Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00136

Research Article

Optimizing textual sentiment recognition through LASSO-based feature selection and ensemble voting technique

NISHA[1,*], Rakesh KUMAR[1]

[1]*Department of Computer Science & Applications, Kurukshetra University, Haryana, 136119, India*

ARTICLE INFO

Article history
Received: 10 September 2024
Revised: 04 November 2024
Accepted: 14 January 2025

Keywords:
Grid Search; LASSO; Multidomain Textual Sentiment Recognition; Preprocessing; Random Forest; Soft Voting; SVM

ABSTRACT

The nuances of opinion mining across varied datasets demands robust, generic models that can efficiently handle varied emotions in a text. This work handles the stated problem by proposing a novel ensemble-based model aimed at boosting both accuracy and interpretability in sentiment analysis tasks, which is crucial for applications such as customer feedback analysis, public opinion monitoring, and review systems. This article utilizes ensemble soft voting that uses Support Vector Machine and Naive Bayes as base classifiers, leveraging state-of-art feature selection approaches such as grid search optimized LASSO and Chi-square. The rationale of using these strategies due to their proven capacity of dealing high dimensional textual data with reducing the variability.The proposed method was independently evaluated using three publicly available datasets: Sentiment140, US Airlines, and Internet Movie Database, achieving accuracies of 81.75%, 93.25%, and 93.2% respectively. The results depict the proposed model adaptability with both balanced and imbalanced datasets and its strength to identify meaningful features, affirming consistent performance throughout. This work innovation is in fusing ensemble method with grid search-optimized LASSO for selecting the features, surpassing current individual and ensemble models. This work pave the groundwork for future progress, encompassing the extension to larger datasets and the integration of multiple emotion.

Cite this article as: Nisha, Kumar R. Optimizing textual sentiment recognition through LASSO-basedfeatureselectionandensemblevotingtechnique.SigmaJEngNatSci2025;43(6):1915–1929.

INTRODUCTION

Sentiment is the English word but originally rooted from the French word "santement, sentiment", which means "to express" or "to convey feelings". An integral part of language in human communication is sentiment [1]. When faced with a challenging scenario, our emotions often guide us to make important judgments. Artificial Intelligence systems need to consider human emotions. There are three domains where it has a major impact: decision-making, pattern recognition, and Human Computer Interaction (HCI). From the past decade, research on emotion has been flourishing exponentially. Far reaching availability of electronic devices, people largely engaged in spectrum of online activities nowadays such as social media, shopping, video games, online education, and related fields. The COVID-19 epidemic has significantly increased the frequency of these actions, which in turn has boosted their popularity

*Corresponding author.
*E-mail address: nisha.jrf.dcsa@kuk.ac.in

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

and acceptability [2]. Existing literature demonstrated that Machine learning (ML) and Deep learning (DL) techniques are highly feasible to build the similar system, but have substantial limitations such as more computational power is required to build DL model for emotion detection, at the same time ML classifiers can be less precise [3,4], [5]. Further the nature of data could be balanced or imbalanced adding further complexity in detecting emotions. There may be an imbalance in the data if one emotion is underrepresented in Textual Sentiment Classification (TSC) results compared to others. This disparity arises from a variety of causes, including gender and age differences, cultural backgrounds, linguistic diversity, challenges in correctly categorizing, the individual's inherent emotional distribution, data collection constraints, and other related issues [6]. Data preparation is the first stage in preparing the text for further processing. Due to the absence of a defined data model, the unstructured text is unfit for further processing. Therefore, certain data preparation or preprocessing techniques are required to reduce text size, eliminate noise, and find relevant patterns [7]. When dealing with textual data, it is necessary to employ text representations which transformed unstructured text into structured vectors, enabling ML models to analyzed them effeciently. BoW (Bag of Words), TF-IDF (Term Frequency–Inverse Document Frequency), Term Class Relevance (TCR), and phrase representation are commonly employed feature extraction techniques for textual data [8]. The aim of feature selection phase is to feed the classifier with a minimal relevant feature by eliminating noise and redunandant features and thus improving model's performance [9]. Significant scientific endeavors have been devoted to this field in previous years. Nevertheless, the problem of managing a huge feature set from multiple domains continues to provide significant challenges. LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge are regularization techniques but have inbuilt feature reduction capability. This research tackles this difficulty by utilizing a LASSO technique using a grid search. TSC has several use cases: including customer feedback analysis, social media monitoring [4], brand monitoring, product and service reviews, market research [3], political analysis, tourism [10], financial analysis, healthcare [11], government intelligence, E-learning, education, and most recent use case is scientometric analysis where the goal is to examine the writers' feelings as they relate to citations in scientific papers [12]. This research proposes a method for understanding textual sentiment using ensemble ML voting classifier. The reason for this is that ML models are simpler and more direct compared to DL models. The effects of different text representation and preprocessing techniques on English text with stemming and lemmatization, to preprocessing, and to regularization approaches such as LASSO and Ridge with inherent feature selection are explored in this work to check their efficacy in detecting sentiments. Here, the proposed Ensemble Soft Voting Classifier leverages the computational power of three distinct base classifiers, including Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB). This work introduces a grid search optimized feature selection strategy to tackle the short textual dataset with diverse domains. The high-dimensionality of the multidomain textual data results into the wide spectrum of vocabulary, thus demands a strong feature selection method. Employing Grid Search-based CV(Cross-Validation) feature selection using LASSO, this study handles multicollinearity problem in textual datasets by reducing dimensionality and feeding relevant features for a model.

The main contributions of this work are detailed below:

- We introduced a novel grid search optimized LASSO feature selection in conjuction with ensemble soft voting which is capable of handling high dimensional imbalanced textual data.
- The proposed framework is grouping of five unified steps: data preprocessing, feature extraction and selection, training a model, and thus evaluation. The use of multivariate feature selection LASSO in compare to univariate Chi-square making the system more robust and accurate.
- Proposed model leverages the ensemble soft voting for final prediction which constitute three heterogeneous base classifiers: SVM, RF, and NB. Grid search optimization is employed to automatically choose hyperparameters of these ML model.
- The proposed work is independently evaluated on publicly available three textual datasets, such as Sentiment140, US Airlines, and IMDB (Internet Movie Database) aiming to enhance the task of sentiment analysis (SA).
- The experimental evaluation reveals that our proposed ensemble framework in conjuction with LASSO feature selection approach attains better classification accuracies than without feature selection.

The subsequent sections of the paper are organized in the following manner: Section 2, presents the literature survey, which offers a comprehensive summary of the current research in the field. In Section 3, the suggested design is delineated, providing a comprehensive account of the features incorporated and the classifiers under consideration. The outcomes of conducted experiments are detailed and evaluated in Section 4 of this study, which also contains the experimental data. In the concluding section of this paper Section 5, summarize the main results and conduct an in-depth evaluation of the possible conclusions that may be taken from this study.

Literature Survey

ML algorithms have recently made significant improvements, which led to their widespread use cases in multiple fields. Previous studies shown that ML can robustly determine the toughness of Mode-I rock fractures using metaheuristic optimization algorithms [13], as well as optimize the compression of respiratory signals and predict

financial system profits using DL models [14,15]. These instances demonstrate how ML may be tailored to address intricate, industry-specific challenges. Taking advantage of this adaptability, we utilize ensemble techniques to tackle natural language processing (NLP) and classification-specific problems in SA of textual data. Using ML [16–18] and DL [19–22] for textual data classification and analysis, has become an important application area in the field of NLP. There have been a lot of research looking into different ways to make sentiment detection systems better and faster. The methods, datasets, and results of these works are summarized in Table 1, which offers a comparative overview of important contributions to SA. In recent years, academics have made significant advancements in the field of TSC by developing, refining, and comparing various feature extraction and selection methodologies, classification algorithms, and databases. Significant contributions have been made by researchers in the fields of ensemble learning [23–25] and feature selection [16,17,24]. However, there is a dearth of research undertaken on multidomain projects incorporating different domains with large datasets.

