International Journal of Applied Operational Research
Vol. 11, No. 2, pp. 77-97, Spring 2023

Journal homepage: ijorlu.liau.ac.ir

Airline passenger's sentiment analysis for improving the quality of airline services by using a deep learning approach

A. Nourbakhsh*, M. Rezaei Chelkasari

Received: 9 February2022 **;** **Accepted:** 2 July 2022

Abstract Advances in technology have increased the availability and use of smartphones. Customer experience is one of the major concerns in the aviation industry. Twitter is one of the most popular social media platforms where travelers can share their feedback. Tweets' Classification based on user sentiments, is an important and common task which has addressed in many researches. Data mining, text mining, web mining, classification for analysis, and illustrating Twitter comments are some of the activities carried out in this field. Text mining is one of the prominent fields of data mining that able to extract useful information from travelers' tweets. This study presents a machine learning-based method for tweets analyzing to improve customer experience handling. The deep learning algorithm identifies ambiguous tweets and decides based on the level of ambiguity. The proposed method provides the feature vector for classification by extracting the word vector from the text analysis, constructing the added Message polarity feature with the WordNet dictionary from the trained examples. The results obtained from the deep learning algorithm validation show that the proposed method is able to identify passenger sentiments in two-class analysis with 99.97% accuracy and in a three-class analysis with 88.83% accuracy.

Keyword: Deep Learning, Opinion Mining, WordNet Dictionary, Tweet Sentiments Analysis, Machine Learning.

1 Introduction

User satisfaction is the ultimate goal that every service pursues. To achieve this goal, user participation in any service development and planning is essential. User participation can be assured by receiving their feedback, complaints, and suggestions. Decision-makers often use traditional methods such as questionnaires and surveys to obtain user feedback. Traditional methods are time-consuming and require effort and staff. With the development of social media in recent years, this process becomes easier [1]. People post their thoughts and comments on social media accounts such as Facebook, Twitter, Instagram, and other media. Decision-makers gain user feedback through data from these social media platforms and exploring comments. The process of extracting comments is called sentiment analysis, which Sentiment analysis is used in many areas. In the field of tourism, to get users 'opinions about

* **Corresponding Author. (✉)**
E-mail: nourbakhsh@liau.ac.ir (A. Nourbakhsh)

A. Nourbakhsh
Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran

M. Rezaei Chelkasari
Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

restaurants, hotels, airlines, and other things. In the e-commerce field, sentiment analysis is used for receiving customer comments about the product, in education it is used for receiving students feedback, and in transportation it helps by receiving passengers' comments about the transportation network, road conditions, and passenger complaints [2,3]. In analyzing social network tweets, contrary to direct feedback, travelers are comfortable expressing their opinions, but due to the high volume of passengers, the analysis of tweets becomes very confusing. In most cases, the tweets are not clear [4] that requires a problem-solving approach. The deep learning algorithm in the proposed method has the role of identifying ambiguous tweets that are decided based on the level of ambiguity. The purpose of this study is to analyze the comments of airline passengers by comments analysis for being effective in better service of airlines and providing the standard facilities for passengers. Gaining this aim is through reducing the error of identifying positive and negative comments of users' tweets in satisfaction with the flight service, improving detection accuracy by deciding on vague tweets—possibility of analysis for tweets with neutral polarity status in addition to positive and negative analysis-, and improving the service situation of passengers on airlines by identifying negative reasons. These can be achieved by developing a model based on predicting passenger satisfaction, by analyzing free Twitter comments. The most important research variables are travelers' comments and message polarity. That passengers' comments can be exploited after text analysis. Message polarity is a slightly discrete variable that can be measured through classification operations. Accuracy, F-measure, sensitivity and specificity criteria are used to measure classifier performance.

Although many researchers have paid attention to the field of emotion analysis, there are still many challenges that need to be addressed. Unstructured data, volume data analysis, the low performance of classifications, and sentiments analysis in specific areas are some of the challenges in identifying users' sentiments analysis. The low recognition accuracy rate is another challenge in this field due to the dependency of the features extracted from the text negative sentiments analysis of users' and behaviorists' tweets is an important and challenging issue that is caused by the wrong input texts processing. Also the diversity of people thoughts and expressions makes the situation difficult. In this study, Twitter social network data is used to inform users' behavior and comments. This data helps to track the user's interests and communications according to his point of view. Also, an airline passenger comments feedback system is reviewed in which different passenger comments and perspectives is assessed and evaluated to identify flight satisfaction by a deep learning algorithm.

The rest of the article is organized as follows. In the research background section, previous works are examined in several areas. There is also a comparison of the studies. In the research method, the details of the proposed method and the studied data are provided. In the research evaluation part, the research method is implemented step by step, and the results of each stage are interpreted and the approach of other methods related to the proposed method is discussed. At the end of the article, a summary of the study is provided, along with recommendations for improvement and future works.

2 Literature review

In recent years, lots of researches are devoted to the field of classification of social network users' preferences. Fen et al. [5] used Sentiment Analysis (SA) in their study to examine users' perceptions of public transportation in Malaysia. Understanding user feedback is critical for

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

improving the transportation system and providing a good riding experience for passengers. In this case tweeter messages has been used. These Tweets collected from the Twitter API must be processed before analyzing. The sentiments score of each collected tweet is calculated by using three standard Syuzhet, bing, and AFINN dictionaries. By implementing Machine Learning (ML) algorithms such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), tweets are classified into positive, negative, and neutral poles. Then, the confusion matrix is used to examine the performance of each model [5]. Based on the results of performance criteria, combining their specific research glossary with a support vector machine, caused the best class with 76.77% accuracy. Mohan et al. plotted the extraction of sentiments in a graph. Sentiment analysis is created in the form of movie reviews, product reviews, social discussions, travel reviews, and more. This makes the difference between a computer and a human. The aviation industry is a multi-million dollar industry with millions of passengers. Analysis of sentiments comments for reviewing Movie also received, which is based on the Twitter data classification into positive, negative, and neutral sections [6]. The importance of public opinion mining and the used tools are discussed by Prabhakar et al. [7, 8]. In a study by Wan et al. [9], researchers presented the extraction of sentiments that were used to solve these problems. This approach is found by comparing tweets against a predetermined sum of mental words. The Bayesian theorem was used to evaluate the posterior probability. They conducted this case study for three airlines [9]. In a study conducted by Prabhakar et al. [10], the importance of feedback was considered. Hrazi et al. [11], designed and implemented a sentiment analyzer that categorizes real tweets collected by the Twitter API. In this research, a sentiment analysis method based on machine learning is used and several Supervised learning algorithms such as logistic regression, simple BBs, support vector machine and decision trees were applied. Bag-of-word and TF-IDF techniques were used to extract features from pre-processed tweets, and uni-gram, bi-gram and tri-gram were used to rank features to identify the best prediction accuracy of the classifier. The result showed that among the used classification techniques, logistic regression has the best performance in terms of accuracy for validation and test data while using three-gram features with stop words and TF-IDF feature extraction technique.

