A Benchmark Study for Sentiment Analysis of Customer Satisfaction Based on US Airline Service Twitter Data

Al Hasib Mahamud[1], Mohammad Tariqul Islam Tuhin[1], Arnab Mitra Utsab [1], Marwa Nabi[1], Hasan Muhammed Zahidul Amin[1], Mohammed Julfikar Ali Mahbub[1], Md Motahar Mahtab[1], Annajiat Alim Rasel[1]

[1] Department of Computer Science and Engineering, School of Data Sciences (SDS), BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh
{al.hasib.mahamud, tariqul.islam.tuhin, arnob.mitra.utsab, marwa.nabi, muhammed.zahidul.amin, mohammed.julfikar.ali.mahbub, md.motahar.mahtab }@g.bracu.ac.bd, annajiat@gmail.com

Abstract. Internet plays a vital role in decision making as many views and opinions are available on different sites like Twitter, Facebook, and Instagram. These opinions are very important because people preference, choices can be guessed from there. So it may help both customer and service providers as service providers can understand the faults and other customers can easily take decision to choose one service from a lot of choices. Some research works are proposed based on US Airline Service to understand customers sentiment based on Twitter Tweets and other different sources. But some drawbacks can be seen from these research papers. In this article, a survey is done based on published research works to understand the methodology and drawbacks of their proposed works. Also a new research methodology is proposed to classify sentiment from Twitter tweets for US airline service in a target to overcome the drawbacks of the survey papers where performance is compared with existing methodology.

Keywords: Sentiment Analysis, US Airlines, Convolutional Neural Network

1 Introduction

Due to the increasing utilization of different popular Internet sites like Facebook, Twitter, Instagram and others, it is very easy to collect various types of data about different reasons. Customer feedback is very important as it provides the path to improve the quality of services, customer preferences, problems that may arise etc. Sentiment analysis can play a vital role in this regard. After performing sentiment analysis from user's text, an individual can easily understand the sentiment, preferences of the user. Sentiment analysis can be done in different domains like education, entertainment, recommender system etc.

Airline industry is considered one of the largest industry of this world. According to the reports of Federal Aviation Administration Air Traffic, flights are taken almost 2,246,000 passengers in the United States of America [1]. For this reason, it is important to perform sentiment analysis of those customers who have taken flights. Traditional methods which leverage questionnaires and forms seems simple but it takes time and the data from the answers of questions may have different noises [2]. It is better to use Twitter Tweets to perform sentiment analysis. Twitter is an online news and social networking service in America and in Twitter user can post messages and can interact with one another [1]. Jack Dorsey, Noah Glass, Biz Stone and Evan Williams created Twitter in March 2006, yet launching time was in July of that year [3, 4, and 5]. Twitter is considered reliable source because the Tweets are genuine and it is capable to use for investigation [6].

Some research works are seen based on airline industry to perform sentiment analysis where authors have used Twitter Tweets as dataset. Data preprocessing, natural language processing, and machine learning techniques are utilized in the methodology. Satisfactory results are obtained in different research works but some drawbacks are also seen. In this paper, a survey is done based on the published research papers on airline Twitter data to perform sentiment analysis. Also the implementation of the proposed methodology of this research occurs to improve the drawbacks of published research methodology.

2 Literature Review

Ankita et al. [2] discussed a multiclass sentiment analysis procedure using a dataset where the dataset was composed of tweets for six major airlines and the dataset was labeled by three classes- Positive, Negative and Neutral. The methodology of this work starts with data cleaning as twitter data is most of the time inconsistent and missing value can occur. Tokenization and lemmatization took place as data preprocessing. Symbols, punctuations and non-english words were eliminated as sentiment analysis was done based on English vocabulary. After completing the pre-processing, tweets are represented by vectors using Doc2vec. Finally seven traditional supervised machine learning classifiers used to predict the final class of tweets where AdaBoost achieved the highest accuracy 84.5%. The main drawback of this work seems dataset as the dataset is biased to one class.The total number of tweets are 14640 where the tweets belong to Negative sentiment are 9178 which is higher combining to other two classes. Also no deep learning algorithms are used in this work. Authors only utilized traditional machine learning classifiers where using deep learning classifiers may improve the accuracy.