Table 1. Literature review summary

Reference	Dataset	Feature selection /extraction	Classifier	Drawbacks
[16]	Sentiment140 Movie review	POS	NB ME	Only 1000 tweets via random selection were utilized in the experiments. Despite using less no of instances; the model is not precise.
[17]	Rotten Tomatoes	N-gram TF-IDF	SVM NB ME	Potential difficulty in analyzing short Twitter comments.
[18]	IMDB	TF-IDF	SVM	The suggested hybrid feature selection technique does not exhibit enhanced classification performance on specific datasets, perhaps because of the limited availability of high-quality training data.
[19]	SST-1 Restaurant reviews	POS	CNN	There are potential complexities in integrating DL and rule-based methods which possibly affect the scalability and interpretability of the model.
[20]	US-airline, US-presidential election, IMDB, Car reviews	Word2vec Glove BERT	CBRNN	A too complex model with high computational power.
[21]	IMDB	---	Hybrid CBRNN	The scaling of the model to larger datasets or real-world applications may necessitate additional computational and resource resources, which are not taken into consideration.
[22]	Sentiment140 IMDB	Word2Vec BERT	LSTM Bi-GRU	Complex ensemble framework and adaptability issues due to constant legth preprocessing.
[23]	Twitter data	TF-IDF	Ensemble (LR-SGD)	A small single-domain dataset is used and the accuracy achieved is only 79%.
[24]	Tourism reviews	Rule based	Ensemble SVM, NB, & RF	The challenge of efficiently extracting valuable information from voluminous and noisy user-generated data.
[25]	Movie E-commerce	Uni-gram Bi-gram	SVM NB Bagging	The study focused on the Turkish language and a few examples, allowing for potential generalizability to other languages or cultures.
[26]	IMDB, Amazon movie	Word2vec	Hybrid CNN-LSTM	CNNs contain a multitude of convolutional layers to capture long-term dependencies making the model complex. Only one domain is captured during modeling.

Table 1. Literature review summary

Reference	Dataset	Feature selection /extraction	Classifier	Drawbacks
[27]	Twitter US Airline	TF-IDF	KNN, RF, MNB ANN, LSTM, Bi-LSTM	A single-domain dataset is used. Extensive hyperparameter tuning is required
[28]	IMDB SST2 MR	---	Bi-LSTM	While the study highlights the potential of simpler architectures, it lacks a comprehensive exploration of interpretability and robustness across diverse datasets, and a lack of robustness in handling imbalanced datasets.
[29]	IMDB Twitter US Airline SMS Spam Collection	TF-IDF BOW	LSTM	The efficacy of the model could be significantly enhanced through the incorporation of a variety of hyperparameter tuning with the base model, as its current accuracy fails to meet anticipated levels.
[30]	Sentiment140	---	BERT	This research was limited to a singular investigation into the effect of word elongation on sentiment classification, and encountered difficulties in assembling the dataset.
[31]	IMDB Twitter US Airline Sentiment140	GloVe	BERT LSTM	The reliance on GloVe pre-trained word embedding for data augmentation may limit the model's adaptability to domains with specialized vocabularies or dialects.
[32]	IMDB Twitter US Airline Sentiment140	GloVe	RoBERTa BiLSTM GRU	A possible limitation of this research pertains to the computational requirements and complexity associated with the execution and enhancement of the ensemble hybrid DL model.
[33]	Twitter dataset	Word2vec	CLAS (CNN-LSTM-Attention-SVM)	DL models used in this research need substantial computational resources for both training and inference processes, hence imposing limitations on their accessibility and scalability, particularly in contexts with limited resources.

PROPOSED METHODOLOGY

The processing unit is a critical element in TSC systems as it is tasked with extracting essential information from the incoming text. Subsequently, a ML classifier is used to determine the primary text sentence's expressed subjectivity. Figure 1 depicts the proposed methodology used in this work which makes the use of base classifiers as SVM and NB coupled with LASSO as feature selection. This section gives a synopsis of the classification methods used to build the proposed model, as well as the datasets used, feature extraction and selection, and model development.

Datasets

The proposed system's effectiveness and scalability are evaluated on various datasets, including three cutting-edge ones, with both positive and negative evaluations conducted, excluding neutral ones. The datasets for US airlines reviews [34], IMDB review, and Sentiment 140 [35] were obtained from Kaggle, IMDB review, and Twitter API. The US Airlines dataset, which contains 11,517 tweets about six US airlines, has an inherent imbalanced distribution. The IMDB review dataset, which contains 50,000 reviews, has a well-balanced distribution with 25,000 favorable and 25,000 unfavorable evaluations. The Sentiment140 dataset, which contains 160,000 tweets, has an imbalanced distribution with a mean word count below 115 words. The distribution of original samples from the datasets as shown in Figure 2.

Data Preprocessing

The aim of this phase is preprocessing the disorganized social media reviews for further classification. It goes over a lot of different methods, such as Normalization which involves a series of simultaneous activities, such as converting text to lowercase, removing URLs, and eliminating punctuation, hashtags, and whitespace, to improve the preparation process uniformity. Tokenization is the process of breaking down large text fragments into smaller tokens, dividing them into smaller units for efficient data analysis. We used Wordnet tokenizer for this study. Stemming and lemmatization, both destined to do same work. Stemming removes prefixes, suffixes, and definite articles from words,

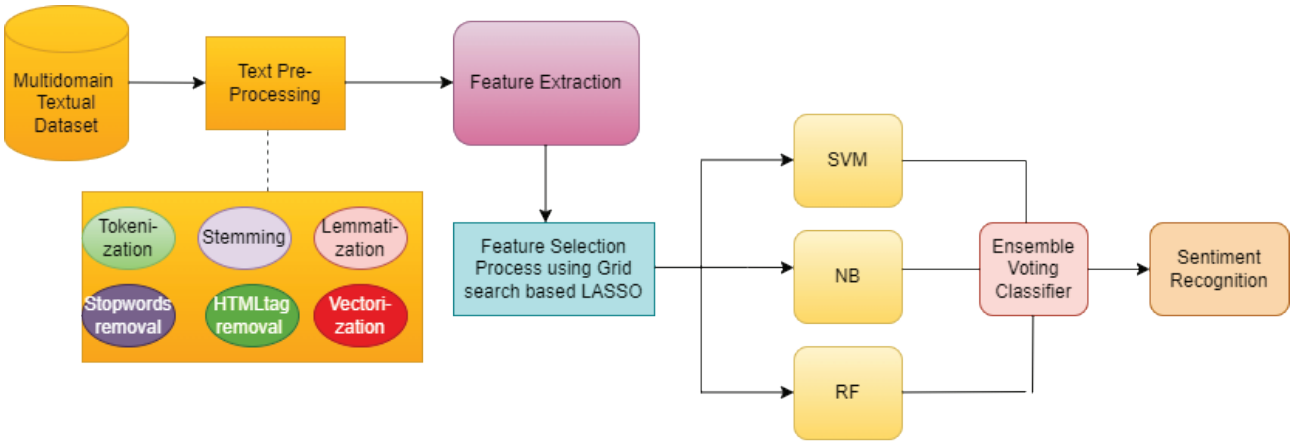

Figure 1. Architecture of the proposed ensemble model.