In a study by Jabbar et al. [12], researchers found that online media has grown so much in recent years that it has also necessitated the analysis of sentiments. The purpose of studying the website is to explore customer views and feedback and determine whether the customer is interested in that particular product or not. This item can automatically extract customer feedback and comments about the product. This case is based on text analysis that uses Natural language processing method. This method is used in the e-commerce, transportation, and automotive industries. In a study by Prabhakar et al., a new classification system based on Twitter data analysis, was proposed for airline service analysis [10, 12]. There are two types of feelings. One is Dictionary-based sentiments, and the other is learning sentiments. Dictionary-based sentiments are done using dictionaries. These sentiments collect passenger feedback about airlines. Researchers have proposed a group approach to improve the accuracy. The authors also introduced an enhanced version of the enhancement approach, for better accuracy and control of unbalanced data [10, 13]. Algur et al.'s research [14] reported good results in analyzing sentiments by identifying speaker polarity in Twitter data. They set up tools for analyzing data sentiments by presenting some tweet data as input and obtaining the corresponding points as output [14]. In their study, Kitaoka et al.[15] tried to examine the relationship between geographic area and the emotional use of Twitter data. They used Geo-Twitter data to analyze the cause of local criminal activity. Location-based Social Networks

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

(LBSNs) data were used to model and understand local motivational behaviors of local criminal activities [15].

Krebs et al. [16] presented their work (based on Convolutional networks and recursive neural) on methods of exploring social sentiments to predict users' reactions to Facebook posts. They proposed and evaluated alternative methods for predicting these reactions to user posts on the public pages of organizations/companies (such as supermarket chains) [16]. Also Mohata et al. [17] described an effective way to explore big data, one with Apache Hadoop Map Reduce and the other with visualization-based methods called Visual Web Mining (VWM). They found that visual web exploration was effective in visualizing and gaining initial insights into big data, while Apache Hadoop and other technologies were effective in supporting later issues such as storage and processing [17]. Alghalibi et al. [18], focused on designing and implementing an in-depth model for illustrating Twitter comments (Mood) based on Deep Learning (DL) network. This research is about sentiment analysis based on Natural Language Processing (NLP) and the deep learning framework for illustrating and classifying Twitter exploration perspectives. The used method is based on sentiments natural language processing analysis on a large number of tweets to score the predicted state of the illustrated tweets, and as a result, public tweets are used to discover knowledge. In addition, it will be useful for detecting fake news. The mechanism consists of several successive stages, like: data collection stage, pre-processing stage, natural language processing stage, sentiments analysis stage, and prediction and classification stage using the deep learning model. The US Airlines sentiments analysis Twitter dataset was used, which was previously provided with data for all. The system, presented through the media and the public, monitors Twitter feeds. The system could visualize and extract meaningful data from tweets in real-time and store it for in-depth model analysis, That is suitable for a wide range of applications, including big data analytics solutions, predicting e-commerce customer behavior, improving marketing strategy, gaining competitive advantage in the market, in addition to visualizing various opinion mining programs [18]. Samonte et al. classified the polarity of tweets through annotations using naive Bayes (NB), support vector machine, and random forest classification, to develop a model. Researchers collected local airline tweets about their experience with services provided by airlines in the Philippines, determined the sentiments of positive, neutral, and negative tweets and provided quantitative and qualitative analyzes, as well as comment analyses to better understand the results of the experiments [19]. For examining Twitter's potential for gaining customer knowledge about airlines, Sreenivasan et al. focused on only five-, four-, and three-star airlines by collecting tweets and analyzing the contents of the messages. They analyzed tweets using content, and welcome tweets and compliments outweighed complaints [20]. The Carnein et al. , collected millions of customer-written posts for 48 airlines, 58 accounts on Facebook, and 66 accounts on Twitter, as well as responses from airlines between January and November 2016. They conducted their analysis on the top 10 airlines with 1.5 million social media posts. The results showed many changes in the level of services provided by airlines on social media [21]. Seo et al. examined the relationship between customers and social media. They examined social media business activities in customer behavior and brand value. Targeting Korea National Airlines, the survey completed by 352 participants, and 302 responses were used for analysis [22]. Their results are consistent with the work of Dijkmans et al. [23] measured passenger complaints on social media and perceived corporate reputation using Dutch airline KLM as a case study. The three main sections were divided into questions related to the reputation of perceived companies, the extent to which participants use social media, and the level of participating in

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

the social media activities of the airlines. The results showed that non-customer users on social media, follow an airline out of interest or curiosity, while customers are more likely to use specific features of social media pages such as customer service, direct feedback, and updates about interests in their flight [23].

Existing sentiment analysis models suffer from low accuracy due to inconsistencies in the tweet text and the assigned tag. From this perspective, [24] proposed a hybrid sentiment analysis approach where vocabulary-based methods are used with deep learning models to improve sentiment accuracy. In this research, the effectiveness of TextBlob was evaluated against AFINN and VADER (Valence Aware Dictionary for Sentiment Reasoning). CNN (Convolutional Neural Network), LSTM (Long Short Term Memory), GRU (Gated Recursive Unit), and CNN-LSTM have been applied as advanced machine learning models. In addition, the efficiency and effectiveness of TF-IDF (idiom frequency-inverse document frequency) and BoW (bag of words) were also investigated. The results show that the models perform better when trained using the emotions assigned to the TextBlob compared to the original emotions in the dataset. LSTM-GRU outperforms all previous models and studies with the highest accuracy of 0.97 and F1 score of 0.96. Among the machine learning models, the support vector classifier and additive tree classifier achieved the highest accuracy score of 0.92 with TF-IDF and BoW, respectively. Al-Qahtani and bint Abdulrahman [25] provided analysis of the content tweets text of users' emotions in US airline company services and investigated the use of two Machine Learning (ML) Logistic Regression (LR) and Naïve Bayes (NB), also four Deep Learning (DL) algorithms such as Convolutional Neural Networks (CNN), BERT, XLNET, and ALBERT to prediction of sentiment from US airline tweets. The results showed that both BERT and ALBERT methods outperform all other algorithms. Also in [26] (ALTamimi, 2022), text and sentiment analysis on extracted airline travel tweets is done, taking into account when the tweet was 'tweeted' and if it had a good or negative impact. In the [27] to measure customer satisfaction with sentiment analysis, relevant tweets were retrieved from the Twitter API and processed through tokenization and vectorization. After that, these processed vectors were passed to a pre-trained machine learning classifier for sentiment prediction. Then, time series methods such as Bollinger bands were used to detect anomalies in sentiment data. Using historical records from January to July 2022, it is proven that the proposed approach is able to capture sudden and significant changes in passenger sentiment. In the next section, a deep learning framework for Twitter opinion mining is proposed, which uses the deep learning model to predict users' comments and an illustration design for examining the quality of Twitter comments.

