An analytic framework for enhancing the performance of big heterogeneous data analysis

Mohamed Salama¹, Hatem Abdul Kader¹, and Amira Abdelwahab¹,²

Abstract
The use of social media networks is becoming a current phenomenon in the world today where people are sharing posts and tweets, connect with different groups, and share their opinions about things. This data is extremely heterogeneous and so it is hard to analyze and derive information from this data that is considered an indispensable source for decision-makers. New techniques are therefore needed to handle these huge amounts of data to find the hidden information thus improve the results of the analysis. We are developing a framework for the analysis of heterogeneous data using machine learning (ML) techniques. In contrast to most of the literature frameworks that focus on a specific type of heterogeneous data for evaluating the proposed framework, we have analyzed 15k tweets data from six American airlines. These tweets are collected from the open stream of Twitter, also predict, classify each tweet as a negative or positive review, and test the ability of deep learning (DL) algorithms by comparing it with traditional ML algorithms. The findings confirmed the validity of the proposed framework and helped to achieve the study objective by providing excellent analysis performance and provide insights into additional aspects of information extraction from heterogeneous data.

Keywords
Heterogeneous data, text classification, data analysis, machine learning, information extraction, natural language processing

Date received: 12 September 2020; accepted: 29 December 2020

Introduction
With the fast growth of heterogeneous data, big data poses new problems for information extraction from this data. The researchers have noted over the years the value of knowledge that emerges from interactions on social networks. There is a need to understand and use this knowledge and analyze their huge volumes rising day after day.

In the present age, social media networks are known as one of the key sources of big data. A growing number of people are using these platforms not only to interact and communicate but also to express their opinions. This has led to a rapid increase in heterogeneous content via the internet.

The traditional methods of data analysis are unable to handle such data are considered heterogeneous data, heterogeneous data are data, which are not organized in a uniform format and therefore difficult to operate and use.¹

¹ Department of Information Systems, Faculty of Computers and Information, Menoufia University, Shbin Al Kawm, Menoufia, Egypt
² Department of Information Systems, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia

Corresponding author:
Mohamed Salama, Department of Information Systems, Faculty of Computers and Information, Menoufia University, Shbin Al Kawm 32511, Menoufia, Egypt.
Email: mohamed.salama@ci.menofia.edu.eg
heterogeneous Data from different sources would rise to 90% in the last few years and IDC analyze future company estimated that heterogeneous data would constitute 95% of global data by 2021.¹ This may include heterogeneous sources of data, including documents, photos, video, audio and social media comments, blogs, metadata, medical information, and web pages.

Text data analysis and Natural language processing (NLP) are two methods for the extraction of information from a textual context. The detection of interesting patterns and the extraction of useful information from a textual context has many uses and a variety of fields.² This approach is intended to organize complex heterogeneous data and get the necessary information.

Research has shown that most of the existing strategies for large data processing in social networking are focused on ML algorithms, DL, and solutions based on lexicons. From the point of view of many academics, ML is considered an area of artificial intelligence for the identification of data patterns in several social media networks.³

There are three main categories for ML, which are supervised learning, unsupervised learning, and reinforcement learning. Supervised learnings are those used for predictive analysis to solve this complicated issue of heterogeneous data analysis³ the right time for using ML methods and NLP in the most efficient way to process and interpret all these data.

In the classification phase of an analysis of sentiment, The classification process plays a crucial role in a pre-classified data sample called a training data set, is used to train and build a classifier using an ML technique this classifier tags the previously unlabeled data.⁴ In addition, the performance of any classifier often depends heavily on the data used to train it. DL is a subset of ML in which neural networks, human brain-inspired algorithms learn from large volumes of data. This is composed of various types of neural networks and neural models that work loose on a brain’s neuronal concept, and the data sets commonly used for such techniques are huge.⁵

DL is often used to construct deep neural networks to help solve complex binary problems, take decisions, or have high accuracy numerical answers. It has a great ability to identify trends in different data types like image, audio, text, etc. Lexicon-based approach² using a compilation of words, known as sentiment lexicon, consisting of predefined expressions of emotion.

A lexicon is used to count on target text the presence of feelings. Then choose the text subjectivity type of text, also known as text polarity, which is divided into two positive and negative main classes. Studies have shown that the use of ML approaches in text content analysis leads to improved performance.