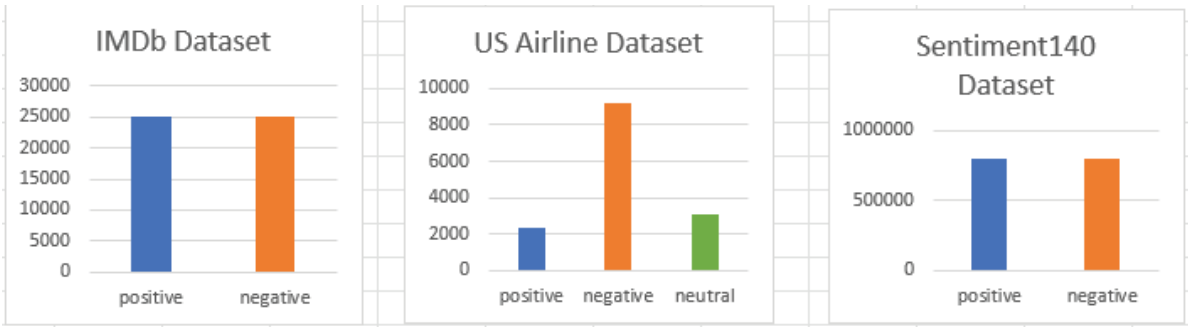

Figure 2. The distribution of samples from the datasets.

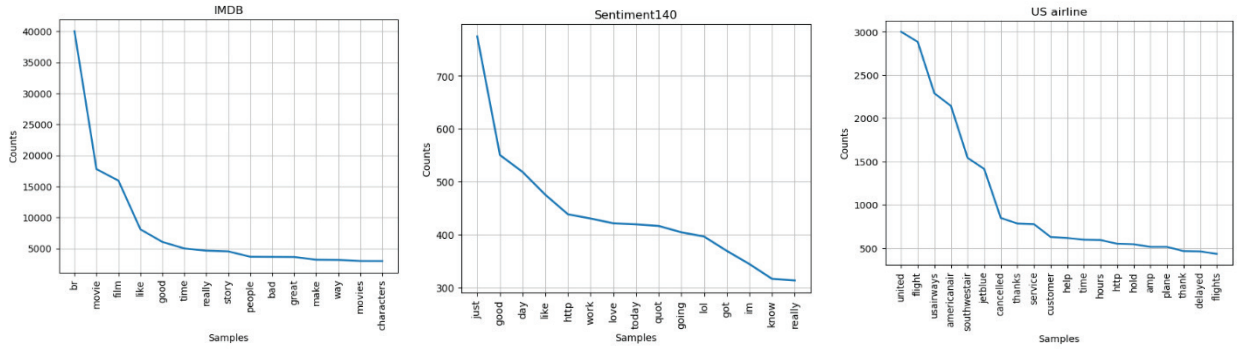

Figure 3. Top 15 most common words after removal of stop words for IMDB, Sentiment140, and US airline dataset.

using hybrid, mild, and root stemming procedures, while lemmatization merges words into a single word, eliminating termination through morphological examination. The conducted experiment utilized lemmatization due to its perceived more meaningful linguistic representations, despite the minimal impact of stemming or lemmatization on accuracy. Stop words, such as "at" "is" and "the" are common in any language but not relevant to SA process. The distribution of the top 15 words in corresponding datasets after removing stop words are depicted in Figure 3.

Figure 4 represent before and after preprocessing results on few reviews of sentiment140 dataset which include removal of hastags, punctuations symbols and others which doesn't add on subjectivity of text. Natural Language Toolkit (NLTK) and Scikit-learn stop words used here to enhance text preparation and reduce noise.

Feature Extraction

Words are too complex for ML techniques to comprehend, therefore feature extraction turns them into vectors

	text	target	cleaned_tweets_w/o_Stopwords	cleaned_tweets_with_SW
0	@ home studying for maths wooot ! im so going to fail this shit	-1	home studying for math wooot i am so going to fail this shit	home studying math wooot going fail shit
1	Pickin up @misstinayao waitin on @sadittysash 2 hurry up...I odeeee missed dem Table talk 2nite...LOL bout to be fat...	-1	pickin up waitin on hurry up i odeeee missed dem table talk nite lol bout to be fat	pickin waitin hurry odeeee missed dem table talk nite lol bout fat
2	@ProudGamerTweet I rather average 32370	-1	i rather average	average
3	@officialnjonas Good luck with that	1	good luck with that	good luck
4	this song's middle change just doesn't want to be born..... arghhhh!!	-1	this song is middle change just doe not want to be born arghhhh	song middle change want born arghhhh
5	im starting my many hours of work now	-1	i am starting my many hour of work now	starting hour work
6	Thunderstorms yesterday, more on the way. Looks like I won't be online much again today. HAPPY FATHER'S DAY	-1	thunderstorm yesterday more on the way look like i will not be online much again today happy father s day	thunderstorm yesterday way look like online today happy father day
7	@cloecouturier Yes, I do have a few .. 4 Girls ... You son is 23 ... all grown up now .. happens fast!!	1	yes i do have a few girl you son is all grown up now happens fast	yes girl son grown happens fast
8	Last free travel at AP-1 http://yfrog.com/0uvu5j	-1	last free travel at ap	free travel
9	TF2 has updated but its way to late to play. looks like I'll be spying and sniping my heart out tomorrow	1	tf ha updated but it way to late to play look like i will be spying and sniping my heart out tomorrow	updated way late play look like spying sniping heart tomorrow
10	@PreternaReviews yeah, but, dude, that's the key here. I'm a CHICK. Chicks wear pink. Especially the ultra cool rocker chicks	1	yeah but dude that is the key here i am a chick chick wear pink especially the ultra cool rocker chick	yeah dude key chick chick wear pink especially ultra cool rocker chick

Figure 4. An illustration showcasing the results of before and after data transformation (Sentiment140 dataset).

in space. After comparing TF-IDF and BoW approaches, TF-IDF was selected because of its better performance. It improves the importance of important terms by quantifying the rank of each term in a document, taking term frequency and overall corpus frequency into account. The following equation illustrates the process of calculating TF-IDF using IDF.

$$TF = \frac{total\ number\ of\ time\ a\ term\ present\ in\ the\ document}{total\ terms\ in\ the\ document} \quad (1)$$

$$IDF\ (t, D) = log\left(\frac{|D|}{|\{d\ \in D; t\ \in d\}|}\right) \quad (2)$$

$$TF - IDF(t, d, D) = tf(t, d) * IDF(t, D) \quad (3)$$

Here *'t'* denotes the term in the document *'d'* and *'D'* represents the full corpora of the document.