3 **Research method**

Sentiments analysis is an important issue in business intelligence applications and suggestion systems, where user input and feedback can be summarized quickly. The proposed method is a multi-step process that identifies the polarity of the tweet. In the first step, the data is entered and pre-processed, the attribute vector is extracted with conceptual and other kinds of attributes. The features extracted from classification, deep learning is used to identify the customer satisfaction status. The most important aspects of research innovation are identifying vague tweets with a deep learning algorithm and deciding on tweets with different levels of ambiguity. The weighting of the word vectors can distinguish between reliable and unreliable samples based on the number of their margins at the decision boundary. This

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

identification allows decisions to be made on a group of samples. In the proposed method, different decisions have been made on the two levels of ambiguity, high and low, to identify the polarity of the tweets, which is new in nature, and can effectively decide for neutral tweets. Social media allows users to express and share their information, ideas, and comments through social networking sites. Airline tweets are popular and are used as a data set to assess customer concerns. Twitter is a fast-growing social networking site, and people like to keep their comments simple and concise. Nowadays Regarding to data volume increasing, text mining has an important role in accessing and extracting meaningful insights from the text for specific purposes. The research method measures customer satisfactionarry by analyzing the sentiments of tweets for finding descriptive words, and interpreting them quantitatively. The new factor in this work is to check the vague tweets and neutralize them. The overall task is to extract and obtain the most accurate results based on the Twitter dataset for US Airlines and the implementation of different levels. Figure 1 shows the effectively of identifying the polarity of air traveler tweets from their personal comments on the Twitter network. Initially, the travelers' comments are received in the form of text tweets. The deletion of ineffective words is provided by the text-processing process, which includes tokenization. The deep learning algorithm is used to determine the polarity of the word vector after preparing the final feature vector. In the first step, after analyzing the text and extracting the effective words, the identification of ambiguous tweets is made by the naive Bayes classification algorithm. The classification results based on the reliability parameters of each class show the identification ambiguity of the tweet's polarity, so the level of ambiguity is decided. Unambiguous tweets will be separated, and ambiguous samples will have zero weight. For high-ambiguity samples, the value of the low-weight words is replaced by zero to smooth over low weights. This method leads to the decision about each tweet's polarity based on the keywords. Optimal and decided samples are merged together and selected for validation operations. In the next step, weighted words are extracted and stored as effective words for each tweet. By using the WordNet dictionary, the selected words are weighted and added as the Message polarity attribute from the word vector to the original vector. Configuring a deep learning algorithm for the final vector will lead to effective learning and optimal model construction. The results of the K-Fold validation test data shows that the research method for deciding ambiguous data is effective in reducing passenger polarization error.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

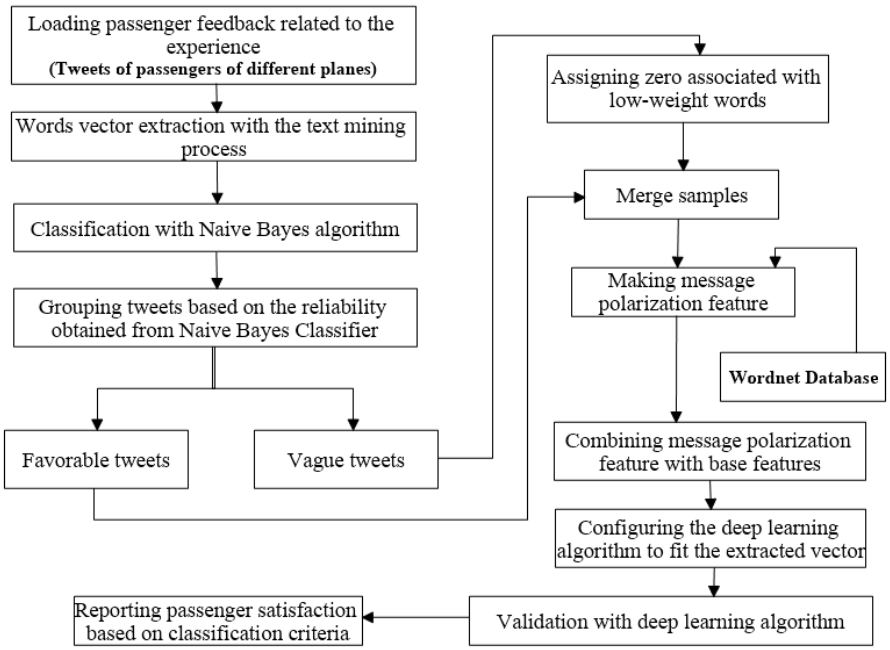

Fig. 1 A proposed method for predicting passenger satisfaction with flight experience

3.1 Classification comments approach based on Convolutional neural network

A convolutional neural network is a g function that is represented by mapping x data to another output vector denoted by y. Then, the g function will be a compact sequence of simpler functions f_l, called computational blocks, or denoted by $g = f_1 \dots f_L$. It is assumed that the network input is $x = 0$, and the network outputs are x_1, x_2, . . . , x_L, in which each output of the previous output is $x_l - 1$ and is calculated by taking the function f_l with the parameters w_l as shown in Equation 1 [28].

$$x_l = fl(x_l - 1; w_l) \quad (1)$$

The data flowing through the network represents the feature area $x_1 \in \mathbb{R}^{H_l \times W_l \times D_l}$; because x data has a spatial structure, H_l and W_l are spatial coordinates and D_l are the depth of channels. Function f_l acts as fixed interpretation and local operators, hence the network is called Convolutional. Convolutional neural networks are used to determine the distinction between classes by generating the probability vectors shown in Equation 2 , for all the images [29].

$$\hat{y} = f(x) \quad (2)$$

Where y is the actual data tag in the Convolutional neural network function. The real y tag of the x data is computed by the loss function, which assigns a penalty to classification errors (Koushik, 2016) [29]. As shown in Figure 2, where the weights specify the Convolutional filter, the Convolutional layer combines the result of the previous layer with a set of learning filters [30]. Each filter is sliding in width and height, producing a 1D-filter activation map. The filters have the same depth as the input. The output size can be controlled with three cloud parameters: depth, step, and zero paddings. The depth of the Convolutional layer is actually the number of filters applied to the input data. At the same time, the number of Convolutional filters steps allows the filter to jump when sliding on the dimensions of the

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

data. Finally, zero padding: The cover zeros are around the input boundaries to maintain their size. Figure 2 shows an example of processing.

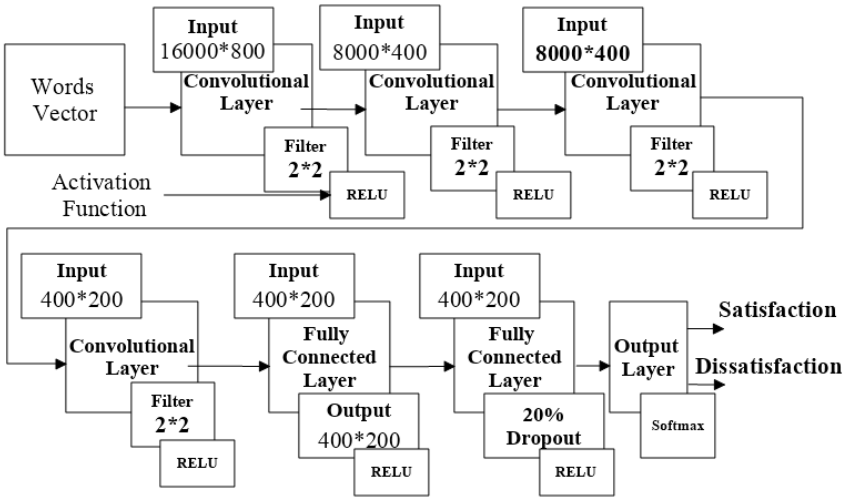

Fig. 2 Structure of the convolutional deep neural network

The Pooling layer parameter reduces its input size and allows multi-scale analysis. Maximum and average aggregation are the most popular aggregation operators. These operators calculate the maximum or average value in a small spatial block. Integration with two filter sizes with step 1 is considered ideal. Figure 2 shows the maximum aggregation performance with two filters. Also, an example of a 2 × 2 aggregation layer, is shown to be used to reduce the spatial size of the representation to reduce the number of parameters and computations in the network. The fully connected layers connect to all the neurons in the previous layer. Fully connected layers are usually considered as the last layer of the network and perform classification. An example of a Convolutional neural network is shown in Figure 2, which shows both previous layers.

