Multimedia Tools and Applications (2024) 83:42207–42231
https://doi.org/10.1007/s11042-023-17278-6

Sentiment analysis using a deep ensemble learning model

Muhammet Sinan Başarslan[1] · Fatih Kayaalp[2]

Received: 23 May 2022 / Revised: 9 May 2023 / Accepted: 25 September 2023 /
Published online: 16 October 2023

Abstract

The coronavirus pandemic has kept people away from social life and this has led to an increase in the use of social media over the past two years. Thanks to social media, people can now instantly share their thoughts on various topics such as their favourite movies, restaurants, hotels, etc. This has created a huge amount of data and many researchers from different sciences have focused on analysing this data. Natural Language Processing (NLP) is one of these areas of computer science that uses artificial technologies. Sentiment analysis is also one of the tasks of NLP, which is based on extracting emotions from huge post data. In this study, sentiment analysis was performed on two datasets of tweets about coronavirus and TripAdvisor hotel reviews. A frequency-based word representation method (Term Frequency-Inverse Document Frequency (TF-IDF)) and a prediction-based Word2Vec word embedding method were used to vectorise the datasets. Sentiment analysis models were then built using single machine learning methods (Decision Trees-DT, K-Nearest Neighbour-KNN, Naive Bayes-NB and Support Vector Machine-SVM), single deep learning methods (Long Short Term Memory-LSTM, Recurrent Neural Network-RNN) and heterogeneous ensemble learning methods (Stacking and Majority Voting) based on these single machine learning and deep learning methods. Accuracy was used as a performance measure. The heterogeneous model with stacking (LSTM-RNN) has outperformed the other models with accuracy values of 0.864 on the coronavirus dataset and 0.898 on the Trip Advisor dataset and they have been evaluated as promising results when compared to the literature. It has been observed that the use of single methods as an ensemble gives better results, which is consistent with the literature, which is a step forward in the detection of sentiments through posts. Investigating the performance of heterogeneous ensemble learning models based on different algorithms in sentiment analysis tasks is planned as future work.

Keywords Sentiment Analysis · Text Representation · Word Embedding · Ensemble Learning · Deep Learning · Machine Learning · Deep Ensemble Learning

✉ Muhammet Sinan Başarslan
muhammet.basarslan@medeniyet.edu.tr

[1] Department of Computer Engineering, Natural Science and Engineering Faculty, Istanbul Medeniyet University, 34720 Istanbul, Turkey

[2] Department of Computer Engineering, Faculty of Engineering, Duzce University, 81620 Düzce, Turkey

1 Introduction

With the development of social media applications, people instantly share their experiences of various social events such as films and concerts. Especially in the last two years, these practices have been the field of socialisation with the effect of the pandemic. They communicate their ideas about all developments in the world, from sharing personal information, to other people through social media. Social media has become a means of communication for people. As a result, people's sharing of ideas and experiences in many areas, such as current affairs, business, health, sports and the products they buy, has led to the emergence of a field such as sentiment analysis in the field of natural language processing. Sentiment analysis is the study of whether the comments people share contain positive, negative or neutral sentiments.

The local pneumonia epidemic that started in China in December 2019 has become a pandemic that has affected the whole world for more than two years. After spreading throughout China, the epidemic reached Europe on 1 March 2020. The World Health Organization then reported that the pandemic process had begun [1]. Quarantine decisions were made in country after country. During this time, people began to conduct their education, work and daily activities remotely. The lack of face-to-face activities led people to social media applications. The use of information technology in the tourism sector has had a positive impact on both the players in the sector, such as hotels and resorts, and those who want to spend their holidays there. People write about their experiences, especially on social media and TripAdvisor websites, and give ideas to those planning their holidays. Websites such as TripAdvisor and social media applications are the first port of call for holidaymakers, especially in the tourism sector, which has been closed due to the pandemic and is expected to peak in the summer of 2022.

Research into natural language processing is growing by the day. The main reason for this is the development of social media applications and internet technologies. Especially during the pandemic, usage rates have increased many times over. So much so that people are using social media applications to share their experiences of hotels, restaurants and commercial products, and to give advice to other people. These comments on social media and websites have given rise to sentiment analysis, a branch of natural language processing. According to people's feelings, many commercial institutions, especially in tourism, are more careful in their services and products. In this study, a sentiment analysis study was conducted on positive, negative and neutral labelled Corona tweets from Twitter and hotel reviews from the TripAdvisor application.

The fact that social media is a tool out of necessity and that it constantly expresses opinions on concepts such as social, economic, health, product and brand is called sentiment analysis, which is also done in this study. Sentiment analysis studies try to predict the emotional expressions in texts. The inference process in texts is made possible by natural language processing, a branch of artificial intelligence. In other words, sentiment analysis can be defined as one of the most common problems developed with natural language processing techniques.

Natural language processing, which is a subset of artificial intelligence, enables the inference process in text. In other words, sentiment analysis can be defined as one of the most common problems developed using natural language processing techniques. Studies on sentiment analysis began in the early 2000s. Some ideas such as perspective effects, emotion adjectives, metaphor interpretation were put forward before the 2000s, but in-depth studies in the field were carried out in the following years. Although sentiment analysis is commonly used in the studies, opinion mining, opinion extraction, sentiment mining, etc. have also been used.

The reason for the abundance of studies in this area is that people create large amounts of data using social media and forum-like sites where opinions can be expressed. In addition to studies on emotion detection [2, 3] on these large datasets, new methods based on deep learning are being developed. Neuro-symbolic AI [4], prompt-based analysis [5, 6] and graph-based methods [7, 8], which can be mentioned as new trends in emotion analysis studies, have been studied recently.

Using different algorithms from machine learning and/or deep learning together in a model to achieve a higher performance is called ensemble learning. There are many presented studies in literature which report that the results of using algorithms in ensemble format results better than using single methods. Because of this, we have decided to use the algorithms in ensemble way.

An ensemble system consists of ensemble building and ensemble integration. In ensemble building, different base classifiers are used to create different sets of models. In ensemble integration, many integration methods are used to combine the decisions of the base classifiers to obtain the final decision of the system. In addition to the classifier selection and ensemble integration methods required for ensemble generation. Two other important factors determine the success of the ensemble. These are. The sole performance of the core learners of the ensemble (low error) and the independence of the results of the core learners (high diversity). As the diversity of the core learners increases. The classification success of the system is also increased, variation, bagging for homogeneous systems random subspaces provided by traditional ensemble algorithms such as random forests and rotation forests. Variety of heterogeneous systems. While providing different learning algorithms, it ensures that the results are mixed with different decision techniques such as majority voting, stacking and bagging. In addition, the study has built models using the deep learning algorithms RNN and LSTM. In addition to the single use of machine learning and deep learning classifiers, models have been built using ensemble learning methods (stacking, majority voting), where they are used together in a heterogeneous way. The experiments investigated the performance of ensemble learning methods where deep learning and machine learning algorithms are used together in a heterogeneous way.

The main contributions of this study can be listed as follows:

- The impact of using deep learning and machine learning methods single and heterogeneously in ensemble methods on model performance was investigated.
- The effects of frequency-based text representation (TF-IDF) and predictive text representation (word2Vec) methods on the performance of sentiment analysis tasks on data from different domains were investigated.
- By selecting the dataset for the most affected hotel and corona virus in the pandemic Deep and machine learning methods were investigated, as well as ensemble learning methods that allow these methods to be used together.
- The impact of frequency-based text representation (TF-IDF) and predictive text representation (word2Vec) methods on learning models will be investigated on data collected from different domains.

The rest of the paper is organised as follows: Section 2 presents the related work. Section 3 consists of six subsections introducing the datasets, preprocessing, word embedding and text representations, ensemble learning, machine learning and deep learning methods, respectively. Section 4 presents the numerical results. Finally, the conclusions of the study are presented in Section 5.

2 Related works

The opinions and tweets shared by people on websites such as TripAdvisor and Twitter are the main sources used in sentiment analysis studies. Social media has become more widespread with the development of internet technologies, especially smartphones. Thanks to these developments, the field of categorising people's emotions has emerged with sentiment analysis studies within natural language processing.

