International Journal of Information Technology and Web Engineering
Volume 17 • Issue 1

Sentiment Analysis on Movie Reviews Dataset Using Support Vector Machines and Ensemble Learning

Razia Sulthana A., Birla Institute of Technology and Science, Pilani, Dubai, UAE
Jaithunbi A. K., R. M. D. Engineering College, India
Haritha Harikrishnan, University of Dubai, UAE
Vijayakumar Varadarajan, University of New South Wales, Australia
https://orcid.org/0000-0003-3752-7220

ABSTRACT

The internet makes it easier for people to connect to each other and has become a platform to express ideas and share information with the world. The growth of the internet has indirectly led to the development of social networking sites. The reviews posted by people on these sites implies their opinion, and analysis over reviews is required to understand their intent. In this paper, natural language processing technique and machine learning algorithms are applied to classify the text data. The contributions of the proposed approach are three-fold: 1) chi square selector is applied to select the k-best features, 2) support vector machines is executed to classify the reviews (hyperparameters of the SVM classifier are tuned using GridSearch approach), and 3) bagging algorithm is applied with the base classifier over the newly built SVM classifier. The number of base classifiers of the bagging algorithm is varied accordingly. The results of the proposed approach are compared to the similar existing work, and hence, it is found to achieve better results as compared to the existing systems.

KEYWORDS

Count Vectorizer, Emotion Analysis, Gridsearch, Hyperparameters, Natural Language Processing

1. INTRODUCTION

Sentiment Analysis, otherwise called as opinion-mining applies natural language processing techniques to schematically recognize, extract, enumerate, and study the subjective information. It is extensively applied to reviews given by the customers; responses to the surveys; reviews in online and social media; reviews given over products in online e-stores, to create AI based bots or assistants. Sentiment analysis classifies the opinion as positive or negative. Lexicon-based and Machine-learning (ML) based approaches are applied to identify the sentiment of any sentence. The former approach uses a vocabulary which contains pre-defined negative and positive words and the latter approach uses training and testing data to identify the positive and negative words. Sentiment analysis can be applied to classify emotions based on subjective parameters (Liu, 2010). It is known as emotion AI and has a variety of purposes in different fields like analyzing sentiments in emails, comments and

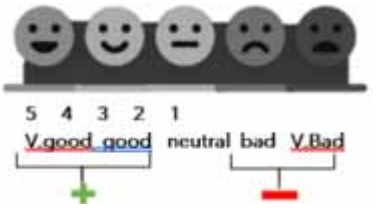
DOI: 10.4018/IJITWE.311428 *Corresponding Author

survey feedback. It plays an imperative role in the domain of Artificial Intelligence (Mäntylä et al, 2018; Poria et al, 2018).

The textual datasets that are applied for sentiment analysis are first subjected to preprocessing. Most of the datasets require removal or fixing of missing values, null values or redundant values. Data pre-processing step includes sampling, cleaning and transformation of data. The type of data pre-processing needed by a particular dataset depends on the type of datasets (textual/image/numerical dataset). In the proposed approach, the type of dataset is a textual dataset.

Movies are one of the finest forms of entertainment and it's a very common thing that the people watch movies and share their opinions on the social media platforms. By analyzing the reviews on the movies, the positive and negative opinion over the movie can be found. Thus, sentiment analysis can help in knowing the public opinion of that movie. Twitter, another platform where a huge perception of the user's opinion is posted every day and these opinions can be over any generic content. Few of the recent research articles focus over detecting the hatred words in tweets. A number of emotional labels is used largely in tweets and is given in Figure 1.

Figure 1. Labels used to classify the sentiments of the comments



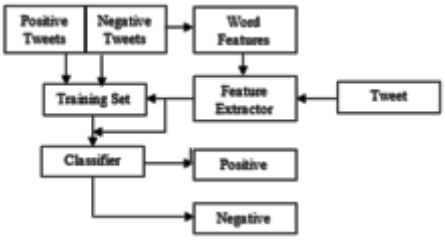
The section split of this paper is given here: Section 2 details the terminologies, tasks, levels and open challenges in sentiment analysis; Section 3 does a detailed analysis about the literature work done in this field; Section 4 explains the step by step procedure of implementing review analysis using SVM; Section 5 tabulates all the experimental outcomes and compares with the results of existing works; Section 6 concludes the research work.

2. TERMINOLOGIES

- **Natural language processing (NLP):** It is applied in sentiment analysis to review the marketing strategies and has reshaped the business approach. The steps of applying NLP (Chowdhury, 2003) in analyzing a review includes the process of tokenization; applying Part Of Speech (POS); text lemmatization; stop word identification, etc
- **Tokenization:** Tokenization (Webster & Kit, 1992) is splitting a phrase or sentence or paragraph, or an entire text document into smaller units or terms. Each of these smaller units are called tokens. Tokenization is important because the meaning of the text could easily be interpreted by analyzing the words present in the text. Tokenization is a critical step in NLP and jumping into the model-building is not possible without applying tokenization (Pentheroudakis et al, 2006).
- **Bag-of-words:** It's a way of representing text data as a group of words. The bag-of-words model is applied in language and document classification (Voorhees, 1999).

A Model for the sentiment analysis is as given in Figure 2.

Figure 2. General model of sentiment analysis



2.1 Q&A to be Evaluated Before Taking up Sentiment Analysis Process Over Dataset

The most probable questions related to the proposed system is given here:

Q1: What are the repeatable words in the dataset?
Q2: The quotable words found in positive and negative comments in the dataset.
Q3: Are the tweets quoted with hashtags? And how many of them are in average in a tweet?
Q4: The possible type of trends found in the dataset?
Q5: In what way they are related with the sentiments?

2.2 Machine Learning Applied in Sentiment Analysis

A number of learning techniques under Learning Models (Supervised and Unsupervised) are applied for emotion recognition (Figure 1). Supervised Learning Model (SLM) also known as Spoon Fed Algorithm has both input and output. In addition to SLM, the model uses techniques like Phrase and Topic Modelling. Phrase Modelling is essentially extracting the frequently used phrases and grouping them together. Topic modelling determines the most often used topic identified by adopting the technique of Latent Dirichlet Allocation Mechanism (LDA). Extracting the suitable data model (Araque et al., 2019) and choosing the appropriate lexical model technique (Cachola et al., 2018) is an important step in ML.

The two main machine learning techniques used for sentiment analysis are: Supervised Learning, Unsupervised Learning. Unsupervised Learning: It works over unlabeled dataset ex. Clustering. Supervised Learning: It works over the labeled dataset. This supervised dataset is first trained and the significant outputs are obtained in the decision-making phase.