3.2 Research data set

The data set used in the research is about the personal experiences of airline passengers, which were extracted from the social network tweets. Features include passenger tweets or comments about the airline, Message polarity, reply to the tweets, tweets location and tweets time zone. The most important analytical feature of the research is the textual feature of the tweet, which was obtained by the data collection tool from the Twitter social network. The attribute tag of the same Message polarity shows the sentiments of the users in positive and negative directions. The collection of research data includes 14,640 tweets from 7700 passengers. These tweets can be discussed for specific purposes such as quality of service or discussion of delays. The first experimental results for visualizing the entire Twitter comments are based on the distribution of the data set classes (modes). Table 1 shows the statistical analysis (comments chart) and statistical calculation for the entire Twitter airline data set based on observations and using statistical calculations for each state class.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

Table 1 Total statistical observations of airlines

#	class Comments	Frequency of each class's samples
1	Negative	9178
2	Neutral	3099
3	Positive	2363

Then, the tweets of each airline were extracted based on the statistical observations. Figure 3 shows the analysis of sentiments (chart of comments) Twitter mode based on Twitter comments for each airline.

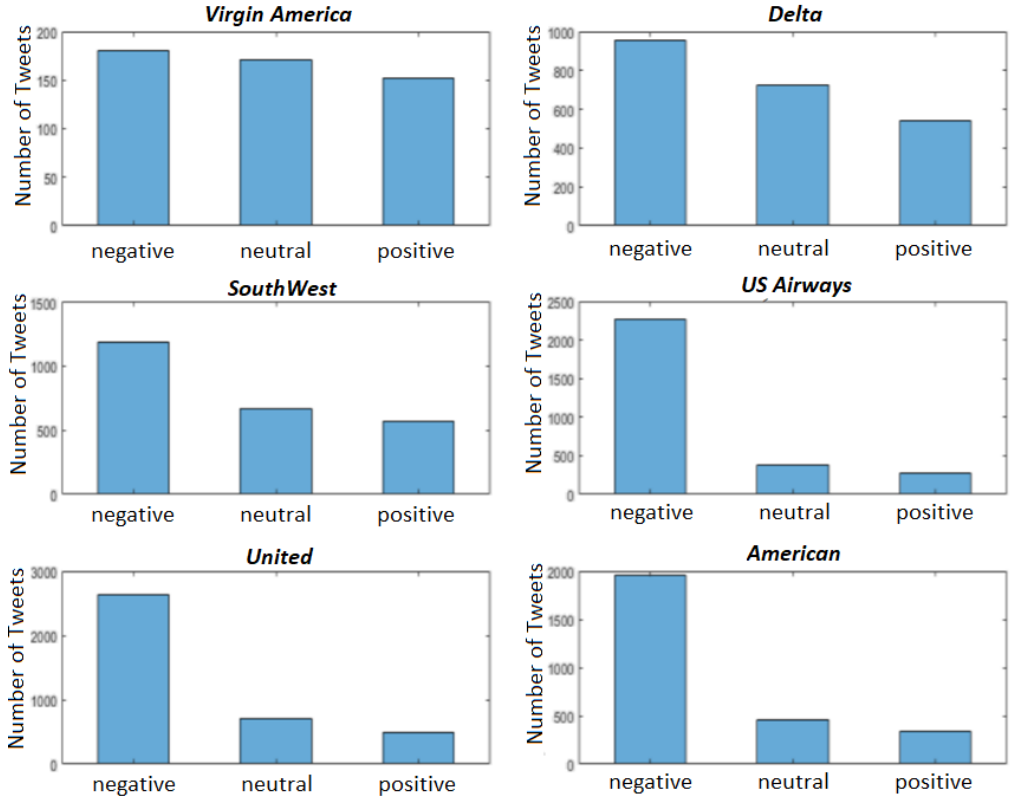

Fig. 3 Analysis of Twitter mode sentiments for each airline.

Table 2 shows the statistical observations for each airline in the entire data set. United has the most negative comments, SouthWest has the most positive comments, and Virgin USA has the best position in terms of the ratio of positive to negative comments.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

Table 2 Statistical view by each airline

Statistical view	Airline name	Statistical view	Airline name
negative 2263 neutral 381 positive 269	US Airways	negative 181 neutral 171 positive 152	Virgin America
negative 2633 neutral 697 positive 492	United	negative 955 neutral 723 positive 544	Delta
negative 1960 neutral 463 positive 336	American	negative 1186 neutral 664 positive 570	SouthWest

3.2 Measurement criteria

There are different criteria for evaluating the obtained results. Some important research criteria are mentioned below. The total accuracy is obtained from the number of correctly predicted samples to the total predicted samples. The amount of precision, which is the same as the accuracy of the positive category, is obtained from the class variation according to the status of the predicted labels. The recall criteria are obtained from the ratio of correctly identified positive cases to the total positive cases class. The three accuracy, recall, and precision criteria, which are more applicable in the classification results, are used for research results' evaluation.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad \text{(E.q. 3)}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \quad \text{(E.q. 4)}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \quad \text{(E.q. 5)}$$

4 Evaluation and results

In the previous section, a hybrid method based on feature extraction from text analysis and deep classification learning was proposed. In this section, the various stages of the proposed method are evaluated. The performance of machine learning algorithms is compared with that of the deep learning algorithm. Extracting effective words from travelers' tweet texts, scoring by analyzing WordNet sentiments and deep classification learning is among the steps evaluated in the research. Classification accuracy, precision, and recall related to sentimental classes are the most important criteria to be reported. Classifier analysis requires computational features. The study set, as shown in Table 3, contains 15 features and 14,848 initial records. 7 out of 15 attributes are suitable for analysis and are marked in the table with a checkmark. Some data are not labeled. The number of labeled data for analysis is14,616, so unlabeled data will be deleted before analysis. The airline sentiment feature of the dependent variable and other properties are considered as independent variables of the study set.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

Table 3. Selected variables of research data set for evaluation

Selection status	Description	Feature
√	Tweet Number	tweet_id
√	User Emotions (Negative, Positive, Neutral)	airline_sentiment
×	Confidence of users' emotions	airline_sentiment_confidence
√	The reason for the negative tendency	negative reason
√	Reliability of the reason for users' feelings	negativereason_confidence
√	Airline name	Airline
×	Analysis of Golden Aviation Tweets	airline_sentiment_gold
×	Tweet sender name	Name
×	The reason for the negative tendency of Golden Tweets	negativereason_gold
√	Number of open tweets	retweet_count
√	Tweet text	Text
×	Location of the tweeter	tweet_coord
×	Time to send tweets	tweet_created
×	Tweeter location	tweet_location
×	Tweeter position	user_timezone

The features of negative reasons of travelers are nominally available in the collection of research data. By reading all the negatives and deleting duplicates, and coding the remaining ones, the negatives mentioned by the passengers can be extracted. The names of airline winners are among the other features that are nominally mentioned in the data set, such as the negative reasons of passengers. Brand names are coded as negative reasons. Next, three algorithms, naive Bayes, K-Nearest Neighbor, and deep learning, were used for comparison. Naive Bayes shows the most efficient in terms of speed. KNN is the closest to the opposite of the deep neural model.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