Hotels are one of the most affected areas, especially during a pandemic. Therefore, hotel and COVID-19 datasets were selected. The datasets were used to investigate model performance using ensemble models, which allow machine learning and deep learning models to be used together rather than in isolation.

Ensemble learning algorithms have been shown to improve the performance of machine learning and deep learning models in many areas, such as natural language processing tasks, as in this study, and in the healthcare domain.

Ensemble learning algorithms have been shown to improve the performance of machine and deep learning models in many areas, such as natural language processing tasks [9–11], and in the healthcare domain [12, 13].

In this section, we describe related studies on the TripAdvisor [14] and Coronavirus [15] datasets used in the study, which were created using machine learning, deep learning and ensemble learning methods.

Mostafa collected a total of eleven thousand four hundred and fifty-eight reviews on the TripAdvisor website for five hotels in Aswan, Egypt. He used TF-IDF for text representation of these reviews. Among SVM, NB and DT models for sentiment analysis, NB gave the highest accuracy value of 0.85 [16]. Dehkharghani et al. used a TripAdvisor dataset consisting of approximately two hundred and fifty thousand reviews provided by customers for one thousand eight hundred and fifty hotels. In this study, they proposed subjectivity-based feature extraction. They built a model using SVM, NB and DT and SVM gave the best result with 0.8751 [17]. Raut and Londhe collected two thousand reviews from TripAdvisor website, five hundred positive and five hundred negative reviews for training sets and five hundred positive and five hundred negative reviews for test sets. They built a model for sentiment analysis on these reviews using SVM, NB and DT. They obtained an accuracy value of 0.87 in the model they created with SVM [18]. Tiwari et al. used n-gram method and different combinations of n-grams (bigram, unigram, trigram) for text representation in their sentiment analysis study on Rotten Tomatoes dataset. In their study, they built models using SVM, Maximum Entropy (ME) and NB and found that unigram, bigram and trigram methods each gave better results with ME. The best result was 0.8965 accuracy with ME after unigram and bigram [19]. Zhou created text representations with GLoVe and FastText after dividing one hundred thousand hotel reviews into 70% training set, 10% validation set and 20% test set. For sentiment analysis, models were built using BiLSTM and CNN. The BiLSTM model with GLoVe text representation technique gave the best performance with a test accuracy of 0.7373 [20]. Sahu and Ahuja created a dataset by collecting eight thousand movie reviews from the Rotten Tomatoes website. They built models using RF, KNN, NB and Bagging on this dataset. In their study, RF gave the highest accuracy of 0.95 [21]. Virginia et al. used different machine learning methods such as

KNN, NB, DT, SVM in their sentiment analysis studies on Rotten Tomatoes movie review dataset. KNN gave the highest accuracy value of 0.95 [22]. Mostafa conducted a sentiment analysis study using machine learning models following the word2Vec method to analyse the emotions of Egyptian students during the learning process with the pandemic. In the models built with NB, SVM and DT after word2Vec text representation, NB gave the best result with 0.91 accuracy [23].

Machuca et al. achieved an accuracy of 0.785 in their study on sentiment analysis using LR on tweets posted on Twitter during the COVID-19 pandemic in 2020 [24]. Siddiqua et al. achieved the best result with an accuracy of 0.877 in their experimental study on sentiment analysis using SVM, NB and Sequential Minimum Optimisation methods combined with MV ensemble learning on Twitter data [25]. Rahman and Islam built models using different ensemble learning methods for sentiment analysis on approximately twelve thousand tweet data in the UK. Among these models, they obtained 0.835 with stacking, 0.833 with voting and 0.832 with bagging, the last model [26].

Barkur et al. obtained 0.7429 accuracy in the model they built with LR after TF-IDF text representation on COVID-19 tweets [27].

Isnain et al. obtained 0.84 accuracy in the model they built with NB using TF-IDF and n-gram text representation methods on tweets collected to study the economic and social impact of COVID-19 [28].

Imran et al. obtained an accuracy value of 0.82.4 in the model they built with RNN and RNN variants using Bidirectional Encoder Representations from Transformers (BERT) and GLoVe text representation methods for sentiment analysis on COVID-19 data collected from Reddit [29].

Buntoro et al. obtained 0.8438 accuracy value in the model they built with SVM after TF-IDF and word2Vec text representation methods for sentiment analysis on COVID-19 data [30]. Vernikou et al. obtained 0.65 accuracy value in the model they built with LSTM neural network after TF-IDF and BOW text representation methods for sentiment analysis on COVID-19 data [31]. Antonio et al. obtained a 0.65 accuracy value in the model they created with Stochastic Gradient Descent (SGD) after TF-IDF text representation method for sentiment analysis on COVID-19 data [32].

As seen in these studies, experiments were conducted on text representations or classification models to improve the performance of models created in sentiment analysis studies. Similarly, in this study. Ensemble learning based models were created using traditional and prediction based text representation methods together with different classification models and their effects on classification were investigated. In addition, the results of the proposed method were investigated by combining the ensemble learning approach with deep learning techniques..

3 Proposed system

Feature extraction methods are used in the study in this section. Word embedding techniques, text representation techniques, methods for combining ensemble decisions, and the details of the proposed system are mentioned. The proposed system is shown in Fig. 1.

In our system, shown in Fig. 1, single classifier models (DT, KNN, LR, NB, SVM) were built after holdout separation. In addition, stacking (DT, KNN, LR, NB, SVM) and majority voting (DT, KNN, LR, NB, SVM) ensemble models of machine learning

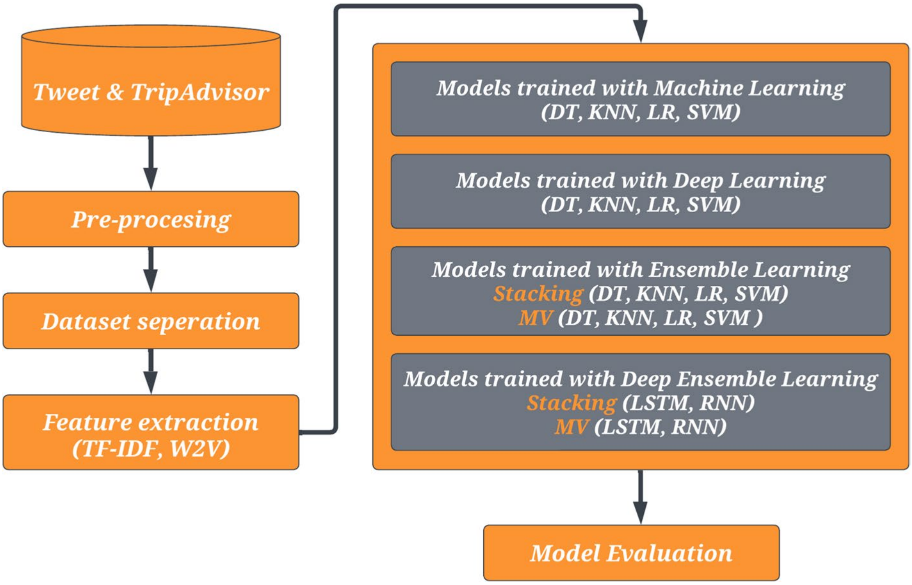

Tweet & TripAdvisor
Pre-procesing
Dataset seperation
Feature extraction (TF-IDF, W2V)
Models trained with Machine Learning (DT, KNN, LR, SVM)
Models trained with Deep Learning (DT, KNN, LR, SVM)
Models trained with Ensemble Learning
Stacking (DT, KNN, LR, SVM)
MV (DT, KNN, LR, SVM)
Models trained with Deep Ensemble Learning
Stacking (LSTM, RNN)
MV (LSTM, RNN)
Model Evaluation

Fig. 1 Flowchart of the proposed system

classifiers were also built. Deep ensemble models have been created with our proposed methods Stacking (RNN, LSTM) and Majority Voting (RNN, LSTM).

