AN EFFICIENT MACHINE LEARNING APPROACH FOR TWITTER SENTIMENTAL ANALYSIS FOR LOW CLASSIFICATION ERROR RATES

Shahid Ul Islam[1], Dr. Pooja Sharma[2]

[1]Dept. of Computer Science and Engineering, I.K. Gujral Punjab Technical University

Kapurthala Road Jalandhar-144603, India.

[2]Assistant Professor Dept. of Computer Science and Engineering, I.K. Gujral Punjab Technical University Kapurthala Road Jalandhar-144603, India.

---***---

Abstract: Sentiment analysis is one of the popular fields in the data mining process in which the individual sentiments are analysed and evaluated based on the actions and reactions. In recent days a lot of work is done for the sentiment analysis which is used to achieve appropriate accuracies of the machine learning models which needs low classification error rates but still the problem is not yet resolved which needs to be taken care because the selection of appropriate model is also one of the important task in the statistical analysis of every model. This paper put some light on the twitter based sentiment analysis through which the sentiments are classified as the positive or negative sentiment and on the basis of which the performance will be evaluated and analysed. The research work shows the classification models of different classifiers i.e. Random Forest, Decision Trees, Naïve Bayes and Support Vector Regressions. It can be seen from the result and discussion that the random forest and decision tree performance is near about same in terms of sensitivity and specificity which should be high for the efficient performance of the sentiment classifications.

Key Words: Sentiment Analysis, Twitter, Polarity, Text, Classification.

I. INTRODUCTION

Sentiment analysis is a process of text mining which categorizes and extracts subjective data in source material, and serving a business to recognize the social sentimentality of their brand, invention or service while checking online discussions. Yet, study of social media is usually controlled to just basic feeling analysis and tally metrics based on Contextual Semantic Search [1]. Now this is where effects get really motivating. To derive unlawful insights, it is imperative to understand what feature of the product a user is discussing approximately. For illustration: Amazon would need to segregate communications that related to: late transfers, billing matters, promotion linked queries, product evaluations etc. Additionally, Starbucks would need to classify communications based on whether they recount to control behaviour, new coffee smacks, hygiene reaction, online guidelines, store name and position etc. [2]

Sentiment analysis is the development of detecting constructive or negative emotion in text. It's repeatedly used by productions to detect feeling in social data, brand reputation, and understand businesses. Since trades express their views and feelings more willingly than ever earlier, sentiment analysis is suitable to monitor and realize that sentiment. Automatically investigating customer comment, such as opinions in examination responses and social media discussions, allows varieties to learn what makes businesses happy or upset, so that they can modify products and amenities to meet their clients' needs [3].

1.1 Sentiment Analysis Types

Sentiment analysis representations focus on polarization but also on emotional state and emotions, urgency and even objectives. Dependent on how a person wants to interpret purchaser feedback and questions, you can outline and tailor your groupings to meet your sentiment study needs. In the intervening time, here are most widespread categories of sentiment breakdown:

1. Fine-grained Sentiment Analysis

If divergence is important to corporate, we might study expanding our polarity groupings to comprise:

- Very optimistic
- Confident
- Nonaligned
- Destructive
- Very destructive

This is frequently raised to as fine-grained sentiment study, and could be used to take full ratings in an appraisal, for example:

Very Constructive = 5 stars
Very Destructive = 1 star

2. Emotion exposure

These sentiment goals to detect sentiments, like happiness, hindrance, anger, grief, and so on. Many emotion recognition systems use dictionaries or difficult machine learning systems.
One of the problems of using monolingual dictionary is that people direct emotions in diverse ways. Some arguments classically express irritation which might also express joy. [4][5]

3. Aspect-based Sentiment Analysis

Generally, when analysing thoughts of texts, let's say manufactured goods reviews, you'll want to distinguish particular characteristics or features individuals are mentioning in a constructive, neutral, or undesirable way.

4. Multilingual sentiment analysis

This type of analysis of the sentiment can be hard. It contains a lot of pre-processing and properties. Most of these properties are available operationally while others need to be produced such as translated quantities or noise discovery processes but we need to distinguish then and also how to use them.
Otherwise, it could detect morphological texts automatically with language classifier, and then train a tradition sentiment analysis prototype to classify manuscripts in the any language [6].