A good learning algorithm learns the selection of certain features to detect sentiments. A number of machine learning algorithms are applied in sentiment analysis. The frequently used feature selection approach in sentiment classification are:

1. Term frequency; Inverse Document frequency
2. POS
3. Negations
4. Opinion analysis

2.2.1 Sentiment Analysis Tasks

Sentiment Analysis is a multifaceted task and includes the following tasks:

1. **Subjectivity Classification:** The type of classification in which the sentences are classified as opinionated or not opinionated.

2. **Sentiment Classification:** Once after subjectivity classification is done, sentiment classification identifies the polarity as positive or negative. The Classification can be two class or multi-class with class labels.
3. **Complementary Tasks:** This includes object extraction and feature Extraction. The former discovers opinion holders or sources and the latter discovers the target entity.

2.2.2 Levels of Sentiment Analysis

There are different sentiment analysis techniques namely Sentence, Aspect or Document. The sentence-based analysis finds the sentiment of a sentence. Aspect based analysis finds the property of an entity in a sentence. The third approach finds the sentiment of the whole document.

There are various levels of sentiment analysis namely word-based, sentence-based, document-based, and feature-based.

2.3 Open Challenges and Proposed System

There are number of challenges in the sentiment analysis (Mozetič et al, 2016) techniques followed in literature:

1. A lack of specification still persists while classifying the words or documents. A word can have different meaning in different realm. Often there can be inconsistencies amongst human annotators (Kralj Novak et al., 2015). Though the task seems rather simple, at times it can be difficult to classify annotators.
2. Dealing with special characters is still a challenging and demanding task.
3. Graphics and symbols example emoticons etc. are tricky to classify.
4. Data problem: Textual data is highly inconsistent.
5. Language Problem: Yet minimal work is still left undone in languages other than English.
6. Availability of inadequate labelled dataset.
7. The ways of dealing with complex sentences (many adverbs, adjectives) is still a challenge (Mendes et al., 2020).
8. Recognizing the subjective parts of the text.
9. Domain Dependence.
10. Sarcasm Detection.
11. Thwarted Expressions.
12. Explicit Negation of sentence.
13. Entity Recognition.
14. Classifiers for subjective and objective sentiments.
15. Handling the comparative sentences.

It's well known fact that a number of customized processing engines for document analysis are proposed and implemented by researchers. However, not all models work with all dataset. Naive Bayes algorithm works well with huge dataset and Logistic regression sounds good with binary dataset. In general, there are 2 approaches to analyze Sentiments:

1. One way is to either use SaaS (Software as a Service).
2. Other way is to design an algorithm (Astya et al., 2017).

3. LITERATURE STUDY

Recently, there have been a lot of developments in the domain of sentiment analysis and numerous challenges left unaddressed. Hence a very detailed analysis is done over the algorithms applied by

the researchers, their methodology, advantages of the approach and the concerns. A number of latest works done by researchers in this field is taken for study and it is found that a wide variety of them uses SVM or SVM in combinations with other algorithms. Many of these either work on the reviews posted by customers in online e-stores or with the tweets posted by them in twitter. These works are differentiated and tabulated as (Table 1 and Table 2) for better understanding. Table 1 signifies in detail the objective, methodology, dataset used, pros and cons of the recent research works that focus on applying SVM in detecting the sentiments from the reviews posted by customers. The dataset taken up by the authors referenced in the Table 1 are either a standard dataset or self-created dataset formed by scraping reviews from specific e-stores websites. Table 1 details 7 recent papers falling in the range since 2017 to 2020 that are narrowed down based on the proposed problem statement.

Table 1. Review analysis using SVM

Ref/ Year	Objectives	Methodology	Dataset	Advantage	Disadvantage	Performance Measure Value
(Kumari et al., 2017)	Sentiment analysis of the various reviews on smart phones using SVM.	Text pre-processing, transformation, clustering, SVM classification	Phone brands reviews	SVM sounds good with overlapping points	The NLP methods cannot understand sarcasm, emoticons etc.	Accuracy 90.90%
(Korovkinas et al., 2017)	To build a good classifier for sentiment-analysis	Naïve Bayes and SVM classifier	Movie dataset, Amazon customer reviews	Can handle lengthy sentences, better accuracy obtained, reduced data pre-processing time	Minimal accuracy, the NLP methods cannot understand sarcasm.	Naïve bayes. 0,14% - 0.67% increment for SVM and 2,84%-6,99% Naıve Bayes.
(Ahmad et al., 2017)	Sentiment analysis to classify tweets using SVM	Convert text files into ARFF format and apply SVM.	Self-driving cars dataset, dataset containing tweets about apple products	Gives a prediction independent of the type of dataset.	Minimal accuracy, the NLP methods cannot understand sarcasm.	Accuracy with two datasets: 59.91%, 71.2% respectively
(Ahmad et al., 2018)	To measure the performance of SVM using grid search technique.	Dataset insertion into WEKA, Pre-processing (TF-IDF,Stop words, Stemming, words to keep, n gram tokenizing), Classification (SVM using grid search technique, k-fold cross validation,) and Results.	Tweets related to Apple, Google, Microsoft and Twitter, dataset containing tweets related to U.S. airlines, (IMDB) reviews dataset	Improved performance compared to (Ahmad et al., 2017)	The underlying meaning of the sentence is yet to be learnt.	Highest Accuracy, fscore, recall = 87.2,86.8,88.7

Table 1 continued on next page

Table 1 continued

Ref/ Year	Objectives	Methodology	Dataset	Advantage	Disadvantage	Performance Measure Value
(Ali et al., 2016)	One-Class SVM and binary class SVM for Spam Classification	Data cleaning, pre-processing, Feature reduction using the Gain Ratio algorithm and SVM	Spam-base dataset	Classifies the mails with a high accuracy	Minimal number of features are used.	Near to around 100%
(Mohammed et al., 2019)	Compare Naïve Bayes and SVM algorithm for spam email classification	Data cleaning, pre-processing, Feature reduction and SVM	Spam Data and SPAMBASE.	100% precision was achieved which means that out of all the positive predictions, all the spam mails were classified correctly.	Recall of SVM (87%) is lesser than Naïve Bayes (92%)	SVM algorithm (98.3% accuracy, 100% precision) outperforms Naïve Bayes Algorithm in precision.
(Xia & Deng, 2020)	Apply Cluster Intelligence and classify the email as spam	Data pre-processing, Particle swarm optimization and SVM	Network public dataset with 5000 samples	Particle swarm optimization could find optimal values of the SVM parameters	Improper handling of sarcastic sentences	Accuracy has been improved by 12.841% as compared to (Mohammed et al., 2019)

Table 2 is organized focusing on different machine learning algorithms based on clustering, classification, CNN, NLP, RNN and fuzzy based approach. These works are applied over tweets for detecting their sentiments. It includes around 20 latest works taken from 2015 to 2020 explaining the algorithm used with pros and cons of each of them. Every single research work either expands or updates the results as compared to its referential work. However, a common challenge that is faced by all of them is the risk of overfitting or the implementation to be more expensive or increased training time or minimized accuracy.

