Sentiment classification for employees reviews using regression vector-stochastic gradient descent classifier (RV-SGDC)

Babacar Gaye, Dezheng Zhang and Aziguli Wulamu

School of Computer and Communication Engineering, University of Science and Technology, Beijing, China

Submitted 17 May 2021
Accepted 22 August 2021
Published 23 September 2021

Corresponding author
Aziguli Wulamu, aziguli@ustb.edu.cn

Academic editor
Binh Nguyen

Additional Information and Declarations can be found on page 23

DOI 10.7717/peerj-cs.712

ABSTRACT

The satisfaction of employees is very important for any organization to make sufficient progress in production and to achieve its goals. Organizations try to keep their employees satisfied by making their policies according to employees' demands which help to create a good environment for the collective. For this reason, it is beneficial for organizations to perform staff satisfaction surveys to be analyzed, allowing them to gauge the levels of satisfaction among employees. Sentiment analysis is an approach that can assist in this regard as it categorizes sentiments of reviews into positive and negative results. In this study, we perform experiments for the world's big six companies and classify their employees' reviews based on their sentiments. For this, we proposed an approach using lexicon-based and machine learning based techniques. Firstly, we extracted the sentiments of employees from text reviews and labeled the dataset as positive and negative using TextBlob. Then we proposed a hybrid/voting model named Regression Vector-Stochastic Gradient Descent Classifier (RV-SGDC) for sentiment classification. RV-SGDC is a combination of logistic regression, support vector machines, and stochastic gradient descent. We combined these models under a majority voting criteria. We also used other machine learning models in the performance comparison of RV-SGDC. Further, three feature extraction techniques: term frequency-inverse document frequency (TF-IDF), bag of words, and global vectors are used to train learning models. We evaluated the performance of all models in terms of accuracy, precision, recall, and F1 score. The results revealed that RV-SGDC outperforms with a 0.97 accuracy score using the TF-IDF feature due to its hybrid architecture.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Mining and Machine Learning, Natural Language and Speech
Keywords Employees classification, Machine learning, Sentiment analysis, Hybrid model, Text classification

INTRODUCTION

In recent years, people have adopted new ways of communication and interaction with each other due to the development of information and communication technologies and the exponential growth of Word Wide Web applications. People communicate their experiences, opinions, interests, etc. on online platforms such as Facebook, Twitter, Reddit, and other web forums provided by organizations for employees. Employee online reviews

How to cite this article Gaye B, Zhang D, Wulamu A. 2021. Sentiment classification for employees reviews using regression vector-stochastic gradient descent classifier (RV-SGDC). *PeerJ Comput. Sci.* **7**:e712 http://doi.org/10.7717/peerj-cs.712

are marked as the most prevalent form of user-generated content which has the potential to convey critical and useful meanings to their readers along with the satisfaction of employees of the company's services. Other than social media platforms, many companies such as Samsung, Microsoft, Netflix, etc. provide a platform for their employees to express their personal opinions regarding the company's policies, working environment, and products. Considering the reviews and comment assists in decision-making processes for business owners to redesign policies or make changes to products, it is important to have the means to collect data effectively. Similarly, employee online reviews regarding their job experience are a major contribution in defining a company's reputation for its employee-centeredness positivity, and supportiveness. Being the voice of employees, mining the sentiment of these reviews offers new opportunities for company owners and managers to acquire valuable insights by the satisfaction and dissatisfaction of employees with their job (*Feng, 2020*). The main aim of this research is to help organizations to make good progress in production. Organizations try to keep their employees happy and make their policies according to employees' demands which help to create a good environment between the organization and employees. The proposed research is based on Sentiment Classification for Employees Reviews and categorized into positive or negative reviews so the organization can easily find the satisfaction or dissatisfaction of employees about their products.

Sentiment analysis (SAs) is a natural language processing (NLP) task which deals with the computation of people's emotions, opinions, and sentiments directed towards a specific entity (*Basiri et al., 2021*). Different studies (*Medhat, Hassan & Korashy, 2014*) used sentiments analysis to solve different problems. SAs aims to reveal the positive or negative sentiment in data examined and, in some cases, neutral sentiment can also be considered which was not the case in this study as we consider this study as a binary classification problem. There are several relevant methods in SAs that can be integrated in analyzing sentiments of employees' reviews, lexicon-based method, and supervised machine learning method (*Özçelik et al., 2021*). The machine learning method requires annotated data for training which is often difficult to acquire (*Huang et al., 2020*). Comparatively, the lexicon-based method utilizes a pre-defined dictionary, also known as, sentiment lexicon for exploring sentiments of data under analysis (*Khan et al., 2021b*). This study utilizes a dataset in which employees' reviews are unannotated, making the lexicon-based method more appropriate for the current analysis.

A sentiment lexicon is a database of sentiment bearing words with their corresponding polarity scores (*Mohamed, Moussa & Haggag, 2020*). A sentiment lexicon works by assigning respective polarity score ($P_{s}core$) to the sentiment bearing words (W_{n}) in a whole text, which is further aggregated into an overall polarity score (Overall $P_{s}core = \sum_{i=1}^{n} P_{(score_{n})}$). Each sentiment lexicon has a relative overall polarity score range($R_{(P_{s}core)} = [R_{+}, R_{-}]$) in which R_{+} corresponds to the positive sentiment and R_{-} corresponds to negative sentiment value. The resulting overall polarity score is then compared with a threshold value (th), explicitly set by the programmer, which discloses the nature of the text under analysis.

Algorithm 1 Algorithm to find sentiments

```
**Input:** Text under analysis ($W\_n$)
**Output:** Sentiment
**Assigning Polarity Score** ($P\_score$)
$W\_1 \Longleftarrow P\_score\_1$
$W\_2 \Longleftarrow P\_score\_2$
$W\_n \Longleftarrow P\_score\_n$
**Aggregating Polarity Score**
Overall $P\_score = \sum W\_1 + W\_2 + + W\_n$
**Comparing with Threshold Value**
**if**
 Overall $P\_score > 0$
return
 Sentiment (Positive)
**else-if**
 Overall $P\_score < 0$
return
 Sentiment (Negative)
```

The foundation of this study is that utilizing polarity-based sentiments of employees' online reviews as the ground truth for prediction of job satisfaction or dissatisfaction, produces more accurate results as compared to previous studies (*Rustam et al., 2021a*; *Rehan et al., 2021*). For this purpose, the current study incorporates preexisting SLs for exploiting the sentiment score of employees' reviews which are further classified by machine learning (ML) classifiers. In addition to this, a voting classifier is proposed and evaluated on employees' reviews dataset containing employee reviews from Facebook, Google, Microsoft, Apple, and Amazon (*Rehan, 2020*).

The main contributions for this study are listed below:

- TextBlob is utilized for extracting the sentiment score of the reviews which are further categorized as positive and negative sentiments based on the threshold value. These sentiments are incorporated as the ground truth for further analysis in this study.
- Seven supervised machine learning algorithms including logistic regression (LR), random forest (RF), AdaBoost Classifier (AC), Multi-layer Perceptron (MLP), Stochastic Gradient Descent Classifier (SGDC), Support Vector Classifier (SVC), and Extra Tree Classifier (ETC) are examined to evaluate their performance on a dataset containing employee reviews.
- A voting classifier referred to as Regression Vector Stochastic Gradient Descent Classifier (RV-SGDC) is devised to classify employees' reviews as positive and negative. RV-SGDC is based on LR, SVC, and SGDC under hard voting. The performance of the proposed voting classifier is compared with selected state-of-the-art ML models.

- The efficacy of three feature extraction techniques including term frequency-inverse document frequency (TF-IDF), bag of words (BoW), and global vectors for word representation (GloVe) is compared.
- Selected classification models along with a proposed voting classifier are evaluated and compared in terms of accuracy, precision, recall, and f1-score.
- Performance of results is analyzed on another data set Twitter US Airline Sentiment.
- Performance of proposed framework is analyzed against the previous study conducted concerning job satisfaction prediction which considered users' rating as the ground truth for carrying out a sentiment analysis of employees' reviews (*Rustam et al., 2021a*; *Rehan et al., 2021*).

The rest of the paper is structured as 'Related work' provides a deep insight into the previous researches carried out related to the current study. Additionally, this section describes the contribution of the present study to the literature. 'Materials and Methods' discusses the dataset in detail, data preparation techniques, feature extraction techniques, machine learning models, and evaluation metrics used in the proposed framework. 'Proposed Approach' describes the proposed architecture in detail. 'Experimental Results' explores the results of the proposed approach along with its comparison with state-of-the-art methods. This section also compares the current study with the previous study. 'Conclusion' concludes the study along with limitations and future work.