NLP is an important process for analyzing social media data and is the ability of machines to understand and interpret human language, enabling machines to handle and comprehend heterogeneous text to improve social media data analysis.⁵ This method is a great example of an NLP that enables computers to manage heterogeneous texts such as retrieval in ways such as the extraction of artifacts, extraction of relationships, and identification of named persons, automated summing of texts, and stemming.⁶

In this study, we test the ability of the DL algorithms to enhance the analysis of heterogeneous data by contrasting it with conventional ML algorithms.

This paper is structured according to the following: Literature reviews and concerning studies are presented in the second section. The third section describes the framework proposed. The fourth section displays the experiments and outcomes. The discussion of the experiment is presented in the fifth section. Finally, the sixth section summarizes the results and discusses future work.

Literature review

Over the last years, the study of social media data has gained much attention this is large because of the recent expansion using social media. Since heterogeneous data analysis is an interesting topic for many researchers, a good number of articles in this area are published each year and the number of articles is growing over the years. Some literature reviews related to our work are presented here.

Rehman et al.⁶ present a strategy for obtaining improved data collection results. The model is proposed to minimize data in the early phases. In this study, the suggested approach focused solely on data reduction in this research and does not consider the use of ML approaches and data processing. Another research to test approaches that can enhance corporate decision-making is undertaken by Song and Ying.⁷ Through this research, multiple ML approaches were found to be used in data analysis. Therefore, the data review process suggested focused primarily on determining the best outcome for the open data according to the priorities of this study.

Furthermore, the study by Lismont et al.⁸ suggested a model for the classification of tweets as thematic, positive, and negative. Using Twitter API, they built a Twitter corpus to capture tweets and automatically annotate certain tweets using emoticons. They created a sentiment classifier based on the Naive Bayes multinomial technique, and the training set they used was less successful because it contains only emoticon tweets.

Many of the analytical techniques used by Qamar et al.⁹ concentrated on the review of tweets about telecommunications firms. They have used surveyed ML algorithms to classify data so that they can perform opinions mining on it.

They also used TF-IDF to determine how important certain each tweet is, findings showed an accuracy of 62%. Furthermore, Jimenez-Marquez et al.¹⁰ presented a
methodical attitude for comprehension of the approaches and techniques used for social media data in big data analytics.

It is used to analyze tourism data; in particular, reviews of hotel users, where user reviews are marked positively or negatively in the dataset. Results were used in many information combinations. Initially, learning classifiers with a multiclass classification were considered accessible for the five stars. The accuracy result was about 57%. Following, one star was then chosen as negative, two, three, four as neutral, and five as positive. It resulted in 67%.

Tripathy et al.11 used ML to collect feedback on a corpus of movie reviews. Several algorithms were used for the classification. By implementing support vector machine (SVM) while implementing the Naive Bayes (NB) classifier, the accuracy was 72.9%, resulting in 73.7%.

Another model is used for evaluating a wide corpus of Twitter posts to evaluate population mood.12 They find that the Twitter user’s mood could forecast the DJIA. A flagging detection scheme has been developed.

In addition, Chen and Zhang13 proposed a model combining SVM and CNN for the analysis of sentiment. The findings of their study show that their approach for text sentiment analysis provides greater precision than CNN or SVM alone. The data set used has been processed to filter it first and then Word2vec trains the filtered array. The CNN learns the features while the identification is handled by the SVM.

Ramadhani and Goo14 suggest a model for the classification of sentences using a combined CNN and RNN method. The words from sentences will be converted to feature vectors in the first stage. To lower the number of parameters, a small alternative of a CNN is then used. The result is then moved to an RNN of multilayered LSTM a classification layer is a 72% accuracy logistic regression classifier.

The method by Day and Lin15 analyzes the emotions of Google Play Chinese consumer reviews. Then the information was preprocessed, and the dictionary integration was done. Using the SVM classifiers, Naive Bayes, and LSTM, and comparing the results, the results show that the LSTM’s 71% accuracy yielded much better results than the SVM and Naive Bayes.

Ghosh et al.16 suggest a hybrid approach for classifying feelings using (RBM) and (PNN) in five data sets that performed better than traditional deep architecture and reduce dimensionality using a Contrastive Divergence Algorithm.