Fig 1: Different Sentiments

A. Sentiment Analysis Significance

Sentiment analysis is enormously imperative because it helps productions to grow faster and understand the overall thoughts of their clients. By repeatedly sorting the emotion behind evaluations, social media discussions, and more, we can make quicker and more precise decisions.
It's projected that 90% of the ecosphere's data is shapeless, in other words it's disorganized. Huge volumes of shapeless business data are fashioned every day: emails, funding tickets, discussions, social media talks, studies, trainings, brochures, etc.). But it's tough to analyze for emotion in a timely and efficient fashion [12].

1.2 Sentiment Analysis Benefits

1. Sorting Data

Can we visualize manually sorting done thousands of cheeps, customer support discussions, or investigations? There's just too considerable business information to practice manually. Sentiment study helps industries process huge quantities of data in an effectual and cost-effective method [13].

2. Real-Time Investigation

Sentiment analysis can classify critical problems in real-time, for illustration is a PR crisis on public media increasing? Is an angry consumer about to agitate? Sentiment analysis representations can help you instantaneously identify these categories of conditions, so we can take action accurately [14].

3. Consistent measures

It's appraised that people only agree about 60-65% of the period when influential sentiment of a precise text takes place. Tagging version by emotion is highly personal, influenced by private experiences, feelings, and principles. By using a consolidated sentiment analysis classification, companies can relate the same principles to all of their statistics, helping them recover accuracy and increase better understandings [15][16].

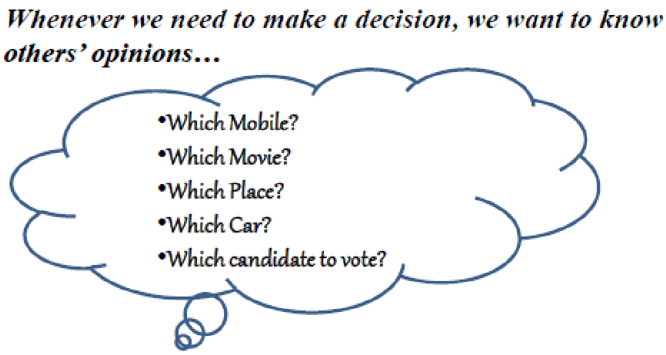


Fig 2: Opinion Mining

2. RELATED WORK

This section covers various related works for the sentimental analysis and opinion mining using machine learning processes. Munezero et al. [7] in their research paper proposed a word correction system based on Twitter feeds. This looks into ill formed or out of vocabulary (OOV) words such as "sweeet" instead of "sweet", abbreviations and slangs such as "imo" for "in my opinion", "b4" for "before" and corrects them. Out of the 449 tweets analyzed, 254 consisted of such OOV words. With the ill formed words that were analyzed, they found out that 72.44% of the ill formed words in tweets had words with either missing or extra letters, 12.2% of the words were slang and 5.12% of the ill formed words were specified in either numbers or a combination of letters and numbers.