Feature Selection

Finding useful features in pre-processed text input is the goal of feature selection, a method that improves the performance of ML models. Thus, selecting the essential features making model more transparent, maintain overfitting and results into much better computional efficient. We experminted with three feature selection methods here: Chi-square, LASSO- feature selection method which proven to be better accuracy than filter and reduced computional overhead than wrapper method. It is done by adding a regularizer term which create sparsity to less important feature thus refines the model's foreseeability and transparency [37]. When it comes to feature selection in SA with textual data, LASSO works incredibly well since it automatically removes redundant features, resulting in a model that is easier to understand. LASSO helps concentrate on the most relevant words or phrases by reducing the coefficients of less important features to zero by incorporating an L1 penalty into the model. This improves model generalizability and lessens overfitting, both of which are essential for managing big and frequently noisy text input in SA tasks. This work assesses the efficacy of LASSO on balanced and imbalanced textual datasets, despite the fact that its use in feature selection for textual datasets has received less attention. By punishing big regression coefficients and deleting extraneous characteristics from the model, the multivariate approach LASSO performs exceptionally well on high-dimensional datasets such as text. Accordingly, we used grid search-based cross-validation to determine the ideal regularization value, which allowed us to accomplish our goal in both balanced and imbalanced datasets. Excluding certain words or keywords is a common way to do this with textual data. Applying LASSO to textual data that has been processed using TF-IDF produces effective results, as shown in Equation 4.

$$minimize\left(\frac{1}{2n}\sum_{i=1}^{n}\left(yi - \sum_{j=1}^{p}\beta j\ xij\right)^2 + \alpha\sum_{j=1}^{p}|\beta j|)\right) \quad (4)$$

The Equation above represents the objective function of LASSO which build upon two components. First is loss function which calulate mean squared error (MSE) between the predicted (yi) and actual values ($\sum_{j=1}^{p}\beta j\ xij$) and also shows how well the model fits the data. The second is Regularization Term ($\alpha\sum_{j=1}^{p}|\beta j|$) which applies a penalty that is directly proportionate to the total of the absolute values of the coefficients ($|\beta j|$), facilitates sparsity by reducing

some coefficients to zero, hence executing feature selection. The other variables used in Lasso equation are explained below:

- n represents the number of text documents in your dataset.
- p signifies the number of unique words extracted from the text documents, which serve as feature space.
- xij signifies the presence or frequency of the j-th word or term in the i-th text document.
- yi signifies the output variable pertinent with the i-th text document.
- βj represents the coefficient linked with the j-th word or term, signifying its importance in projecting the target variable yi.
- α is the regularization parameter that controls the strength of the regularization (penalty term) and model fit (MSE term). Higher values of α result in more coefficients (βj) being shrunk towards zero, effectively removing less important words or terms from the model.

A balanced relationship between sparsity (the simplicity of the model) and predictive performance is guaranteed by an ideal α, for this reason the ideal regularization strength is found by methodically here by exploring several values of α in Grid Search CV. Less important feature coefficients are reduced to zero by the regularization process in LASSO, which imposes a penalty on their magnitude. To encourage sparsity through the removal of irrelevant features and simplify the model for improved interpretability, features with lower correlation with the target variable are penalized more severely. By keeping one feature from a set of correlated variables and rejecting the others, LASSO efficiently handles multicollinearity and improves the model's efficiency and resilience. By improving generalization, decreasing overfitting, and concentrating on the most important features, regularization greatly boosts model performance. This is especially important when dealing with high-dimensional datasets, such as textual SA, because the model can easily get overwhelmed by irrelevant information. To avoid underfitting due to over-regularization and overfitting from under-regularization, it is necessary to properly tune the regularization parameter (α). Another benefit of regularization is that it makes the model faster and easier to understand by lowering the number of features which is very helpful in case of high dimensional textual features.

SENTIMENT CLASSIFIER

SVM Classifier

The most widely used algorithm for SA provides high precision for extensive datasets [38]. The SVM strategy use hyperplanes to assess data and establish decision boundaries in this technique. SVM are a form of deterministic supervised learning method commonly employed for classification purposes. The fundamental SVM is to identify the hyperplane that may effectively split the data into multiple groups. SVM aims to identify the hyperplane that has the maximum possible margin.

Naïve Bayes

NB is a Bayesian classification method using Baye's theorem to determine the probability of a label from a set of features. To classify texts based on presence/absence features, Bernoulli NB (BNB) variant of NB is employed here, which excels at representing features as binary or Boolean [39].

Random Forest Classifier

RF is as an ensemble classifier, meaning it combines the predictions of multiple DTs to get its final prediction as shown in Figure 5. The model uses decision trees (DT) trained on various datasets, with the final output determined by majority vote. It exhibits high resilience, stability, and reduced overfitting. The ideal split is determined by selecting features and building DT.

Ensemble Voting Classifier

Ensemble model works on "wisdom of crowds" that capitalizes the power of multiple base classifiers, aiming to build more accurate and robust predictive model. The potency of ensemble method relies on the accuracy of their constituent participating classifiers, necessitating that these classifiers perform better than the top-performing individual classifier. The investigation involves examining the top six classifiers to construct a customized ensemble technique known as the voting classification model. Included in the consideration are SVM, RF, BNB, SVM-Chi-square, SVM-LASSO, RF-Chi-square, RF-LASSO, BNB-Chi-square, and BNB-LASSO. However, the Ridge feature selection method displayed inadequate performance and is therefore not

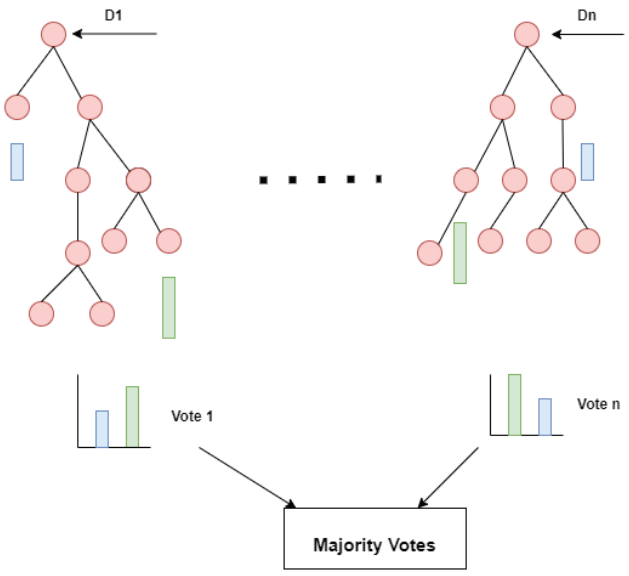

Figure 5. Random Forest classification procedure with n= tree count.

being considered for the final classifier selection. Soft voting has been used in this work indicating that the predicted class labels are determined based on the argmax of the sums of predicted probabilities, suitable for combining probabilities from multiple classifiers. Since in this work, a soft voting classifier is used, the probabilities from each base classifier are combined, and the class with the highest combined probability is chosen as the final prediction. Subsequently, the voting strategy achieved the best accuracy in sentiment prediction out of all the approaches examined by aggregating the classifier's outputs using an average weighting methodology. Here's the corrected equation:

$$P(voting_{soft}) = argmax\,(\sum_{i=1}^{n} w_i P_i(class))$$

In this equation:

- $P(voting)$ represents the probability of the voting classifier.
- wi represents the weight allocated to the ith base classifier.
- Pi (class) represents the probability predicted by the ith base classifier for a particular class.
- n represents the total number of base classifiers.

In the soft voting process, the final class prediction is determined by computing the weighted average of the predicted probability for each base model. Its performance on binary classification is enhanced by this ensemble strategy. However, ensemble model can be computationally expensive sometime, but in this work the main aim is to improve the performance like accuracy and F1 measure. So, the capacity of this technology to assess text content in a balanced and sophisticated way makes it unbeatable by complex sentiment patterns.

The ensemble voting classifier's procedure

1. Gather the data using a suitable group of features.
2. Dataset is splitted into two parts: one for training and one for testing. Eight to two is the ratio that was noted.
3. Individually train SVM, RF and BNB classifiers using a designated training set with or without feature selection methods. Chi, LASSO, and Ridge were used in this study.
4. All classifiers should be utilized to make predictions for the labels of the testing set.
5. Utilize the soft voting procedure to merge the predictions of classifiers. SVM and BNB with LASSO and Chi-square as these two are giving promising result in this work.
6. After the optimization of the ensemble classifier, it gains the ability to make predictions on novel and unobserved data.