Table 2. Tweet Analysis using ML Algorithms

Reference	Methodology	Dataset	Algorithm	Advantage	Disadvantage
(Zhang et al., 2020)	Classification	Twitter	Support Vector Machine	Filters useful mails and messages.	Requires much more time to process
(Chen et al., 2018)	Clustering	Microblogs	Naive Bayes	Prevents the loss of valuable time and information if a server fails	Zero probability problem
(Madisetty & Desarkar, 2018)	Convolutional Neural Networks	Twitter Tweets	Deep Learning and feature-based methods	Gives input to the clients about messages subsequently decreasing misclassification.	Risk of overfitting
(Rajapaksha et al., 2017)	Clustering	Twitter Tweets	Hierarchical clustering	Recognizing oddities, considering the fleeting progression of information streams in web-based media.	Cannot analyze high-dimensional data
(Liu & Chen, 2019)	Convolutional Neural Networks	Patient descriptions from social media	Deep learning-based methods	Detects incorrect medical information.	High error susceptibility
(Feizollah et al., 2019)	Convolutional Neural Networks	Word2Vec and GLoVe models.	Deep learning-based methods	Detects all Illegal, boycotted items	Mediocre accuracy value
(Amin et al., 2020)	Recurrent Neural Network	A collection of tweet dataset	Naïve Bayes	High in accuracy and performance.	High error-susceptibility.
(Tocoglu et al., 2019)	N-gram method	Turkish tweets from twitter	Naïve Bayes classifier	Works only for positive and negative emotions.	Very expensive, and requires a large amount of memory and computational resources
(Liu et al, 2019)	Fuzzy Approach Methodology	Tweets from twitter	Support Vector Machine	Requires less computing Power	The results are perceived based on assumption
(Leis et al., 2019)	Natural Language Processing	Sequential groupings of posts and remarks from twitter	Neural Network	The model is robust to missing data	Requires high computational power
(Razzaq et al., 2019)	Natural Language Processing	Taobao tweets	Support Vector machine	Predictions are faster	Not suitable for large data sets
(Jelodar et al., 2020)	Natural Language Processing	Reddit Comments	Kernel SVM	More efficient and less cost	Long training time

Table 2 continued on next page

Table 2 continued

Reference	Methodology	Dataset	Algorithm	Advantage	Disadvantage
(Chiroma et al., 2018)	Prism	Suicide related tweets	Prism and SVM	Cost efficient	Time consuming
(Imran et al., 2020)	Deep Neural Network	Twitter covid related tweets	Logistic Regression	Less time consuming	Extremely expensive to train
(Alnajran et al., 2018)	Twitter preprocessing methodology	Brexit twitter Dataset	Clustering Algorithm	Preprocessing methodology	High time complexity
(Hmeidi et al., 2015)	Convolutional neural network	Free Arabic news text documents	Support Vector Machine	Requires minimal preprocessing procedure	Loss of internal data
(Blasch et al., 2019)	Convolutional neural network	Twitter Dataset	Support vector machine and logistic regression	Higher Accuracy	Assumption of linearity between the dependent and independent variable
(Abualigah, 2020)	Natural Language processing	Twitter and youtube comments	Support vector machine	Scales relatively well to high dimensional data.	Does not perform well if there is very large dataset
(Dias et al., 2018)	Supervised Learning	Corpus data	Naive bayes	Requires less training data	Overrated decision boundary
(Dastanwala & Patel, 2016)	Natural language processing	Twitter dataset	Text Mining	Efficient analysis of data	Invasion of privacy

3.1 Contributions of the Proposed Work

The literature work done in section 3 focused on applying SVM in different strengths for sentiment analysis. Though sentiment analysis is a huge research area with several challenges, the proposed article aims to apply SVM on analyzing tweets by adopting different ways of preprocessing, feature extraction and hyperparameter (k) tuning.

The contributions of the proposed work includes:

1. Applying optimal NLP and feature selection techniques in preprocessing the IMDB reviews.
2. Builds a precise SVM classifier by tuning with optimal hyperparameters and choosing the appropriate kernel for the proposed problem.
3. Proposing the usage of Chi square selector to identify the optimal value of k.
4. Applying the k-cross validation GridSearch method for Hyperparameter tuning.
5. Creating an ensemble bagging model to fetch the correct count of base estimators leading to higher accuracy.

Our proposed approach uses SVM learning models with modified preprocessing, feature selection and kernel methods. The experimental results prove that the precision and accuracy obtained are better as compared to existing models.

4. IMPLEMENTATION METHODOLOGY OF REVIEW ANALYSIS USING SVM

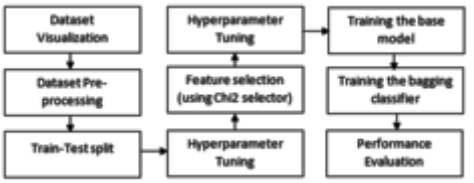
This section explains the steps of implementing the proposed system, a flowchart explaining the sequence of operations done, hyperparameter tuning procedure and model building procedure. The experimental results of each of them is shown the Section 5.

4.1 Steps Followed in Proposed Methodology

The very first procedure of review analysis is a detailed study done on the structure of the dataset, the type of contents in the dataset. Subsequently the dataset is visualized to know the proportional split up of words in it. Based on the understanding of all these, the data preprocessing procedure is done.

Figure 3 details the step by step implementation of the proposed methodology.

Figure 3. Step by step methodology of the proposed system



4.2 Understanding the Dataset and its Structure

The dataset is used in the paper is IMDB reviews downloaded from the Kaggle datasets. It contains 50,000 highly polarized IMDB reviews with less than thirty reviews for each movie. It is a 2-dimensional dataset which contains 2 attributes namely reviews and sentiment. The review column has the reviews and sentiment column tells whether the review is positive or negative. A negative review is identified as the review with score lower than 4/10 and a positive review has review higher than 7/10. There are 25000 records available for each class of reviews. It does not contain null values and is a balanced dataset.

4.3 Dataset Visualization

The dataset can be visualized using the word clouds and matplotlib library:

1. **Summary statistics:** The summary of the columns in dataset is printed using the describe() function of pandas library. As it can be seen from the same, 50000-49582 = 418 rows are duplicate. Thus, the duplicates were removed in the data pre-processing process. In the Figure 4, only the sentiment column is analysed and it looks very uniform in values.
2. **Positive vs. Negative class proportion:** The ratio of negative to positive class was found out by plotting the same. The count of all positive and negative reviews is plot in the pie chart. As it can be inferred from the pie chart in Figure 5, this is a balanced dataset.
3. **Wordclouds:** Wordclouds help us to know the frequent words used in a specific class. Wordclouds show the more frequent word in bigger font compared to the less frequent words. Figure 6 shows the wordclouds of both the classes. Negative reviews wordcloud contains words like bad, least, better, never, little etc whereas the positive reviews wordcloud contains words like good, well, really, great etc.