K. Sailunaz and R. Alhajj [8] experimented on a Twitter dataset to detect and analyze sentiment and emotions demonstrated by people in their twitter posts and use them for generating results by applying different machine-learning algorithms. They collect tweets and replies on some keyword search topics and created a dataset with username, text, emotion, sentiments. They compared the results of simple positive-negative word counting, Naïve Bayes, Support Vector machine, and Random Forest classifiers are used. The Naïve Bayes machine learning classifier performed better than the other methods in terms of accuracy for detecting correct sentiment. For ISEAR dataset, the accuracy level of classification was higher than that of the twitter dataset. The dissemination of sentiments in dataset was another factor for lower accuracy. Unlike this approach, emoticons in Twitter text were used for emotion analysis, where the author discussed sentiments and emotions in detail while mentioning existing emotion models. Under these three classifiers, Naïve Bayes gives the higher accuracy for both full text and NAVA words for both sentiments and emotions of the Dataset. S. Elbagir & J. Yang et al. [9] proposed a Twitter Sentiment Analysis based on Ordinal Regression using Machine Learning Classifiers. The Dataset created by using API and then performed pre-processing to remove noise such as using NLP. The tweets or text are classified into High Positive, Moderate Positive, Neutral, High Negative, and Moderate Negative. Afterward, feature extraction performs export sentiment-relevant features using term frequency-inverse document frequency (TF-IDF). Finally, the model trained using machine learning approaches such as Support Vector Regression, Decision Tree, Random Forest and Multinomial Logistic Regression (SoftMax). The proposed framework performed sentiment analysis using Scikit-Learn for experiments, is an open source of ML software packages in Python. Under these four classifiers, Decision Tree gives the higher accuracy; the overall accuracy obtained by the 10-fold cross validation is used in each of the classifiers. K. Garg et al. [10] examined several steps for sentiment analysis on Twitter data of TV Show "Maan Ki Baat" using lexicon learning algorithms. They also delivered details of the planned approach for sentimentality. The approach collected data and then preprocessed tweets using term frequency technique and convert created two files into unigrams occurred that helps in understanding the human emotions, views, and thoughts and assigned Positive and Negative values. Afterward, feature extraction using tokenization was performed to export sentiment-applicable features. Finally, a model was trained, Sentiment analysis using Term Frequency approach performed best results. Kumar, Garg et al. [11] scrutinized several phases for sentiment study on Twitter data by machine learning processes. They also provided details of the proposed approach for sentiment analysis. The approach collected data and then preprocessed tweets using NLP-based techniques. A model was trained using machine learning classifiers, such as naive Bayes classifiers, support vector machine (SVM), and decision tree. The proposed framework performed sentiment analysis using multinomial naive Bayes and decision tree algorithms. Results showed that decision trees perform effectively, showing 100% accuracy, precision, recall, and F1-score.