EXPERIMENTAL SETUP

The model described in this study was implemented using the following modeling environments: The browser-based interface of Jupyter Notebook an open-source application was utilized to provide freedom for act equally in local and cloud contexts. Google Colaboratory, a Google Research tool, was used for model implementation using Python version 3.7.13, along with various libraries like NumPy, Pandas, Scikit-learn, and Matplotlib. The proposed model was developed and evaluated in a thorough and flexible environment made possible by the combination of various technologies. Additionally, three ML algorithms—SVM, RF, and BNB that are widely used in text categorization were included in the study for a more thorough evaluation. These three methods made use of the BoW and TF-IDF vectorization approaches to extract characteristics from the text input. The ML algorithms are then fed the characteristics or vectors specified earlier. Token density determines how many features each vectorization approach produces for a certain review or text. For these ML algorithms, the TF-IDF method outperformed the BoW method in terms of accuracy. Therefore, evaluation of proposed framework is done with only TF-IDF technique outputs.

RESULTS AND DISCUSSION

The selection of three separate datasets was based on their individual qualities and how well they fit the research goals. Because it does not impose any particular domain restrictions, the sentiment140 dataset was selected for its generalizability. Abbreviations and emojis abound in this dataset which include short messages, usually no more than 100 words long. On the other hand, the IMDB dataset was

Table 2. Specification of the sentiment140, IMDB, and twitter US airline datasets

Sentiment140	IMDB	Twitter US Airline	
Input length	150	500	200
Mean word count	90	231	110
Standard deviation	35	170	40
Vocabulary size (before removing stop words)	45977	158898	23032
Vocabulary size (after removing stop words)	29821	45977	13235
Train-test ratio	80:20		
No of classes	2	2	2

incorporated because of its association with the film industry; it contains far more extensive communications, with an average length of about 500 words. The last reason the US airlines dataset was chosen is because of the imbalance in sentiment distribution and the fact that the text review durations are all over the place. The overall specification of discussed datasets is presented in Table 2 below.

The construction of model is both complicated and exciting by virtue of the eclectic mix of datasets. This study primarily aims to create a model that can successfully analyze and comprehend various types of data inputs. The goal is to build a strong model that works well with a variety of inputs by taking use of the unique features of each dataset. The fundamental objective of this study is to establish a model that is highly effective in dealing with the complicated issues presented by the specific datasets that have been chosen. Thus, improved the text cleaning process in the initial step of preprocessing by eliminating a larger variety of frequently used words from both libraries using a combined set of stop words from scikit-learn and NLTK. When you compare the sizes of these words as shown in Table 2, its highly observable how stop words affect the complexity and richness of dataset's language. This comes especially handy when dealing with English text data, which can include many words that are regularly used yet don't really tell us anything.

The feature set in this study typically has many dimensions, noisy, and may contain irrelevant or redundant information due to the nature of textual features. So, there is dire need of allowing only relevant features feed into the model. In this work, filter-based feature selection method, specifically $\chi 2$ known for their potency in processing categorical features is used. These methods are applied both independently and in conjunction with ML classifiers, including SVM, RF, BNB, and proposed ensemble voting classifier to examine their effect on feature relevance and model outcome. This study introduces LASSO a feature selection methodology integrated with an ensemble voting classifier, a regularization technique used which has been rarely studied in case of text SA, that is specifically tailored for a dataset consisting of different domains. Additionally, Grid Search CV is used to identify the optimal hyperparameters for LASSO regularization and other methods. Furthermore, experiments are run with different values of n in the $\chi 2$ method, in order to choose a value that give promisable accuracy. Experimental outcomes with datasets demonstrate that LASSO which is multivariate analysis consistently outperforms the $\chi 2$ which is univariate approach when integrated with the proposed voting classifier. We explored various train-test ratios in this study, but finally concluded that 80:20 ratio yielding the most promising outcomes managing bias-variance tradeoff. In addition, precise hyperparameter tuning was done to refine model performance however, for brevity, only the most accomplished configuration for this study is presented in Table 3.

The experiment was limited to unigram and bigram approaches, and random sampling was used to ensure equal representation of classes. Potential data augmentation

Table 3. Chosen hyperparameter by grid search CV

Model	Chosen Hyperparameter by grid search CV
LASSO	LASSO_model_best_alpha {'alpha': 0.0001}
Chi-sq	K=2000
RF	Random state = 42, estimators =100
SVC	Kernel ='linear', C=1
BNB	Alpha =2.0

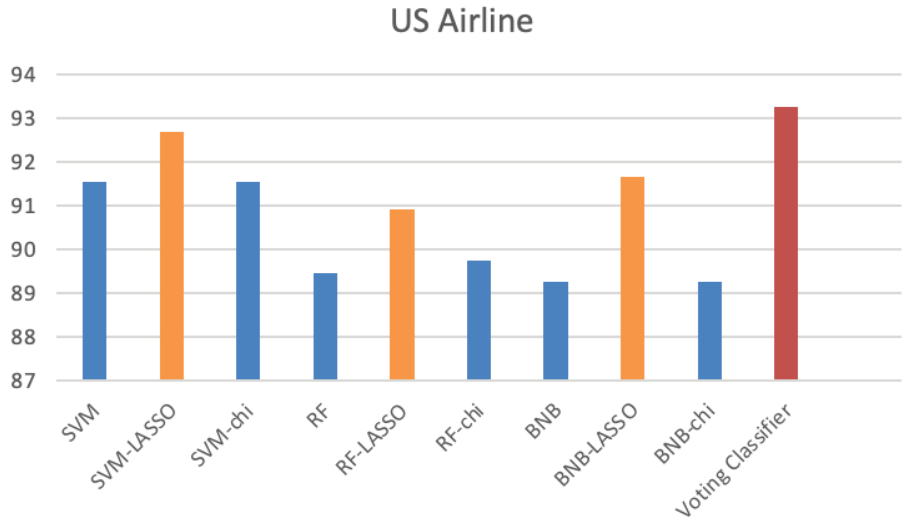

Figure 6. Performance of SVM, RF, BNB, and ensemble voting classifiers on the US Airline dataset using LASSO.

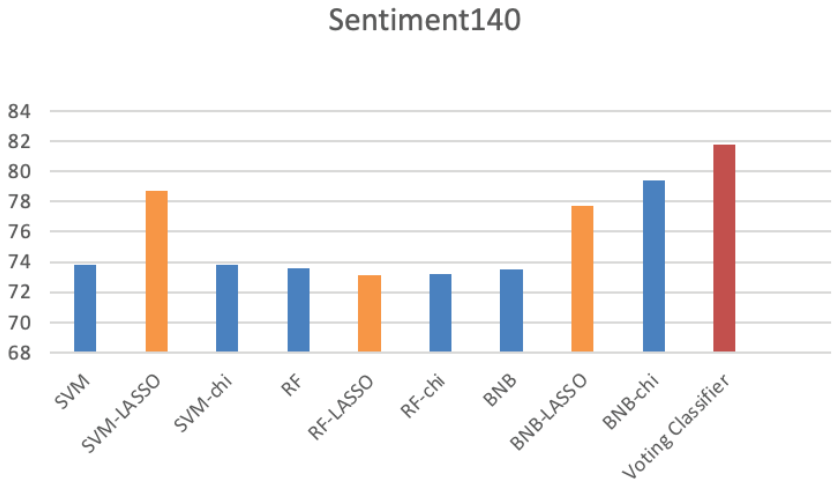

Figure 7. Performance of SVM, RF, BNB, and ensemble voting classifiers on the Sentiment140 dataset using LASSO.