With the aid of Google’s Word2Vec used to predict the current word and use RNN for data training and subsequent clustering Zharmagambetov and Pak17 present an in-deep learning approach to sentiment analysis. The results showed that deep learning was better than the traditional method but was not the performance difference.

The study by Dey et al.18 presents a method for integrating heterogeneous and structured data into enterprise analytics. In their method, structured data were processed in the form of time series that captured enterprise performance data such as daily progress reports, purchases figures, income, and stock prices, while heterogeneous data were taken from customer reviews and comments, discussion forums, blogs, and social media.

This study utilizes text processing and mining techniques to extract information from heterogeneous sources and enables multiple heterogeneous inputs to automate the knowledge discovery process through the correlation of data extracted components.

An efficient Smart Big Data (3D, Ultra HD) algorithm on Cloud Computing systems based on advanced scalable media was suggested by Psannis et al.19 This study contributes to the understanding and survey of technology such as BD and Cloud to define their typical use characteristics and suggest an operation that will assist with the Big Data problem. The proposed encoding algorithm exceeds the conventional HEVC standard demonstrated by efficiency evaluations.

Furthermore, in the study of Jeba et al.,20 the practice of server consolidation for improved programming abilities and lower power and cooling costs was adopted by a virtual machine migration-based algorithm to decrease power consumption. The analysis relies on inferred complex resource planning based on three search algorithms—sequential search, random search, and optimum justice search, with a view to competently using data center services, findings have shown that about 30% of energy savings have been achieved.

Plageras et al.21 suggested a survey of the Internet of Things, Cloud Computing, Big Data, and Sensors technologies to define their common operations and integrate their functionality in order to provide useful scenarios for their use in gathering and handling sensor data in a smart building that operates in the IoT environment. With the use of an Arduino board and compatible sensors, the proposed system has been introduced to decrease energy and the cost of energy used in the construction.

In addition, the study by Gupta and Gugulothu22 presented the security problems with NoSQL databases MongoDB, HBase, and Cassandra and proposed a framework to achieve security for the web crawler applications using Cassandra, NoSQL. It used amazon Web Services a familiar cloud platform, this led toward the design of the scalable architecture for the NoSQL datastore and Cassandra where the data is analyzed and converted to a structured format suitable to evaluate the performance.

**Proposed framework**

**Data gathering**

The framework application-programming interface (API) provides data from various sources, including chat rooms, web forums, user reviews and reports, social networks, phone conversations, web conferencing, and so on. Other
frameworks tend to focus on one type of format, in particular textual data, our approach integrates all four types of heterogeneous data into one framework shown in Figure 1 (text, image, audio, and video).

We implement a framework for heterogeneous data collection and processing to generate outputs capable of providing insight and new information patterns.

The type of data will determine the techniques. For example, if we collect textual data for preprocessing from forums, discussion groups, and customer reviews from social media we will now use textual data analytics to walk through the techniques in the Framework shown in Figure 2.

The proposed framework for the analysis of big heterogeneous text data is shown in Figure 2. This approach suggests a sequence of phases, from data entry to the creation of ML models. This helps also to provide a structured understanding approach to understanding social media strategies and technologies. The suggested mechanism also helps to maintain a structured exploration approach.

### Data entry phase

It is possible to collect heterogeneous data from text data on social media Chats, blogs, or microblogging. In this study, we used a Twitter US Airline Sentiment dataset from Crowdflower’s Data for everyone library. This dataset contains a collection of more than 15k tweets from 7700 users. These Tweets are classified as positive, neutral, or negative for Six US airlines shown in Figure 3. The graph shows the difference between the emotions of different airlines. United Company had the most negative tweets of emotion, but also the highest number of tweets. The number of airline tweets can be compared to the number of airlines run by the airline, after applying the preprocessing to the data we need.

### Data preparation phase

Preprocessing is an important phase of Big Data analytics and a very important process for data preparation. This is also a crucial step in the method of data cleaning and is

---

**Figure 1.** A proposed heterogeneous data capture framework.
required to train every ML and DL model, Preprocessing and integration of data is performed to enhance the quality of big data analysis. This approach reduces information uncertainty and gives better analytical criteria, this increases comprehension of the complexity of the information and makes the data analysis process more robust and efficient, and without cleaning and preprocessing the results differ significantly.