3. PROPOSED WORK

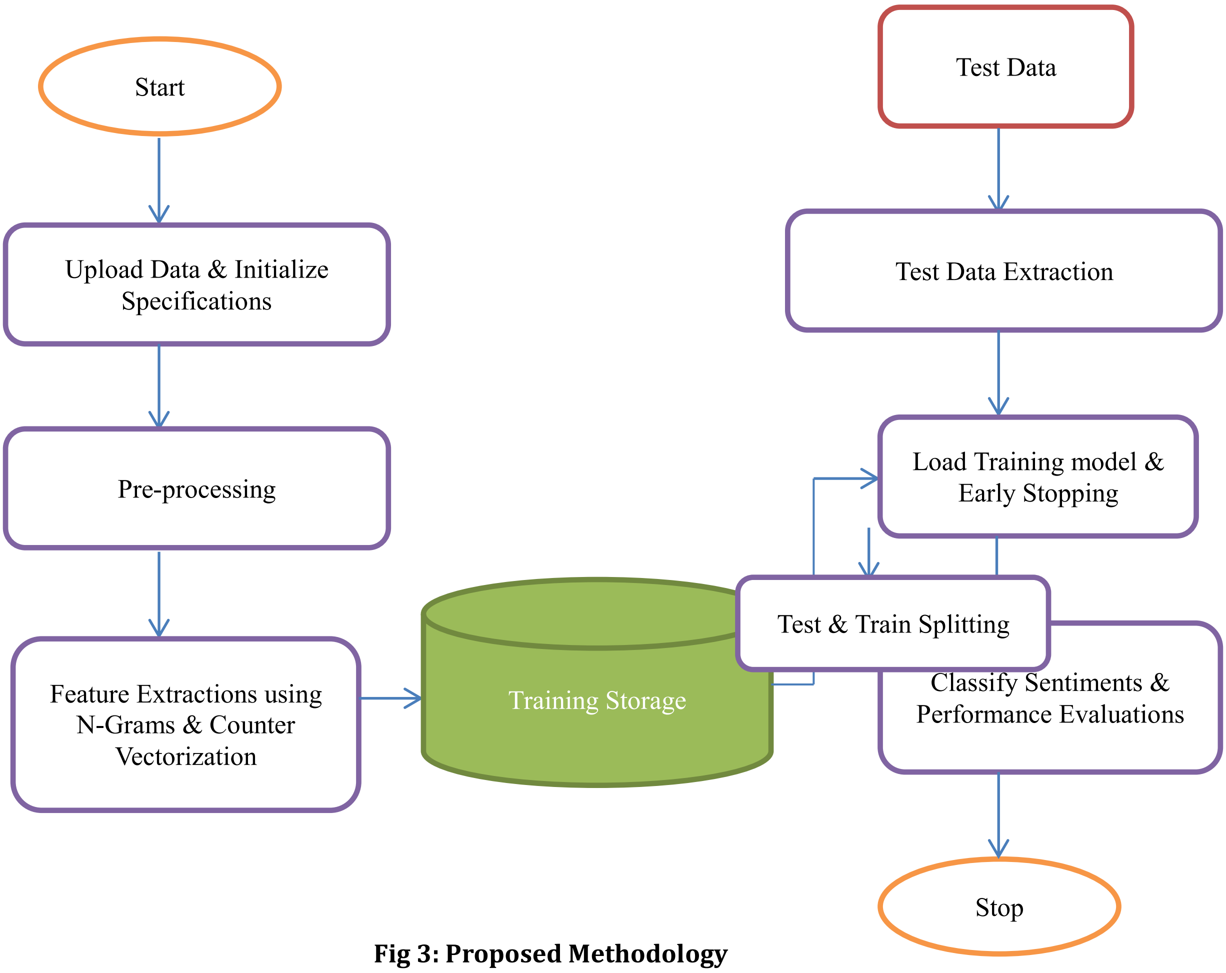


Fig 3: Proposed Methodology

4. RESULT & DISCUSSIONS

The proposed work deals with the sentiment analysis include the comparison among different classifiers. The classifier includes Random forest, Decision tree, Naive Bayes and Support Vector Regression. The performance is evaluated using accuracy, precision, recall, mean square error rate and mean square error rate. In the proposed approach the sentimental analysis is done in an efficient way that the classification error rates are low and recognition accuracy is high.

The fig 4 shows the classification model performance for the detections of sentiments. The detections are made in terms of positive and negative sentiments. It can be seen from the fig 3 that the random forest classifications are having high counts in terms of the detection rate and are having high capability because of its ensemble learning process. The better performance is achieved by the decision tree which is achieve high true positive rates and true negative rates to achieve low loss functions to perform high detections of accurate sentiments.