3.1 Datasets

In the study, comments from tweets and hotel review sites were used and shared as open source by those who had collected them from these sites.

The TripAdvisor hotel reviews dataset [14] consists of data from 20490 hotel reviews. Reviews with 1-2 stars are marked as negative. 3 are neutral and 4-5 stars are positive. Information about the dataset is given in Table 1.

It has been studied on the open source shared label Twitter Coronavirus dataset [15]. The dataset contains 6 features and 44955 data, as shown in Table 2. In the study, only the label attributes with the class Tweet and Sentiment were used. The other 4 attributes were not used.

Table 1 TripAdvisor Dataset information

Dataset Attribute	Attribute Description
Review	Review Text
Rating	Review Rating (stars)

Table 2 Twitter Coronavirus Dataset information

Dataset Attribute	Attribute Description
Location	Tweet location
Tweet at	When the tweet was sent Original tweet
Label	Sentiment class
Username	Twitter users can tag Name (integer)
Location	Location of the tweet
Screen Name	Name (integer) shown to other users on Twitter

3.2 Preprocessing

Twitter data was pre-processed prior to classification. This included removing punctuation and symbols in comments, converting characters to lower case and normalising to root words. The ’mentions’ used as ’@username’ and stop words were removed using the Python NLTK library..

3.3 Word embedding and text representations

Twitter data was pre-processed before classification. This included removing punctuation and symbols in comments, converting characters to lower case and normalising to root words. Mentions used as ’@username’ and stop words were removed using the Python NLTK library.

Word embedding attempts to find better representations of the words in a document. It attempts to extract semantic and syntactic information from documents into a dataset. Word embedding Each word is a distributed representation of words, in which they are represented as vectors of actual value in a predefined vector space. This notation is used in several natural language processing applications.

3.3.1 TF-IDF

TF-IDF is a statistically calculated weighting factor that indicates the importance of a term in a document. TF; Used to calculate the weight of terms in the document. IDF; Calculates the status of a word as a term by finding the number of occurrences of a word in more than one document. This is done by taking the absolute value of the logarithm divided by the number of documents divided by the number of documents in the period [33].

3.3.2 Word2Vec

Word2Vec is an unsupervised, prediction-based model that attempts to represent words in vector space. It was invented in 2013 by Google researcher Tomas Mikolov and his team. There are 2 types of submethods: Continuous Bag of Words (CBOW) and Skip-gram [33].

CBOW is an architecture for creating word embeddings that uses both future words and past words to create a word embedding. The objective function of CBOW is given by equation (1) [33]:

$$J_\theta = \frac{1}{T} = \sum\nolimits_{t=1}^{T} \log p\left(w_t \middle| w_{t-n,\dots\dots\dots,} w_{t-1,} w_{t+1,\dots\dots\dots} w_{t+n}\right) \tag{1}$$

In the CBOW model, the distributed representations of context are used to predict the word in the centre of the window. This is in contrast to skip-gram word2Vec, where the distributed representation of the input word is used to predict the context. The CBOW model is shown in Fig. 2 [33].

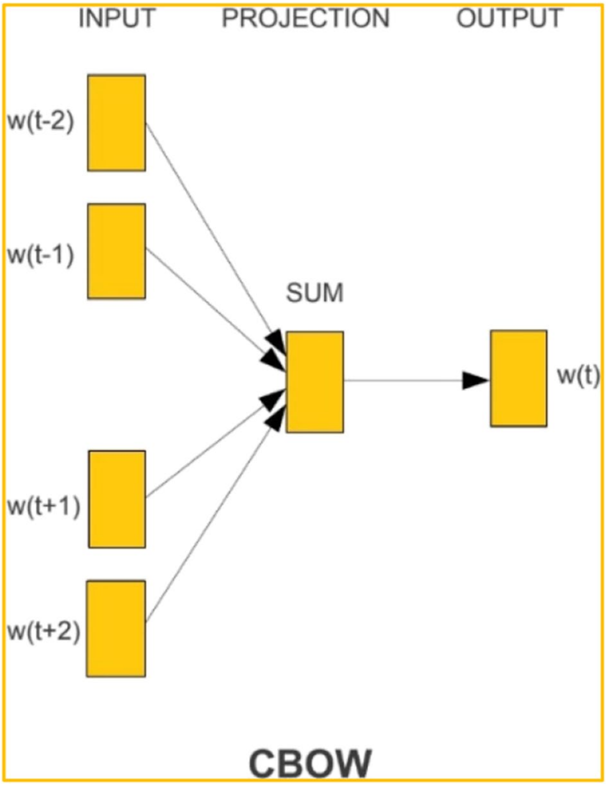
Skip-gram is an architecture for computing word embeddings. Instead of using the surrounding words to predict the centre word, as in CBOW word2Vec, Skip-gram word2Vec uses the centre word to predict the surrounding words [18]. The Skip-gram objective function sums the log probabilities of the surrounding n words to the left and right of the target word w_t to produce the following objective for Skip-gram, given by equation (2) [33]:

$$J_\theta = \frac{1}{T} = \sum\nolimits_{t=1}^{T} \sum\nolimits_{-n\leq j\leq n,\neq 0} \log p\left(w_{j+1} \middle| w_t\right) \tag{2}$$

The skip-gram model is shown in Fig. 3 [33].

In the study, the parameter settings for word2Vec were vector size 100 and 200, window size 5, sub-sampling rate 1e-3 and min-count 5.

Fig. 2 CBOW model [33]



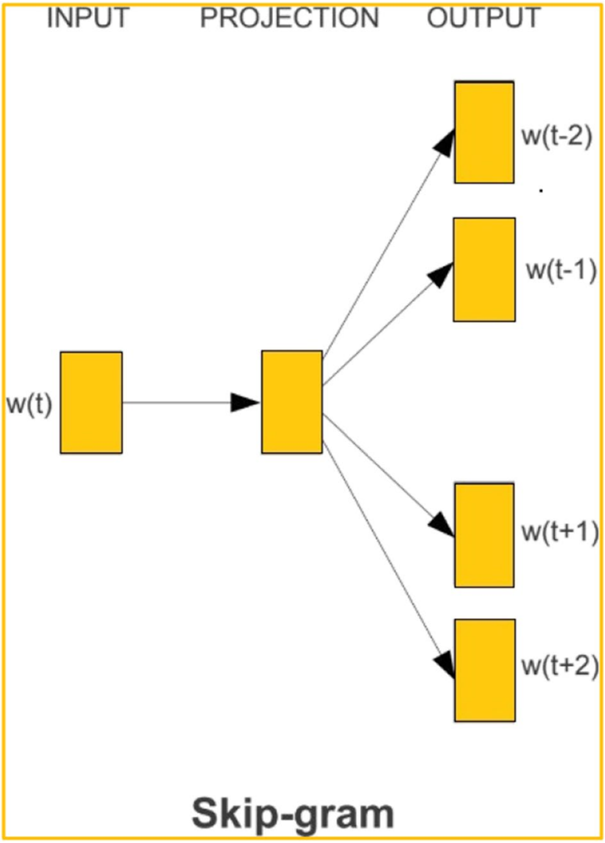


Fig. 3 Skip-gram model [33]

3.4 Ensemble learning

Ensemble learning refers to the use of machine learning algorithms together to solve mainly classification and regression problems. These algorithms can be of the same type (homogeneous ensemble learning) or of different types (heterogeneous ensemble learning). Ensemble learning performs a strategic combination of different expert or machine learning models to improve the effectiveness obtained with a single weak model [34].

There are different types of ensemble learning techniques, which mainly differ in the type of models used (homogeneous or heterogeneous models), the data sampling and the decision function (voting, averaging, metamodel, etc.). Stacking and majority voting were used in the study.

3.4.1 Voting

Voting is perhaps the simplest ensemble algorithm and is generally very effective. It can be used for classification or regression problems. Voting works by creating two or more submodels. Each submodel makes predictions that are somehow combined. The average or modes of the predictions are taken and each submodel is allowed to vote on what the result should be [34].