RELATED WORK

The performance of an employee is believed to be directly correlated with job satisfaction as it plays a significant role in the success and development of an organization (*Kuzey, 2012*). Job satisfaction is associated with a variety of variables such as higher productivity, lower turnover, and customer satisfaction. Regarding the platforms provided by organizations, employees communicate their personal opinion concerning the organization which depicts the satisfaction of employees with the organization. In recent years, researchers have more focused on exploiting employees' reviews using machine learning models to deduce useful insights for organizations. Previous researches concerning the exploitation of job satisfaction and dissatisfaction from employees' reviews are briefly described in Table 1.

Contribution to literature

Exploring the literature review in the previous section illustrates that recent approaches used for predicting job satisfaction from employees' reviews target the rate as the ground truth, whereas *Rehan et al. (2021)* concluded that overall ratings are not entirely correlated with corresponding textual reviews rendering misleading information for the organizations. The present study contributes to literature proposing an efficient system for evaluation of employees' reviews as satisfied and unsatisfied by incorporating textual features. The primary goal is to exploit the sentiment of employees' reviews using a sentiment lexicon which is further regarded as the target class or ground truth for the machine learning models to perform classification tasks. Another limitation in employee classification is the low accuracy score in previous studies which is a motivation for other researchers to

Table 1 Brief description of previous work related to present study.

Ref.	Summary
Moniz & de Jong (2014)	Forecasting of organizations' earnings is correlated with the satisfaction level of employees of an organization. As stated by the authors of this study as they identified one significant aspect of employees' reviews using Latent Dirichlet Allocation (LDA) which was associated with the organizations' outlook. Authors integrated the General Inquirer dictionary to explore sentiment of the reviews which was further utilized for exploring job satisfaction of employees when combined with the outlook aspect of the organization.
Costa & Veloso (2015)	Contributed to employee analytics by focusing on improving text representations by computing vector representations of fixed lengths. The authors implemented word2vec for extracting features of employees' reviews and used SVR and SVM as the machine learning classifiers for sentiment-based classification of reviews.
Jung & Suh (2019)	Derived 30 factors and their corresponding keywords, impacting the degree of job satisfaction, using LDA. Afterward, the authors conducted importance and sentiment analysis of these factors at various levels. They calculated the sentiment based on the frequency of nouns correlated with the description of job satisfaction factors. They also analyzed the most dominant factors impacting job satisfaction.
Bajpai et al. (2019)	Proposed ELM (extreme learning machine); an ensemble model which integrated SentiWordNet and SenticNet for exploiting sentiment of the reviews and machine learning models for classifying the satisfaction and dissatisfaction of employees. In addition to this, the authors also evaluated the proposed approach using SVM. Results showed that ELM obtained a 74.09% accuracy score whereas, SVM yielded an accuracy score of 74.85%.
Stamo lampros et al. (2019)	Explored the employee turnover and determinants of job satisfaction by incorporating. They exploited factors impacting job satisfaction from employees' reviews obtained from Glassdoor.com using an unsupervised approach of structural topic modeling such as LDA. In addition to this, the authors also explored the impact of the overall rating given by employees on the satisfaction of employees with the job. The study concluded that positive reviews are more correlated to the factors such as leadership, culture, career opportunities, etc. Contrarily, negative reviews correspond to topics such as communication with management, managerial behavior, etc.
Kashive, Khanna & Bharthi (2020)	Conducted sentiment analysis of employees' reviews along with text analysis using SAS analyzer tool. Authors carried out their research on employees' reviews from four sectors including info-tech, manufacturing, FMCG, and Pharmaceuticals. Sentiment analysis of reviews showed that factors concerning the job satisfaction level of an employee are social value, work-life value, interest value, economical value, brand value, development value, management value, and interest value.
Dina & Juniarta (2020)	Proposed aspect-based sentiment analysis of employees' reviews crawled from Glassdoor.com using Webharvy web-scrapper to exploit employees' satisfaction and dissatisfaction. Crawled reviews were then preprocessed to produce keywords using RapidMiner. Afterward, authors assigned the weights to each keyword using term frequency to which Stanford POS tagger was used to tag nouns only. Nouns were then classified based on their aspect as negative and positive.

(*continued on next page*)

Table 1 (*continued*)

Ref.	Summary
Rajendran (2020)	Proposed a four-fold framework for analysis of employees' reviews concerning an improvement in the performance of the delivery services industry. The proposed approach integrates employee reviews related to four delivery companies which are then preprocessed to eliminate unnecessary data and separated as positive, negative, and neutral based on the overall rating. Afterward, bigrams, trigrams are extracted which are then analyzed by SWOT technique. The study concluded that most of the negative reviews by the employee were related to health issues, quality planning of routes, etc.
Rustam et al. (2021a)	Integrated machine learning and deep learning models for classifying employee reviews into satisfied and unsatisfied. Authors referred to overall ratings (1-5) as the ground truth for carrying out classification tasks. For this purpose, they assigned rating >2.5 to satisfied class and rating <2.5 to unsatisfied class. Afterward, the preprocessed the reviews and extracted features using TF-IDF which were then utilized for the training of MLP. Evaluation of the proposed model yielded an accuracy score of 83%.
Rehan et al. (2021)	Exploited the correlation between reviews written by employees and overall ratings. The authors proposed the ERCE model which combined two modules with AND gate logic. Module-1 utilized an aggregate of all numeric ratings as the ground truth, whereas, module-2 utilized overall rating as the ground truth, both modules were trained and evaluated on the employees' review dataset. The study concluded that 76% of reviews and ratings correlate whereas, 24% of the reviews and ratings do not correlate. In addition to this, authors also classified employees' reviews as satisfied and unsatisfied which yielded 100% accuracy by module-1 and 79% accuracy by module-2.

work. Concerning all the presented works, we focused on the correlated target class and the improvement in the machine learning model's performance.

MATERIALS AND METHODS

The proposed framework is five-fold *i.e.*, pre-processing of the reviews to eliminate unnecessary data, sentiment labeling, feature extraction, training of prediction models, and evaluating the performance of trained predictors. These steps involved in carrying out job satisfaction prediction are briefly discussed in this section.

Dataset description

Employee reviews dataset contains reviews from Facebook, Microsoft, Netflix, Apple, Amazon, and Google employees and were obtained from Kaggle (*Rehan, 2020*). The dataset contains a total of 67,529 records and sixteen variables among which seven are numeric variables which include numeric and star ratings given by employees from 1 to 5, and four are text variables including summary, pros, cons, and advice to management by the reviewer employee. Rest variables correspond to information regarding the reviewer employee. These variables are discussed in Table 2. Current research work concatenates four text variables for carrying out job satisfaction prediction which is illustrated in Table 3.

To make a comparison another dataset is used for the experimental process. Another dataset is based on sentiments of US airlines names ”Twitter US Airline Sentiment” (*Figure Eight, 2019*). Twitter US Airline Sentiment dataset is based on 8473 records with 18 features.

Table 2 **Description of dataset variables.**

Variable	Description
Index	Index of record
Company	Name of company
Location	Company location
Date	Date on which review was written
Job-Title	Job title of employee
Summary	Summary of review written by employee
Pros	Benefits regarding employees of company
Cons	Drawbacks regarding employees of company
Advice to Management	Advice given by employee to management of company
Overall Rating	Overall rating given by employee from 1–5
Work/Life Balance Rating	Rating given by employee on balance between work and life 1–5
Culture and Values Rating	Rating given by employee on culture of the company 1–5
Career Opportunities Rating	Rating given by employee showing how many career opportunities does company provide 1–5
Senior Management Rating	Rating given by employee according to management of company 1–5
Helpful Review Count	Count of how many people found this review useful
Link to Review	Link which will redirect to the particular review

Table 3 **Sample of data utilized for research work.**

Summary	Pros	Cons	Advice to management	Review
good but i rreally dont kniow about that	i think its may be even better than any other ...	hmmmm i think all the downsides working at exc...	there is only one advice to the management tha...	good but i rreally dont kniow about that i think its may…
social media marketing	creating engaging social media strategies and...	monitoring the success of social media campaig...	assisting the company's business development t...	social media marketing creating engaging social media …

Data preprocessing

Preprocessing techniques are used to clean the data to increase the ML model's efficiency. Preprocessing of the data involves eradicating unreliable, noisy, and irrelevant data which affect the performance of machine learning models (*Kotsiantis, Kanellopoulos & Pintelas, 2006*). The current study incorporates preprocessing steps such as:

- Tokenization: Reviews are split into tokens of single words on the criteria of white-space.
- Lowercase Conversion: Each token was converted into lowercase as ML models are case sensitive.
- Spelling Correction: Misspelled words are corrected. The spell checker is used to check the misspelled words and convert them into correct words.
- Remove Numeric: Numeric values are removed such as contact number, value, or date which is not included in another review

- Removing unnecessary data: Username and stop-words do not add meaning to current analysis thus are removed.
- Remove Punctuation: Punctuations marks are removed from data, to increase the computation time. The punctuation signs like , #, $, %, &, etc., have been removed. Removal of these signs is done using regular expressions.
- Stemming: Tokens are converted into their root forms for reducing the complexity.