E. Prabhakar et al. [1] proposed a research methodology introducing a sentiment analysis technique by improving Adaboost. Proposed methodology starts with dataset preprocessing to eliminate unwanted data. After utilizing data mining techniques, performance is analyzed by calculating precision, recall and F-score where the New Adaboost approach achieved highest precision and F-score respectively 0.78 and 0.68

compared to other traditional machine learning techniques. The main drawbacks of this paper seem that data processing and data mining techniques are not defined, authors did not write using methods they used for preprocessing and data mining.

Yun Wan at el. [11] shows a procedure for airline service analysis using twitter data where Weka is utilized to calculate each attribute's Gain Ratio and rank them in decreasing order. The supervised filter Attribute Selection is used and for the search option in the Attribute Selection filter InforGainAttributeEval algorithm is chosen. But this paper should do some research on basic ML approach and also can give a broad descriptions on methodology. For Lexicon-based approach their accuracy rate is too low.

Sachin Kumar et al. [12] studied emotion utilizing features extracted from tweets of some of the most well-known major airline companies' twitter accounts using the n-gram methodology and Glove dictionary method, along with machine learning approaches SVM, ANN, and CNN. Authors have created connections between different challenges that travelers confront using the Apriori algorithm [13].

In 2009, Prem Melville et al. [14] explored integrating word knowledge with text categorization to perform sentiment analysis on blogs. The majority of previous sentiment analysis research used dictionaries that define the sentiment-polarity of words and simple linguistic patterns to characterize the sentiment. Lately, some researchers have used a machine learning approach [15, 16]. In this paper the authors described a new approach using multinomial Naive Bayes classifier and for document classification, the abundance of positive and negative keywords is employed. In this process, they used data from various blogs that focus on enterprise software like the IBM Lotus software brand, movies to political matters.

Table 1. Information Extraction after Survey Papers.

Article	Methodology and Procedure	Result
Ankita et al. [2]	Preprocessing: Tokenization, Lemmatization, Removing Punctuations and Symbols Method Used: Doc2Vec, SVM, Decision Tree, Random Forest, AdaBoost, Logistic Regression, Gaussian Naïve Bayes	Six major US airline related Tweets are collected and 84.5% accuracy is achieved using AdaBoost.
E. Prabhakar et al. [1]	Method Used: SVM, Decision Tree, Random Forest, Boosting and Bagging	Based on top US airline service carrier related Tweeter Tweets, AdaBoost approach achieved highest precision 0.78 and F-score 0.68.
Sachin Kumar et al. [12]	Preprocessing: Json to CSV, Collecting hashtag from company Method Used: SVM, ANN, CNN, n-gram and Glove	Major airline companies Twitter follower's tweets are used and using CNN a drastic improvement is observed.

Yun Wan at el. [11]	Dictionary approach Methods: Supervised Attribute Filters, InforGainAttribute Eval, Ensemble Classifier	Using original tweets and retweets, highest result is obtained by Ensemble Classifier where Precision, Recall and F1-score all are 84.2%.

3 Methodology

The proposed methodology of this paper is based on the methodology proposed by Ankita Rane et al. [2] in a target to improve authors' method and to improve the accuracy of the proposed model. A comparison can be seen from the Fig. 1 and Fig. 2 where Fig. 1 shows the proposed methodology by Ankita Rane et al. [2] and Fig. 2 shows the proposed methodology of this paper.