This phase includes the finding of outlier values that do not correspond to the rest of the data. The preprocessing and cleaning phases are composed of different methods, for example:

- Delete noisy data: refers to modifying the values contained errors or outliers and removing the unnecessary data
- Switch to lower case: The lower case is important to ensure that the word translated into the corresponding function, i.e. “HELLO” and “HELLO” is converted into “hello.”
- Delete stop words: these are words that contribute little and give a little data. So, taking them as features like “The,” “an,” “in” and so on is meaningless. Terms to avoid text should be excluded.
- Gives every word back to its roots.
- Delete punctuation: such as commas, quotations, and symbols.
- Tokenization: This is the process by which the text content is divided into names, terms, and symbols.

Additionally, the removal of duplicates, numbers, URLs, usernames, and hashtags marks its significant steps. Any failure to detect and modify suspicious data will result in a considerable difference in the results of the analysis.

**Feature extraction and split data phase**

This Phase includes analyzing the text to recognize specific features described as a vector that can be interpreted by the classifier. Text representation is a very powerful and important stage in the processing of natural languages. In this model, the Term frequency-inverse document frequency has emerged to be a powerful word representation technique until today. It gives a representation of the word based on the word context. On the other hand, it has some limitations like vocabulary, but as we are using a small dataset, it is not going to be a problem right now the words in every text tweet will be our features. Each word will have to be tokenized for this reason. The 2000 most common words are used as features. After that, we need to split our data to train and test. The model uses 80% of the dataset to train and using the remaining 20% to test. This model will concentrate on the processing of natural language with ML and DL algorithms for text classification.

**ML and NLP integration phase**

This phase suggests the integration of ML techniques with NLP techniques. The model starts with training the classifier on the training dataset. NLP and ML techniques work together to understand the attitude of the text author. Analysis of data can be performed more easily compared with data using ML and NLP approaches. Production of this phase is a data analysis technique, which incorporates ML algorithms and NLP.

**ML and DL classification model apply phase**

This stage includes ML and DL classifiers used in this model. The ML works as good classifiers when the training set is relatively small, and they usually surpass the techniques of deep learning. We used the Scikit-learn library to incorporate these methods. For ML techniques, we used TF-IDF as a representation of the term. This model is used to evaluate current data to gain insights such as user interests and dislikes, extract features, and analyze data, depending on the ML and NLP methods, the predictive precision of the results can be different. This model uses a classification algorithm for classifying text with Different ML and DL classifiers for classification Like Logistic Regression, SVM Classifier, Naïve Bayes, Gradient Boosting, Random Forest, FFNN, CNN, and LSTM.

**Evaluation metrics phase**

Various measures defined based on the results of the confusion matrix, such as the accuracy of the correctly classified cases, can be defined as the proportion of the instances correctly classified \((TP + TN)/(TP + TN + FP + FN)\). Where \(TP\), \(FN\), \(FP\), and \(TN\) represent the number of true positives, false negatives, false positives, and true negatives, respectively, and certain measures used to validate each model and to test its accuracy, recall, and F1 score for each classifier.

Our evaluation measures are Precision (1), recall (2), and F1-measure (3).

\[
\text{Precision} = \frac{TP}{(TP + FP)} \quad (1) \\
\text{Recall} = \frac{TP}{(TP + FN)} \quad (2) \\
F1 = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)
\]

where \(TP\), \(FN\), \(FP\), and \(TN\) refer to the number of true positive cases, the number of false-negative cases, the number of false negatives, and the number of true negative cases, respectively.

(a) **Recall**: Correct positive ones are defined, separated by veritable positives and false negatives.
(b) **Precision**: Positive correctness divided by true and false positive ones
(c) **F-measure (F1)**: This is a calculation of the accuracy of the model using accuracy and memory, so analyzing the text is a great measure.

**The results and experiments**

The data set obtained from the open stream on Twitter is used to improve comprehension of the performance of their services and their users’ positive and negative reviews to validate the proposed model. We analyzed 15k tweet data from six American airlines, and trained ML and DL models to predict and identify each tweet is a negative or a positive review.

**Machine learning techniques evaluation results**

We present our experimental results in this section and evaluate performance based on different ML classifications. We run the classifiers for the dataset and compare performance based on, recall, f-score, and precision.

(a) **Logistic regression classifier**: is a simple technique classification, but at times powerful, for multiclass classification. Table 1 present the results of the classifier and Confusion Matrix results are presented in Table 2.