Majority voting In majority voting, each single classifier votes for a class and the majority wins. Statistically, the predicted target label of the ensemble is the mode of the distribution of the single predicted labels. We predict the class label by majority (plurality) voting of each classifier C this equation (3) is given [35].

$$\hat{y} = mode\left\{C_1(x).C_2(x).\dots\dots C_m(x)\right\} \tag{3}$$

Suppose, as in the following equation (4) [35], we combine three classifiers that classify a training sample:

$$\hat{y} = mode\{0.0.1\} = 0 \tag{4}$$

By majority vote. We would classify the sample as "Class 0" [35].

Stacking Stacking is an extension of the voting method used in classification problems. You can choose more than one submodel. It also allows you to use another model for the best combination of predictions. [34].

3.5 Machine learning

Machine learning is one of the areas of artificial intelligence that has been studied in many fields in recent years. It is an area of computer science that includes various learning algorithms that allow inferences to be made by developing models of data. SVM, NB, KNN, LR, DT classifier algorithms used in the study are explained in this section.

3.5.1 Naïve bayes

The algorithm is named after the English mathematician Thomas Bayes. Bayesian algorithms are among the statistical classification methods and are based on the statistical Bayes theorem [36].

Naïve Bayes works with discrete data instead of continuous data. Therefore, continuous dependent or independent variables are transformed into categorical data. For example, the numerical independent variable age should be made discrete such as "16-25", "25-44", "45-65", "65+" [36].

Let y = { $y_1, y_2, y_3, \dots, y_n$ } be the sample set and $C_1, C_2, C_3, \dots, C_n$ be the class set. The probability calculation of the set to be subjected to the classification process is calculated as shown in equation (5) [36].

$$P\left(y|C_i\right) = \frac{P\left(y|C_i\right)P\left(C_i\right)}{P(y)} \tag{5}$$

3.5.2 Support vector machine

Support vector machines are a classification technique first used by Joachims for text classification [37]. SVM is a powerful supervised learning instance based on the principle of

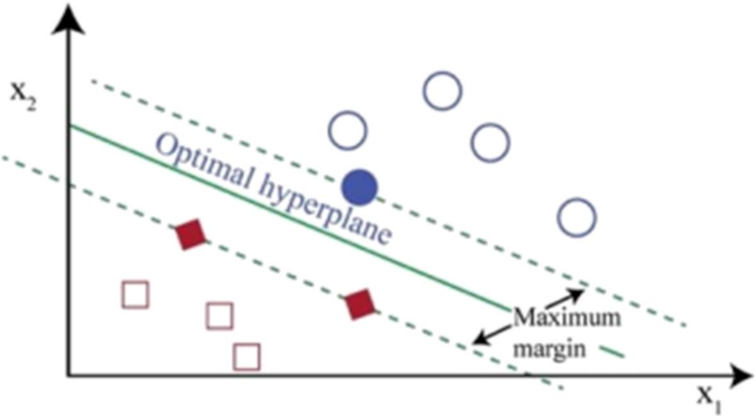


Fig. 4 SVM architecture [38]

least structural risk. During training, the algorithm creates a hyperplane to separate positive and negative examples. It classifies new examples by specifying where each example should be placed in the hyperplane. Figure 4 shows how the SVM architecture works [38].

SVM performs both linear and non-linear classification. Non-linear classification is implemented using the kernel function. In non-linear classification, kernels are defined as homogeneous polynomial, complex polynomial, Gaussian radial basis function and hyperbolic tangent function [38].

3.5.3 Logistic regression

Logistic Regression is a statistical technique used to analyse a dataset containing one or more independent variables that determine an outcome. The outcome is measured by a binary variable.

The purpose of logistic regression is to find the most appropriate model to explain the relationship between two directional characteristics (the dependent variable) and a set of related independent variables. Logistic regression produces the coefficients (and their standard errors and significance levels) of a formula for estimating the probability of the existence of the characteristics of interest by the logit transformation in equation (6) [37].

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots b_kXk_0 \tag{6}$$

In Equation (7), p is the probability of finding the feature [38].

$$rate = \frac{p}{1-p} = \frac{\text{the probability of finding the feature}}{\text{the probability that the feature is absent}} \tag{7}$$

The p value found in Equation (7) is logit transformed in Equation (8).

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{8}$$

In logistic Regression, instead of choosing parameters that minimise the sum of square root errors, parameters are chosen that maximise the probability of observing sample values [38]

3.5.4 Decision tree

It is a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes according to the feature and target. The decision tree algorithm is developed by dividing the dataset into smaller or even smaller parts. A decision node can have one or more branches. The first node is called the root node. A decision tree can consist of both categorical and numerical data [39]. A Decision Tree is a supervised algorithm that uses a set of rules to make decisions, similar to the way humans make decisions [40].

3.5.5 K-Nearest neighbor

K-Nearest Neighbour is one of the easiest supervised learning algorithms to implement. Although it is used to solve both classification and regression problems, it is most commonly used in industry to solve classification problems [41].

KNN algorithms have been proposed by Cover and Hart. The algorithm is used by using data from a sample set with known classes. The distance of the new data to be included in the sample set is calculated according to the existing data and the number of k nearest neighbours is checked. In general, 3 types of distance functions are used for distance calculations [38]:

- "Euclidean" Distance
- "Manhattan" Distance
- "Minkowski" Distance

3.6 Deep learning

Deep learning originated in the 1940s with the use of multiple hidden layers of neural networks. It is a machine learning algorithm that includes the feature selection process. It has regained prominence as hardware and data resources have increased. While neural networks are now relatively easy to train, they were almost impossible to train when they first appeared [42]. Deep learning emerged as an attempt to mathematically express the neurons in the human brain. It is also used in sentiment analysis, as it is used in many fields. In this study, LSTM and RNN deep learning algorithms were used.

3.6.1 Recurrent neural network

The Recurrent neural network, also called Elman network [43]. RNN is a type of artificial neural network that utilises sequential or time series data. These deep learning algorithms are widely used for sequential or temporal problems such as language translation, natural language processing, speech recognition, and image captioning. RNNs use the training dataset to learn. They are distinguished by "memory" as they take information from previous inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural

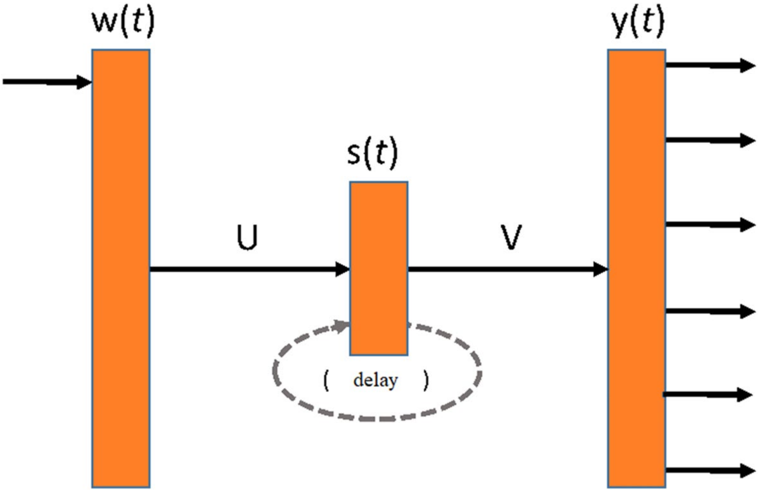


Fig. 5 RNN architecture

networks depends on the previous elements in the sequence. A simple RNN architecture is given in Fig. 5 [44].

The RNN network is trained using back propagation. It includes input, hidden and output layers. The value calculations in these layers are shown in equations 9, 10 and 11.

$$x(t) = \left[w(t)^T s(t-1)^T\right]^T \tag{9}$$

$$s_j(t) = f\left(\sum\nolimits_i x_i(t) u_{ji}\right) \tag{10}$$

$$y_k(t) = g\left(\sum\nolimits_j s_j(t) v_{kj}\right) \tag{11}$$

The vector $x(t)$= is formed by concatenating the vector $w(t)$, which represents the current word when using 1 of N encodings, and the vector $s(t-1)$, which represents the output values in the hidden layer from the previous time step. $s(t)$ represents the hidden layer and $y(t)$ the output vector [45].