Preprocessing minimizes computation overhead and significantly influences predictive accuracy (*Crone, Lessmann & Stahlbock, 2006*). Data preprocessing is carried out using Python NLTK libraries.

Sentiment labeling

Sentiment refers to an individual's feelings, point of view, or opinion towards an entity. These opinions can be expressed by writing reviews or giving numeric or star ratings. Sentiments are broadly categorized as positive or negative based on the written review or rating (*Khan et al., 2021b*). The current study utilizes sentiment orientation of reviews written by employees' which is explored by sentiment lexicon namely TextBlob (*Rustam et al., 2021b*).

TextBlob

TextBlob is a well-known python library that provides sentiment orientation of the text under analysis (*Chaudhri, Saranya & Dubey, 2021a*). It assigns a P_score which is a float value within the range, $R_{(P_score)} = [+1.0, -1.0]$, in which $P_score = +1.0$ refers to positive polarity and $P_score = -1.0$ refers to negative polarity. In addition to this, TextBlob also provides a subjectivity score which quantifies the text under analysis as factual information or a personal opinion (*Chaudhri, Saranya & Dubey, 2021b*). This analysis is only concerned with the polarity score thus we did not integrate the subjectivity score. The threshold for TextBlob was set to 0, thus categorizing reviews with Overall $P_score > 0$ as positive reviews (satisfied) and Overall $P_score < 0$ as negative reviews (unsatisfied). Sample of polarity score assigned by TextBlob to employees' reviews is shown in Table 4.

Feature extraction

ML models require pre-defined features for training to carry out predictive tasks (*Peng, 2021*). In order to do so, features are extracted from data by using several feature extraction techniques. Feature extraction is a method that represents the text into a set of features relative to the current analysis. In this study, three feature extraction techniques such as TF-IDF, BoW, and GloVe are used to extract features from preprocessed employees' reviews. We select these feature extraction methods because of their importance in the literature review. These methods are mostly used by the research for text classification and have good results (*Rupapara et al., 2021*; *Khan et al., 2021b*).

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is an information retrieval algorithm that statistically measures the relevance of a term in a document (*Jiang et al., 2021*). It works by weighing a term t's frequency (tf) and its inverse document frequency (idf). Term frequency refers to the number of occurrences of a

Table 4 TextBlob score assigned to review text and their corresponding sentiments class.

Review text	Polarity score	Sentiment
pockets incompetence overall google…	−0.05	Negative
bad free food bad management…	−0.291	Negative

term in a document and inverse document frequency refers to the significance of that term in the whole dataset (D). In this study, we used the scikit-learn library, TfidfVectorizer, in which TF-IDF assigns weight to each term t in correspondence to its significance in the review R under analysis in the following manner:

$$tfidf(t,R,D) = tf(t,R) \times idf(t,D) \tag{1}$$

where,

$$tf(t,R) = log(1+freq(t,R)) \tag{2}$$

$$idf(t,D) = \frac{logN}{(df+1)}. \tag{3}$$

Bag of Words (BoW)

The BoW is a well-known feature extraction technique that represents text data disregarding word order and grammatical details (Sahin, 2021). It is widely used for NLP tasks, text classification, topic modeling, and information retrieval. This approach is easy to implement. Each instance in this feature extraction technique is tokenized to construct a vocabulary then, frequency of each token in the vocabulary is calculated (Boag et al., 2021). In simple words, BoW converts text of variable lengths into a vector of fixed length thus making it easy for ML models to work with data. For the implementation of BoW in our experiments, we integrated CountVectorizer which is a sci-kit-learn library in python.

Global Vector Representation of Words (GloVe)

GloVe was released by a group of NLP researchers from Stanford for vector representation of words in continuous space (Li, He & Chen, 2021). It maps words into relevant space by the distance between words and their semantic similarity. The GloVe is widely used for entity name recognition, machine translation, and many other NLP tasks. It mainly works by constructing a co-occurrence matrix M, in which M_ij represents the frequency of word i appearing in some context of word j, in which co-occurrence is calculated by moving a context size window over each sentence in the text. The study import GloVe embedding from 'zeugma' library to employ in our experiments.

Supervised machine learning models

We have incorporated eight machine learning models including LR, RF, AC, MLP, ETC, SVC, SGDC, and RV-SGDC in our experiments. These ML models are briefly described in this section. Best hyper-parameters were used for the optimization of ML models by the hit and trial method. During tuning each time we split the dataset and change the model's hyper-parameters values. These hyper-parameters settings are shown in Table 5.

Table 5 **Hyper-parameter settings for supervised machine learning models.**

Model	Hyper-parameter Settings
LR	random_state=100, multi_class='ovr', $C=3$
RF	n_estimators=100, random_state=50, max_depth=250
AC	n_estimators=300, random_state=50
MLP	random_state=20, max_iter=300
ETC	n_estimators=200, random_state=50, max_depth=150
SGDC	max_iter=1000, tol=1e−3
SVC	kernel='linear', $C=1.0$, random_state=500
RV-SGDC	LR, SVC, and SGDC, Voting=Hard

Logistic regression

LR is a statistical ML model which models a set of input features (X: input) into target variables (Y: output) by means of a sigmoid function which is an 'S' shaped curve and restricts Y in the range of 0 and 1 (De Cock et al., 2021). Sigmoid function $\sigma : R \longrightarrow (0,1)$ is defined as:

$$\sigma(X) = \frac{1}{1+e^{(-X)}} \tag{4}$$

where e is the base of the natural log. LR with its easy implementation produces efficient results in the case of binary classification (Rupapara et al., 2021). In our experiments, we have integrated various parameters in LR such as multi_class, random_state, and C. Multi_class is set to ovr which is the best choice for binary classification, random_state is set to 100, and C is set to 3 for strong regularization.

Random forest

RF is a meta estimator which utilizes the bagging method for learning patterns from data. It is an ensemble of decision trees that are built on numerous sub-samples of the dataset. It produces optimized prediction results by pooling outputs obtained from each decision tree in the ensemble (Chen et al., 2021). It controls over-fitting by adding additional randomness to the model, it does so by considering the most appropriate features from a random subset of features while splitting a node instead of utilizing the most important feature. One of the most significant tasks in the construction of a decision tree in RF is the selection of a significant attribute as a root node. For this purpose, two techniques including Gini impurity and Information Gain are used Baranauskas, Oshiro & Perez (2012). The current study involves Gini impurity as the criterion to quantify the split of a node. Gini impurity for a node can be calculated as:

$$\sum_{x=}^{N} f_x(1-f_x) \tag{5}$$

where, f_x is the number of occurrences of target attribute x at a node and N is the number of unique attributes. Several hyper-parameters are utilized for tuning of RF to acquire the best results. In our experiments, we set n_estimators, which defines the number of decision trees to be built in the ensemble, to 100. The randomness of samples bootstrapped in the construction of decision trees in the forest is set to 50, and the depth of each decision tree is set to 250 as shown in Table 5.

AdaBoost classifier

AC, also referred to as Adaptive Boosting, is an ensemble method that reassigns weights to each classified instance of the learner and higher weights are assigned to instances that had been classified incorrectly (*Wu et al., 2020*). AC constructs an ensemble of learners in a sequential manner in which each successive learner is constructed from preceding learners which results in less variance and bias of the models. It employs boosting for transforming a weak learner into a strong learner by concentrating on instances that are wrongly classified by a preceding weak learner. For a given dataset $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ where each instance x_i has a corresponding target variable $y_i\varepsilon\{-1,+1\}$, weak learners $\{k_1,k_2,\ldots,k_m\}$ are combined in the form:

$$C_{(t-1)}(x_n)=w_1k_1(x_n)+,,,+w_{(t-1)}k_{(t-1)}(x_n) \tag{6}$$

where, $w_1,,,w_{(t-1)}$ is the weight assigned to each predicted instance and $C_{(t-1)}(x_n)$ represents the target variable at $(t-1)th$ iteration which is further extended to construct a strong learner by adding another weak learner k_t with another weight w_t.

$$C_t(x_n)=C_{(t-1)}(x_n)+w_tC_t(x_n). \tag{7}$$

The hyper-parameter setting for AC used in our experiments is shown in Table 5. We set n_estimators to 300 and random_state to 50 to acquire optimized results.

Multilayer perceptron

MLP classifies an input vector $'x'$ by multiplying it with a weight 'w' and adding bias 'b' to it, such as, output:w ×x + b. MLP is restricted to the linear mapping of input and output variables, whereas, perceptron in MLP is capable of carrying out the classification of data which is not linearly separable (*Car et al., 2020*). MLP is an extended feed-forward neural network that maps a non-linear relationship ($f:R^m\longrightarrow R^n$) between the input layer of m dimensions and output vectors of n dimensions. Additionally, MLP consists of an arbitrary number of hidden layers in which neurons are trained by integrating the back-propagation technique. Computationally, a set of $\{n_i|n_1,n_2,\ldots,n_m\}$ input features from input layer are assigned a weight in the hidden layer such as: $w_1n_1+w_2n_2+,,,+w_mn_m$ in relation to a non-linear function (f: R → R). Current study utilizes rectified linear unit function (relu: f(x) =max(0,x)) as an activation function. The hyper-parameter setting for MLP in our experiments is shown in Table 5.

