into supervised methods and lexicon-based methods. Supervised methods are based on training classifiers (such as Naive Bayes, Support Vector Machine, Random Forest) using various combinations of features such as Part-Of-Speech (POS) tags, word N-grams, and tweet context information features, such as hashtags, retweets, emoticon, capital words etc. Lexicon-based methods determine the overall sentiment tendency of a given text by utilizing pre-established lexicons of words weighted with their sentiment orientations, such as SentiWordNet.

These methods rely on the presence of lexical or syntactical features that explicitly express the sentiment information. Though, in a lot of cases, the sentiment of a tweet is implicitly associated with the semantics of its context. In this work, we present semantic feature for sentiment analysis, which is word vector contextual representation of a word in tweet, which can capture the deep and implicit semantic relation information in the words of tweets.

III. PROPOSED WORK

A. DATA EXTRACTION

In this work, the dataset contains various tweets that were taken from the standard Kaggle Dataset: Twitter Indian Airlines Sentiment released by CrowdFlower. A total of 14640 tweets were extracted which formed the experimental dataset. 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, testing and numerical computing environment. The following table gives the tweets sentiment distribution.

B. 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

Table 1. SENTIMENT DISTRIBUION OF TWEETS

Sentiment Tweet	Count
Positive	2363
Negative	9178

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. 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 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.

© 2018, IJCSE All Rights Reserved

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:

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

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

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

F. Gaussian Naïve Bayes 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.

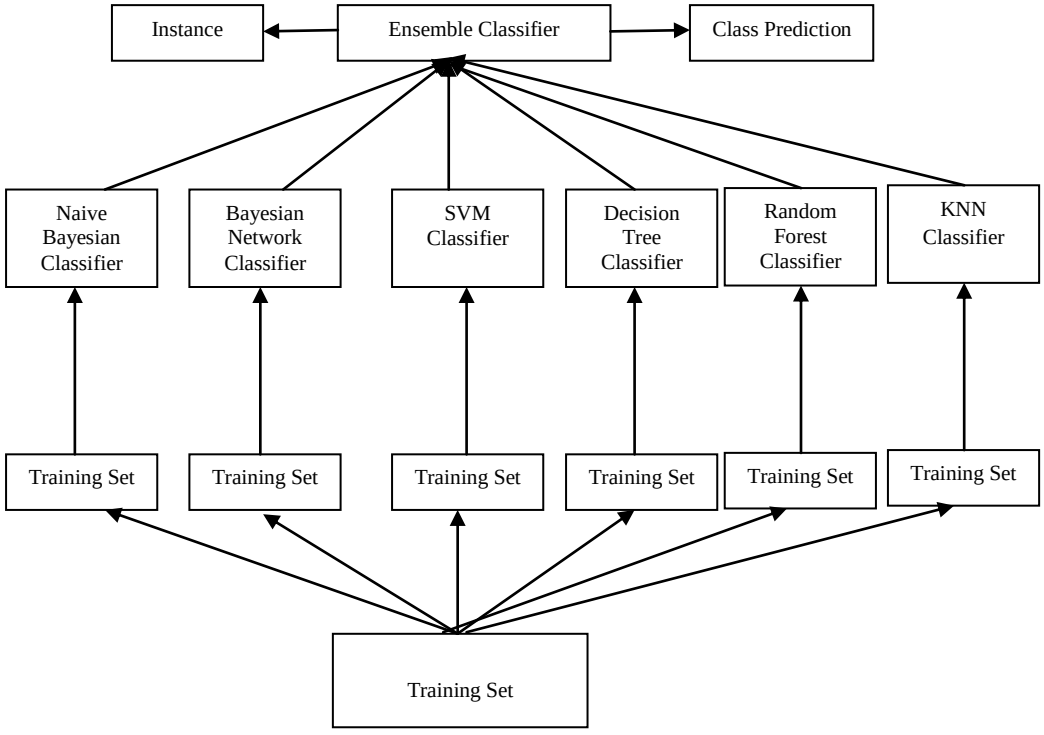

Figure 1. Advanced Classification System (Ensemble Classifier)