3.6.2 Long short term memory network

RNN architectures have an approach based on the use of prior knowledge. LSTM networks are no different from RNN networks. LSTMs were developed by Hochreiter and Schmidhuber in 1997 due to the need to estimate contextual gaps, known as the drawback of RNNs [46]. LSTM networks use a structure to compute hidden states. LSTMs contain memory cells. These memory cells are the cells that hold the input information with the previous state. These cells in the network architecture decide which data to keep and which data to delete. In the next step, they combine the previous state with the current memory and the input data. Such an approach eliminates long-term dependencies, making it possible to preserve datasets. The LSTM cell contains these three gates [46]:

- The input gate; it controls the flow of input activations to the memory cell.
- The output gate; it controls the output flow of cell activations.
- The forget gate; it filters the information in the input and previous output and decides what to remember or forget. In addition to the three gates, the LSTM cell contains a cell update, the tanh layer, which is usually part of the cell state.

Each LSTM cell contains three variables:

- X_t, current input,
- h_{t-1}, previous output,
- C_{t-1}, previous cell state On the other hand, two variables come out of the cell:
- h_t current output
- C_t current cell state

The structure of the LSTM is shown in Fig. 6 [47].

4 Experimental results

In the experiments, two datasets were split using a holdout (80%-20% and 70%-30%) test training separation. The experiments are implemented in Colaboratory (Colab Pro) [48].

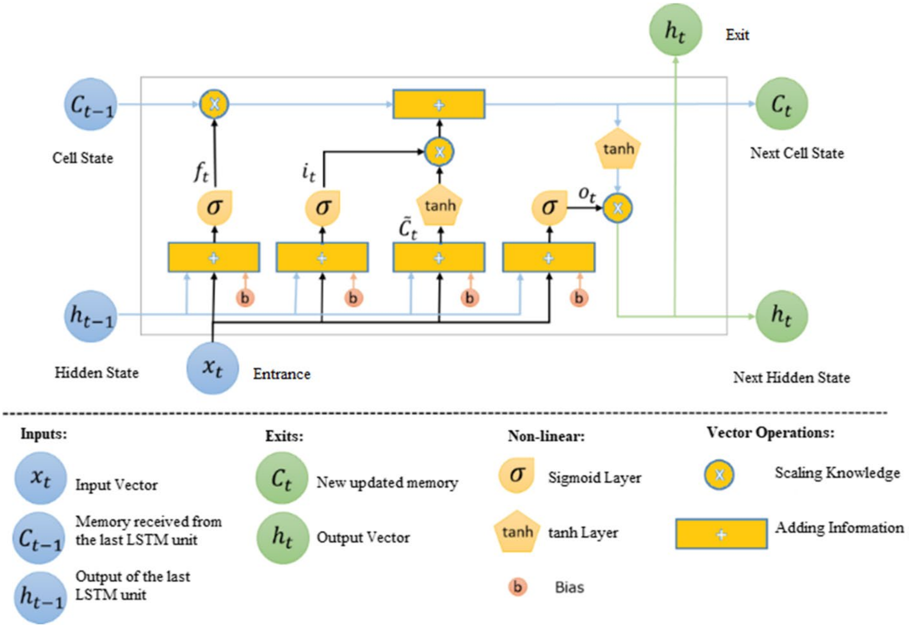


Fig. 6 LSTM architecture

4.1 Evaluation measure

The performance of the sentiment analysis models was evaluated using the accuracy criterion. This criterion is obtained by the ratio of true negatives (TN) and true positives (TP) to the total number of samples in equation (12) [45].

$$ACC = \frac{\mathrm{T_P} + \mathrm{T_N}}{\mathrm{T_P} + \mathrm{T_N} + \mathrm{F_P} + \mathrm{F_N}} \tag{12}$$

The deep learning parameters are given in Table 3.

Experiments on the model were performed on Google Colaboratory (Colab Pro) [48] using the TensorFlow 2.9.0 [49] and Keras 2.9 [50] libraries and Python version 3.9.13 [51] .

Coronavirus and TripAdvisor datasets, word representation methods (TF-IDF) and word embedding method (word2Vec), followed by DT, KNN LR, NB, SVM machine learning and holdout with Majority Voting and Stacking ensemble learning, and RNN, LSTM deep learning classifiers Sentiment analysis models were built by making train-test separations (70%-30% and 80%-20%). Table 4, Table 5 show the results obtained on the TripAdvisor dataset, Table 6, Table 7 on the Coronavirus dataset.

The background colour of the classification results of single machine learning is green, the background colour of the classification results of single deep learning is orange, the background colour of the classification results of heterogeneous use of machine learning algorithms in ensemble learning is grey, and the background colour of the classification results of heterogeneous use of deep learning algorithms in ensemble learning is blue in all tables.

According to Table 4, in the models created with TripAdvisor after TF-IDF, Stacking (LSTM-RNN) gives the best results in classification models, while Majority Voting (LSTM-RNN) comes second with very close results. This is followed by the deep learning models LSTM and RNN. The single machine learning models, Majority Voting (DT, KNN, LR, NB, SVM), lag behind Stacking (DT, KNN, LR, NB, SVM). While the single machine learning models lagged behind the single deep learning and ensemble models, SVM performed best alone.

Table 4 shows that prior to TripAdvisor's 70%-30% separation of training and testing, classification models were built using word2Vec word embedding models with two types of CBOW and Skip-gram.

Table 3 Deep Learning Parameters

Parameter	Value / Setting
Train size	70%, 80%
Test size	30%, 20%
Epochs	30
Activations	Tanh
Optimizer	Adam
Loss	Binary Cross Entropy
Metrics	Accuracy
Batch size	256
Verbose	1

Table 4 Accuracy Performance of Classification Algorithms on the TripAdvisor Dataset (70:30 training and test set)

Text Representation Methods	TF-IDF	word2Vec			
		CBOW		Skip-gram	
Classifier Models		100	200	100	200
LR	0.841	0.858	0.859	0.851	0.857
SVM	0.850	0.861	0.869	0.866	0.869
NB	0.833	0.855	0.862	0.861	0.862
KNN	0.821	0.843	0.846	0.847	0.831
DT	0.820	0.825	0.821	0.823	0.824
Stacking (NB, SVM, LR, DT)	0.858	0.858	0.868	0.866	0.867
Majority Voting (NB, SVM, LR, DT)	0.856	0.873	0.865	0.865	0.863
LSTM	0.872	0.874	0.876	0.872	0.872
RNN	0.868	0.868	0.872	0.869	0.870
Majority Voting (LSTM-RNN)	0.875	0.872	0.873	0.874	0.873
Stacking (LSTM-RNN)	**0.876**	**0.898**	**0.888**	**0.876**	**0.876**

- Among these models, CBOW's 200-dimensional vector size stacking (LSTM-RNN) classification models give the best results. This is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. Majority Voting (DT, KNN, LR, NB, SVM) ranks fifth in the heterogeneous use of single machine learning models, while Stacking (DT, KNN, LR, NB, SVM) ranks sixth. Among the single machine learning models, SVM produced the best results.
- These models lead to CBOW's 200-dimensional Vector Size Stacking (LSTM-RNN) classification models. This is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM).
- Among these models, it gives the 100-dimensional vector size stacking (LSTM-RNN) classification models of Skip-gram. This is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM).
- Among these models are Skip-gram 200-dimensional Vector Size Stacking (LSTM-RNN) classification models are given. This is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM).