The confusion matrix (Table 2) is useful for gaining insight into the results. We have plotted the confusion matrix for the expected sentiments and the real sentiments.
(negative, neutral, and positive) are the likely outcomes of the category.

(b) Support Vector Machines classifier: is a very common classification technique. It just tries to find the line (vector) that best divides the groups. Table 3 shows the results for the classifier and Confusion Matrix results are shown in Table 4.

c) Naive Bayes classifier: The technique of NB is completely dependent on the probability theory for the classification of Bayes. It is a powerful classifier, as it is simple, fast, can work on large datasets. Table 5 present the results of the classifier and Confusion Matrix results are shown in Table 6.

d) Gradient Boosting classifier: is a type of ensemble model, fast, accurate, and parallel to the algorithm. It combines a set of weak learners and provides enhanced accuracy of prediction. Table 7 present the results for the classifier and Confusion Matrix results are shown in Table 8.

e) Random Forest classifier: Random forest is predictably dependent on decision trees, it’s accurate, robust, and rarely overfits, but it’s slow because it has so many predictability decisions to make. Table 9 present the results of the classifier and Confusion Matrix results are shown in Table 10.

Table 1. Evaluation results of logistic regression accuracy of logistic regression is 0.7950819672131147.

| class | precision | recall | fl-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.83      | 0.93   | 0.88     | 924     |
| 1     | 0.75      | 0.66   | 0.70     | 238     |
| 2     | 0.67      | 0.48   | 0.56     | 302     |

Accuracy: 0.80
Macro avg: 0.75
Weighted avg: 0.78

Table 2. The confusion matrix of logistic regression.

| Predicted | negative | positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 862      | 22       | 40      |
| positive  | 50       | 158      | 30      |
| neutral   | 127      | 31       | 144     |

Table 3. Evaluation results of SVM accuracy of SVM is 0.7903005464480874.

| class | precision | Recall | fl-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.84      | 0.92   | 0.88     | 924     |
| 1     | 0.72      | 0.63   | 0.70     | 238     |
| 2     | 0.63      | 0.49   | 0.55     | 302     |

Accuracy: 0.78
Macro avg: 0.71
Weighted avg: 0.78

Table 4. The confusion matrix of SVM.

| Predicted | Negative | Positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 846      | 25       | 52      |
| positive  | 44       | 161      | 33      |
| neutral   | 115      | 38       | 149     |

Table 5. Evaluation results of naive Bayes classifier accuracy of naive Bayes is 0.7867213114754098.

| Class | precision | Recall | fl-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.83      | 0.93   | 0.89     | 924     |
| 1     | 0.71      | 0.74   | 0.70     | 238     |
| 2     | 0.64      | 0.51   | 0.56     | 302     |

Accuracy: 0.79
Macro avg: 0.79
Weighted avg: 0.79

Table 6. The confusion matrix of naive Bayes.

| Predicted | negative | positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 920      | 0        | 4       |
| positive  | 178      | 55       | 5       |
| neutral   | 253      | 4        | 45      |

Table 7. Evaluation results of gradient boosting accuracy of gradient boosting is 0.7774590163934426.

| class | precision | Recall | fl-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.81      | 0.91   | 0.83     | 924     |
| 1     | 0.73      | 0.73   | 0.69     | 238     |
| 2     | 0.62      | 0.50   | 0.55     | 302     |

Accuracy: 0.78
Macro avg: 0.78
Weighted avg: 0.78

Table 8. The confusion matrix of gradient boosting.

| Predicted | negative | Positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 890      | 27       | 7       |
| positive  | 95       | 138      | 5       |
| neutral   | 241      | 24       | 37      |

Table 11 displays the ML classifier’s precision results and compares the output of each classifier used during model development.

According to the evaluation results, it is found that Logistic Regression and SVM are better for the given datasets than the other classifiers. They are the most accurate of all ML algorithms.
Deep learning techniques evaluation results

We present our experimental results in this section and evaluate performance based on different classifications of deep learning. We run the data set classifiers and compare the performance based on accuracy, precision, recall, and f-score.

(a) **Feed-Forward Neural Network classifier:** FFNN, given a moderate amount of data, can perform solid classifications. In this model, used trained word2vec. The key idea of Word2Vec is to define words with the aid of vectors so that semantic relationships between words are preserved as simple linear algebra operations.\(^{26}\) Table 12 present the results of the classifier and Confusion Matrix results are shown in Table 13.