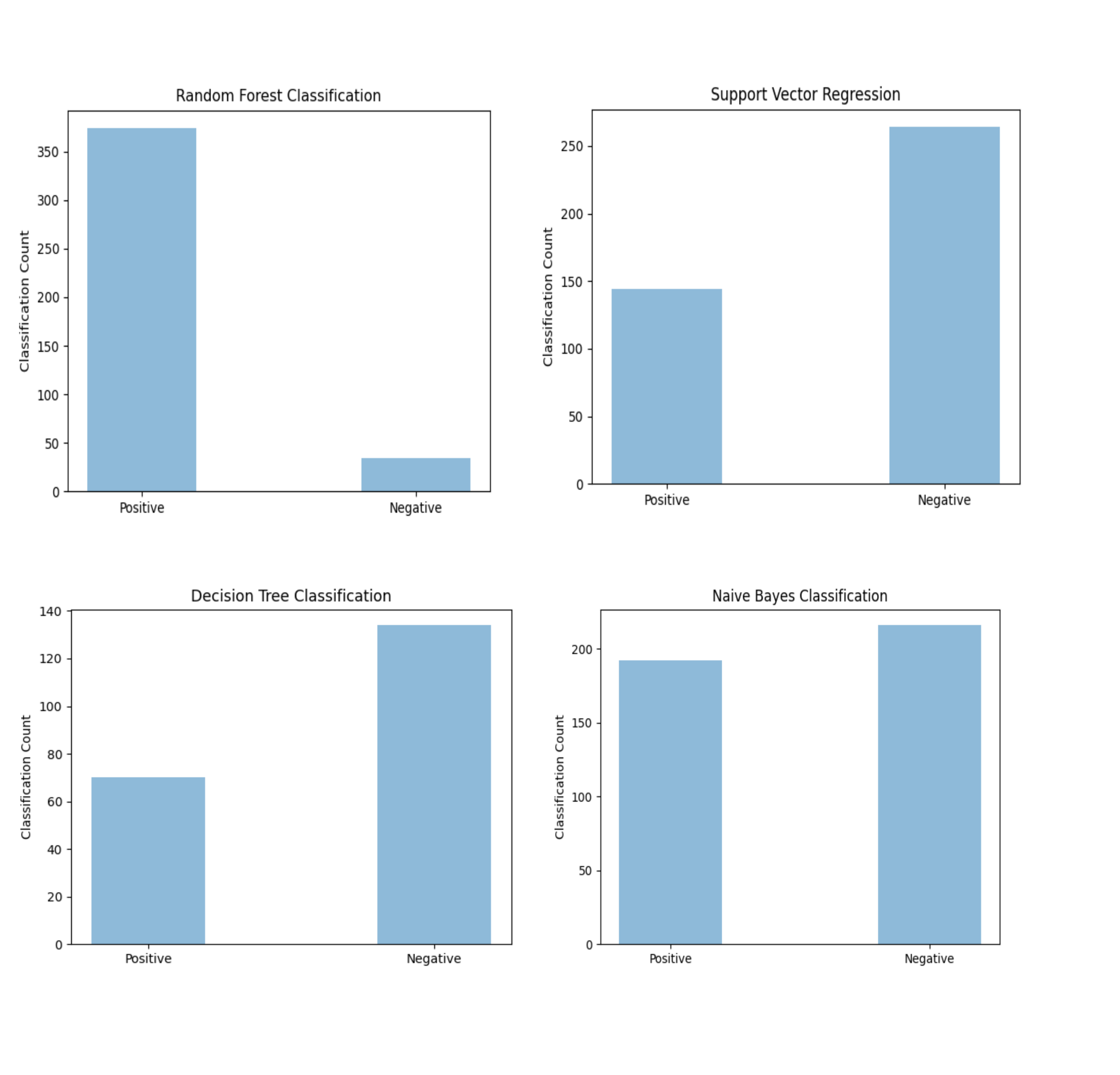


Fig 4: Implemented Classifications

The fig 5 shows recall performance of all the implemented classifiers and it can be seen from the figure that the decision random forest and decision tree is achieving sufficient recall rates to have high retrieval information for the accurate classifications for the high positive rates. This shows that the proposed approach is achieving high information gaining process which increases the high sensitivity of the data to increase classification rates.

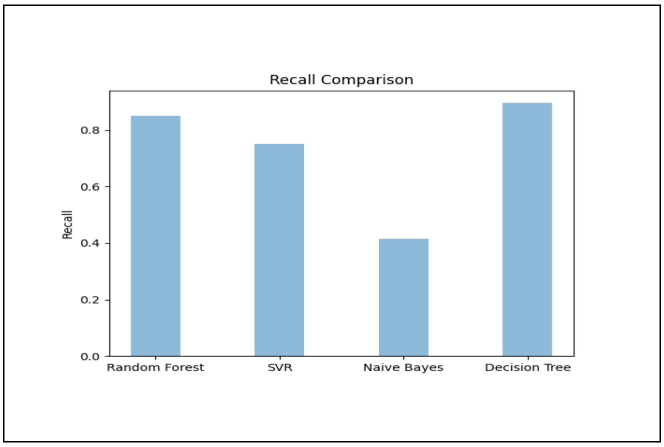


Fig 5: Recall Comparison

The fig 6 shows precision performance in terms of comparative analysis through which it can be noticed that the specificity of the decision tree and random forest are too close to each other and is having minute difference to achieve high true positive rates. The precision must be high which shows the relative instances for the high classifications of the positive and negative sentiments and polarities.

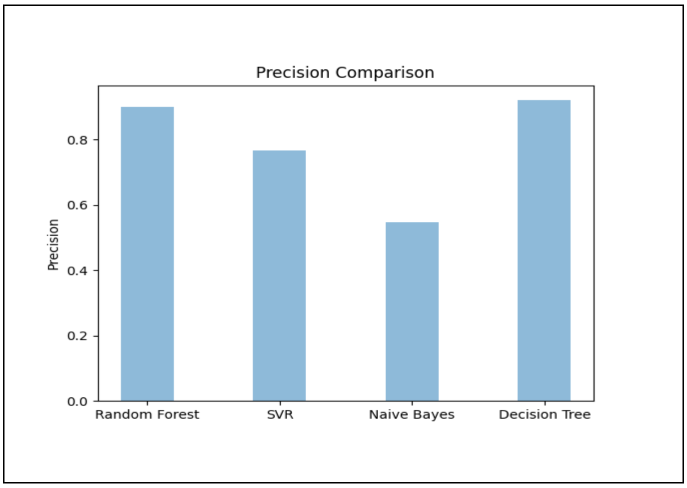


Fig 6: Precision Comparison

The fig. 7 shows accuracy performance in terms of comparative analysis among different classifiers and it can be seen that the decision tree is achieving high accuracy for the detection of the sentiments. It is also noticed that the recognition accuracy of the random forest is having minute difference from the decision tree. Also the naive Bayes depends on the probability distributions and are fluctuated which degraded the performance in terms of the classification rates. The accuracy must be high for the sentiment analysis in terms of positive and negative sentiments.