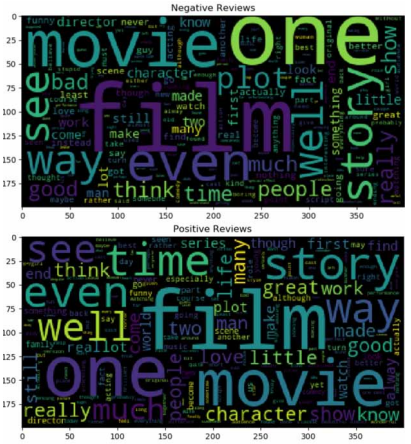
Figure 4. Summary of dataset attributes

	review	sentiment
count	50000	50000
unique	49582	2
top	Loved today's show!!! It was a variety and not...	negative
freq	5	25000

	sentiment
count	50000.000000
mean	0.500000
std	0.500005
min	0.000000
25%	0.000000
50%	0.500000
75%	1.000000
max	1.000000

Figure 5. Positive versus negative reviews

Figure 6. Word clouds for the negative and positive reviews



4.4 Dataset Pre-Processing

The steps done in pre-processing is given here: converting the text into lower case removing the non-english words and non-ascii characters from the reviews. The review is tokenized into words. Stopwords (list present in nltk python library), punctuations (list in string module) and emoticons are removed from the reviews. The words are then stemmed using PorterStemmer of nltk library, appended and finally returned as shown in Figure 7.

Figure 7. Preprocessed dataset

	sentiment	words_only
25172	1	note saw sell dead international film festival...
25173	1	one several leave usual sailor girl ensue sort...
25174	0	envy best jack black ben stiller amy great cas...
25175	1	ben jo day travel morocco holiday meet mysteri...
25176	1	love seen every single episode ever come must ...

Based on the nature of the dataset, feature extraction, dimensionality reduction, normalization and sampling are done. Increase in the number of features, increases the possibility of overfitting and results in a poor performance on the dataset. Once after the feature selection is done, the data is split into a training set and testing set following which the model is fit using an SVM classifier. The sentence level, aspect level and document level analysis is done in the forthcoming sections. The PorterStemmer class of NLTK library performs the sentence level analysis, the count vectorizer handles the work or aspect level analysis. The document level analysis is one by forming a word cloud that discerns the positive and negative aspect of the context in the document.

From the dataset summary obtained, it's found out that the dataset contains duplicate values and the duplicate values are dropped. The dataset doesn't contain null values and does not require to be handled explicitly. Since the dataset is balanced, techniques for handling imbalanced datasets like under sampling, SMOTE, oversampling etc is not applied.

4.5 Train-Test Split and Vectorization

A train-test split is done and following which the dataset is divided into two parts. 75% of the data is taken as the training part and 25% of the dataset is taken as the testing part. As machine learning models only accept numerical values as inputs, the words of reviews need to be converted to floating point or integer values using Count Vectorizer. Count Vectorizer is applied which converts the collection of word documents into vector of token counts (Figure 8).

Figure 8. Dataset after CountVectorization

```
(0, 1149)     1
(0, 1427)     1
(0, 2322)     1
(0, 2592)     1
(0, 2760)     1
(0, 2811)     1
(0, 4480)     1
(0, 4826)     1
(0, 5309)     1
(0, 5404)     1
(0, 5561)     1
(0, 5588)     1
(0, 6120)     1
(0, 7410)     1
(0, 7688)     1
(0, 8125)     1
```

4.6 Feature Selection

The results of count vectorizer produces a matrix of document-word-count. The word-count vs document vector is analysed to identify the appropriate words which can act as features and this feature selection is done using the Chi square selector test. The chi square selector identifies the relationship between the chosen word and the predictor variable. The words that contributes much to the predictor variable are taken up as best features or attributes. Chi square test builds its own hypothesis, constructs the contingency table for analysis, applies the heuristic principles and calculates the chi square statistic values. Finally it decides upon the right hypothesis either null or alternative for every individual word. Thus, this test is used to select the best attributes or features and are taken forward for training. This procedure avoids overfitting. Also, during the execution of the model, the number of attributes to be selected are varied to get the best accuracy. Alternative to chi square in literature TF-IDF, fisher's exact test, are applied. However the chi square feature selection method choses the words that are highly dependent on each other and it is a robust feature selection method.

4.7 Parameter Tuning of SVM Model

The hyperparameters of the SVM classifier i.e C, gamma and kernel are tuned by using GridSearch. GridSearch takes a set of values that each parameter can be tested for and identifies the possible combinations of the set of values. GridSearch trains the SVM classifier in pairs. It finds the (a,b) pair for each classifier, where a represents the regularization constant and b represents the kernel hyperparameter. The k-cross validation was performed to test multiple models with the same parameters. The training data is divided into k-subsets of same size. On an interim basis, one subset is tested using the classifier built on the rest of the k-1 subsets. K-cross validation overcomes the overfitting problem and finally the best score and hyperparameters drawn from the best score is found out. In section 4.5 the dataset is split into training and test sets to apply Vectorization and to perform chi square analysis. The chi square feature selection is done on the training data in order to measure the hypothesis value. This differs from the GridSearch which applies cross validation for training the SVM classifier.

4.8 Training the SVM Model

The SVM classifier is then used to build the model. The accuracy of the classifier is measured. To further improve on the metrics, bagging technique is used. Bagging generates various models having their training dataset formed by performing sampling with replacement on the main training dataset. The final prediction i.e the output class is then determined using the voting method. The base estimators chosen during execution of the bagging algorithm decides the accuracy of classification. An increase in number of base estimators will increase of probability of allocating the review to the right class. Let say there are n estimators of which m chose the review to be positive and n-m chose it to be negative them the review would be allocated to class positive if m>n-m else negatice if n-m>m.

5. EXPERIMENTS AND RESULTS

5.1 Performance Metrics Applied

The proposed model is evaluated using metrics like accuracy, precision and recall measures. The better the metrics value of the algorithm, the better the model. The four metrics used are given in Eq. 1, Eq. 2, Eq. 3, Eq. 4:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ samples} \tag{1}$$

$$Precision = \frac{True\ Positive}{True\ Positive + True\ Negative} \tag{2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

The values of evaluation metrics were improved by varying various parameters like: Dataset is trained using different train-test split ratios; the number of features selected by chi2 selector is varied; Grid Search Algorithm is used on various hyperparameter values to get the best hyperparameters; bagging is done by varying the number of estimators and finally compared with existing literature.

The performance of the model is tabulated in Table 3. The precision, recall and F1-score values indicate that the model prediction is not biased as almost all of them lie closer to each other. This signifies that the true positive value is high comparatively with other parameters.