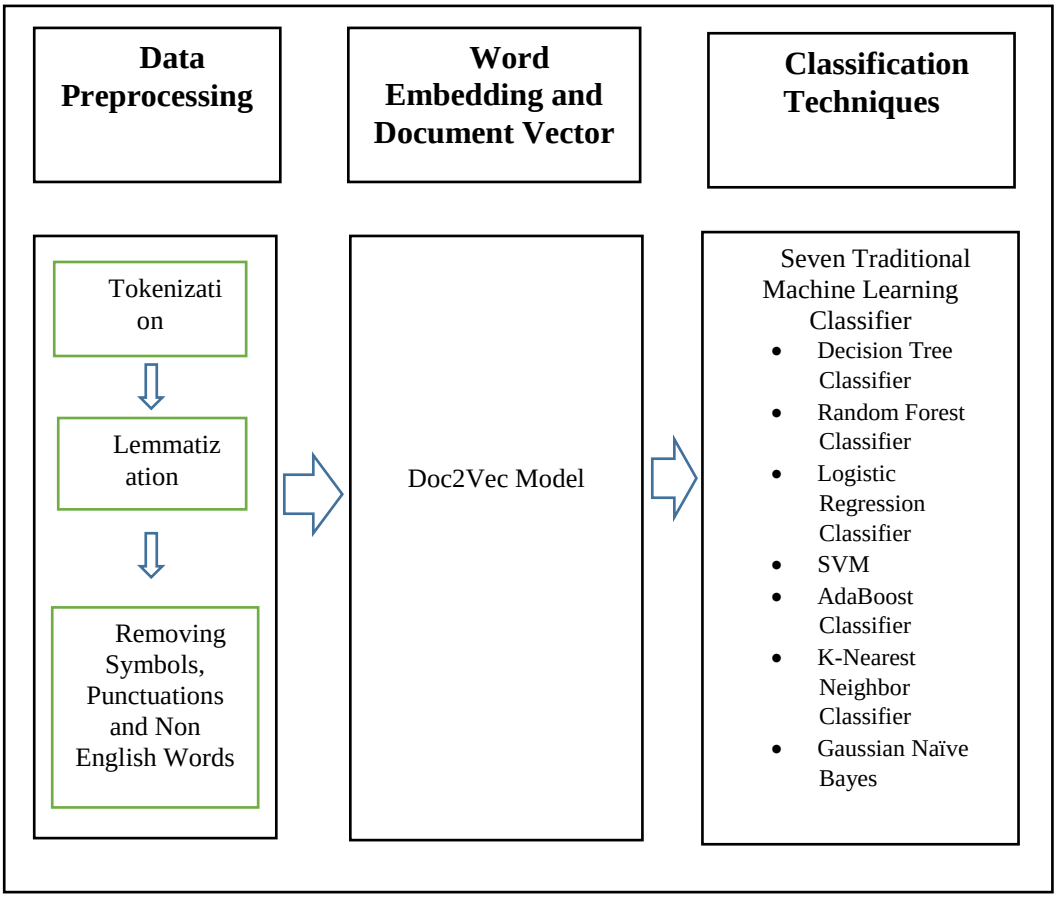

Fig. 1. Proposed Methodology by Ankita et al. [2]