(b) **LSTM classifier:** is an artificial recurrent neural network model used in deep learning, LSTM networks are well suited for detecting, processing, and forecasting based on data from time series. Table 14 shows the results for the classifier and Confusion Matrix results are shown in Table 15.

(c) **CNN classifier:** CNN is well known in image processing tasks, but it also performs quite well in the NLP field. It is also well known for its parallel computing capabilities. Table 16 presents the results of the classifier and Confusion Matrix results are shown in Table 17.

Table 9. Evaluation results of random forest accuracy of random forest is 0.7807377049180327.

| class | precision | recall | f1-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.81      | 0.93   | 0.86     | 924     |
| 1     | 0.76      | 0.59   | 0.67     | 238     |
| 2     | 0.66      | 0.48   | 0.55     | 302     |
| accuracy |        |        |          |         |
| 0.74  | 0.67      | 0.69   | 1464     |
| weighted avg | 0.77 | 0.77   | 1464     |

Table 10. The confusion matrix of random forest.

| Predicted | negative | Positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 864      | 18       | 42      |
| positive  | 72       | 140      | 26      |
| neutral   | 134      | 25       | 143     |

Table 11. Accuracy of ML classifiers results summary.

| Technique         | Accuracy | Precision | Recall | F1-score |
|-------------------|----------|-----------|--------|----------|
| Logistic Regression | 0.80     | 0.79      | 0.80   | 0.79     |
| SVM               | 0.80     | 0.79      | 0.80   | 0.79     |
| Naive Bayes       | 0.79     | 0.78      | 0.79   | 0.78     |
| Gradient Boosting | 0.78     | 0.77      | 0.78   | 0.77     |
| Random Forest     | 0.79     | 0.78      | 0.79   | 0.78     |

Table 12. Evaluation results of FFNN accuracy of FFNN: 0.7807377049180327.

| class | precision | recall | f1-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.81      | 0.93   | 0.86     | 924     |
| 1     | 0.76      | 0.59   | 0.67     | 238     |
| 2     | 0.66      | 0.48   | 0.55     | 302     |
| accuracy |        |        |          |         |
| 0.74  | 0.67      | 0.69   | 1464     |
| weighted avg | 0.77 | 0.77   | 1464     |

Table 13. The confusion matrix of FFNN.

| Predicted | negative | Positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 813      | 35       | 97      |
| positive  | 37       | 170      | 37      |
| neutral   | 74       | 33       | 168     |

Table 14. Evaluation results of LSTM accuracy of LSTM: 0.7780765027322405.

| class | precision | recall | f1-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.86      | 0.86   | 0.86     | 919     |
| 1     | 0.69      | 0.66   | 0.67     | 250     |
| 2     | 0.55      | 0.56   | 0.56     | 295     |
| accuracy |        |        |          |         |
| 0.77  | 0.77      | 1464   |
| weighted avg | 0.77 | 0.77   | 1464     |

Table 15. The confusion matrix of LSTM.

| Predicted | Negative | Positive | neutral |
|-----------|----------|----------|---------|
| Actual    |          |          |         |
| negative  | 800      | 23       | 93      |
| positive  | 42       | 179      | 45      |
| neutral   | 82       | 36       | 164     |

Table 16. Evaluation results of CNN accuracy of CNN: 0.7670765027322405.

| class | precision | recall | f1-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.84      | 0.86   | 0.84     | 860     |
| 1     | 0.69      | 0.65   | 0.66     | 271     |
| 2     | 0.54      | 0.56   | 0.55     | 333     |
| accuracy |        |        |          |         |
| 0.76  | 0.76      | 1464   |
| weighted avg | 0.76 | 0.76   | 1464     |
accurate of all DL algorithms. The results show that the deep learning model Feed Forward and RNN classifier with LSTM gives the highest accuracy.

The experiment discussion

This work used the available data and experimented with the proposed framework to analyze and understand data. The used dataset contains a collection of more than 14500 tweets from 7700 users. These Tweets are classified as positive, neutral, or negative. The collected tweets were for six big US airlines, which are US Airways, Virgin America Airways, Southwest Airways, United Airways, and Delta Airways. The tweets are a combination of positive sentiment, negative and neutral. The tweets classified with a variety of sentiment that assisted us to use the supervised ML approach. The programming was implemented in the anaconda environment using a Jupyter notebook, which is a versatile Python language development environment with advanced editing, checking, and numerical computing environment.