© 2018, IJCSE All Rights Reserved

V. EXPERIMENT AND EVALUATION

The dataset consists of 12120 tweets on which we perform a train-test split using the 80-20 rule where data get 80% for training and will used 20% data 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 4 Airlines was visualized in Figure.2

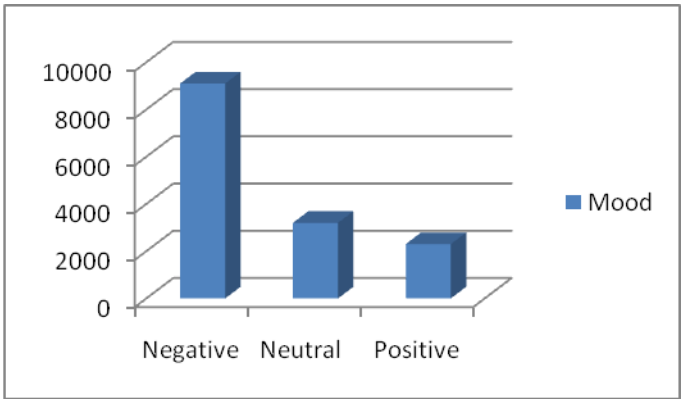

Figure 2. Overall Sentiment Count

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.

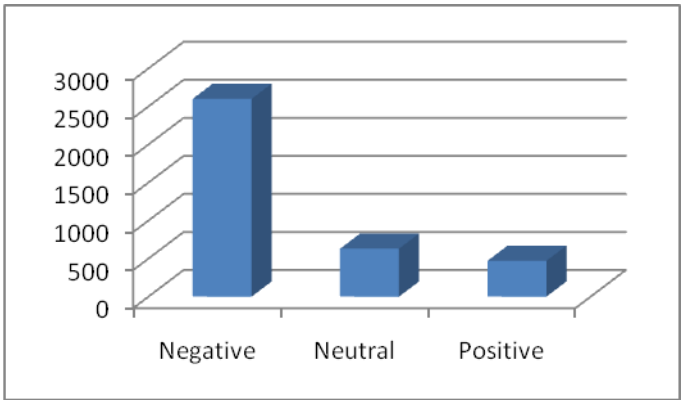

Figure 3. Sentiment Count for All India AirLines

The sentiment distribution for Indian Airlines is also plotted in Figure.3. The classifiers listed in the earlier section were trained using the training data and tested on the test set for their accuracies. In accuracy evaluation, 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. F- Measure is a comprehensive evaluation which integrates both precision and recall. The Table II 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 Figure.4.

Table 2. ACCURACY OF CLASSIFIER FOR 3- CLASS DATASET

Classifier	Precision	Recall	F- Measure
Decision Tree	63%	64.60%	64.50%
Random Forest	82.60%	82.50%	82.50%
SVM	81.20%	84.40%	84.80%
Gaussian Naïve Bayes	64.20%	64.70%	64.60%
Logistic Regression	81%	81.60%	81.90%
KNN	59%	59.20%	59.30%
AdaBoost	84.50%	83.50%	83.50%