Table 5 Accuracy Performance of Classification Algorithms on the TripAdvisor Dataset (80:20 training and test set)

Text Representation Methods	TF-IDF	word2Vec			
		CBOW		Skip-gram	
Classifier Models		100	200	100	200
LR	0.828	0.832	0.834	0.835	0.837
SVM	0.833	0.835	0.836	0.837	0.842
NB	0.830	0.833	0.835	0.843	0.849
KNN	0.824	0.826	0.830	0.826	0.821
DT	0.821	0.822	0.824	0.830	0.843
Stacking (NB, SVM, LR, DT)	0.836	0.841	0.838	0.839	0.847
Majority Voting (NB, SVM, LR, DT)	0.835	0.837	0.837	0.838	0.844
LSTM	0.852	0.854	0.857	0.857	0.854
RNN	0.851	0.852	0.855	0.852	0.853
Majority Voting (LSTM-RNN)	0.860	0.866	0.862	0.864	0.862
Stacking (LSTM-RNN)	**0.862**	**0.878**	**0.88**	**0.868**	**0.866**

Table 6 Accuracy Performance of Classification Algorithms on the Corona Virus Dataset (70:30 training and test set)

Text Representation Methods	TF-IDF	word2Vec			
		CBOW		Skip-gram	
Classifier Models		100	200	100	200
LR	0.722	0.760	0.768	0.753	0.761
SVM	0.762	0.774	0.777	0.773	0.772
NB	0.755	0.741	0.747	0.752	0.754
KNN	0.761	0.771	0.771	0.771	0.771
DT	0.691	0.668	0.671	0.679	0.673
Stacking (NB, SVM, LR, DT)	0.766	0.776	0.775	0.772	0.772
Majority Voting (NB, SVM, LR, DT)	0.756	0.777	0.778	0.778	0.773
LSTM	0.783	0.814	0.784	0.789	0.781
RNN	0.781	0.806	0.782	0.784	0.779
Majority Voting (LSTM-RNN)	0.787	0.819	0.788	0.793	0.795
Stacking (LSTM-RNN)	**0.788**	**0.827**	**0.841**	**0.832**	**0.842**

Table 7 Accuracy Performance of Classification Algorithms on the Corona Virus Dataset (80:20 training and test set)

Text Representation Methods	TF-IDF	word2Vec			
		CBOW		Skip-gram	
Classifier Models		100	200	100	200
LR	0.7760	0.779	0.777	0.782	0.784
SVM	0.7803	0.796	0.798	0.793	0.791
NB	0.7664	0.773	0.786	0.788	0.789
KNN	0.6924	0.704	0.706	0.716	0.717
DT	0.6947	0.718	0.710	0.701	0.710
Stacking (NB, SVM, LR, DT)	0.817	0.813	0.814	0.821	0.825
Majority Voting (NB, SVM, LR, DT)	0.813	0.825	0.824	0.824	0.828
LSTM	0.822	0.820	0.841	0.854	0.837
RNN	0.813	0.814	0.837	0.843	0.828
Majority Voting (LSTM-RNN)	0.809	0.815	0.838	0.857	0.836
Stacking (LSTM-RNN)	**0.834**	**0.842**	**0.853**	**0.864**	**0.743**

In general, Table 4 shows that heterogeneous deep ensemble models performed best for word representation and post-embedding classification models. This was followed by deep learning, while heterogeneous machine learning models outperformed single machine learning models.

According to Table 5, Stacking (LSTM-RNN) gave the best results in the models built acording to TripAdvisor's 80%-20% training and testing split, while Majority Voting (LSTM-RNN) came second after TF-IDF in the classification models. This is followed by the LSTM and RNN deep learning models. In the heterogeneous use of single machine learning models, Majority Voting (DT, KNN, LR, NB, SVM) lagged behind Stacking (DT, KNN, LR, NB, SVM), although the ensemble model was very close. The single machine learning models lagged behind the single deep learning and ensemble models, while SVM performed best among the single machine learning models.

According to Table 5, two types of word2Vec word embedding and two vector dimensional models were created after TripAdvisor's 80%-20% training and testing distinction.

- Among these models, CBOW's 100-dimensional vector size stacking (LSTM-RNN) classification model gives the best accuracy, while Majority Voting (LSTM-RNN) comes second. LSTM and RNN, the only deep learning model, ranked third and fourth respectively. Stacking (DT, KNN, LR, NB, SVM) ranks fifth in the heterogeneous use of single machine learning models, while Majority Voting (DT, KNN, LR, NB, SVM) ranks sixth. Among the single machine learning models, SVM produced the best results.

- These models CBOW's 200-dimensional vector size stacking (LSTM-RNN) classification model gives the best accuracy, while Majority Voting (LSTM-RNN) comes second. LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, Stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming Majority Voting (DT, KNN, LR, NB, SVM).
- Among these models, Skip-gram's 100-dimensional vector size stacking (LSTM-RNN) classification model gives the best accuracy, while Majority Voting (LSTM-RNN) comes second. LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, Stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming Majority Voting (DT, KNN, LR, NB, SVM).
- Among these models, Skip-gram's 200-dimensional vector size stacking (LSTM-RNN) classification model gives the best accuracy, while Majority Voting (LSTM-RNN) comes second. LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, Stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming Majority Voting (DT, KNN, LR, NB, SVM).

As shown in Table 5, Stacking (LSTM-RNN), which is the recommended method in all word representation and word embedding methods, performed better than other models in the 80%-20% training test separation of coronavirus. While Majority Voting (RNN-LSTM) and LSTM are closest to the proposed method, heterogeneous machine learning community models gave better results than single machine learning methods.

According to Table 6, Stacking (LSTM-RNN) gave the best results, while Majority Voting (LSTM-RNN) took second place in the models built with 70%-30% training test separation of coronavirus tweets after TF-IDF. Deep learning models LSTM and RNN ranked third and fourth respectively. Stacking of single machine learning models (DT, KNN, LR, NB, SVM) ranks fifth, ahead of majority voting (DT, KNN, LR, NB, SVM). Single machine learning models lag behind single deep learning and ensemble models. SVM is the best performing single machine learning model.

According to Table 6, post-embedding models of coronavirus tweets were built using CBOW and Skip-gram.

- Stacking (LSTM-RNN) gives the best results for the CBOW size 100 models. This is followed by majority voting (LSTM-RNN). LSTM and RNN, the only deep learning model, rank third and fourth respectively. Voting (DT, KNN, LR, NB, SVM) ranks fifth in the heterogeneous use of single machine learning models and outperforms stacking (DT, KNN, LR, NB, SVM). Among the single machine learning models, SVM produced the best results.
- Among the models in CBOW's 200-dimensions, Stacking (LSTM-RNN) performs best. This is followed by Majority Voting (LSTM-RNN). The only deep learning models, LSTM and RNN, rank third and fourth respectively. In the heterogeneous use of single machine learning models, Majority Voting (DT, KNN, LR, NB, SVM) ranks fifth, ahead of Stacking (DT, KNN, LR, NB, SVM). Among the single machine learning models, SVM produced the best results.

- Stacking (LSTM-RNN) classification models give the best results among the models in the 100-dimension of Skip-gram. This is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, rank third and fourth respectively. In the heterogeneous use of single machine learning models, Majority Voting (DT, KNN, LR, NB, SVM) ranks fifth, ahead of Stacking (DT, KNN, LR, NB, SVM). Among the single machine learning models, SVM gives the best results.
- It gives the most Stacking (LSTM-RNN) classification models among the models in the 200-dimension of Skip-gram. It is followed by Majority Voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, Majority Voting (DT, KNN, LR, NB, SVM) ranks fifth, ahead of Stacking (DT, KNN, LR, NB, SVM). Among the single machine learning models, SVM gave the best results.

As seen in Table 6, Stacking (LSTM-RNN), which is the recommended method in all word representation and word embedding methods, gave better results than other models in the 70%-30% training test separation of coronavirus. Although Majority Voting (RNN-LSTM) and LSTM are closest to the proposed method, heterogeneous machine learning models in ensemble models gave better results than single machine learning methods.