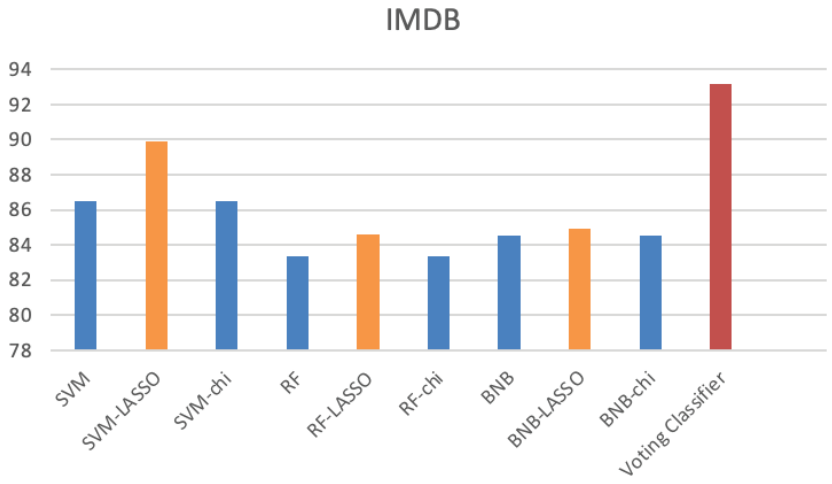

Figure 8. Performance of SVM, RF, BNB, and ensemble voting classifiers on the IMDB dataset using LASSO.

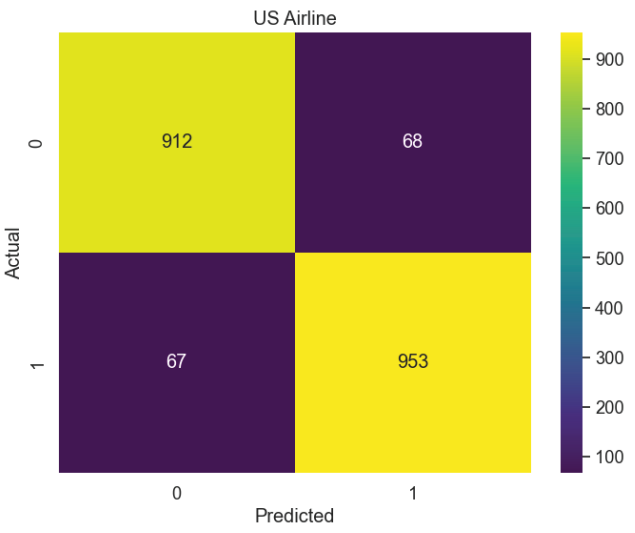

Figure 9. Error matrix of ensemble voting classifier for US Airline using Grid Search based LASSO filtering.

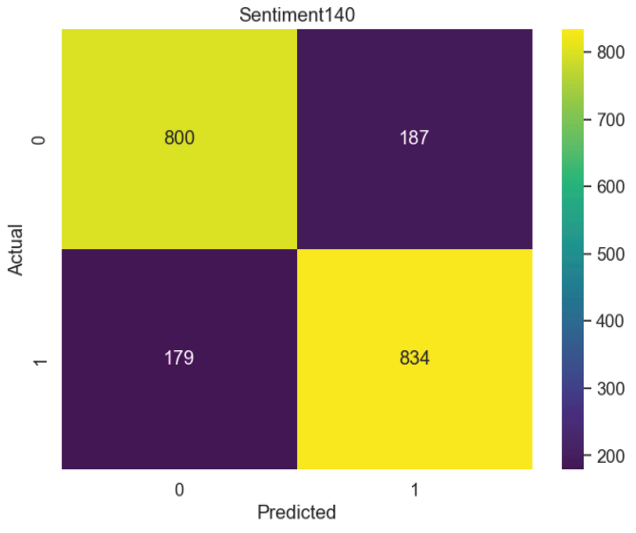

Figure 10. Error matrix of ensemble voting classifier for Sentiment140 using Grid Search based LASSO filtering.

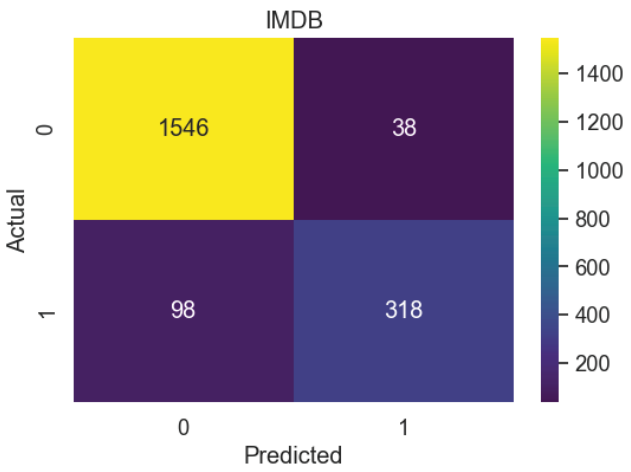

Figure 11. Error matrix of ensemble voting classifier for IMDB using Grid Search based LASSO filtering.

techniques could address class imbalances. Figure 6, 7, and 8 show the variation in classifier accuracy across datasets which shows that our suggested ensemble voting classifier continuously beats all existing classifiers by fusing powerful base classifiers like SVM and BNB with efficient feature selection algorithms like LASSO and Chi-square (Fig. 9-11).

Accuracy, Precision, Recall, and the F-measure (for imbalanced dataset) were some of the performance indicators used to analyze the model's efficacy during this review since. The results for different datasets utilizing the ensemble voting classifier and LASSO are shown in Tables 4, 5, and 6.

Tables 7, 8, and 9 show that our method yields better classification accuracy with least complexity than the work done in previous literature. In comparison to the US Airlines dataset (0.839) and the Sentiment140 dataset (0.800), the IMDB dataset has an MCC of 0.634.

Figure 12, 13, and 14 show the box plots across various metrics for each dataset. For Sentiment140 dataset, box plot shows many outliers indicating that the model's performance varies significantly throughout the dataset which could be due to linguistic features of dataset as it contains very short text with emoticons and varying customer's sentiments on different topics.

The suggested ensemble-voting SA framework shows strong transferability to many domains. Its success on all

Table 4. Precision of all three datasets

Dataset	SVM	RF	BNB	SVM-Chi-square	SVM-LASSO	RF-Chi-square	RF-LASSO	BNB-Chi-square	BNB-LASSO	Voting
IMDB	88.1	84.03	84.01	87.4	88.7	84.03	85.6	85.3	85.8	93.0
Sentiment-140	74.7	74.4	73.3	74.0	74.7	75.0	73.0	73.0	72.0	81.7
US airline	91.9	88.4	85.1	92.1	91.7	88.3	87.4	82.8	87.2	92.8

Table 5. Recall of all three datasets

Dataset	SVM	RF	BNB	SVM-Chi-square	SVM-LASSO	RF-Chi-square	RF-LASSO	BNB-Chi-square	BNB-LASSO	Voting
IMDB	88.4	84.2	84.0	87.4	87.8	84.2	85.0	84.2	85.7	93.0
Sentiment-140	74.6	74.4	73.0	74	75.0	75.0	73.30	73.0	72.0	81.8
US airline	81.6	79.5	87.9	84.4	84.8	78.9	81.3	88.1	87.5	93.2