The results from the experiment are concluded as follows:

- The bulk of tweets have a negative feeling greater than 62%, as shown in Figure 4.
- Customer Service Issues and Late Flights are the main reasons for the negative feelings shown in Figure 5.
- Many American airlines, United Airlines had negative feelings about Airways tweets. United has the bulk of negative tweets, but because of customer service concerns, Southwest had the highest amount of negative feedback as shown in Figure 6.

| Table 17. The confusion matrix of CNN. |
|----------------------------------------|
| Predicted                              |
| negative | positive | neutral |
| Actual |
| negative | 777 | 38 | 95 |
| positive | 60 | 165 | 61 |
| neutral | 87 | 35 | 146 |

| Table 18. Accuracy of DL classifiers results summary. |
|-----------------------------------------------|
| Technique | Accuracy | Precision | Recall | F1-score |
| Feed Forward | 0.79 | 0.78 | 0.78 | 0.78 |
| LSTM | 0.78 | 0.79 | 0.79 | 0.79 |
| CNN | 0.77 | 0.76 | 0.77 | 0.77 |

Figure 4. Negative, neutral, and positive sentiments.

Figure 5. Reason for negative comments sentiments.

Figure 6. Reason for negative comments sentiments by airline.
• Negative tweets toward Delta are based primarily on old flights and not so much on issues of customer service as on the Rest of the airlines.
• The majority of tweets are aimed at United Airlines, followed by American and US Airways.
• Virgin airlines are getting very few tweets.
• Delta, Virgin, and Southwest tweets have approximately the same proportion of negative, neutral, and optimistic emotions.
• The majority of tweets are not replaced. Many tweets come from the United States & Canada.

Conclusions and future work
This study proposed a framework for solving heterogeneous data analysis problems and reducing the volume of data and focused mainly on the preparation and discovery of the optimum ML model for analyzing heterogeneous text data. In this study, several ML and DL algorithms have been compared throughout our case study to more than 14,500 tweets. We used supervised ML algorithms and deep learning algorithms to improve the precision of the positive, negative, and neutral classification of tweets.

The outcomes of the case study confirm that the framework is acceptable and useful for achieving the research objective by achieving excellent performance and adequate performance for broad heterogeneous classification of texts. We applied both ML and DL techniques and compare the results. ML techniques in essence are more simple and easier to implement.

DL approaches are better and more complex, producing better results than traditional ML algorithms in most cases. However, the gap in the accuracy of the two methods, even in these situations, is not very high.

The process of deep learning also increases the difficulty of the solution. The classification of sentences was much stronger, consisting of clear words describing certain emotions, such as great, bad, amazing horrible, and the list goes on. Overall, the performance of the models on small sentences was better than the long ones, but in classifying long sentences, deep learning techniques were better than ML algorithms.

ML classifiers especially logistical regression and SVM performed at the same level as DL classifiers, because the dataset is considered small, DL classifiers have been better at handling words that may vary in meaning depending on the context, especially LSTM, and have more trainable and learnable parameters, thus requiring more data for better performance. In classifying the negative type of reviews, the accuracy of all the implemented techniques was better because there are more negative reviews of the training data than neutral or positive reviews.

For the future, the expansion of this concept testing to the larger dataset is expected concentrating on test development to investigate the conceptual transformation accomplished by heterogeneous text analysis methods for larger datasets and attempting to identify sentences by Bert, replacing word2vec with Quick Text, which can manage both context and sub words.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Mohamed Salama https://orcid.org/0000-0002-6273-6748