Extra tree classifier

ETC is an ensemble of unpruned decision trees that involves an arbitrary subset of features for splitting of nodes (*Bhati & Rai, 2020*). Unlike RF, it integrates whole data in the construction of a decision tree instead of bootstrapping data. It involves two main parameters such as the number of randomized input features selected at each node(K), the minimum size of sample required for splitting a node n_{min}, and the number of decision trees in the ensemble (M). Decision trees in ETC are less likely to be correlated due to the randomized selection of points of the split. ETC averages predictions of decision trees in the ensemble to produce a final prediction in case of regression (*Alsariera et al., 2020*). This study concerns binary classification thus pertaining to majority voting of decision trees' predictions to output a final prediction. There are several hyper-parameters involved in the tuning of ETC. In our experiments we have set n_estimators which states the number of trees in the ensemble to 200, random_state to 50, max_depth of each decision tree to 150, and Gini impurity as the criteria for node split.

Support vector classifier

SVC is a pattern-based classifier which utilizes linear kernel function to map input vectors x_i into a high dimensional vector space (*Rezaeinia et al., 2019*). It then creates a linear hyper-plane with the optimized margin between the target classes y_i. The linear kernel function integrates the following mathematical computation for pattern recognition.

$$K(x_i, y_i) = x_i' y_i. \tag{8}$$

Hyper-parameters concerning SVC in this study are shown in Table 5. Regularization parameter C is set to 1.0 for strong regularization and random_state, to shuffle the data for estimating probability, is set to 500 for optimized results.

Stochastic gradient descent

Gradient descent (GD) is a well-known optimization technique that learns the optimized values of models' parameters at each iteration to minimize the cost function (c^f) (*Rustam et al., 2019*). SGD is a variant of GD which concerns itself with random probability (stochastic) such that, at each iteration, a single sample is selected for the training of the model (*Tan et al., 2021*). It requires significantly less training time for finding c^f of only one training sample x^i at each iteration to reach local minima. It does so by updating the parameters of the model for each x^i and corresponding target class y^i.

$$\theta_j = \theta_j - \alpha(y^{i'} - y^i)x_j^i \tag{9}$$

where θ_j is the parameter and α is the learning rate of the model. SGD employs several hyper-parameters which supports its performance on data under analysis. In this study, loss function hinge was selected as default with $l2$ regularization. Maximum epochs were set to 1000 and the criterion of stopping iteration was set to $1e-3$ as shown in Table 5.

Regression Vector Stochastic Gradient Descent Classifier (RV-SGDC)

This study proposed an RV-SGDC model which is a combination of three individual models LR, SVC, and SGDC. The RV-SGDC outperforms all other models because of its ensemble architecture. We combined these models using majority voting criteria which mean that individual models will give their predictions for the target class and then the most predicted class by the models will be considered as the final target class. We combined LR, SVC, and SGDC because of their better performance individually as compare to other models and these models produce a more efficient and accurate model with the hard voting combination. We can describe RV-SGDC mathematically as:

$$LR_{prediction} = LR_{trained}(Data) \tag{10}$$

$$SVC_{prediction} = SVC_{trained}(Data) \tag{11}$$

$$SGDC_{prediction} = SGDC_{trained}(Data) \tag{12}$$

And

$$RV - SGDC_{prediction} = model\{LR_{prediction}, SVC_{prediction}, SGDC_{prediction}\}. \tag{13}$$

Here $LR_{prediction}, SVC_{prediction}, SGDC_{prediction}$ are the predictions by the individual models and $RV - SGDC_{prediction}$ is final prediction using majority voting between individual models' predictions. The architecture of proposed model is shown in Fig. 1.

PROPOSED APPROACH

The current study proposes a framework that is focused on mining sentiments of employees' reviews. The proposed framework involves an SL for sentiment labeling and ML models for classifying employees' reviews as positive and negative.

Employees' reviews are first acquired by concatenating four textual features such as summary (f_1), pros (f_2), cons (f_3), and advice to management (f_4) from the dataset into a single text feature ($f : f_1 + f_2 + f_3 + f_4$). Preprocessing techniques including tokenization, lowercase conversion, numeric removal, punctuation removal, stop-word removal, and stemming were employed to remove unnecessary data for low computation. For each preprocessed f, sentiment score s is assigned using TextBlob. Then, we set threshold value to 0 which referred $s > 0$ as a positive sentiment: f_+, and $s < 0$ as a negative sentiment: f_- resulting in 62,465 f_+ instances and 5,064 f_- instances from a total of 67,529 records. This causes an imbalance in data which might lead to ambiguous results. Therefore, to address the problem of imbalanced data, we included 5,064 records randomly from f_+ thus balancing both classes as shown in Table 6. For further process, 10128 records were used.

Afterward, features were extracted from f by using TF-IDF, BoW, and GloVe. ML models and proposed RV-SGDC are then trained on feature sets extracted from feature extraction techniques and labels extracted by TextBlob. In the last, the performance of trained predictive models is evaluated in terms of accuracy, precision, recall, and f1-score on the test set. The architecture of the proposed model is illustrated in Fig. 2.

Evaluation

Evaluation of models focuses on estimation of the performance of the model on unseen data.

Carrying out the classification of reviews into positive or negative produced four outcomes described below (*Rustam et al., 2020*):

- True positive (TP): Instances that are actually positive and also predicted positive.
- True Negative (TN): Instances belonging to the negative class which are correctly predicted as negative.

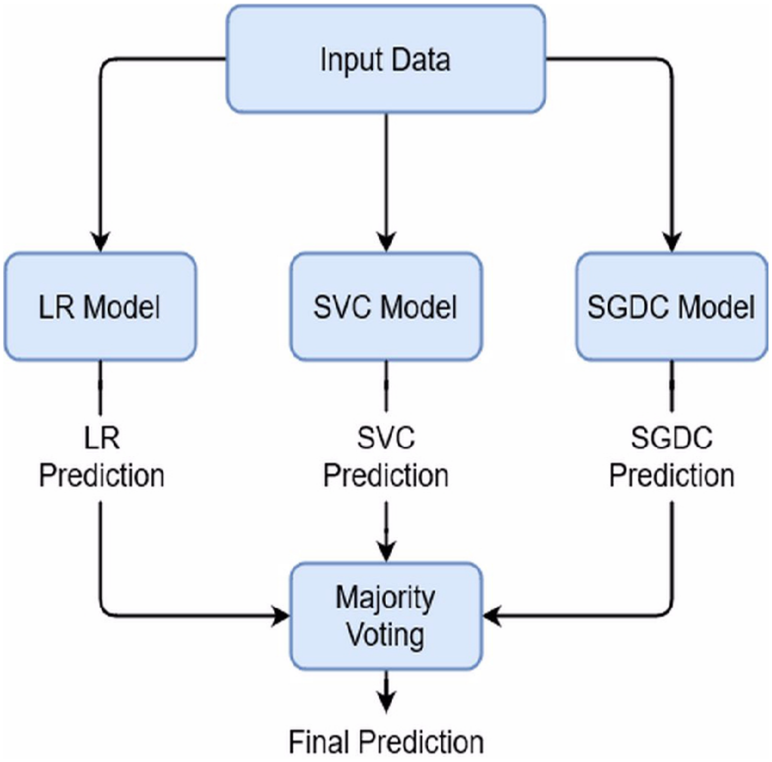


Figure 1 **Proposed model architecture.**

Full-size DOI: 10.7717/peerjcs.712/fig-1

Table 6 **Dataset count for experiments.**

Dataset	Positive	Negative	Total
Original dataset	62,465	5,064	67,529
Used dataset	5,064	5,064	10128

- False Positive (FP): Instances that belong to negative class but are predicted positive.
- False Negative (FN): Positive instances predicted as negative instances.