According to Table 7, Stacking (LSTM-RNN) gives the best results in the models built with 80%-20% training test separation of coronavirus tweets after TF-IDF, while Majority Vote (LSTM-RNN) takes the second place. These two models were followed by LSTM and RNN respectively in the deep learning models. In the heterogeneous use of single machine learning models, Stacking (DT, KNN, LR, NB, SVM) ranked fifth and sixth respectively, followed by Majority Voting (DT, KNN, LR, NB, SVM). SVM performed better among the single machine learning models.

According to Table 7, word embedding models were created using CBOW and Skip-gram.

- Stacking (LSTM-RNN) gave the best results in the 100-dimensions of CBOW. This is followed by Majority Vote (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. Voting (DT, KNN, LR, NB, SVM) ranks fifth in the heterogeneous use of single machine learning models. This is followed by stacking (DT, KNN, LR, NB, SVM). Among the single machine learning models, SVM gave the best results.
- CBOW's Size 200 Stacking (LSTM-RNN) gave the best results. This is followed by Majority Vote (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM).
- Stacking (LSTM-RNN) gave the best results at the 200 Skip-gram size. This is followed by majority voting (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM). Of the single machine learning models, SVM produced the best results.
- Among these models, the 200-dimensional stacking (LSTM-RNN) classification models of Skip-gram give the best results. This is followed by Majority Vote (LSTM-RNN). LSTM and RNN, the only deep learning model, ranked third and

Table 8 Coronavirus dataset studies in the literature

Authors	Model	Accuracy
Machuca et. al. [24]	LR	0.785
Siddiqua et. al. [25]	MV	0.877
Rahman and Islam [26]	Bagging	0.832
	Voting	0.833
	Stacking	0.835
Barkur et. al. [27]	LR	0.7429
Marga et. al. [28]	NB	0.84
Imran et. al. [29]	RNN	0.824
Buntoro et. al. [30]	SVM	0.8438
Vernikou et. al. [31]	LSTM	0.65
Antonio et. al. [32]	SGD	0.65
The Presented Model	Stacking (LSTM-RNN) Table 7	0.864

Table 9 TripAdvisor dataset studies in the literature

Yazar	Model	Accuracy
Mostafa. [16]	NB	0.85
Dehkharghani et. al	SVM	0.8751
Raut and Londhe	SVM	0.87
Tiwari et. al.	ME	0.8965
Zhou	BiLSTM -CNN	0.7373
The Presented Model	Stacking (LSTM-RNN) Table 4	0.898

fourth respectively. In the heterogeneous use of single machine learning models, stacking (DT, KNN, LR, NB, SVM) ranks fifth, outperforming majority voting (DT, KNN, LR, NB, SVM).

Overall, as can be seen in Table 7, Stacking (LSTM-RNN), the recommended method for all word representation and word embedding methods, gave better results than other models. While Majority Voting (RNN-LSTM) and LSTM are closest to the proposed method, heterogeneous machine learning models in ensemble models gave better results than single machine learning methods.

In the study, the model created by using LSTM-RNN together in the word2Vec stacking method gave good results on both the TripAdvisor and Coronavirus datasets. Table 8 shows the comparison of the Coronavirus dataset and Table 9 shows the comparison of the TripAdvisor dataset with the literature.

Table 8 shows that the best performing model is competitive with other studies on the coronavirus dataset in the literature.

Table 9 shows that the best performing model can compete with other studies in the literature on the TripAdvisor dataset.

When Table 8 and Table 9 are evaluated together, it can be seen that the combination of single machine learning algorithms in heterogeneous ensemble learning algorithms provides a significant advantage in terms of classification performance over single classifiers. Of all the representation methods, stacking after word2Vec gave the best performance of all the models.

5 Conclusion

In this study, sentiment analysis was performed on the coronavirus tweets posted on Twitter and the TripAdvisor hotel comments. In the classification process, a sensitivity analysis study was conducted using single machine learning and single deep learning, heterogeneous ensemble learning methods, and deep and machine learning methods. It was proposed to use single machine learning and single deep learning algorithms together heterogeneously in ensemble models. In the proposed system, datasets belonging to two different domains are enriched in terms of context and syntax by using different text representation techniques. Furthermore, it is shown that the combination of the ensemble learning approach with deep learning techniques provides a significant advantage in terms of classification performance and is presented as the recommended method. Word embedding (word2Vec) with word representation (TF-IDF) in Coronavirus and TripAdvisor datasets, followed by stacking of LR, SVM, NB, KNN, DT, machine learning and RNN, LSTM deep learning classifiers in single and heterogeneous ensemble learning methods and majority voting models for sentiment analysis were created. In both datasets, 70%-30% and 80%-20% train-test separations were performed randomly.

As can be seen from the accuracy results of the models obtained in Table 4, Table 5, TripAdvisor, Table 6, Table 7, Coronavirus datasets, the stacking (LSTM-RNN) ensemble classification method gave better results than the other models in the whole word representation and word embedding methods.

Ensemble models using heterogeneous combinations of deep learning methods and heterogeneous combinations of machine learning methods showed better accuracy performance than using these methods alone. In addition, the word2Vec word embedding method gave better results than TF-IDF in terms of accuracy performance in all models. As a result, the use of multiple methods as a whole, rather than a single machine and deep learning method, led to an increase in performance. Therefore, it is recommended to use models in combination with ensemble methods.

Future work will investigate the performance of different deep learning algorithms in ensembles with different word representations and pre-trained state-of-the-art word embedding methods. It has been observed that using single methods as an ensemble gives better results, which is consistent with the literature and represents a step forward in the detection of sentiment from posts. Investigating the performance of heterogeneous ensemble learning models based on different algorithms in sentiment analysis tasks is planned as future work.

Data availability Public datasets are used. https://www.kaggle.com/datatattle/covid-19-nlp-text-classification, https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

Code availability Custom Code

Declarations

Conflicts of interest/Competing interests No Conflicts of interest or Competing interests between authors