References
1. Hariri RH, Fredericks EM, and Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 2019; 6(1): 44.
2. Alqasemi F, Abdelwahab A, and Abdelkader H. Constructing automatic domain-specific sentiment lexicon using KNN search via terms discrimination vectors. Int J Comput Appl 2019; 41(2): 129–139.
3. Ghani NA, Hamid S, Hashem IA, et al. Social media big data analytics: a survey. Comput Hum Behav 2018; 101(12): 417–428.
4. Jain K and Kaushal S. A comparative study of machine learning and deep learning techniques for sentiment analysis. In: 7th international conference on reliability, inforcom technologies, and optimization (trends and future directions) (ICRITO), Noida, Uttar Pradesh, India, 29–31 August 2018, pp. 483–487.
5. Mariel WC, Mariyah S, and Pramana S. Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naive Bayes for Indonesian text. J Phys 2018; 971: 012049.
6. Rehman MH, Chang V, Batool A, et al. Big data reduction framework for value creation in sustainable enterprises. Int J Inform Manage 2016; 36: 917–928.
7. Song YY and Ying LU. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015; 27(2): 130.
8. Lismont J, Vanthienen J, Baesens B, et al. Defining analytics maturity indicators: a survey approach. Int J Inform Manage 2017; 37(3): 114–124.
9. Qamar AM, Alsuhibany SA, and Ahmed SS. Sentiment classification of Twitter data belonging to Saudi Arabian telecommunication companies. Int J Adv Comput Sci Appl 2017; 8(1): 395–401.
10. Jimenez-Marquez JL, Gonzalez-Carrasco I, Lopez-Cuadrado JL, et al. Towards a big data framework for analyzing social media content. Int J Inform Manage 2019; 44(1): 1–12.
11. Tripathy A, Agrawal A, and Rath SK. Classification of sentiment reviews using machine learning techniques. Procedia Comput Sci 2015; 57(1): 821–829.
12. Bollen J, Mao H, and Zeng X. Twitter mood predicts the stock market. *J Comput Sci* 2011; 2(1): 1–8.

13. Chen Y and Zhang Z. Research on text sentiment analysis based on CNN and SVM. In: *13th IEEE conference on industrial electronics and applications (ICIEA)*, Wuhan, China, 31 May–2 June 2018, pp. 2731–2734.

14. Ramadhani AM and Goo HS. Twitter sentiment analysis using deep learning methods. In: *International annual engineering seminar*, Yogyakarta, Indonesia, 1–2 August 2017, pp. 1–4. IEEE.

15. Day MY and Lin YD. Deep learning for sentiment analysis on Google Play consumer review. In: *IEEE international conference on information reuse and integration*, San Diego, CA, USA, 4–6 August 2017, pp. 382–388.

16. Ghosh R, Ravi K, and Ravi V. A novel deep learning architecture for sentiment classification. In: *3rd international conference on recent advances in information technology*, Dhanbad, Jharkhand, India, 3–5 March 2016, pp. 511–516.

17. Zharmagambetov AS and Pak AA. Sentiment analysis of a document using deep learning approach and decision trees. In: *12th international conference on electronics computer and computation (ICECCO)*, Almaty, Kazakhstan, 27–30 September 2015, pp. 1–4.

18. Dey L, Verma I, Khurdiya A, et al. A framework to integrate unstructured and structured data for enterprise analytics. In: *Proceedings of the 16th international conference on information fusion*, Istanbul, Turkey, 9–12 July 2013, pp. 1988–1995. IEEE.

19. Psannis KE, Stergiou C, and Gupta BB. Advanced media-based smart big data on intelligent cloud systems. *IEEE Trans Sustain Comput* 2018; 4(1): 77–87.

20. Jeba JA, Roy S, Rashid MO, et al. Towards green cloud computing an algorithmic approach for energy minimization in cloud data centers. *Int J Cloud Appl Comput* 2019; 9(1): 59–81.

21. Plageras AP, Psannis KE, Stergiou C, et al. Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings. *Future Gener Comp Syst* 2018; 82: 349–357.

22. Gupta S and Gugulothu N. Secure NoSQL for the social networking and e-commerce based big data applications deployed in cloud. *Int J Cloud Appl Comput* 2018; 8(2): 113–129.

23. Stieglitz S, Mirbabaie M, Ross B, et al. Social media analytics—challenges in topic discovery, data collection, and data preparation. *Int J Inform Manage* 2018; 39(4): 156–168.

24. Wu JY, Hsiao YC, and Nian MW. Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment. *Interact Learn Environ* 2020; 28(1): 65–80.

25. Shafqat S, Kishwer S, Rasool RU, et al. Big data analytics enhanced healthcare systems: a review. *J Supercomput* 2020; 76(3): 1754–1799.

26. Hammou BA, Lahcen AA, and Mouline S. Towards a real-time processing framework based on improved distributed recurrent neural network variants with fast text for social big data analytics. *Inform Process Manage* 2020; 57(1): 102–122.