Table 6. F-1 Score of all three datasets

Dataset	SVM	RF	BNB	SVM-Chi-square	SVM-LASSO	RF-Chi-square	RF-LASSO	BNB-Chi-square	BNB-LASSO	Voting
IMDB	87.6	84.3	84.0	87.4	87.6	84.5	85.0	84.3	85.7	93.0
Sentiment-140	74.6	74.3	72.9	74.0	75.0	75.0	73.0	73.0	72.0	81.7
US airline	85.5	82.8	86.4	87.5	86.9	82.4	83.9	85.0	87.4	93.1

Table 7. Comparison of model complexity and performance on the IMDB dataset

Sr. no	Source & publication year	Feature extraction	Feature filtering	Classifiers	Algorithm complexity	Classification accuracy
[18]	Naïfs & Awang, 2021	TF-IDF	Hybrid SVM+RFE	SVM	O (M×N) M is size of data; N is no of features.	83.16%
[20]	Kokab et al, 2022	Word2vec, Glove	---	CNN, LSTM, CBRNN	O(P∗F∗T) P stands for the parameter count. F represents the computational complexity of the forward pass through the network, T represents the training time required to train the model on a given dataset.	93%
[22]	Subba & Kumari, 2022	GloVe, Word2vec, Bert	---	Ensemble LSTM, GRU	O ((P1+P2+p3) ∗ (F1+F2+ F3) ∗ T) Here, P combines all parameters from Word2Vec, GloVe, BERT, and LSTM/GRU models; F combines their respective forward pass complexities; and T reflects the overall training time, influenced by data size and training epochs.	92%
[26]	Rehman et al., 2019	Word2vec	---	CNN + LSTM	O (M × K) + O (M × H2) M is size of input sequence, H is no of input unit, K is kernel size.	91.78%
--	Proposed work	TF-IDF	LASSO	Ensemble (SVM, BNB, RF)	O(P*Q) + T*log(n) P is size of data, Q is count of features, T is count of trees and n is total samples count in dataset.	93.20%

Table 8. Comparison of model complexity and performance on the US airlines dataset

Sr. no	Source & Publication Year	Feature extraction	Feature filtering	Classifiers	Algorithm Complexity	Classification Accuracy
[29]	Bataineh & Kaur, 2021	Keras embedding	---	CSA-LSTM	O (P * F * T)	92.25%
[31]	Tan et al, 2022	GloVe	---	RoBERTa-LSTM	$O(L*N^{2*}d) + O(N*d*h)$ N is the sequence length, d is the input dimension h is the hidden size of the LSTM.	85.89%
[32]	Tan et al, 2022	GloVe	---	Ensembling of RoBERTa with LSTM, BiLSTM, GRU	$O(L*N^{2}*d) + O(N*d*h_{LSTM}) + O(2*N*d*h_{BiLSTM}) + O(N*d*h_{GRU})$	91.77%
--	Proposed work	TF-IDF	LASSO	Ensemble Voting (SVM, RF BNB)	O(P*Q) + T*log(n) P is size of data, Q is count of features, T is count of trees and n is total samples count in dataset.	93.25%

Table 9. Comparison of model complexity and performance on the sentiment140 dataset

Sr. no	Source & Publication Year	Feature extraction	Feature filtering	Classifiers	Algorithm Complexity	Classification Accuracy
[40]	Singla et al, 2022	Word2vec	---	BERT	O(M*2)∗ d Here M is input sequence length; d is dimensionality of the model.	81.03%
--	Proposed work	TF-IDF	LASSO	Ensemble Voting (SVM, RF BNB)	O(P*Q) + T*log(n) P is size of data, Q is count of features, T is count of trees and n is total samples count in dataset.	81.75%

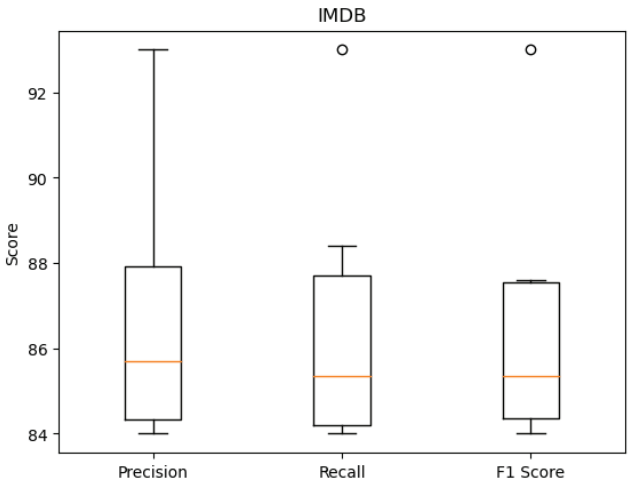

Figure 12. Visualization of metrics across IMDB using whisker plot.

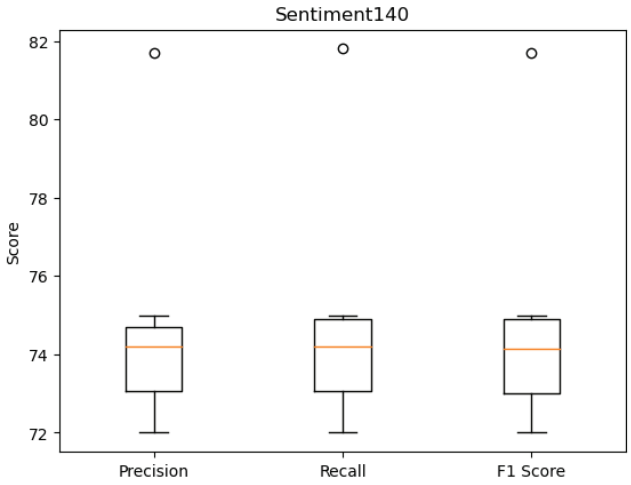

Figure 14. Visualization of metrics across sentiment140 using whisker plot.

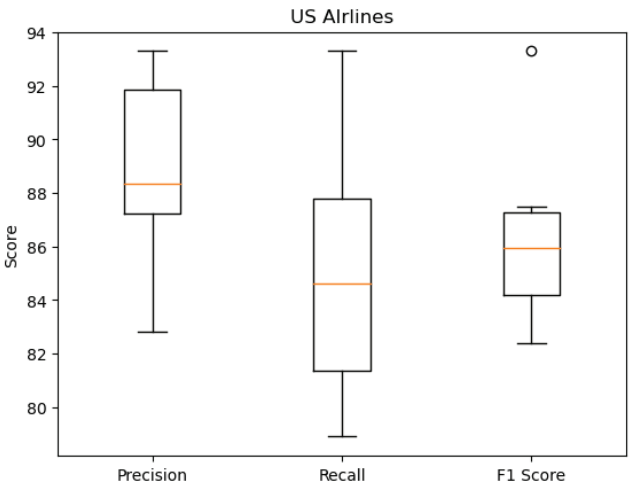

Figure 13. Visualization of metrics across US Airlines using whisker plot.

three datasets depicts that it might be useful to analyze trends, public opinion monitoring, and wide-scale consumer input in fields like social media, entertainment, and travel. Although the proposed model's result proven better detection of sentiments, but there are still some limitations of this study. In high-dimensional text data, nuanced predictors may be useful, but LASSO has tendency to enforce sparsity by zeroing out less important features, can lead to their permanent removal and additionally ensemble model can be complex and computationally heavy, especially when balancing base classifiers share. Sometimes, this extra complexity makes things harder to understand and makes real-time analysis less efficient.