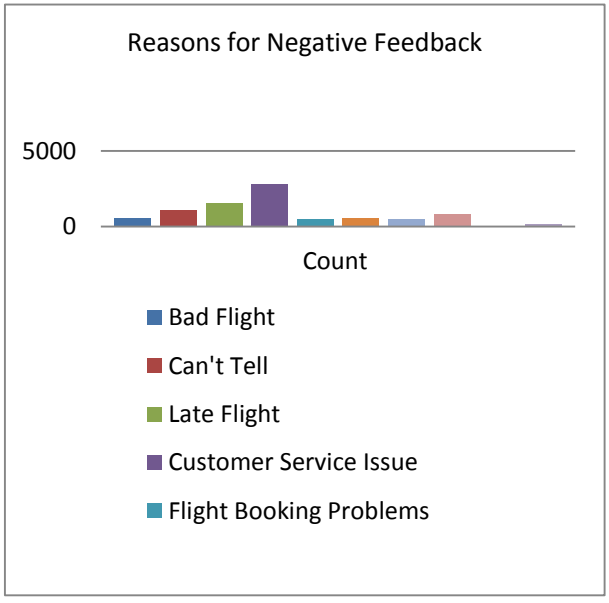

Figure 4. Reasons for Negative Feedback

© 2015, IJCSE All Rights Reserved

VI. CONCLUSION

In this paper, we compare various traditional classification techniques and compare their accuracies. In the area of SA 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 have done a multi-level analysis of tweets using Advanced Classifier System. The classification techniques used include ensemble approaches such as AdaBoost (Ensemble) 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] Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Min Knowl Discov 2012;24:478–514.

[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.

[3] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[4] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl-Based Syst 2013;41:89–97.

[5] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.

[6] Tao Xu, Peng Qinke, Cheng Yinzhao. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. KnowlBased Syst 2012;35:279–89.

[7] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 2012;53:680–8.

[8] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[9] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21.

[10] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:82–9.

[11] Montoyo Andre´ s, Martı´nez-Barco Patricio, Balahur Alexandra Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.

[12] Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun. DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Syst Appl 2010;37:6182–91.

[13] Lu Cheng-Yu, Lin Shian-Hua, Liu Jen-Chang, Cruz-Lara Samuel, Hong Jen-Shin. Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Syst Appl 2010;37:1643–53.

[14] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru.Recognition of Affect, Judgment, and Appreciation in Text. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Beijing; 2010. p. 806–14.

[15] Bai X. Predicting consumer sentiments from online text. Decis Support Syst 2011;50:732–42.

[16] Zhao Yan-Yan, Qin Bing, Liu Ting. Integrating intra- and interdocument evidences for improving sentence sentiment classification. Acta Automatica Sinica 2010;36(October'10).

[17] Yi Hu, Li Wenjie. Document sentiment classification by exploring description model of topical terms. Comput Speech Lang 2011;25:386–403.

[18] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the ''helpfulness'' of online user reviews: a text mining approach. Decis Support Syst 2011;50:511–21.

[19] He Yulan, Zhou Deyu. Self-training from labeled features for sentiment analysis. Inf Process Manage 2011;47:606–16.

[20] Tan Songbo, Wu Qiong. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Syst Appl 2011:12094–100.

[21] Tan Songbo, Wang Yuefen. Weighted SCL model for adaptation of sentiment classification. Expert Syst Appl 2011;38:10524–31.

[22] Qiong Wu, Tan Songbo. A two-stage framework for crossdomain sentiment classification. Expert Syst Appl

[23] 2011;38:14269–75.

[24] Jiao Jian, Zhou Yanquan. Sentiment Polarity Analysis based multi-dictionary. In: Presented at the 2011 International Conference on Physics Science and Technology (ICPST'11); 2011.

[25] Lambov Dinko, Pais Sebastia˜ o, Dias Ga˜ el. Merged agreement algorithms for domain independent sentiment analysis. In: Presented at the Pacific Association for, Computational Linguistics (PACLING'11); 2011.

[26] Xu Kaiquan, Liao Stephen Shaoyi, Li Jiexun, Song Yuxia. Mining comparative opinions from customer reviews for competitive intelligence. Decis Support Syst 2011;50:743–54.

[27] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis Support Syst 2011;50:755–68.

[28] Fan Teng-Kai, Chang Chia-Hui. Blogger-centric contextual advertising. Expert Syst Appl 2011;38:1777–88.

© 2018, IJCSE All Rights Reserved