- Accuracy refers to the percentage of correctly predicted instances by ML model. It can be calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \tag{14}$$

- Precision refers to the percentage of predicted instances relevant to a certain class among all the instances. It can be calculated as:

$$Precision = \frac{TP}{(TP + FP)}. \tag{15}$$

- Recall is the fraction of relevant instances which are successfully predicted by ML model. It can be calculated as:

$$Recall = \frac{TP}{(TP + FN)}. \tag{16}$$

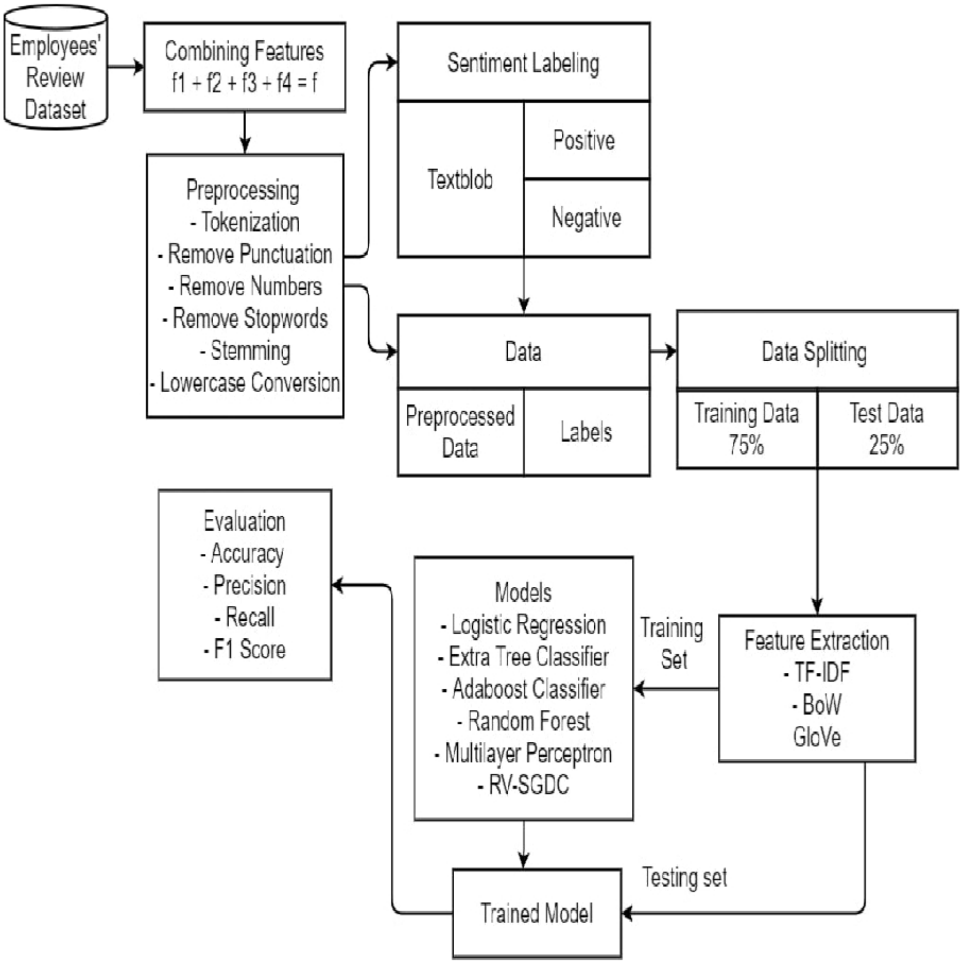


Figure 2 **Proposed methodology.**

Full-size DOI: 10.7717/peerjcs.712/fig-2

- F1-Score is the harmonic mean of recall and precision.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{17}$$

This study evaluates the proposed approach in terms of accuracy, precision, recall, and f1-score.

EXPERIMENTAL RESULTS

This section provides details of experimental results as well as a discussion of the results. Experiments are implemented in Python language. Experiments are conducted by integrating ML models such as LR, RF, ETC, AC, SGDC, SVC, MLP, and RV-SGDC for classifying employees' reviews into positive (satisfied) and negative (unsatisfied) classes. The efficacy of the proposed approach is tested over a 3:1 train-test set by using three different feature extraction techniques such as TF-IDF, BoW, and GloVe. The presented results of this study were achieved from sentiments, whereas the experiments were also done using the other rating attributes such as work balance, culture-values, carrier-opportunities.., as a target class but the achieved results are not as good or significant shows where the classification accuracy goes down because all these features contain the data of single departments so it is difficult to train the models whereas the review is about the whole company. Results obtained by integrating TF-IDF are shown in Table 7.

Experimental results of ML models using TF-IDF features demonstrate that LR, SVC, and SGDC perform well with a maximum accuracy score of 0.96. In terms of positive and negative classes, these three models have also yielded the highest precision, recall, and f1-score of 0.96 in comparison to other ML models. Except for SVC which yielded 0.95 precision in prediction of negative class and 0.95 recall in the prediction of the positive class. LR works best for problems on binary classifications due to its powerful working architecture involving sigmoid functions. In our experiments, LR worked well because instead of making presumptions about the categorization of positive and negative classes in vector space, it integrates maximum likelihood for the estimation of accurate results. On the other hand, SVC provides optimum generalizability to the proposed approach with minimization of structural risk along with exploiting a maximum boundary between target classes. Whereas, SGD with its capability of computing any sample at a time and frequently updating parameters assist in better learning of data under consideration and thus producing better results. Table 8 shows results of ML models obtained when trained and tested on features extracted by BoW.

Empirical results show that ML models have shown better performance with BoW features. MLP, SVC, SGDC, and AC have yielded maximum accuracy of 0.96 while ETC and RF acquired a 0.94 accuracy score on test data. Whereas, the performance of LR was not affected by changes in the feature set as can be observed by the results. TF-IDF produces weighted features which as a result reduces the feature size, on the other hand, BoW extract features regardless of their weight in the document thus creating a larger feature set. This is leveraged by MLP, since, being a neural network, it requires a larger feature set for better training. Therefore, producing better results with the BoW feature set when evaluated on the test set. Similarly, AC has shown outstanding performance as well in classifying data into positive and negative classes. Although RF and AC are both ensembles of decision trees the main difference between them is that RF assigns equal weight to decisions made by each decision tree, whereas, in AC the errors made by the preceding decision tree influence the weight-age given to the successive decision tree. This results in minimization of error and improvement in accuracy results of AC rather than RF. Consequently, ETC being a meta estimator adds randomization in choosing the split of nodes instead of selecting the optimized one. This causes low accuracy results as shown by the experiments in this study.

GloVe generates word embedding in a vector space by measuring the distance between words, but sentiment information of words is ignored by this feature extraction technique (*Rezaeinia et al., 2019*). For instance, it might create a co-occurrence between words good and bad which limits its performance in sentiment analysis. Results with GloVe features reveal a drop in classification accuracy of ML models in classifying employees' reviews as shown in Table 9.

Table 7 Accuracy, precision, recall and F1-score results of all ML models using TF-IDF.

Model	Accuracy	Negative class			Positive class		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	0.94	0.95	0.92	0.93	0.92	0.96	0.94
LR	0.96	0.96	0.97	0.96	0.97	0.96	0.96
ETC	0.94	0.95	0.92	0.93	0.92	0.95	0.94
MLP	0.95	0.95	0.95	0.95	0.95	0.95	0.95
AC	0.95	0.95	0.95	0.95	0.95	0.95	0.95
SVC	0.96	0.95	0.98	0.96	0.98	0.95	0.96
SGDC	0.96	0.96	0.97	0.96	0.96	0.96	0.96

Table 8 Accuracy, precision, recall and F1-score results of all ML models using BoW.

Model	Accuracy	Negative class			Positive class		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	0.94	0.96	0.92	0.94	0.92	0.96	0.94
LR	0.96	0.96	0.97	0.96	0.97	0.96	0.96
ETC	0.94	0.96	0.92	0.94	0.92	0.96	0.94
MLP	0.96	0.96	0.96	0.96	0.96	0.96	0.96
AC	0.96	0.95	0.96	0.96	0.96	0.95	0.96
SVC	0.96	0.96	0.96	0.96	0.96	0.96	0.96
SGDC	0.96	0.95	0.96	0.96	0.96	0.95	0.95