CONCLUSION

Using LASSO, an inherent feature selection method based on multivariate analysis, this work aims to detect textual data polarity. To generalize the effectiveness of proposed ensemble voting classifier, model is test on dataset with different domain and different length of input text like short in sentiment140 and long for Internet Movie Database and also with balanced and imbalanced in case of US airline dataset. Results showing satisfactory accuracy as compared to DL models which are more complex and black

box in nature. It must be emphasized that Grid Search-based LASSO is a noteworthy feature selection method. For the Sentiment140, Internet Movie Database, and US airlines dataset, the ensemble voting model with Grid Search-based LASSO achieves an accuracy of 81.75%, 93.20%, and 93.25% respectively. A proposed voting classifier which is an ensemble model based on soft voting was built in this study using the Grid Search-based multivariate feature selection method LASSO in conjunction with the classifiers SVM, RF, and BNB. In our forthcoming research, we intend to assess the efficacy of the model in several tasks, including the identification of distinct emotions (e.g., sadness, happiness) and the analysis of emotions based on certain aspects. Furthermore, we will also look into data imbalances and tweets that include non-textual elements such as emojis. In addition, methods for pre-trained text embedding will be utilized to construct sophisticated emotion model. To handle class imbalance by dynamically prioritizing cases from underrepresented emotions, future study could investigate the possibility of using ensemble classifiers from adaptive weighted neural networks for real-time text emotion identification.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Ligthart A, Catal C, Tekinerdogan B. Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev 2021;54:1–57. [CrossRef]

[2] Krakovsky M. Artificial (emotional) intelligence. Commun ACM 2018;61:18–19. [CrossRef]

[3] Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 2022;55:5731–5780. [CrossRef]

[4] Kumar R. Twitter sentiment analysis for polarity detection: a survey. In: 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE; 2023. p. 559–564. [CrossRef]

[5] Altınel Girgin AB, Gümüşçekiççi G, Birdemir N. Turkish sentiment analysis: a comprehensive review. Sigma J Eng Nat Sci 2024;42:1292–1314.

[6] Ren R, Wu DD, Liu T. Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst J 2018;13:760–770. [CrossRef]

[7] Alhaj YA, Xiang J, Zhao D, Al-Qaness MA, Abd Elaziz M, Dahou A. A study of the effects of stemming strategies on Arabic document classification. IEEE Access 2019;7:32664–32671. [CrossRef]

[8] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Based Syst 2015;89:14–46. [CrossRef]

[9] Rasool A, Tao R, Kamyab M, Hayat S. GAWA – a feature selection method for hybrid sentiment classification. IEEE Access 2020;8:191850–191861. [CrossRef]

[10] Ye Q, Zhang Z, Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst Appl 2009;36:6527–6535. [CrossRef]

[11] Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. J Biomed Inform 2016;62:148–158. [CrossRef]

[12] Shaik T, Tao X, Dann C, Xie H, Li Y, Galligan L. Sentiment analysis and opinion mining on educational data: a survey. Nat Lang Process J 2023;2:100003. [CrossRef]

[13] Mahmoodzadeh A, Nejati HR, Mohammadi M, Ibrahim HH, Khishe M, Rashidi S, et al. Prediction of Mode-I rock fracture toughness using support vector regression with metaheuristic optimization algorithms. Eng Fract Mech 2022;264:108334. [CrossRef]

[14] Mosavi MR, Aghababaie M, Naseri MJ, Khishe M. Compression of respiratory signals using linear predictive coding method based on optimized algorithm of humpback whales to transfer by sonobouy. Iran J Mar Technol 2020;7:1–13.

[15] Tang W, Yang S, Khishe M. Profit prediction optimization using financial accounting information system by optimized DLSTM. Heliyon 2023;9:e19431. [CrossRef]

[16] Appel O, Chiclana F, Carter J, Fujita H. A hybrid approach to the sentiment analysis problem at the sentence level. Knowl Based Syst 2016;108:110–124. [CrossRef]

[17] Tiwari P, Mishra BK, Kumar S, Kumar V. Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. In: Cognitive analytics: concepts, methodologies, tools, and applications. IGI Global; 2020. p. 689–701. [CrossRef]

[18] Nafis NSM, Awang S. An enhanced hybrid feature selection technique using TF-IDF and SVM-RFE for sentiment classification. IEEE Access 2021;9:52177–52192. [CrossRef]

[19] Ray P, Chakrabarti A. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Appl Comput Inform 2022;18:163–178. [CrossRef]

[20] Kokab ST, Asghar S, Naz S. Transformer-based deep learning models for the sentiment analysis of social media data. Array 2022;14:100157. [CrossRef]

[21] Soubraylu S, Rajalakshmi R. Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews. Comput Intell 2021;37:735–757. [CrossRef]

[22] Subba B, Kumari S. A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings. Comput Intell 2022;38:530–559. [CrossRef]

[23] Yousaf A, Umer M, Sadiq S, Ullah S, Mirjalili S, Rupapara V, et al. Emotion recognition by textual tweets classification using voting classifier (LR-SGD). IEEE Access 2020;9:6286–6295. [CrossRef]

[24] Saraswathi N, Rooba TS, Chakaravarthi S. Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews. Meas Sens 2023;29:100888. [CrossRef]

[25] Catal C, Nangir M. A sentiment classification model based on multiple classifiers. Appl Soft Comput 2017;50:135–141. [CrossRef]

[26] Rehman AU, Malik AK, Raza B, Ali W. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimed Tools Appl 2019;78:26597–26613. [CrossRef]

[27] Gouthami S, Hegde NP. Feature selection based sentiment analysis on US airline twitter data. Delta 2023;955;723:544.

[28] Hameed Z, Garcia-Zapirain B. Sentiment classification using a single-layered BiLSTM model. IEEE Access 2020;8:73992–74001. [CrossRef]

[29] Al Bataineh A, Kaur D. Immunocomputing-based approach for optimizing the topologies of LSTM networks. IEEE Access 2021;9:78993–79004. [CrossRef]

[30] Rafae A, Erritali M, Roche M. Fusion of BERT embeddings and elongation-driven features. Multimed Tools Appl 2024;83:80773–80797. [CrossRef]

[31] Tan KL, Lee CP, Anbananthen KSM, Lim KM. RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access 2022;10:21517–21525. [CrossRef]

[32] Tan KL, Lee CP, Lim KM, Anbananthen KSM. Sentiment analysis with ensemble hybrid deep learning model. IEEE Access 2022;10:103694–103704. [CrossRef]

[33] Baqach A, Battou A. CLAS: A new deep learning approach for sentiment analysis from Twitter data. Multimed Tools Appl 2023;82:47457–47475. [CrossRef]

[34] Wan Y, Gao Q. An ensemble sentiment classification system of Twitter data for airline services analysis. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE; 2015. p. 1318–1325. [CrossRef]

[35] Nisha, Kumar R. Comparison between feature extraction algorithms for sentiment recognition from text. In: International Conference on Intelligent Systems Design and Applications. Springer; 2023. p. 301–310. [CrossRef]

[36] Ahmad SR, Bakar AA, Yaakub MR. A review of feature selection techniques in sentiment analysis. Intell Data Anal 2019;23:159–189. [CrossRef]

[37] Tan M, Tsang IW, Wang L. Matching pursuit LASSO part I: Sparse recovery over big dictionary. IEEE Trans Signal Process 2014;63:727–741. [CrossRef]

[38] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 2020;408:189–215. [CrossRef]

[39] Badik ŞT, Akar M. Machine learning classification models for the patients who have heart failure. Sigma J Eng Nat Sci 2024;42:235–244. [CrossRef]

[40] Singla C, Al-Wesabi FN, Pathania YS, Alfurhood BS, Hilal AM, Rizwanullah M, et al. An optimized deep learning model for emotion classification in tweets. Comput Mater Contin 2022;70:6365–6380. [CrossRef]