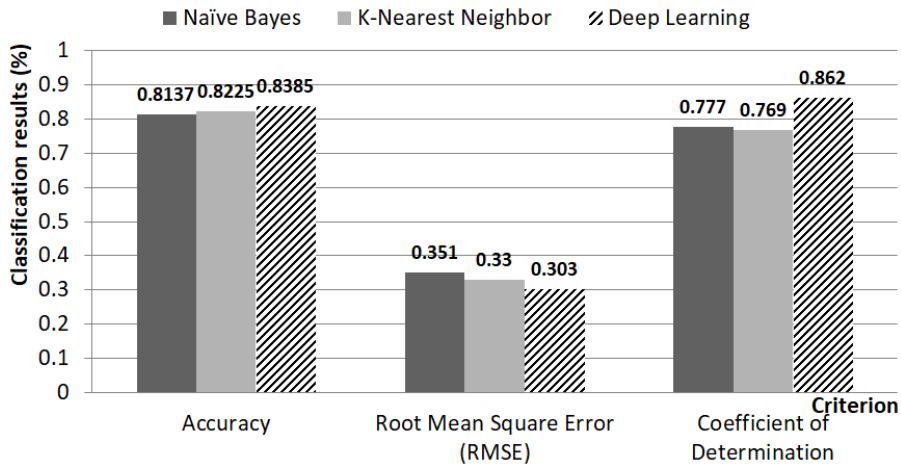

Fig. 4 Comparison of widely used algorithms' performance in identifying the status of user sentiments on the primary vector

In other words, the deep neural model has the most operations in the training process, but the K-Nearest Neighbor is only a simple mapping, and in the experimental stage, Nearest Neighbor calculations are performed. Figure 4 shows that the deep learning algorithm is higher in both accuracy and coefficient of determination than the other two algorithms and has a lower mean square error (RMSE). The analysis is based on the baseline variables, and the feature of the tweet text is not considered. A complete report of the classification results can be provided. The precision and recall criteria associated with neutral, positive, and negative sentiments show the extent of each classifier success in identifying a particular tendency. Table 4 presents the results of precision and recall of tendencies. Precision and recall of positive class do not have favorable conditions in all algorithms, and this issue is considered as an optimization challenge in the continuation of the research. However, the deep learning algorithm, regardless of the positive class, has better conditions in the precision and recall results in the other two classes than the other algorithms.

Table 4 classification challenge with precision and recall criteria

Criterion	Naïve Bayes	K-Nearest Neighbor	Deep learning
precision (neutral sentiments)	53.21	56.88	56.76
precision (positive sentiments)	33.33	43.56	0
precision (negative sentiments)	100	99.98	99.99
recall (neutral sentiments)	99.94	67.18	100
recall (positive sentiments)	0.04	33.22	0
recall (negative sentiments)	96.05	99.97	100

Negative class' Accuracy is favorable for all machine learning algorithms. Therefore, algorithms examining without negative classification can show the behavior of algorithms in identifying other sentiments. As Figure 5 shows, a significant loss of accuracy is achieved by removing negative sentiments. It also shows that the naive Bayes algorithm is more efficient

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

than other algorithms in identifying the other two classes. Therefore, the proposed method uses the naive Bayes algorithm to extract the desired samples before classification.

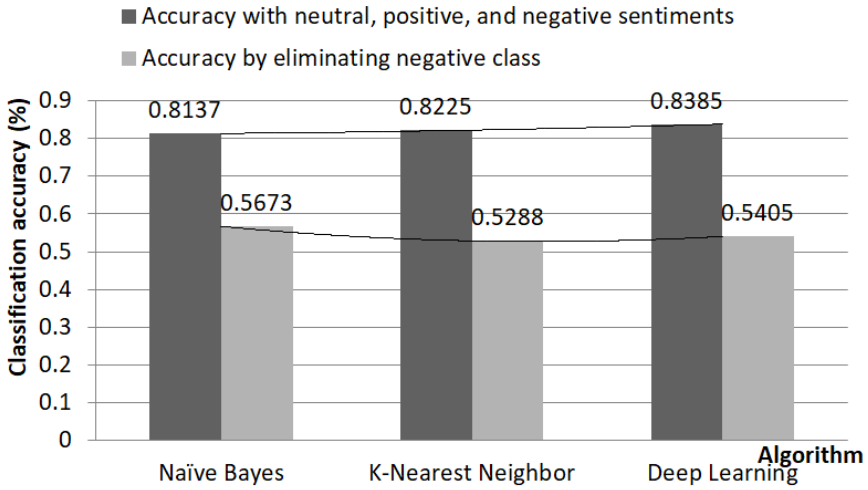

Fig. 5 Classification accuracy difference by the elimination of the majority class

In the tweets text analyzing, after pre-processing and words extracting, words can be scored based on well-known dictionaries such as Wordnet. But efficient words extracting can help summarize the scores associated with the sent tweets and reduce the error of analyzing user preferences. By removing stop words, numbers and symbols, we can expect a more favorable vector of words. In this research, effective words, in addition to applied filters, are created from repetitive words. In other words, words will remain in the attribute vector that is equal to or greater than a number of threshold repetitions (threshold = 3). Repetitive words are re-texted and saved as a tweet of effective words. In the rewriting process, words are marked and weighted by a known WordNet database. In the WordNet database, each word is rated as positive and negative between 1 and -1, and in the current analysis, the effective words are replaced by their weights. In each tweet, the total weight is obtained by calculating the sum of the used weights. The total weight of the tweet represents the result of the positive and negative weights of a tweet, which is determined due to the high value of a threshold In order to achieve maximum performance, the potential feature of the text tweet needs to be processed. The text analysis process begins with the technique of calculating the frequency of words after marking. The text analysis process begins with the technique of calculating the frequency of words after marking. First, all the letters are typed in lower case to have the same shape in terms of uppercase and lowercase letters. Extracted polysyllabic words are then marked with non-characters. Stop words, including words that do not have a specific meaning in the sentence, are removed. Finally, the words are filtered based on the number of characters. At the end, there are 14,616 samples and 382 features available for analysis. For extracting ineffective words by the word frequency calculation technique, each word has a weight of the number of repetitions in all tweet texts. Figure 6 shows some of the words extracted from the tweets

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

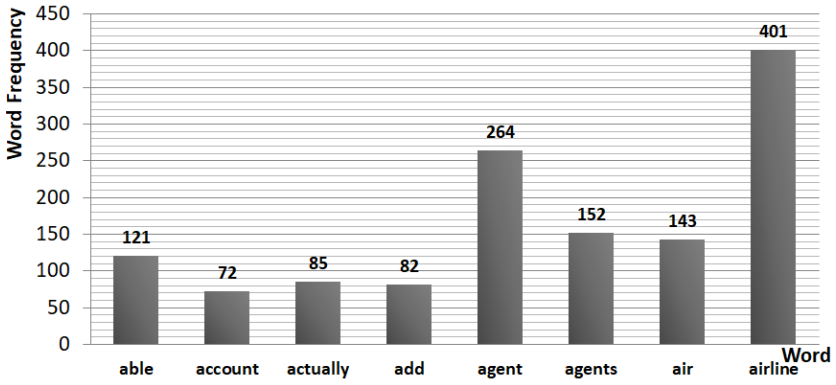

Fig. 6 Repetitive words extracted from users' tweets

The aim is creating a sentiments polarity feature that can be used to improve classification conditions. Providing favorable conditions for determining the tweets polarity is done in this and the next stages. 14,616 records and 382 attributes are selected for validation by naive Bayes Algorithm. After validation, the error-free samples are equal to 10,632, which the number of error-free samples of each class is given independently, in Table 5. As it turns out, the number of positive sentiment errors is higher than the total instances of this class compared to other classes. In the following, it will be analyzed based on the remaining optimal samples, to extract the polarity property. At the end, validation is done based on all samples.