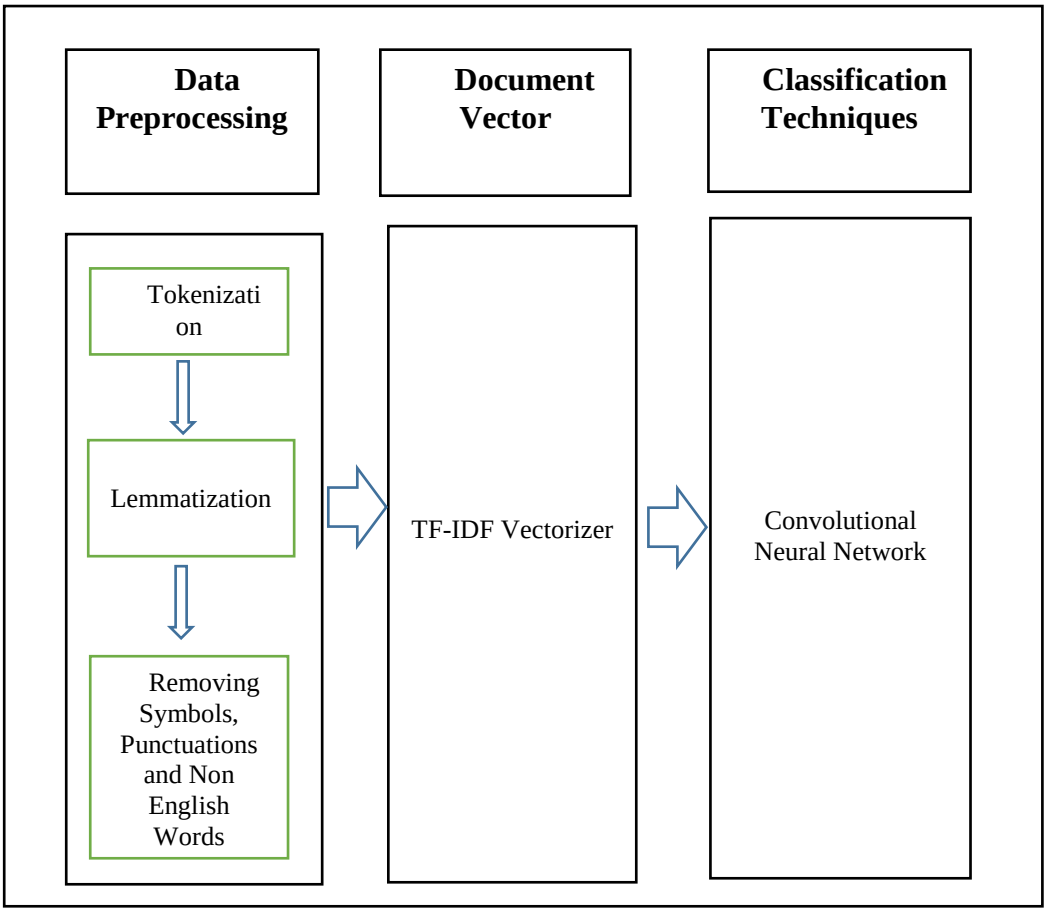

Fig. 2. Proposed Methodology of this research paper

3.1 Dataset Formulation

The dataset which is leveraged in this work is same of the dataset which is used in the research work of Ankita et al. [2]. The dataset is available at Kaggle [7]. The dataset is formulated of Twitter Tweets and the tweets are pre-labeled as the label describes the sentiment for that tweet so the dataset can easily employable for supervised machine learning techniques [8]. Class distribution of the dataset is shown in Table 2.

Table 2. Class Distribution of the dataset.

Sentiment	Number of Tweets
Positive	2363
Negative	9178
Neutral	3099

3.2 Data Preprocessing

For this proposed approach data preprocessing steps remain same as the Ankita et al. [2] proposed in their methodology, that is- tokenization, lemmatization and removing symbols, punctuations, non-English words. As Tweeter Tweets can be formed of inconsistent and missing values, data preprocessing is performed to remove noises. Tokenization is performed to convert the Tweets into tokens so after performing Tokenization unnecessary words and punctuations are removed from the Tweets [9]. Lemmatization is performed to make the common base form of words from Tweets [10].

3.2 TF-IDF-Vectorizer

To perform word embedding TF-IFD Vectorizer is used in this proposed approach so the semantic meaning of the sentences are created. To create verctor term frequency and inverse term frequency is needed to count. For this reason frequency of words are count firstly. After calculating TF and IDF, the value of TF and IDF is multiplied (TF *IDF) to create the vector.

TF= No of times a word appears in the sentence/ Total No of words in the sentence [17]

IDF=log (No of sentences/ Total No of sentences containing words) [18]

3.2 Classification Techniques

To finally classify the class of sentiment, Deep Neural Network (DNN) and Convolutional Neural Network (CNN) is used in this methodology. CNN is considered one of the deep neural network model. Though CNN is utilized for image data, CNN can be utilized in other sectors like speech recognition, text classification etc. For text classification, a word vector is used to represent the tweets [19]. The dimension of tweet matrix becomes $l*d$ where l is the length of the tweet and d is the dimension of word vector [20]. Having the filter size same to the region size, text matrix can be defined as image matrix [21].

4 Results and Discussion

The Scikit learn package, as well as the keras and tensor flow modules, are used in the experiments using Python 3.7. The analyses' balanced data set was methodically produced with three types of emotions: positive, natural, and negative. Because our dataset only has three sentiments classes. DNN and CNN classifiers are used in the analysis. Three methods of word embedding were used to create the dataset: trained, pre-trained, and hybrid. The data was used to test DNN in a variety of settings in order to find the optimal one. We chose DNN setups that performed well in terms of f1-score, accuracy, macro average, and weighted average, as shown in Table 3. We've got an 88 percent accuracy rate in this case.