Table 3. Performance metric value of the proposed model

Metric	Value
Accuracy of the model using SVC	0.934
Precision	0.93
Recall	0.95
F1-Score	0.94

The detailed results of the proposed model for class 0 and 1 with support values is given in Table 4. The work focusses on identifying a series of approach that increases the classification accuracy of predicting the positive reviews as positive and negative reviews as negative. However, in future the work can be extended to further distill the incorrectly predicted review samples uses other higher end machine learning algorithms.

Table 4. Resultant Metrics obtained for both the positive and negative class of reviews – from the running model

Accuracy of the model using SVC = 0.934			F1-score	Support
	Precision	Recall		
0	0.93	0.95	0.94	1065
1	0.94	0.92	0.93	1026
Accuracy			0.93	2091
Macro avg	0.93	0.93	0.93	2091
Weighted avg	0.93	0.93	0.93	2091

5.2 Train-Test-Split Ratio

Various different Train-test split ratios were tried as tabulated in Table 5. The train-test split was done using the sklearn library's function. It was done by choosing records randomly from the dataset. The training set should neither be very large nor biased towards one class compared to the testing set. This prevents overfitting. It is found that testing accuracy was highest for 75-25 percent split and slightly decreases as we increased or decreased the ratio. Thus, 75-25 percent was chosen for further analysis.

Table 5. Accuracy values in different trials of train-test split ratios

Train-Test split ratio	Accuracy (%)
70-30	92.87
75-25	93.40
80-20	93.18

5.3 Chi Square Selector Optimal K-Value

Chi square Selector was used to select the k-best features from the processed dataset. The value of k was varied to maximize the accuracy. A small value of k ignores important features and a large value of k may include unnecessary features resulting in a poor accuracy. On repeatedly building the model with different values of k, the accuracies obtained are tabulated in Table 6. A count of 2000 features yields maximum accuracy. This accuracy is further increased in forthcoming steps (Figure 9).

Figure 9. Accuracy obtained with different number of features i.e k value in Chi square selector

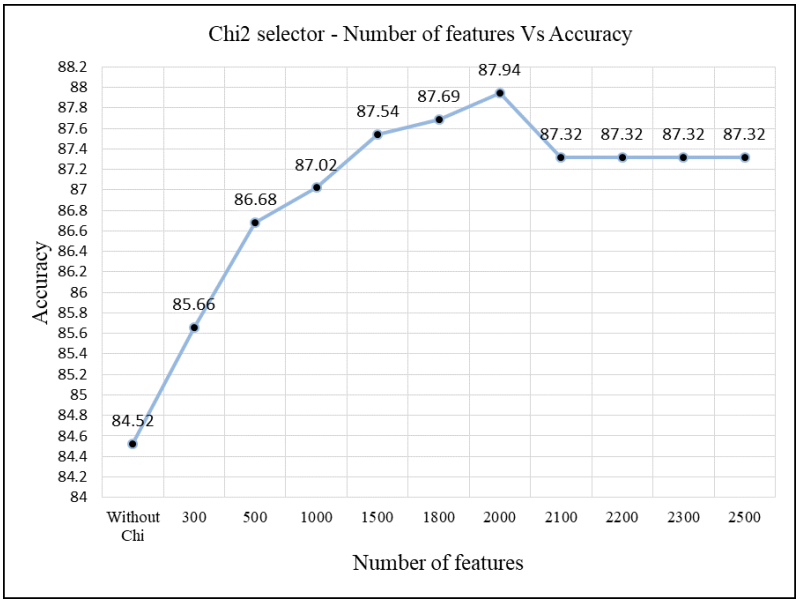


Table 6. Accuracy value with varying k value

Number of features	Accuracy (%)
Without Chi	84.52
300	85.66
500	86.68
1000	87.02
1500	87.54
1800	87.69
2000	87.94
2100	87.32
2200	87.32
2300	87.32
2500	87.32

5.4 Hyperparameter Tuning by K-Cross Validation GridSearch Method

GridSearch was used to optimize the hyperparameters of the SVM Classifier. The values given to GridSearch are given in Figure 10. GridSearch takes a set of values that each parameter can be tested for and identifies the possible combinations of the set of values. GridSearch trains the SVM classifier in pairs. It finds the (a,b) pair for each classifier, where a represents the regularization constant and b represents the kernel hyperparameter.

Figure 10. Parameters to GridSearch

```
parameters_svm = {'C': [0.1, 1, 10, 100, 1000],'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
'kernel': ['linear', 'poly', 'rbf', 'sigmoid']}
```

K-cross validation GridSearch applies the algorithm to each of the possible combinations of the parameters provided in the list parameter_svm. As seen in the Figure 11, the optimum value from the values are passed as parameters to the function, C=100, gamma-0.01 and kernel= RBF gives the best output accuracy of 92.66%. k-fold cross validates the samples in the data and the SVM classifier is applied over different folds of dataset. The normalization of the textual data is done during the hyperparameter tuning and the optimal hyperparameter extracts only the required content from the reviews and thereby reducing the dimension.

Figure 11. Output by GridSearch algorithm

```
[CV] .... C=1000, gamma=0.0001, kernel=rbf, score=0.902, total=   4.7s
[CV] C=1000, gamma=0.0001, kernel=rbf ................................
[CV] .... C=1000, gamma=0.0001, kernel=rbf, score=0.911, total=   4.7s
[CV] C=1000, gamma=0.0001, kernel=rbf ................................
[CV] .... C=1000, gamma=0.0001, kernel=rbf, score=0.905, total=   4.8s
[CV] C=1000, gamma=0.0001, kernel=sigmoid ............................
[CV]  C=1000, gamma=0.0001, kernel=sigmoid, score=0.896, total=   4.8s
[CV] C=1000, gamma=0.0001, kernel=sigmoid ............................
[CV]  C=1000, gamma=0.0001, kernel=sigmoid, score=0.903, total=   4.9s
[CV] C=1000, gamma=0.0001, kernel=sigmoid ............................
[CV]  C=1000, gamma=0.0001, kernel=sigmoid, score=0.905, total=   4.8s
[CV] C=1000, gamma=0.0001, kernel=sigmoid ............................
[CV]  C=1000, gamma=0.0001, kernel=sigmoid, score=0.904, total=   4.8s
[CV] C=1000, gamma=0.0001, kernel=sigmoid ............................
[CV]  C=1000, gamma=0.0001, kernel=sigmoid, score=0.902, total=   4.9s
[Parallel(n_jobs=1)]: Done 500 out of 500 | elapsed: 69.1min finished
0.9266490020778132
{'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}
```

5.5 Varying the Number of Base Estimators in Bagging Algorithm

The accuracy is further increased by using the bagging algorithm with the base classifier as our newly built SVM classifier. The number of base classifiers of the bagging algorithm were varied to find the most optimal solution. Bagging in general opts voting method. The predicted result of majority of classifier is considered than the result produced by marginal number of classifiers. Bagging with SVM on the other hand overcomes overfitting as if avoids increasing the weightage of incorrectly predicted reviews as like boosting.