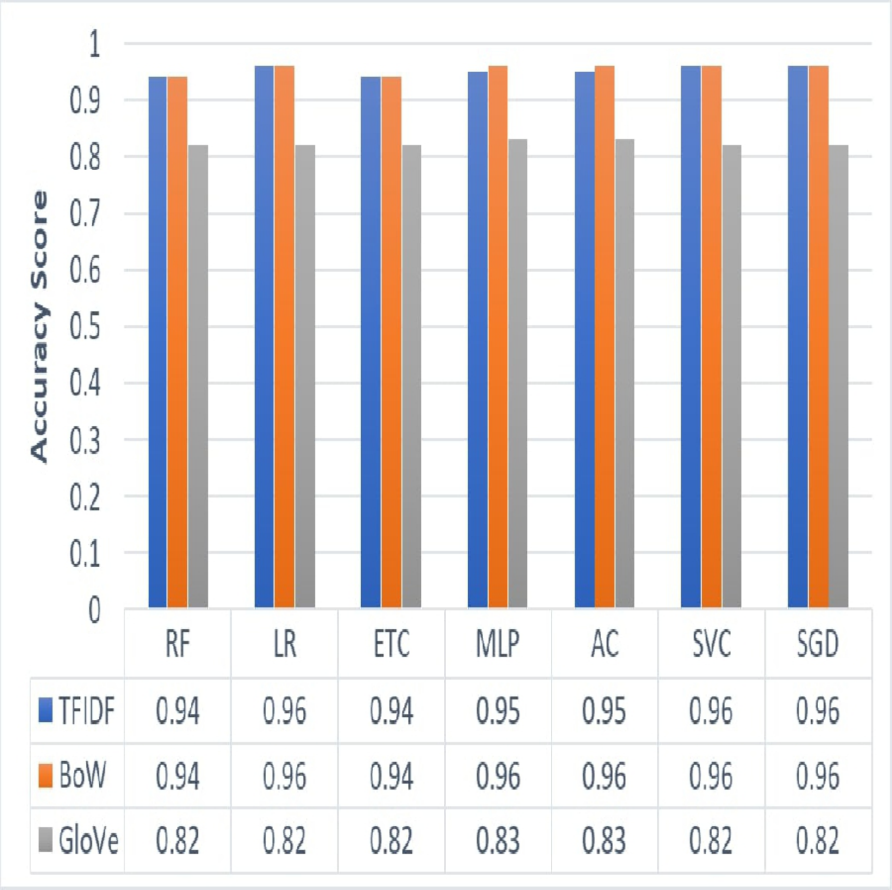
Performance comparison of ML models using TF-IDF, BoW and GloVe

We employed three feature extraction techniques to analyze the most appropriate technique for the classification of employees' reviews data under analysis. Figure 3 shows results yielded by the proposed approach using three different feature extraction techniques such as TF-IDF, BoW, and GloVe. We can see that the overall performance of BoW is observed to be best as compared to TF-IDF and GloVe. BoW tends to produce a feature set based on word count regardless of the grammar or structure of a sentence. TF-IDF produces weighted features and eliminates common words thus reducing feature set size. On the other hand, GloVe extracts features by generating a co-occurrence matrix which sometimes is not able to learn words that are out of vocabulary. This causes a reduction in the accuracy of classification results. In terms of ML models, LR with its effective and efficient architecture along with SVC and SGD has performed well with TF-IDF and BoW. Similarly, MLP and AC showed maximum accuracy with the BoW feature set. On the other hand, RF, ETC have shown poor performance in the case of all three feature sets of the employees' reviews data under consideration.

Figure 4 demonstrates the confusion matrix of ML models with maximum accuracy for classification of employees' reviews data as positive and negative. We can see that LR with BoW predicted a total of 2,433 correct instances among 2,532 total instances in which 1217 instances were correctly predicted from negative class and 1216 instances were correctly predicted from positive class. Whereas, it predicted 43 instances as negative which belonged

Table 9 Accuracy, precision, recall and F1-score results of all ML models using GloVe.

Model	Accuracy	Negative class			Positive class		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	0.82	0.81	0.84	0.82	0.83	0.80	0.82
LR	0.82	0.81	0.83	0.82	0.83	0.81	0.82
ETC	0.82	0.81	0.84	0.82	0.83	0.81	0.82
MLP	0.83	0.81	0.87	0.84	0.86	0.80	0.83
AC	0.81	0.82	0.81	0.81	0.81	0.82	0.81
SVC	0.82	0.82	0.84	0.83	0.83	0.81	0.82
SGDC	0.82	0.84	0.81	0.82	0.81	0.84	0.82



	RF	LR	ETC	MLP	AC	SVC	SGD
🟦 TFIDF	0.94	0.96	0.94	0.95	0.95	0.96	0.96
🟧 BoW	0.94	0.96	0.94	0.96	0.96	0.96	0.96
⬜ GloVe	0.82	0.82	0.82	0.83	0.83	0.82	0.82

Figure 3 Graphical representation of accuracy results yielded by ML models using TF-IDF, BoW and GloVe.

Full-size DOI: 10.7717/peerjcs.712/fig-3

to the positive class and 56 instances belonging to the negative class were predicted positive. There can be seen a minor difference in the performance of LR with TF-IDF as it predicted 2,439 correct instances among a total of 2,532 instances. It succeeds in predicting positive instances more accurately. As for MLP and AC, we can see that they lagged behind LR

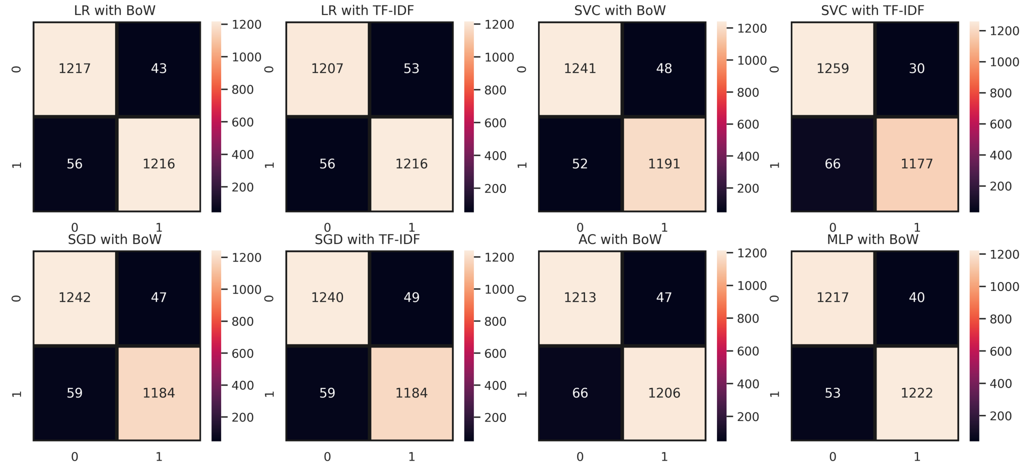


Figure 4 **Confusion matrices of ML models with maximum accuracy (0: Negative Class; 1: Positive Class).**

Full-size DOI: 10.7717/peerjcs.712/fig-4

in classifying employees' reviews correctly as positive or negative. MLP predicted 2,423 correct instances and AC predicted 2,419 correct instances which are somewhat less than instances correctly predicted by LR. In the case of SVC, it predicted a total of 2,436 correct predictions and 96 wrong predictions with TF-IDF whereas, with BoW it predicted 2,432 correct predictions and 100 instances were predicted wrong which demonstrates that SVC works comparatively better with TF-IDF. In the case of SGD, out of 2,532 instances, it predicted 2,426 correct instances, and 106 instances were predicted wrong in terms of BoW features. Conversely, SGD with TF-IDF carried out 2,424 correct predictions and 108 wrong predictions showing that it performed better with weighted features of TF-IDF. In summary, we can say that LR, SGD, and SVC with TF-IDF performed comparatively better in our proposed approach which is the base of our proposed voting classifier.

Performance analysis of proposed RV-SGDC

The proposed model RV-SGDC outperforms all other individual models in terms of all evaluation parameters. The study performed experiments in the same environment using RV-SGDC as we did with individual models and find the ensemble model is significant with all features. RV-SGDC results are shown in Table 10 and according to the results, the model achieved the highest accuracy of the study 97% with the TF-IDF features. This significant performance shows the efficiency of the proposed ensemble architecture and weighted TF-IDF features. RV-SGDC also achieved the highest F1 score on both negative and positive target classes with TF-IDF features.

Figure 5 shows the confusion matrix for the evaluation of RV-SGDC with TF-IDF, BoW, and GloVe features. According to the confusion matrices, RV-SGDC gave the lowest wrong predictions ratio of the study with TF-IDF features which are 81 out of 2532 and gave 2451 correct predictions. Performance with BoW feature was also good in comparison to GloVe which gave 103 wrong predictions out of 2532 predictions and 2429 correct predictions.

Table 10 Accuracy, precision, recall, and F-score results of proposed RV-SGDC model with TF-IDF, BoW and GloVe.