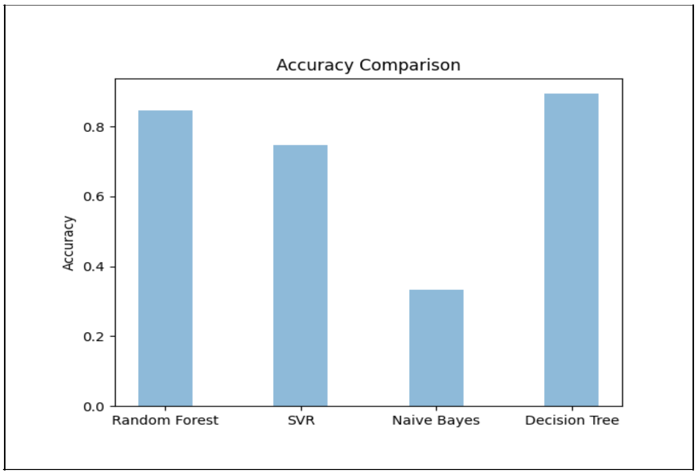
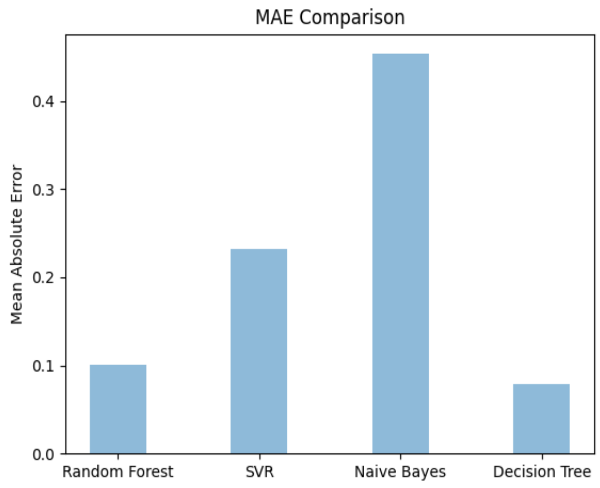


Fig 7: Accuracy Comparison

Fig 8: MSE and MAE Comparison

MSE and MAE fig.(8) are the popular error terms in the statistical analysis and must be low for the high accuracy rates. The proposed approach shows the comparative analysis of the negative and positive sentiment. MSE and MAE show the difference among the true predictions with the actual predictions and these should be low for high classification accuracies. It can be seen that the decision tree and random forest are outperforming classifiers for the high recognitions. It can be seen that the naive Bayes is not performing well than other classifiers because of under fitting or over fitting which can be controlled through regularization techniques.



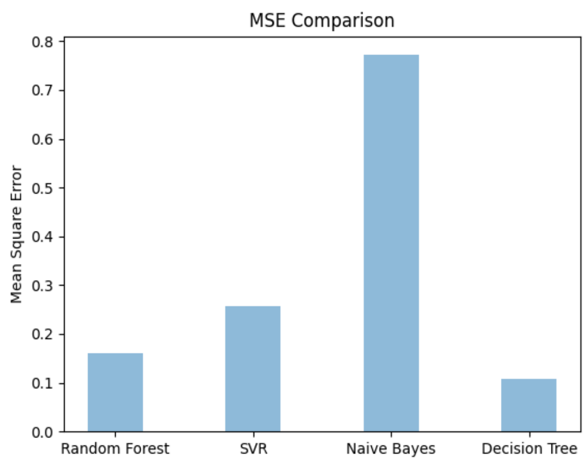


Table 1: Performance Comparison

Classifiers	Accuracy	MSE	MAE
Random Forest	0.93	0.131	0.112
SVR	0.78	0.278	0.245
Decision Tree	0.92	0.134	0.459
Naïve Bayes	0.39	0.781	0.115