Consent to participate No information from human beings are used in the study

Consent for publication No information from human beings are used in the study

References

1. World Health Organization, “WHO statement regarding a cluster of pneumonia cases in Wuhan.” https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china
2. Puertas E, Moreno-Sandoval LG, Redondo J, Alvarado-Valencia JA, Pomares-Quimbaya A (2021) Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities. Cognit Comput 13(2):518–537. https://doi.org/10.1007/s12559-021-09818-9
3. Moreno-Sandoval LG, Puertas E, Pomares-Quimbaya A, Alvarado-Valencia JA (2020) Assembly of polarity, emotion and user statistics for detection of fake profiles Notebook for PAN at CLEF. [Online]
4. Gaur M, Gunaratna K, Bhatt S, Sheth A (2022) Knowledge-Infused Learning: A Sweet Spot in Neuro-Symbolic AI. IEEE Internet Comput 26(4):5–11. https://doi.org/10.1109/MIC.2022.3179759
5. Dang E, Hu Z, Li T (2022) Enhancing Collaborative Filtering Recommender with Prompt-Based Sentiment Analysis, [Online]. Available: http://arxiv.org/abs/2207.12883
6. Li C et al. (2021) SentiPrompt: Sentiment Knowledge Enhanced Prompt-Tuning for Aspect-Based Sentiment Analysis, [Online]. Available: http://arxiv.org/abs/2109.08306
7. Garg M, Gupta AK, Prasad R (2022) Graph Learning and Network Science for Natural Language Processing. CRC Press, Boca Raton. https://doi.org/10.1201/9781003272649
8. Liang B, Su H, Gui L, Cambria E, Xu R (2022) Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowl Based Syst 235:107643. https://doi.org/10.1016/j.knosys.2021.107643
9. Khurram Iqbal M, Abid K, Fuzail M, Ayubi SD, Aslam N Omicron Tweet Sentiment Analysis Using Ensemble Learning, https://doi.org/10.56979/402/2023
10. Shah S, Ghomeshi H, Vakaj E, Cooper E, Mohammad R (2023) An Ensemble-Learning-Based Technique for Bimodal Sentiment Analysis, Big Data and Cognitive Computing, vol. 7, no. 2, p. 85, doi: 10.3390/bdcc7020085
11. M. S. Başarslan ve F. Kayaalp, Sentiment analysis with ensemble and machine learning methods in multi-domain datasets, Turkish Journal of Engineering, c. 7, sayı. 2, ss. 141–148, Nis. 2023, https://doi.org/10.31127/tuje.1079698
12. Zolfaghari B, Mirsadeghi L, Bibak K, Kavousi K (2023) Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning. ACM Comput Surv 55(12):1–34. https://doi.org/10.1145/3580218
13. Mardikoraem M, Woldring D (2023) Protein Fitness Prediction Is Impacted by the Interplay of Language Models, Ensemble Learning, and Sampling Methods. Pharmaceutics 15(5):1337. https://doi.org/10.3390/pharmaceutics15051337
14. Md.H. Alam, Ryu W-J, Lee S (2016) Joint multi-grain topic sentiment: modeling semantic aspects for online reviews, Inf Sci (N Y), vol. 339, pp. 206–223, doi: 10.1016/j.ins.2016.01.013
15. Miglani A, Coronavirus tweets NLP- Text Classification. htttps://www.kaggle.com/datatattle/covid-19-nlp-textclassification
16. Mostafa L (2020) Machine Learning-Based Sentiment Analysis for Analyzing the Travelers Reviews on Egyptian Hotels, 405–413. doi: 10.1007/978-3-030-44289-7_38
17. Dehkharghani R, Yanikoglu B, Tapucu D, Saygin Y (2012) Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment Classification, in 2012 IEEE 12th International Conference on Data Mining Workshops, IEEE, 669–673. doi: https://doi.org/10.1109/ICDMW.2012.121
18. Raut VB, Londhe DD (2014) Opinion Mining and Summarization of Hotel Reviews, in 2014 International Conference on Computational Intelligence and Communication Networks, IEEE, 556–559. doi: https://doi.org/10.1109/CICN.2014.126
19. Tiwari P, Mishra BK, Kumar S, Kumar V (2020) Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis, in Cognitive Analytics, IGI Global, 689–701. doi: 10.4018/978-1-7998-2460-2.ch036
20. Zhou Y (2019) Sentiment Classification With Deep Nural Networks. Faculty of Information Technology and Communication Sciences, (Master’s thesis)
21. Sahu TP, Ahuja S (2016) Sentiment analysis of movie reviews: A study on feature selection & classification algorithms, in 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, 1–6. doi: https://doi.org/10.1109/MicroCom.2016.7522583

22. O. R. H., G. Virginia, and A. R. C. (2016) Sentiment Classification of Film Reviews Using IB1, in 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), IEEE, 78–82. doi: https://doi.org/10.1109/ISMS.2016.38
23. Mostafa L (2021) Egyptian Student Sentiment Analysis Using Word2vec During the Coronavirus (Covid-19) Pandemic, 195–203. doi: 10.1007/978-3-030-58669-0_18
24. Machuca CR, Gallardo C, Toasa RM (2021) Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach, J Phys Conf Ser, vol. 1828, no. 1, p. 012104, doi: 10.1088/1742-6596/1828/1/012104
25. Siddiqua UA, Ahsan T, Chy AN (2016) Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog, in 2016 19th International Conference on Computer and Information Technology (ICCIT), IEEE, 304–309. doi: https://doi.org/10.1109/ICCITECHN.2016.7860214
26. Md.M. Rahman, Islam MN (2022) Exploring the Performance of Ensemble Machine Learning Classifiers for Sentiment Analysis of COVID-19 Tweets, 383–396. doi: 10.1007/978-981-16-5157-1_30
27. Barkur G, Vibha, Kamath GB (2020) Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India, Asian J Psychiatr, vol. 51, p. 102089, doi: 10.1016/j.ajp.2020.102089
28. Isnain AR, Marga NS, Alita D (2021) Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm, IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 1, p. 55, doi: 10.22146/ijccs.60718
29. Imran AS, Daudpota SM, Kastrati Z, Batra R (2020) Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. IEEE Access 8:181074–181090. https://doi.org/10.1109/ACCESS.2020.3027350
30. G. A. B. and M. B. S. R. Mahendrajaya (2019) Gopay User Sentiment Analysis Using the Lexicon Based Method and Support Vector Machine (Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based dan Support Vector Machine), Komputek, vol. 3, no. 2, p. 52
31. Vernikou S, Lyras A, Kanavos A (2022) Multiclass sentiment analysis on COVID-19-related tweets using deep learning models. Neural Comput Appl 34(22):19615–19627. https://doi.org/10.1007/s00521-022-07650-2
32. Dwi Antonio V, Efendi S, Mawengkang H (2022) Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent, Int. J. Nonlinear Anal. Appl, vol. 13, no. 1, pp. 2008–6822, [Online]. Available: 10.22075/ijnaa.2021.5735
33. Mikolov JT, Sutskever I, Chen K, Corrado GS, Dean (2013) Distributed representations of words and phrases and their compositionality, in Advances in Neural Information Processing Systems. [Online]. Available: https://proceedings.neurips.cc/paper
34. Dogan G, Cay I, Ertas SS, Keskin ŞR, Alotaibi N, Sahin E (2020) Where are you?, in Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, New York, NY, USA: ACM, 301–304. doi: https://doi.org/10.1145/3410530.3414354
35. Mohammed A, Kora R (2023) A comprehensive review on ensemble deep learning: Opportunities and challenges. J King Saud Univ Comput Inform Sci 35(2):757–774. https://doi.org/10.1016/j.jksuci.2023.01.014
36. Ağbulut Ü, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renew Sustain Energy Rev 135:110114. https://doi.org/10.1016/j.rser.2020.110114
37. Bakay MS, Ağbulut Ü (2021) Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. J Clean Prod 285:125324. https://doi.org/10.1016/j.jclepro.2020.125324
38. Basarslan MS, Kayaalp F (2020) Sentiment Analysis with Machine Learning Methods on Social Media, ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, vol. 9, no. 3, pp. 5–15, doi: 10.14201/ADCAIJ202093515
39. Uzun E, Agun HV, Yerlikaya T (2012) Web content extraction by using decision tree learning, in 2012 20th Signal Processing and Communications Applications Conference (SIU), IEEE, Apr. 1–4. doi: https://doi.org/10.1109/SIU.2012.6204476
40. Başarslan MS, Kayaalp F (2021) Sentiment Analysis on Social Media Reviews Datasets with Deep Learning Approach, Sakarya University J Comput Inform Sci, doi: 10.35377/saucis.04.01.833026
41. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964
42. Young T, Hazarika D, Poria S, Cambria E (2018) Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Comput Intell Mag 13(3):55–75. https://doi.org/10.1109/MCI.2018.2840738

43. Karita S et al. (2019) A Comparative Study on Transformer vs RNN in Speech Applications, in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 449–456. doi: https://doi.org/10.1109/ASRU46091.2019.9003750
44. Yang H et al (2022) Deep learning neural networks for emotion classification from text: enhanced leaky rectified linear unit activation and weighted loss. Multimed Tools Appl 81(11):15439–15468. https://doi.org/10.1007/s11042-022-12629-1
45. Başarslan MS, Kayaalp F (2023) MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis, J Cloud Comput, 12, no. 1, p. 5, doi: 10.1186/s13677-022-00386-3
46. Huang Z, Yang F, Xu F, Song X, Tsui K-L (2019) Convolutional Gated Recurrent Unit–Recurrent Neural Network for State-of-Charge Estimation of Lithium-Ion Batteries. IEEE Access 7:93139–93149. https://doi.org/10.1109/ACCESS.2019.2928037
47. Canli H, Toklu S (2021) Deep Learning-Based Mobile Application Design for Smart Parking. IEEE Access 9:61171–61183. https://doi.org/10.1109/ACCESS.2021.3074887
48. Google LLC, “Colab.” https://colab.research.google.com/
49. Google Brain Team, “Tensorflow,” 2021. https://tensorflow.org
50. Keras, “Keras.” https://keras.io
51. Python, “Python.” https://www.python.org/downloads/

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.