As can be seen from the Figure 12, the best accuracy was at 90 base estimators (Table 7).

Figure 12. Accuracy achieved by changing number of base models in Bagging

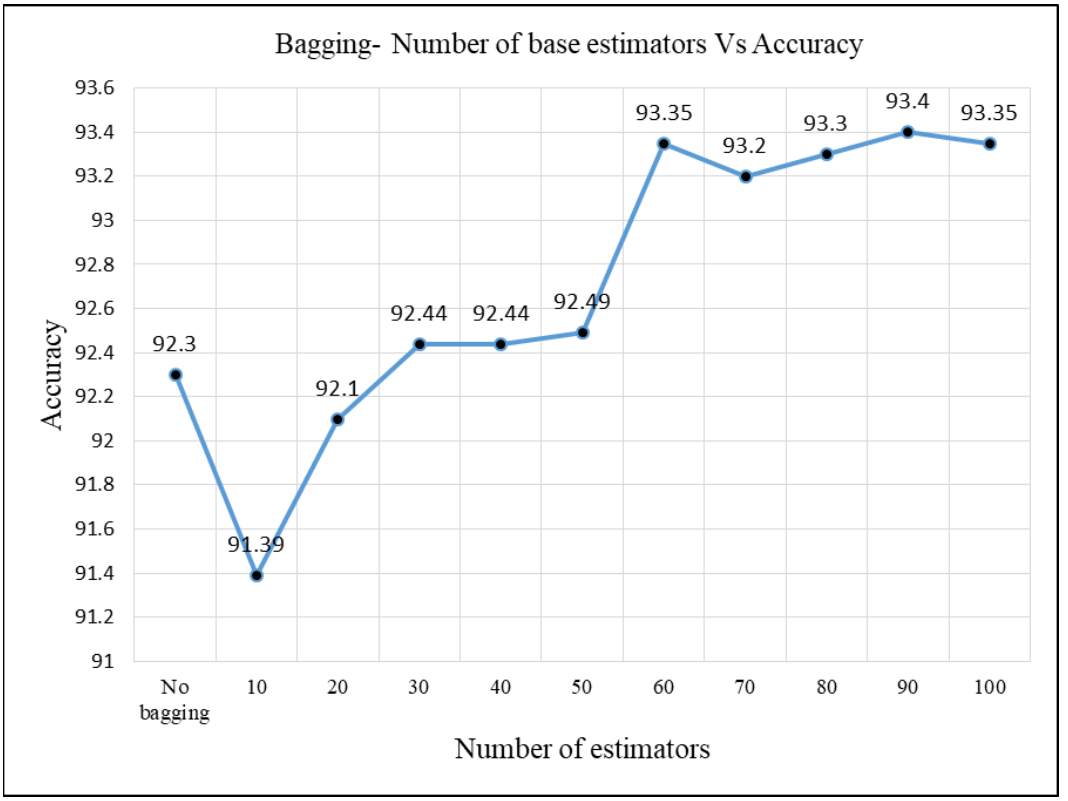


Table 7. Accuracy value with varying base models

Number of estimators	Accuracy (%)
No bagging	92.3
10	91.39
20	92.10
30	92.44
40	92.44
50	92.49
60	93.35
70	93.20
80	93.30
90	93.40
100	93.35

5.6 Comparison With Existing Literature

Comparison of accuracy with existing literature for sentiment analysis problem is tabulated in Table 8. It shows that the proposed model gives a better accuracy compared to the existing models.

(Mullen & Collier, 2004) built an SVM for sentiment analysis over a smaller music review dataset. They applied linear kernel which proved not better than other kernels as the preprocessing done was very minimal. Whereas in the proposed wok a number of preprocessing procedures were employed over the text which increased the accuracy to higher values.

(Sahu & Ahuja, 2016) built an ensemble model over rotten tomatoes review dataset. It applied a number of preprocessing procedures over the reviews. Following which a 10-fold cross validation was applied on the dataset. A number of ML algorithms were applied and finally Random forest ensemble model proved with higher accuracy level of 88.57. However, the proposed work puts together the ensemble model with SVM as a classifier, hence raising the performance still better than the former.

Though the work by (Kumari et al., 2017) applies text preprocessing procedures followed by SVM over reviews of smart phones, the NLP does not greatly handle sarcastic sentences whereas the proposed method builds the bag of clouds, trains the model using GridSearch approach and so handles even complex sentences.

The work proposed by both (Korovkinas et al., 2017) and (Ahmad et al., 2017) were not able to handle sarcastic review sentences and used SVM with NB and SVM respectively and accuracy values were minimal as compared in Table 8. (Ahmad et al., 2018) analyzed individual words in the IMDB reviews. It applied SVM using Grid Search technique. Yet, it was unable to understand and process the context of any full sentence. The proposed method applies an ensemble model which takes multiple iterations in identifying the right number of base estimators thereby recognizes the context of any complex sentence.

(Yasen & Tedmori, 2019) employed the use of 8 classifiers in processing the IMDB reviews and generated an accuracy of 87.45 using SVM. Also it does not use any specific preprocessing or feature selection methods while the proposed ones shows very good results as compared to the former.

Table 8. Comparison of accuracy with existing works using SVM

Existing Literature	Objective of the paper	Accuracy achieved (%)
(Kumari et al., 2017)	To do a sentiment analysis of smart phone product review using SVM classification technique	90.9
(Korovkinas et al., 2017)	To use Naive Bayes and SVM classifier, Ensemble for sentiment analysis	89.19
(Ahmad et al., 2017)	To perform sentiment analysis on tweets using SVM	71.2
(Ahmad et al., 2018)	To measure the performance of SVM using grid search technique	87.2
(Yasen & Tedmori, 2019)	To perform movie review sentiment analysis and classification	87.45
(Sahu & Ahuja, 2016)	To perform movie review sentiment analysis and classification	88.57
(Mullen & Collier, 2004)	sentiment analysis using SVM	86
	Proposed model	93.40

Table 9 features the articles that has taken up the IMDB dataset along with the procedure and methods adopted. Accuracy is the statistical measure that is used to predict the accuracy of the classifier in majority of the research works and the same measure is used in the proposed model for comparison purpose.