Table 5 Results of sample separation after validation process with naive Bayes Algorithm

Class	Total sample number	Number of samples (no errors)	Sample number (error classification)
sentiments negative	9160	7771	1389
sentiments neutral	3096	1574	1522
sentiments positive	2360	1287	1073

In order to extract the feature of sentiments analysis from tweets, it is necessary to form the information in the text. To do so, in the attribute vector, each attribute is weighted and then they are put together as a text, to form a text summary of the text tweet. Table 6 shows a part of the word and tweet vector in which, in the last column, the weighted words are joined by a space.

Table 6 Sample text extracted from weighted words

fleet	Customer	flight	Fly	help	Summary
0	0.236	0	0.288	0	customer fly
0	0	0.135	0	0	Flight
0	0	0	0	0	-
0	0	0.106	0	0.207	flight help
0	0	0.133	0	0.52	flight help

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

Figure 7 shows the pseudocode of integrating effective words from the word vector. First, the last data set, which is obtained from the previous step, is loaded. The matrix determines the output data. There are two loops for row and column navigation of the input matrix. The set condition checks if the word weight is greater than zero and appends the word to the txt variable. At last, the attached words are saved in the form of abbreviated tweets in a separate file. The sentiments analyzing is done based on the summarized tweets.

```
[Data,header]  = Load('Dataset');
[m_r, m_c] =size(Data);
mat(1,:) = ["id", "text", "airline_sentiment"];
mat(2:(m_r+1),1) = Data(1:end, end);
a= header (1,1:end);
mat(2:(m_r+1),3) = tbl(2:end,end-1);
cnt = m_c -2;
for i=1:m_r
   txt = "";
   for j=1:cnt
      if(Data(i,j)>0)
         txt = txt + a(j)+ " ";
      end
   end
   txt = strtrim(txt);
   mat(i+1,2) = txt;
   end
end
Save('NewDataset');
```

Fig. 7 Pseudocode of extracting the summary of the important words of each tweet from the word vector

Important words are very effective in identifying the analysis. In this section, the important words of each tweet are weighed. The WordNet database consists of a dictionary of words with positive and negative weighting, which considers the weight of words between 1- to 1. This vast collection of words is provided by the University of Preston, USA, for pattern recognition researches. Table 7 shows word scoring word support, and no support status. For example, the third row that has the total points was positive. The probability is high. After scoring the desired samples, the other samples that were identified as error-prone, are given different weights. In this way, full-error and error-free samples are combined to classify all samples.

Table 7. Rate of the most important words of travelers' tweets with Wordnet

Total score	Score based on each word	negative	Positive	Words without cover	Covered words	The tweet's text is summarized
0.072	big (0.09) thing (-0.03) virginamerica (0.02)	0.1214	0.1934	0	3	big thing virgin America
0.1587	flight (0.03) pay (0.07) seriously (0.03) virginamerica (0.02)	0.0347	0.1934	0	4	flight pay seriously virginamerica
0.835	good (0.81) virginamerica (0.02)	0	0.835	2	4	amazing arrived good virgin America
0.22	know (0.21)	0.034	0.262	0	2	know virginamerica

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

81	virginamerica (0.02)	7	8			
0.02	virginamerica (0.02)	0	0.02	0	1	virginamerica
0.0373	amp (0.02) great (0.02) trip (-0.02) virginamerica (0.02)	0.104	0.1414	0	4	amp great trip virginamerica
0		0	0	0	0	-
0.17	thanks (0.15) virginamerica (0.02)	0	0.17	0	2	thanks virginamerica
0		0	0	0	0	-
0		0	0	0	0	-
0.1067	help (0.09) virginamerica (0.02)	0.0347	0.1414	1	3	help sfo virginamerica

All analyzes performed on the desired samples were used to extract effective tweet words. One of the most effective words was the polarity feature sentiments of the users. The production process of this feature is different in previous researches. In this study, optimal samples were extracted by the naive Bayes algorithm. Optimal examples increase the accuracy of the added Message polarity feature. Figure 8 shows the results of the deep learning classification algorithm for the analysis of two classes (positive and negative). In this analysis, neutral sentiments are placed in the category of positive sentiments and a new category called non-negative sentiments is created. These non-negative sentiments are expressed in the category of positive sentiments. As shown in Figure 8, recall of positive class and precision of negative class is completely estimated and global accuracy is estimated to be 99.97% (deep learning algorithm configuration is stated in the research method section).

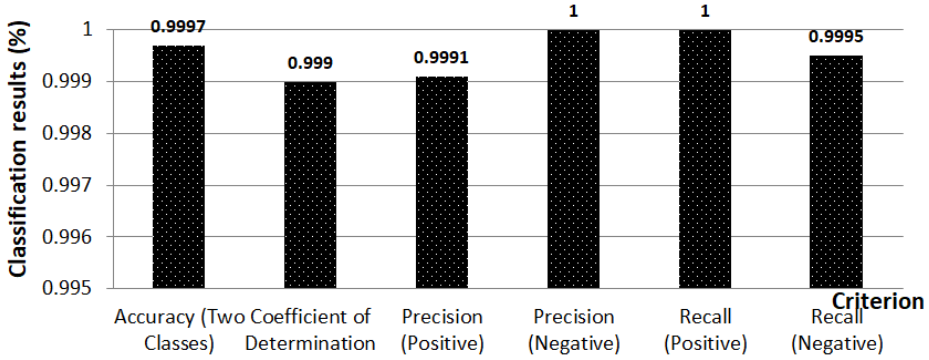

Fig. 8 Detection results in the analysis of two classes, positive and negative, with the deep learning algorithm

In the three classes' analysis, the classification results will be different. Figure 9 shows the distance of the prediction results from the actual data in 100 samples. Blue lines indicate actual data, and red lines refer to forecast data. According to the classification error, measured by root mean square error, the performance of the classifier is evaluated. In the K-Fold validation algorithm, the deep learning Root Mean Square Error of 0.278 was obtained.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

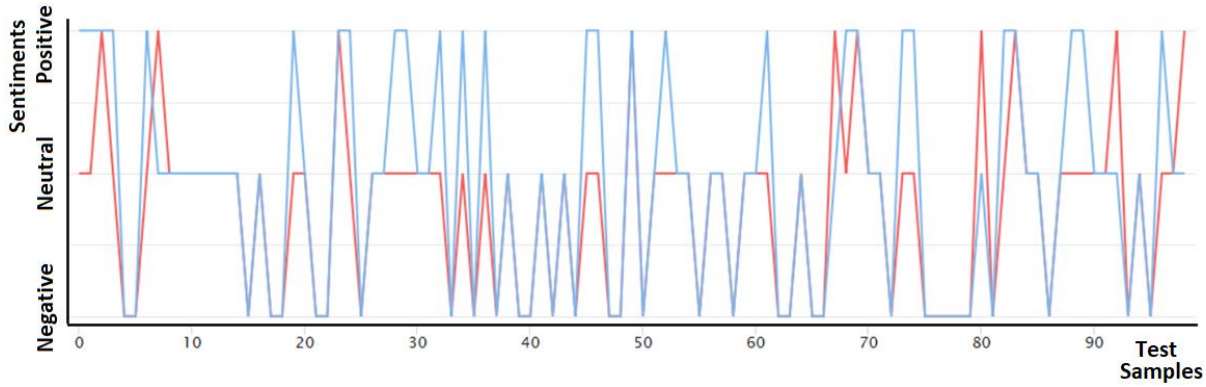

Fig. 9 Distance of prediction results from the original samples in the three-class analysis for 100 experimental samples

The classification results in the three-class analysis are shown in the prediction matrix of Table 8. As it is shown, 2857 samples correctly predicted neutral sentiments, 965 samples correctly predicted sentiments positively, and 9154 samples correctly predicted negative sentiments. Due to the low number of negative class errors, the performance of this class is higher than the other two classes. Based on the results obtained in the prediction matrix, many criteria can be reported as classification results.