Features	Accuracy	Positive class			Negative class		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
TF-IDF	0.97	0.97	0.96	0.97	0.96	0.98	0.97
BoW	0.96	0.96	0.96	0.96	0.96	0.97	0.96
GloVe	0.84	0.83	0.82	0.83	0.83	0.84	0.83

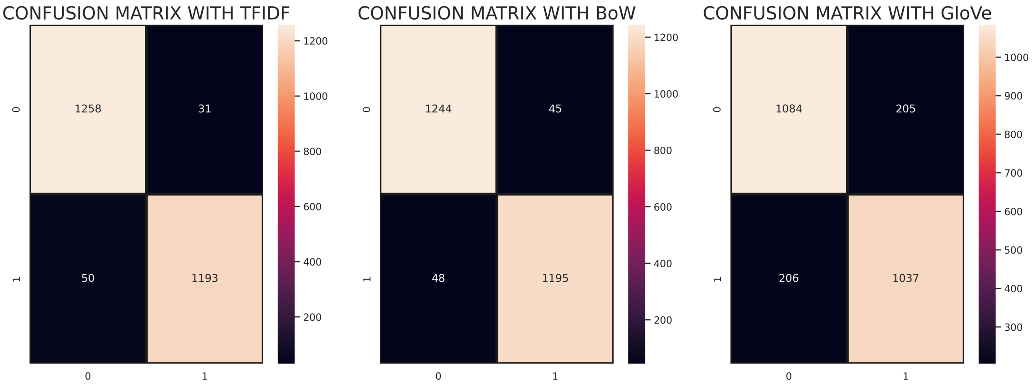


Figure 5 Confusion matrix of RV-SGDC with TF-IDF, BoW, and GloVe features.
Full-size DOI: 10.7717/peerjcs.712/fig-5

RV-SGDC also gave the highest correct prediction ratio using GloVe features compared to other individual models. RV-SGDC gave 2121 correct predictions out of 2531 using GloVe features which show the significance of proposed models on all features.
Further for performance analysis, we have experimented 15 times, each time we run the train and test split method to train and test models on different sets. The mean accuracy and standard deviation (SD) of each model after 15-time experiments are shown in Table 11. There is not too much fluctuation in the results with that method. Still, the RV-SGDC is on top with 0.96 mean accuracy 0.01, 0.02 SD with TF-IDF and BoW respectively. There is some fluctuation with GloVe features as RV-SGDC accuracy goes down 0.84 to 0.82 with 0.02 SD.

K-fold cross-validation is also applied to classify employee reviews and analyze the performance of RV-SGDC. For the experimental process, 10 fold were used. Experimental results of ML models with 10-fold cross-validation using TF-IDF features demonstrate that Rv-SGDC performs well with a maximum accuracy score of 0.96. Whereas LR, SVC, SDGC, and AC also perform well with 0.95 accuracies. The performance of all models shown in Table 12.

Performance comparison of models with another dataset

In this section, the results of all models on another dataset are presented to make a comparison. The US airlines tweets dataset is used to perform the experiments to show the significance of the proposed approach. TF-IDF feature extraction technique is used to

Table 11 Mean accuracy and SD after doing experiments 15 times on different training and testing sets.

Model	TF-IDF		BoW		GloVe	
	Accuracy	SD	Accuracy	SD	Accuracy	SD
RF	0.92	0.02	0.91	0.03	0.80	0.03
LR	0.95	0.01	0.93	0.02	0.82	0.02
ETC	0.93	0.03	0.92	0.05	0.81	0.04
MLP	0.93	0.02	0.94	0.03	0.80	0.02
AC	0.91	0.05	0.92	0.07	0.81	0.07
SVC	0.94	0.02	0.93	0.03	0.81	0.02
SGDC	0.94	0.01	0.95	0.02	0.82	0.03
RV-SGDC	0.96	0.01	0.96	0.02	0.82	0.02

Table 12 Ten fold cross-validation results with TF-IDF.

Model	Accuracy
RF	0.91 (+/- 0.11)
LR	0.95 (+/- 0.06)
ETC	0.91 (+/- 0.11)
MLP	0.94 (+/- 0.07)
AC	0.95 (+/- 0.06)
SVC	0.95 (+/- 0.06)
SGDC	0.95 (+/- 0.06)
RV-SGDC	0.96 (+/- 0.05)

Table 13 Accuracy, precision, recall and F1-score results of all ML models using TF-IDF.

Model	Accuracy	Negative class			Positive class		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	0.88	0.89	0.79	0.84	0.87	0.94	0.90
LR	0.90	0.91	0.84	0.87	0.89	0.94	0.92
ETC	0.88	0.91	0.79	0.85	0.87	0.94	0.91
MLP	0.89	0.87	0.85	0.86	0.90	0.91	0.90
AC	0.89	0.90	0.82	0.86	0.88	0.94	0.91
SVC	0.89	0.91	0.82	0.86	0.88	0.94	0.91
SGDC	0.89	0.90	0.84	0.87	0.89	0.94	0.92
RV-SGDC	0.90	0.91	0.83	0.87	0.89	0.94	0.92

perform analysis. Table 13 shows the performance results of all used models on the US airline tweets dataset.

Performance comparison of RV-SGDC with previous studies

The performance of the proposed framework was compared to previous studies carried out to classify sentiments of employees' reviews. The author proposed multilayer perceptron (MLP); a deep neural network; for classification of employees' reviews as satisfied (positive) and unsatisfied (negative) in *Rustam et al. (2021a)*. They integrate overall numeric rating (1-5) given by employees as the ground truth for training and testing of ML models. Reviews with an overall rating greater than or equal to 2.5 were considered as satisfied reviews and the rest were considered as unsatisfied reviews.

In *Rehan et al. (2021)* researcher's proposed a framework based on combining of two modules by AND gate in which averaging the numeric ratings were considered for classification of reviews as positive and negative in module 1 and module, 2, summary and overall ratings were used as features and classes respectively for classification tasks. Furthermore, the outputs from these modules are unified to classify the reviews as proper and improper. Since this study is concerned with classifying employees' review data into positive (satisfied) and negative (unsatisfied) therefore; we will compare our results with module 2.

The study uses polarity-based sentiments of employees. In this study, TextBlob is utilized for extracting the sentiment scores of the reviews which are further categorized as positive and negative sentiments based on the threshold value. Whereas *Rustam et al. (2021a)* used the overall rating of employees. In this research work to analyze the sentiments of TextBlob and overall rating, the dataset is manually assessed which shows that TextBlob produces more accurate results as compared to the overall rating because in overall rating if the rating is positive and text is negative then it is labeled as positive as well as if the text is positive and the rating is negative then it is also labeled as positive and same for the negative label. So, the dataset is analyzed manually to check the ground truth. Results in Table 14 show that the current study outperformed previous studies in text classification of employees' reviews as positive and negative. The proposed system performs best from previous studies by using RV-SGDC. RV-SGDC performs best because of its ensemble architecture.

Statistical analysis to show significance of RV-SGDC

This study performs a T-test to show the statistical significance of the proposed RV-SGDC model on other used state-of-the-art models (*Fatima et al., 2021*). T-test gives us output in terms of acceptance or rejection of the null hypothesis.

- Null Hypothesis: The proposed model is statistically significant as compared to other models.
- Alternative Hypothesis: The proposed model is not statistically significant as compared to other models.

RV-SGDC accepts the null hypothesis with TF-IDF features in comparison with other models which show that the model is statistically significant. While in the case of the BoW case rejects the null hypothesis and accepts the alternative hypothesis which means that RV-SGDC is not too statistically significant on BoW features. In the case of GloVe features, RV-SGDC shows statistical significance on other models by accepting null hypotheses.

CONCLUSION

This study experimented with sentiment classification using lexicon and machine learning based techniques. Analysis was performed on employee reviews datasets which contains

Table 14 Performance comparison of previous studies with current study.

Ref	Dataset	Features	Ground truth	Features	Model	Results
Proposed	Employees' Reviews	Summary	Overall Rating	TF-IDF	MLP	0.83
	Employees' Reviews	Summary	Overall Rating	TF-IDF	ETC	0.79
	Employees' Reviews	Summary, Pros, Cons, and Advice to Management	TextBlob Sentiments	TF-IDF	RV-SGDC	0.97

six well-known companies' employees' reviews to find the sentiments using text review. First, this study used preprocessing techniques to clean the text dataset, and then TextBlob was used to label the data with sentiments. TF-IDF, BoW, and GloVe features are used to train machine learning models. Performance of machine learning classifiers such as RF, LR, ETC, MLP, SVC, SGD, AC, and proposed RV-SGDC analyzed on these text features for the sentiment classification. Obtained results show that RV-SGDC attained a maximum accuracy score of 0.97 with TF-IDF features followed by similar precision, recall and f1-score whereas from BOW and GloVe RV-SGDC achieved 0.96 and 0.84 respectively. This significant performance of RV-SGDC shows the efficiency of the voting scheme between multiple models for final prediction. TF-IDF also shows its significance because of its approach to compute weighted features for machine learning models. Moreover, the proposed approach is compared with previous work conducted in the same direction. Results show that the proposed approach yielded state-of-the-art results as compared to previously done research on the same dataset concerning sentiment classification of employees' reviews. It can be observed that using TextBlob sentiments has highly impacted the accuracy of ML models individually as well as in a voting classifier as an improvement of $\tilde{0}$.14 accuracy score is observed with our experiments. In the end, we also show the statistical significance of the proposed model RV-SGDC on other models using the T-test which shows that the voting model is statistically significant as compared to individual models using TF-IDF features. However, the obtained results are not irrefutable as accuracy can be impacted by integrating a larger dataset which is the future direction of this work. One of the possible future directions can be the integration of deep learning models for improvement in accuracy results. We will consider ensemble deep learning models such as CNN-LSTM with the latest embedding schema which can be useful to improve the results.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is funded by the University of Science and Technology Beijing. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The University of Science and Technology Beijing.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Babacar Gaye conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Dezheng Zhang conceived and designed the experiments, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.
- Aziguli Wulamu conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The Supplemental File contains the experiment in Python using a machine learning approach.