Table 9. Comparison of accuracy with existing works over IMDB dataset

Existing work and proposed work	Dataset	Procedure	Methods Adopted	Accuracy
(Ahmad et al., 2018)	IMDB	TF-IDF, Tokenization, WordsToKeep Procedure, SVM Classifier	SVM with GridSearch	87.2
(Yasen & Tedmori, 2019)	IMDB	Tokenization, Attribute Selection: Gain Ratio, Classification using classifiers	Eight Classifiers were used namely NB, BN, DT, KNN, RRL, SVM, RF, and SGD	87.45
Proposed Model	IMDB	Term presence and their frequency, Part of Speech, hyperparameter tuning, Chi2 selector	SVM with GridSearch, SVM + Bagging	93.40

An extended future work of sentiment analysis with textual data can be done by applying intent analysis procedures, context analysis procedures using fuzzy rules or linguistic rules. One of the major challenge with English text content is the numerous combination of word-sentence, the different ways of writing pattern. Also, understanding the sense or meaning or the context in which the sentence is written is arduous. Deep learning and deep neural networks can create new frontiers for sentiment analysis.

6. CONCLUSION AND FUTURE WORK

A number of machine learning techniques can be applied to identify the sentiments from textual data. The literature review is done in details in two folds: the first half characterizes the research article that uses SVM for text classification and prediction; the second half details the research articles that employs usage of machine learning algorithm in tweet analysis. This article builds a sentiment analysis model to precisely classify the reviews in IMDB dataset. The analysis of reviews requires extracting appropriate features from the reviews that contributes to accurate prediction. As an initial step, the bag of model is generated and word clouds are visualized to know the pattern of words recurring frequently in the review set. Following which, feature vector is built using Chi square selector approach that selects the needed features from the reviews. . The chi square selector identifies the relationship between the chosen word and the predictor variable. The words that contributes much to the predictor variable are taken up as best features or attributes. The selected features are subjected to SVM using GridSearch Technique. GridSearch applies K-cross validation to the SVM algorithm and tries every possible combination of the parameters to SVM. The SVM classifiers accuracy is learnt by tuning the hyperparameters. The bagging algorithm is implemented on the built model as the base estimator. The number of base classifiers of the bagging algorithm was varied to find the most optimal solution. The proposed system gives an accuracy of 93.40. It is compared with the accuracy of similar featured existing models and yields good accuracy as compared to existing ones. The model is robust and stable because of GridSearch and hyperparameter tuning. As a future direction, accuracy can be further improved by using techniques that can tune the hyperparameters with more precision.

REFERENCES

Abualigah, L., Alfar, H. E., Shehab, M., & Hussein, A. M. A. (2020). Sentiment analysis in healthcare: a brief review. *Recent Advances in NLP: The Case of Arabic Language*, 129-141.

Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment analysis of tweets using svm. *International Journal of Computers and Applications*, *177*(5), 25–29. doi:10.5120/ijca2017915758

Ahmad, M., Aftab, S., Bashir, M. S., Hameed, N., Ali, I., & Nawaz, Z. (2018). SVM optimization for sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *9*(4), 393–398. doi:10.14569/IJACSA.2018.090455

Ali, I., Saad, S., & Ahmed, S. (2016). Using One-Class SVM with Spam Classification. *Iraqi Journal of Science*, *57*(1B), 501–506.

Alnajran, N. N., Crockett, K. A., McLean, D., & Latham, A. (2018, December). A word embedding model learned from political tweets. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)* (pp. 177-183). IEEE. doi:10.1109/ICCES.2018.8639217

Amin, S., Uddin, M. I., Hassan, S., Khan, A., Nasser, N., Alharbi, A., & Alyami, H. (2020). Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 131522–131533. doi:10.1109/ACCESS.2020.3009058

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2019). Depechemood++: A bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*.

Astya, P. (2017, May). Sentiment analysis: approaches and open issues. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 154-158). IEEE.

Blasch, E., Liu, Z., Zheng, Y., Majumder, U., Aved, A., & Zulch, P. (2019, May). Multisource deep learning for situation awareness. In *Automatic Target Recognition XXIX* (Vol. 10988). International Society for Optics and Photonics. doi:10.1117/12.2519236

Cachola, I., Holgate, E., Preoţiuc-Pietro, D., & Li, J. J. (2018, August). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2927-2938). Academic Press.

Chen, Y., Lv, Y., Wang, X., Li, L., & Wang, F. Y. (2018). Detecting traffic information from social media texts with deep learning approaches. *IEEE Transactions on Intelligent Transportation Systems*, *20*(8), 3049–3058. doi:10.1109/TITS.2018.2871269

Chiroma, F., Liu, H., & Cocea, M. (2018, July). Text classification for suicide related tweets. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 2, pp. 587-592). IEEE. doi:10.1109/ICMLC.2018.8527039

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science & Technology*, *37*(1), 51–89. doi:10.1002/aris.1440370103

Dastanwala, P. B., & Patel, V. (2016, March). A review on social audience identification on twitter using text mining methods. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 1917-1920). IEEE. doi:10.1109/WiSPNET.2016.7566476

Dias, D. S., Welikala, M. D., & Dias, N. G. (2018, September). Identifying racist social media comments in sinhala language using text analytics models with machine learning. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 1-6). IEEE. doi:10.1109/ICTER.2018.8615492

Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., & Hazim, M. (2019). Halal products on Twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 83354–83362. doi:10.1109/ACCESS.2019.2923275

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), 114–124. doi:10.1177/0165551514558172

Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 181074–181090. doi:10.1109/ACCESS.2020.3027350 PMID:34812358

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733–2742. doi:10.1109/JBHI.2020.3001216 PMID:32750931

Korovkinas, K., Danėnas, P., & Garšva, G. (2017). SVM and Naïve Bayes classification ensemble method for sentiment analysis. *Baltic Journal of Modern Computing, 5*(4), 398-409.

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS One*, *10*(12), e0144296. doi:10.1371/journal.pone.0144296 PMID:26641093

Kumari, U., Sharma, A. K., & Soni, D. (2017, August). Sentiment analysis of smart phone product review using SVM classification technique. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 1469-1474). IEEE. doi:10.1109/ICECDS.2017.8389689

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting signs of depression in tweets in Spanish: Behavioral and linguistic analysis. *Journal of Medical Internet Research*, *21*(6), e14199. doi:10.2196/14199 PMID:31250832

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2(2010), 627-666.

Liu, H., Burnap, P., Alorainy, W., & Williams, M. L. (2019). A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, *6*(2), 227–240. doi:10.1109/TCSS.2019.2892037

Liu, K., & Chen, L. (2019). Medical social media text classification integrating consumer health terminology. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 78185–78193. doi:10.1109/ACCESS.2019.2921938

Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam detection in Twitter. *IEEE Transactions on Computational Social Systems*, *5*(4), 973–984. doi:10.1109/TCSS.2018.2878852

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. doi:10.1016/j.cosrev.2017.10.002

Mendes, L. F., Gonçalves, M., Cunha, W., Rocha, L., Couto-Rosa, T., & Martins, W. (2020, October). "Keep it Simple, Lazy"--MetaLazy: A New MetaStrategy for Lazy Text Classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1125-1134). ACM.