Table 8 Confusion matrix for three-class classification analysis with the deep learning algorithm

	Neutral (real)	Positive (real)	Negative (real)
Neutral (predicted)	2857	1394	2
Positive (predicted)	238	965	4
Negative (predicted)	1	1	9154

As previously used, the two criteria, precision and recall, were used for different stages to show the improvement of the performance of predicting reclassification of these two criteria. Figure 10 shows the recall and precision of each class. As expected, the negative class has a high recall and precision. In both neutral and positive classes, progress compared to before. The precision neutral class increased from 56.76% to 67.18%, precision of positive class increased from zero to 79.95%, recall of positive class increased from zero to 40.89%. Although the recall of neutral class has a slight decrease compared to the base vector, the sum of the results shows an improvement in performance. Also, classification accuracy changed from 83.85% to 88.78% and coefficient of determination changed from 86.2% to 86.8%.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

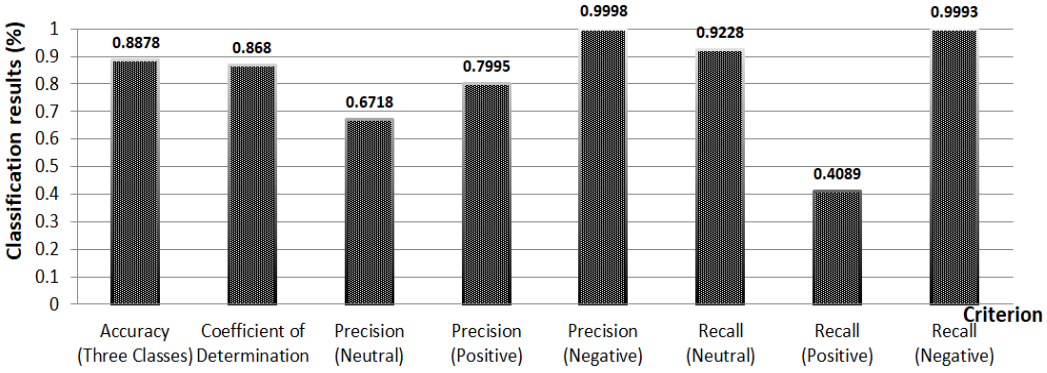

Fig. 10 Detection results in the analysis of three classes, positive, negative, and neutral, with the deep learning algorithm

As it is shown in the previous researches, the K-nearest neighbor algorithm has effective results in classifying words extracted from texts compared to many machine learning algorithms. Also, the Naive Bayes algorithm is faster than other algorithms. Therefore, two high-performance algorithms in this field were compared with the Convolutional recurrent neural network algorithm in [20]. The evaluation results of this study showed that the Convolutional recurrent neural network algorithm is more efficient than the two algorithms (naive Bayes and K-nearest neighbor). Convolutional recurrent neural network algorithm is able to reduce the computations in the learning process by sharing the weight of the nodes and mapping to fewer nodes; so by increasing the volume of tweets, the Deep Neural Network (DNN) with fully connected layers can not achieve the required performance due to increased training time, therefore the proposed method used the Convolutional model for higher performance. The research of Fen et al. [5] will be compared with the research method from different aspects and results. It can be concluded from [5], it is needed to improve the detection of sentiments of airline passengers. Their method was chosen for comparing with the results of the proposed method, because their model was tested in several transportation services. They took a simple and effective approach to process the text of users' comments and used a specific process of applying a machine learning algorithm in detecting the polarity of sentiments, as well as making comparisons with other machine learning algorithms. The results of their performance criteria show an accuracy of 76.77%, which was obtained by combining their specific research glossary with a support vector machine. The research method differs from the previous researches in several aspects. For optimal feature vectors producing, they used an additional pre-processing process, such as word-rooting, to extract words. For classification, they only used the support vector machine algorithm. In the proposed method, an effective distinction between identifying ambiguous tweets is provided and correctly predicted tweets and tries to construct the optimal feature vector by smoothing the low weights of tweets. Finally, deep classifier learning reports predictive metrics. The results obtained in Figure 11 show the efficiency of the proposed method in detecting passengers sentiments.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

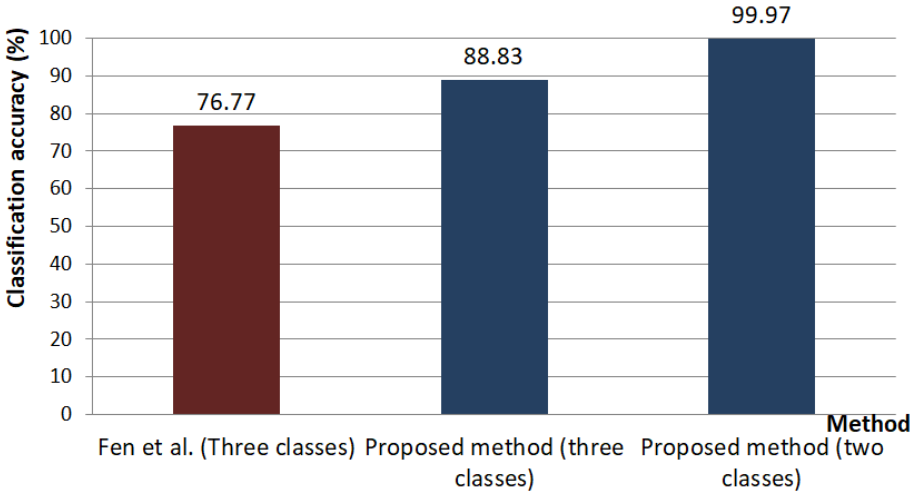

Fig. 11 Comparison of sentiments detection accuracy of the proposed method with Fen et al.

In the final part of the research, some of the most important limitations of the research need to be presented. These constraints contribute to the future recognition and development of the system. WordNet has a little support for names. Therefore, according to the subject under study, it is necessary to add custom words to the database with the desired weight. These weights should be assigned due to the word records in positive and negative sentences. In the process of textualization of the research method, there is a limit for tweets analyzing according to language. All reviewed tweets are in English. The deep learning algorithm configuration is presented according to the suggested word vector extracted from the tweets, and by editing the input tweets, the proposed neural network architecture needs to be changed. Some of the achievemens of this study are mentioned below:

The added Message polarity feature in many studies is done with WordNet analysis. Extracting effective words from prediction-free tweets increases the accuracy of the added Message polarity feature. Multi-syllable words extracting, is effective in conceptual word extraction. Conceptual words have a better relationship between independent variables and dependent variables; also they are more effective in extracting message polarity. The deep learning algorithm configuration is adjusted according to the added feature and the base vector. If the Message polarity method changes, there is no need to change the algorithm configuration. But for other add-ons, the deep neural network needs to be updated.