The dataset is available at Kaggle: https://www.kaggle.com/saqlainrehan/employeesreviews-dataset.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.712#supplemental-information.

REFERENCES

Alsariera YA, Adeyemo VE, Balogun AO, Alazzawi AK. 2020. Ai meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE Access* **8**:142532–142542 DOI 10.1109/ACCESS.2020.3013699.

Bajpai R, Hazarika D, Singh K, Gorantla S, Cambria E, Zimmerman R. 2019. Aspect-sentiment embeddings for company profiling and employee opinion mining. ArXiv preprint. arXiv:1902.08342.

Baranauskas J, Oshiro T, Perez P. 2012. How many trees in a random forest?. In: *Machine Learning and Data Mining in Pattern Recognition*. Berlin Heidelberg: Springer, 154–168.

Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR. 2021. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems* **115**:279–294 DOI 10.1016/j.future.2020.08.005.

Bhati BS, Rai C. 2020. Ensemble based approach for intrusion detection using extra tree classifier. In: *Intelligent computing in engineering*. Singapore: Springer, 213–220.

Boag W, Kovaleva O, McCoy TH, Rumshisky A, Szolovits P, Perlis RH. 2021. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational Psychiatry* **11(1)**:1–6 DOI 10.1038/s41398-020-01158-w.

Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. 2020. Modeling the spread of COVID-19 infection using a multilayer perceptron. *Computational and Mathematical Methods in Medicine* **2020**:5714714 DOI 10.1155/2020/5714714.

Chaudhri AA, Saranya S, Dubey S. 2021a. Implementation paper on analyzing COVID-19 vaccines on twitter dataset using tweepy and text blob. *Annals of the Romanian Society for Cell Biology* **25**(**3**):8393–8396.

Chaudhri AA, Saranya S, Dubey S. 2021b. A survey on analyzing COVID-19 vaccines on twitter dataset using tweepy and text blob. *Annals of the Romanian Society for Cell Biology* **25**(**3**):8579–8581.

Chen Y, Zheng W, Li W, Huang Y. 2021. Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters* **144**:1–5 DOI 10.1016/j.patrec.2021.01.008.

Costa A, Veloso A. 2015. Employee analytics through sentiment analysis. In: *SBBD*, 101–112.

Crone SF, Lessmann S, Stahlbock R. 2006. The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research* **173**(**3**):781–800 DOI 10.1016/j.ejor.2005.07.023.

De Cock M, Dowsley R, Nascimento AC, Railsback D, Shen J, Todoki A. 2021. High performance logistic regression for privacy-preserving genome analysis. *BMC Medical Genomics* **14**(**1**):1–18 DOI 10.1186/s12920-020-00855-1.

Dina NZ, Juniarta N. 2020. Aspect based sentiment analysis of employees review experience. *Journal of Information Systems Engineering and Business Intelligence* **6**(**1**):79–88 DOI 10.20473/jisebi.6.1.79-88.

Fatima EB, Omar B, Abdelmajid EM, Rustam F, Mehmood A, Choi GS. 2021. Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: application to fraud detection. *IEEE Access* **9**:28101–28110 DOI 10.1109/ACCESS.2021.3056285.

Feng SF. 2020. Job Satisfaction, Management Sentiment, and Financial Performance: text Analysis with Job Reviews from Indeed. com. *Authorea Preprints*.

Figure Eight . 2019. Twitter US airline sentiment. Available at *https://www.kaggle.com/crowdflower/twitter-airline-sentiment*.

Huang M, Xie H, Rao Y, Feng J, Wang FL. 2020. Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Information Sciences* **520**:389–399 DOI 10.1016/j.ins.2020.02.026.

Jiang Z, Gao B, He Y, Han Y, Doyle P, Zhu Q. 2021. Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports. *Mathematical Problems in Engineering* **2021**:6619088 DOI 10.1155/2021/6619088.

Jung Y, Suh Y. 2019. Mining the voice of employees: a text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems* **123**:113074 DOI 10.1016/j.dss.2019.113074.

Kashive N, Khanna VT, Bharthi MN. 2020. Employer branding through crowdsourcing: understanding the sentiments of employees. *Journal of Indian Business Research* **12**(**1**):93–111.

Khan R, Rustam F, Kanwal K, Mehmood A, Choi GS. 2021b. US Based COVID-19 tweets sentiment analysis using textblob and supervised machine learning algorithms. In: *2021 international conference on artificial intelligence (ICAI)*. IEEE, 1–8.

Kotsiantis SB, Kanellopoulos D, Pintelas PE. 2006. Data preprocessing for supervised leaning. *International Journal of Computer Science* **1(2)**:111–117.

Kuzey C. 2012. Impact of health care employees' job satisfaction on organizational performance support vector machine approach. *European Journal of Economic & Political Studies* **5(1)**45–68.

Li D, He C, Chen M. 2021. Text sentiment analysis based on glove model and united network. In: *Journal of Physics: conference Series, vol. 1748*. IOP Publishing, 032046.

Medhat W, Hassan A, Korashy H. 2014. Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal* **5(4)**:1093–1113 DOI 10.1016/j.asej.2014.04.011.

Mohamed EH, Moussa ME, Haggag MH. 2020. An enhanced sentiment analysis framework based on pre-trained word embedding. *International Journal of Computational Intelligence and Applications* **19(04)**:2050031 DOI 10.1142/S1469026820500315.

Moniz A, De Jong F. 2014. Sentiment analysis and the impact of employee satisfaction on firm earnings. In: *European conference on information retrieval*, 519–527.

Özçelik M, Arıcan BN, Bakay Ö, Sarmış E, Ergelen Ö, Bayezit NG, Yıldız OT. 2021. HisNet: a polarity lexicon based on wordnet for emotion analysis. In: *Proceedings of the 11th Global Wordnet Conference*, 157–165.

Peng W. 2021. Big data mining and analysis based on convolutional fuzzy neural network. *Arabian Journal for Science and Engineering* 1–11 Epub ahead of print Apr 07 2021.

Rajendran S. 2020. Improving the performance of global courier & delivery services industry by analyzing the voice of customers and employees using text analytics. *International Journal of Logistics Research and Applications* 1–21 DOI 10.1080/13675567.2020.1769042.

Rehan MS. 2020. Employee reviews dataset. Available at *https://www.kaggle.com/saqlainrehan/employeesreviews-dataset*.

Rehan MS, Rustam F, Ullah S, Hussain S, Mehmood A, Choi GS. 2021. Employees reviews classification and evaluation (ERCE) model using supervised machine learning approaches. *Journal of Ambient Intelligence and Humanized Computing* 1–18 Epub ahead of print May 05 2021.

Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H. 2019. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications* **117**:139–147 DOI 10.1016/j.eswa.2018.08.044.

Rupapara V, Rustam F, Shahzad HF, Mehmood A, Ashraf I, Choi GS. 2021. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access* **9(1)**:78621–78634 DOI 10.1109/ACCESS.2021.3083638.

Rustam F, Ashraf I, Mehmood A, Ullah S, Choi GS. 2019. Tweets classification on the base of sentiments for US airline companies. *Entropy* **21(11)**:1078 DOI 10.3390/e21111078.

Rustam F, Ashraf I, Shafique R, Mehmood A, Ullah S, Sang Choi G. 2021a. Review prognosis system to predict employees job satisfaction using deep neural network. *Computational Intelligence* **37(2)**:924–950.

Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. 2021b. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE* **16(2)**:e0245909 DOI 10.1371/journal.pone.0245909.

Rustam F, Mehmood A, Ahmad M, Ullah S, Khan DM, Choi GS. 2020. Classification of shopify app user reviews using novel multi text features. *IEEE Access* **8**:30234–30244 DOI 10.1109/ACCESS.2020.2972632.

Sahin O. 2021. *Develop intelligent IOS apps with swift: understand texts, classify sentiments, and autodetect answers in text using NLP.* New York: Apress.

Stamo lampros P, Korfiatis N, Chalvatzis K., Buhalis D. 2019. Job satisfaction and employee turnover determinants in high contact services: insights from Employees Online reviews. *Tourism Management* **75**:130–147 DOI 10.1016/j.tourman.2019.04.030.

Tan Z, Chen J, Kang Q, Zhou M, Abusorrah A, Sedraoui K. 2021. Dynamic embedding projection-gated convolutional neural networks for text classification. *IEEE Transactions on Neural Networks and Learning Systems* 1–10 DOI 10.1109/TNNLS.2020.3036192.

Wu Y, Ke Y, Chen Z, Liang S, Zhao H, Hong H. 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena* **187**:104396 DOI 10.1016/j.catena.2019.104396.