Table 3. Classification report

F1-score	81%
Accuracy	88%
Macro avg	82%
Weighted avg	88%

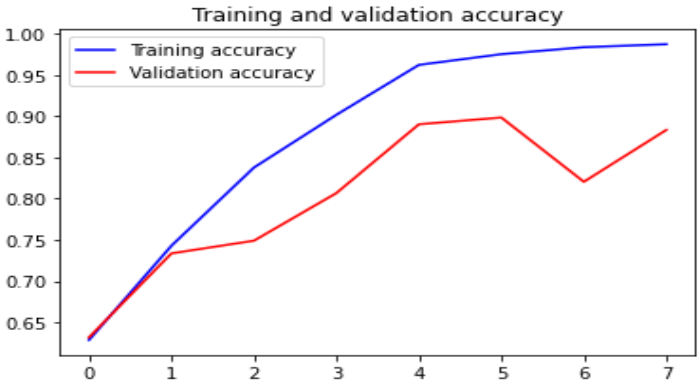

Fig. 3: DNN & CNN – Training and Validation Accuracy

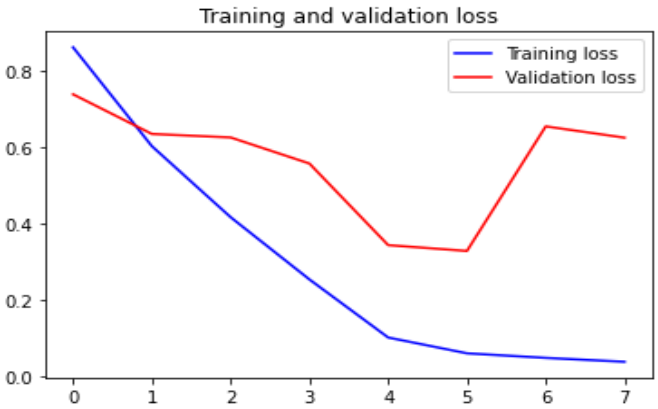

Fig. 4: DNN & CNN – Training and validation loss

Here we can see that with the increment of validation accuracy, training accuracy has also increased. And with the decrement of validation loss training loss has decreased.

To have a comparison between the methodology of Ankita et al. [2] and the proposed methodology of this paper according to classification report, it is clear that this methodology provides a better result. After utilizing seven machine learning classifiers, highest accuracy was obtained by Adaboost in Ankita et al. [2] proposed methodology, the value of accuracy is 84.5% where this proposed methodology provides an accuracy of 88%.

5 Conclusion

The accuracy of this paper is higher comparing to other papers in sentiment analysis for US Airline Services. The other factors of classification report like F1-score, macro average, weight average are also high comparing to existing research articles. In this paper, the main target was to improve the existing methods in sentiment analysis in airline services so that this approach can be utilized in other sectors. Our approach may help US airlines to take decision for improve their services and to understand customers preference and feedback.

References

1. Eswaran, Prabhakar & Santhosh, M & Krishnan, A & Kumar, T. (2019). Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach. International Journal of Engineering and Technical Research. 7. 1-3.
2. A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," *2018 IEEE 42nd Annual Computer Software and*

Applications Conference (COMPSAC), 2018, pp. 769-773, doi: 10.1109/COMPSAC.2018.00114.
3. "Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City". *TechCrunch*. July 30, 2012.
4. Twitter Search Team (May 31, 2011). "The Engineering Behind Twitter's New Search Experience". *Twitter Engineering Blog*. Twitter. Archived from the original on March 25, 2014. Retrieved June 7, 2014.
5. "Twitter turns six" Twitter.com, March 21, 2012. Retrieved December 18, 2012.
6. Kamal, S., N. Dey, A. S. Ashour, S. Ripon, V. E. Balas, and M. S. Kaysar. "FbMapping: An automated system for monitoring Facebook data." *Neural Network World* 27, no. 1 (2017).
7. "Twitter US Airline Sentiment" https://www.kaggle.com/crowdflower/twitter-airline-sentiment Accessed: 14/01/2022
8. Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1–2 (January 2008), 1–135. DOI:https://doi.org/10.1561/1500000011
9. "Tokenization" https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html. Accessed: 14/01/2022
10. "Stemming and Lemmatization" https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. Accessed: 14/01/2022
11. Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1318-1325, doi: 10.1109/ICDMW.2015.7.
12. Sachin Kumar and Mikhail Zymbler. A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6, 12 2019. doi: 10.1186/s40537-019-0224-1.
13. Rakesh A, Ramakrishnan S. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases. September 12–15. 1994. P. 487–99.
14. Melville and Gryc W. (2009) Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, KDD'09, June 28–July 1 2009, France.
15. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In EMNLP, 2002.
16. K. T. Durant and M. D. Smith. Advances in Web Mining and Web Usage Analysis, chapter Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection. Springer, 2007.
17. "What is Term Frequency" https://www.opinosis-analytics.com/knowledge-base/term-frequency-explained/#.YeGn__5BzIU. Accessed: 14/01/2022
18. "TF (Term Frequency)-IDF (Inverse Term Frequency)" https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558. Accessed: 14/01/2022
19. Kumar, S., Zymbler, M. A machine learning approach to analyze customer satisfaction from airline tweets. J Big Data 6, 62 (2019). https://doi.org/10.1186/s40537-019-0224-1
20. Zhang, Ye and Byron C. Wallace. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." *IJCNLP* (2017).
21. Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (ICML '08). Association for Computing Machinery, New York, NY, USA, 160–167. DOI:https://doi.org/10.1145/1390156.1390177