5 Conclusion

In this study, an evaluation performed on the method of recognizing passengers' sentiments on Twitter data. Different evaluation results were extracted based on two different categories. The first category is statistical observation and visualization of Twitter comments. The second category is the prediction and classification of Twitter comments by deep learning method. To predict and classify sentiments of Twitter based on the deep learning model training for the experimental data set (Twitter text data), trained parameters were used to predict the state of each Twitter. The results of deep learning prediction showed better results in comparing to the K-nearest neighbor and naive Bayes algorithms. Naive Bayes results showed highly effective in detecting positive and neutral sentiments. Words were extracted by text analysis. Optimal

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

samples were extracted by a Naive Bayes classifier. By analyzing the sentiments using the WordNet dictionary, the Message polarity feature was created on the desired samples. The proposed method improves the base vector by creating an adding Message polarity feature. The diagnosis was made with a prediction error of 11.17% for the three-class label and less than 1% for the two-class label. The following cases can be considered as the future activities in the development of current research. The first is the use of natural language processing in the Twitter exploration perspective for the most important aspects of sentimental analysis, and the second is the use of different methods for scoring words and calculating scores as an added Message polarity feature, suitable for research development.

References

1. Pournarakis, D. E., Sotiropoulos, D. N., Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. Decision Support Systems, 93, 98-110.
2. Farisi, A. A., Sibaroni, Y., Al Faraby, S. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. Journal of Physics: Conference Series,
3. Haghighi, N. N., Liu, X. C., Wei, R., Li, W., Shao, H. (2018). Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. Public Transport, 10(2), 363-377.
4. Soni, J., Mathur, K., Patsariya, Y. S. (2020). Performance Improvement of Naïve Bayes Classifier for Sentiment Estimation in Ambiguous Tweets of US Airlines. In Data Engineering and Communication Technology (pp. 195-204). Springer.
5. Fen, C. W., Ismail, M. A., Zayet, T. M., Dewi, K. (2020). Sentiment analysis of users perception towards public transportation using twitter. Putrajaya international conference on advanced research (PJIC2020).
6.Mohan, V., Venu, S. H. (2016). Sentiment Analysis Applied to Airline Feedback to Boost Customersâ€™ Endearment. International Journal of Applied and Physical Sciences, 2(2), 51-58.
7. Prabhakar, E., Kumar, V. S., Nandagopal, S., Dhivyaa, C. (2019). Mining Better Advertisement Tool for Government Schemes Using Machine Learning. International Journal of Psychosocial Rehabilitation, 23(4), 1122-1135.
8. Prabhakar, E., Parkavi, R., Sandhiya, N., Ambika, M. (2016). Public opinion mining for government scheme advertisement. International Journal of Information Research and Review, 3(4), 2112-2114.
9. Wan, Y., Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. 2015 IEEE international conference on data mining workshop (ICDMW).
10. Prabhakar, E., Sugashini, K. (2018). New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data. The SIJ Transactions on Industrial, Financial & Business Management (IFBM), 6(1), 7-11.
11. Hrazi, M. M., Althagafi, A. M., Aljuhani A. T., Rahman, J., Rahman M. M., Shorfuzzaman, M., Sentiment Analysis of Tweets from Airlines in the Gulf Region Using Machine Learning, (2021), International Conference of Women in Data Science at Taif University (WiDSTaif), pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430231.
12. Jabbar, J., Urooj, I., JunSheng, W., Azeem, N. (2019). Real-time sentiment analysis on E-commerce application. 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC),
13. Prabhakar, E. (2018). Enhanced AdaBoost Algorithm with Modified Weighting Scheme for Imbalanced Problems. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), 6(4), 2321-2381.
14. Algur, S. P., Patil, R. H. (2017). Sentiment analysis by identifying the speaker's polarity in Twitter data. 2017 international conference on electrical, electronics, communication, computer, and optimization techniques (ICEECCOT),
15.Kitaoka, S., Hasuike, T. (2017). Where is safe: Analyzing the relationship between the area and emotion using Twitter data. 2017 IEEE Symposium Series on Computational Intelligence (SSCI),
16. Krebs, F., Lubascher, B., Moers, T., Schaap, P., Spanakis, G. (2017). Social emotion mining techniques for Facebook posts reaction prediction. arXiv preprint arXiv:1712.03249.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]

17.Mohata, P., Dhande, S. (2015). Web data mining techniques and implementation for handling big data. Computer Science and Mobile Computing, 4(4), 330-334.
18. Alghalibi, M., Al-Azzawi, A., Lawonn, K. (2019). Deep Tweets Analyzer Model for Twitter Mood Visualization and Prediction Based Deep Learning Approach. International Journal of Computer and Communication Engineering, 8, 1-17.
19. Samonte, M. J. C., Garcia, J. M. R., Lucero, V. J. L., Santos, S. C. B. (2017). Sentiment and opinion analysis on Twitter about local airlines. Proceedings of the 3rd International Conference on Communication and Information Processing.
20. Sreenivasan, N. D., Lee, C. S., Goh, D. H. L. (2012). Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines. Program.
21.Carnein, M., Homann, L., Trautmann, H., Vossen, G., Kraume, K. (2017). Customer service in social media: An empirical study of the airline industry. Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband.
22. Seo, E.-J., Park, J.-W. (2018). A study on the effects of social media marketing activities on brand equity and customer response in the airline industry. Journal of Air Transport Management, 66, 36-41.
23. Dijkmans, C., Kerkhof, P., Beukeboom, C. J. (2015). A stage to engage: Social media use and corporate reputation. Tourism management, 47, 58-67.
24. Aljedaani, W., Rustam, F., Wiem Mkaouer, M., Ghallab, A., Rupapara, V., Bernard Washington, P., Lee, E., Ashraf, I. (2022). Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry, Knowledge-Based Systems, Volume 255,109780, ISSN 0950-7051, doi:10.1016/j.knosys.2022.109780.
25. Al-Qahtani, R., bint Abdulrahman, P. N. (2021). Predict sentiment of airline tweets using ML models. (No. 5228). EasyChair.
26. ALTamimi, E., Text Analysis of Airline Tweets, (2022). Thesis. Rochester Institute of Technology.
27.Wu, Sh.,Gao, Y., Happy or grumpy? A Machine Learning Approach to Analyze the Sentiment of Airline Passengers' Tweets. (2023). Accepted by 2023 TRBAM, under review for Transportation Research Record, arXiv:2209.14363 , doi:10.48550/arXiv.2209.14363.
28. Iftene, M., Liu, Q., Wang, Y. (2016). Very high resolution images classification by fine tuning deep convolutional neural networks. Eighth International Conference on Digital Image Processing (ICDIP 2016),
29. Koushik, J. (2016). Understanding convolutional neural networks. arXiv preprint arXiv:1605.09081.
30. Vedaldi, A., Lenc, K. (2015). Matconvnet: Convolutional neural networks for matlab. Proceedings of the 23rd ACM international conference on Multimedia.

[DOI: https://doi.org/10.71885/ijorlu-2023-1-638]