Mohammed, Z., JA, M. F., MP, M. I., Basthikodi, M., & Faizabadi, A. R. (2019). A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm. Academic Press.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS One*, *11*(5), e0155036. doi:10.1371/journal.pone.0155036 PMID:27149621

Mullen, T., & Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 412-418). Academic Press.

Pentheroudakis, J. E., Bradlee, D. G., & Knoll, S. S. (2006). *U.S. Patent No. 7,092,871*. Washington, DC: U.S. Patent and Trademark Office.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). *Meld: A multimodal multi-party dataset for emotion recognition in conversations.* arXiv preprint arXiv:1810.02508.

Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2017, December). Identifying content originator in social networks. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1-6). IEEE. doi:10.1109/GLOCOM.2017.8255074

Razzaq, A., Asim, M., Ali, Z., Qadri, S., Mumtaz, I., Khan, D. M., & Niaz, Q. (2019). Text sentiment analysis using frequency-based vigorous features. *China Communications*, *16*(12), 145–153. doi:10.23919/JCC.2019.12.011

Sahu, T. P., & Ahuja, S. (2016, January). Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)* (pp. 1-6). IEEE. doi:10.1109/MicroCom.2016.7522583

Tocoglu, M. A., Ozturkmenoglu, O., & Alpkocak, A. (2019). Emotion analysis from Turkish tweets using deep neural networks. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 183061–183069. doi:10.1109/ACCESS.2019.2960113

Voorhees, E. M. (1999, July). Natural language processing and information retrieval. In *International summer school on information extraction* (pp. 32–48). Springer. doi:10.1007/3-540-48089-7_3

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*. doi:10.3115/992424.992434

Xia, Z., & Deng, J. (2020, August). Application and Research of Spam Classification Based on Cluster Intelligence Algorithm to Optimize SVM. *Journal of Physics: Conference Series*, *1617*(1), 012050. doi:10.1088/1742-6596/1617/1/012050

Yasen, M., & Tedmori, S. (2019, April). Movies Reviews sentiment analysis and classification. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 860-865). IEEE. doi:10.1109/JEEIT.2019.8717422

Zhang, Z., Hou, R., & Yang, J. (2020). Detection of Social Network Spam Based on Improved Extreme Learning Machine. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 112003–112014. doi:10.1109/ACCESS.2020.3002940

A. Razia Sulthana completed her Bachelor of Technology (B.Tech) in Information Technology from Jaya Engineering College, affiliated under Anna University, TamilNadu, India. She holds 46th Rank under Anna University (all over Tamil Nadu). She then pursued her Master of Engineering in Computer Science from S.A Engineering College, affiliated under Anna University, TamilNadu, India. She holds 1st Rank in Anna University (all over Tamil Nadu). She obtained Master of Business Administration Degree from University of Madras. She then completed her Doctor in philosophy in Computer Science from SRM Institute of Science and Technology. She possesses 9 years of teaching experience in teaching/Research. She is currently working with BITS Pilani, Dubai Campus in the department of Computer Science. She has published a number of research articles in reputed journals and in conferences. Her area of interest includes, Machine Learning, Recommendation System, Data Mining, Big Data, Data Structures and Ontology. She has guided more than 50 projects for undergrads and postgrads in her domain. She has got best paper award in a couple of conferences. She has organized international symposium and workshops. She gave invited lectures over 'Machine learning Paradigms' twice with undergrad students with different universities. She owns IBM Explorer Badge in Artificial Intelligence. A number of certifications is made by her with Microsoft and Cambridge. She has published one patent and 2 book chapters. Scopus Id: Razia Sulthana. ORCID: https://orcid.org/0000-0001-5331-1310. Google Scholar Link: https://scholar.google.com/citations?user=UM_rgP8AAAAJ&hl=en.

A. K. Jaithunbi, B.E., M.E., Ph.D., is Assistant Professor in Department of Computer Science and Engineering. She completed her B.E(CSE) and M.E(CSE) from to Anna University. She obtained her Doctor of Philosophy under the Faculty of Information and Communication Engineering at Anna University, Chennai and the Title of the Thesis is "Secure and Efficient Data Sharing with Integrity Preservation in Public Cloud Storage". She has been in the teaching profession for the past 18 years and has handled both UG and PG programmes. Her areas of interest include Data mining and Cloud Computing. She has participated various Workshops like NS2, OOAD, Intellectual property rights, FOSS etc., She has attended the Mission10x Wipro Tech Ltd, was certified by Dale Carnegie, MxLA training program and also Cambridge Certified. She was the resource person for the FDP to handle the concepts of Data Structure and Overview of the File Subsystem, the Processes concept. She has made a number of publications in National and International journals. She is a active member of many societies like CSI, IAENG etc., She has attended 12 workshops, 20 webinars and 110 FDPs related to her areas of interest.

Haritha Harikrishnan completed her Bachelors of Engineering in Computer Science from Birla Institute of Technology and Science, Pilani, Dubai Campus in 2020. She's currently working as a Research Assistant at the University of Dubai, UAE. Her areas of research involve machine learning and image processing.

Vijayakumar Varadarajan holds a diploma, a BE in Computer Science and Engineering, and an MBA in Human Resources and Development, all with First Class Honours. He also earned an ME in Computer Science and Engineering with a First Rank Award and a PhD in Computer Science and Engineering from Anna University in 2012. Having served as a Professor and Associate Dean for the School of Computing Science and Engineering at VIT University in Chennai, India, he currently hold positions as an Adjunct Professor in the School of Computer Science and Engineering in University of New South Wales in Australia, a Visiting Professor at Kyungpook National University in the Republic of Korea, and a visiting Post Doc Scientist at the Universidade Federal do Piauí in Brazil. He is the Lead Advisor for Data Science Projects and Research Collaborations for BriteYellow Private Ltd in India, and a Cloud Computing Consultant for MIT Square in the United Kingdom. As the team lead for several years in industries such as Satyam, Mahindra Satyam, and Tech Mahindra, he has led research projects and collaborations. Dr. Varadarajan has published articles in national and international journals, made numerous conference presentations, peer-reviewed, and edited chapters in scholarly books and served as a reviewer for IEEE Transactions, Inderscience and Springer Journals. He also initiated many international research collaborations, started joint research projects between VIT University and several industries, and served as a guest editor for journals such as Inderscience, Springer, and IGI Global. He pursued numerous research interests in grid computing, cloud computing, computer networks, cybersecurity and big data and organised several international conferences and special meetings at multiple locations. He is also a member of several national and international learned and professional societies. The focus of Dr Varadarajan's career has been in academia, but he also consults and conduct research for private-sector enterprises. For over 20 years now, he has generated critical contributions to research and development sector through advising organisations, teaching international students, collaborative research, and publications.