Classifiers	Accuracy	MSE	MAE
Random Forest	0.93	0.131	0.112
SVR	0.78	0.278	0.245
Decision Tree	0.92	0.134	0.459
Naïve Bayes	0.39	0.781	0.115

Table 1 shows the different performance analysis of the classifiers and is compared in terms of mean square error rate, mean absolute error rate and accuracy. It can be seen that the decision tree and random forest are closely related to each other in performance evaluations which is the part of the proposed works and is achieving low error rates also. Achieving low error rates or loss functions is the objective of the proposed work in the classification of the opinions and sentiments.

5. CONCLUSION & FUTURE SCOPE

Social media is one of the biggest platforms to express the sentiments for the various contents that are posted on a daily basis. So there are both positive and negative sentiments on the social media platforms in bulk which needs to be analysed in an efficient way. The proposed study aims on comparative analysis of different machine learning algorithms in terms of high accuracy, precision, recall, mean square error rate and mean absolute error rate. The parameters are evaluated in the normalized form. The result shows that the decision tree and random forest are achieving nearly same accuracy for the classification of the sentiments and high true positive and negative rates. It can also be noticed that the Support Vector regression is performing not bad but having low performance than decision tree (0.927) and random forest (0.918). The naive bayes is having less accuracy among all the classifiers for the sentiments recognitions.
The future scope of the research work can be the instance selections for the feature optimizations and performing validations on the data to increase the classification accuracy and reduce the over fitting and under fitting of the model so that less false positive and false negatives will be achieved.

6. REFERENCES

[1] Mäntylä, M. V., Graziotin, D.Kuutila, M., The evolution of sentiment analysis—A review of research topics, venues, and top-cited papers. Computer Science Review, (2018). 27, 16-32.

[2] Yuan, J., Mcdonough, S., You, Q., Luo, J. Sentribute: image sentiment analysis from a mid-level perspective. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, 2013. (p. 10). ACM.

[3] Cao, D., Ji, R., Lin, D., Li, S. Visual sentiment topic model based microblog image sentiment analysis. Multimedia Tools and Applications, 2016. 75(15), 8955-8968.

[4] Katsurai, M., Satoh, S. I. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016. (pp. 2837-2841). IEEE.

[5] Zhang, Y., Shang, L., Jia, X. Sentiment analysis on microblogging by integrating text and image features. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2015. (pp. 52-63). Springer, Cham.

[6] You, Q., Luo, J., Jin, H., Yang, J. *Joint visual-textual sentiment analysis with deep neural networks.* In Proceedings of the 23rd ACM international conference on Multimedia, 2015. (pp. 1071-1074). ACM.

[7] M.D. Munezero, C.S. Montero, E. Sutinen, J. Pajunen, Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text, IEEE Trans. Affect.Comput.5(2)(April 2014) 101–111.

[8] K. Sailunaz and R. Alhajj “Emotion and Sentiment analysis from Twitter Text,” Journal of Computational Science 36 (2019) 101003, pp. 1-18.

[9] S. Elbagir & J. Yang, “Twitter Sentiment Analysis Based on Ordinal Regression,” IEEE Access, vol. 7, pp. 163677–163685, November 20, 2019.

[10] K. Garg, “Sentiment Analysis of Indian PM’s ‘Mann Ki Baat’,” Int. j. inf. Technology, 29 June 2019.

[11] A. Kumar, G. Garg, “Sentiment analysis of multimodal twitter data,” Multimedia Tools and Applications, 78 pp. 24103–24119, 2019.

[12] Singh, Tajinder, and Madhu Kumari. "Role of text pre-processing in twitter sentiment analysis." Procedia Computer Science 89, no. Supplement C (2016): 549-554.

[13] Soleymani, Mohammad, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. [14] "A survey of multimodal sentiment analysis." Image and Vision Computing 65 (2017): 3-14.

[14] Korkontzelos, Ioannis, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H. Gonzalez. "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts." Journal of biomedical informatics 62 (2016): 148-158.

[15] Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. "A review of feature extraction in sentiment analysis." Journal of Basic and Applied Scientific Research 4, no. 3(2014):181-18.