IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Twitter sentiment analysis using hyper tuned machine learning models

Aakash Saraf

Purdue University
West Lafayette, Indiana

Abstract – ***Social media is playing a vital role in communications, and the usage of social media among people has increased dramatically. This growth has paved the way for increased research on sentiment analysis, which helps the individuals and institutions to know about the sentiment on a product, events, topics, politics and more. The research is carried out aiming for sentiment analysis, recommendation systems etc. This paper focuses on sentiment analysis on Twitter posts with the help of machine learning (ML) models. This paper focuses on using different ML models for sentiment analysis. Natural Language processing (NLP) techniques are used in pre-processing by vectorizing the data. In specific, TF-IDF Vectorizer is used. Experimental results showed ML models are more reliable for sentiment analysis. The twitter sentiment classification is performed using algorithms includes Support Vector Machine (SVM), Random Forest, Naive Bayes, XGBoost, and Decision Tree.***

Key Words**:** Social media, Machine learning, Classification, Natural Language processing, Support Vector Machine (SVM), Random Forest, Naive Bayes, XGBoost, and Decision Tree

1.INTRODUCTION

With the emergence of social media such as Facebook, Twitter and Instagram these days people are sharing information and posting their thoughts on social media, which helps institutions and businesses to understand and grow their reputation. In the past decade, the content shared on social media is huge and benefits the researchers to emerge on opinion mining, sentiment analysis and recommendation systems. Individuals utilize social media platforms to express their emotion, opinion on products, events, places and other different contexts. Examining the sentiments of social media posts open gates for further research across various domains such as product evaluation, personality assessment, marketing strategies, business insights and more. Sentiment analysis involves extracting the social media posts and categorizing them as positive, negative and neutral sentiments.

Sentiment analysis

Sentiment analysis involves extracting the social media posts and examining them for sentiment, this is achieved using Natural language processing (NLP) techniques. The pre-processing of text posts from social media are converted to vectorizer by any popular vectorizing techniques like Term Frequency -Inverse Document Frequency (TF-IDF), count vectorization, Word2Vec etc. The polarity of each post is arrived from which the sentiment of posts are categorized as positive, negative or neutral. There are a huge number of applications relying on sentiment analysis, which includes brand monitoring and enhancement, customer satisfaction analysis and improvement, arriving business strategies, market analysis and more.

Artificial intelligence is used nowadays in a wide range of real world applications as it can capture the hidden patterns effectively. Artificial intelligence (AI) is revolutionizing in many real-world applications, sentiment analysis is one of the area, in which its uses are tremendous including applications such as social media monitoring, market research and feedback analysis to improve business, political analysis and opinion monitoring, finding financial sentiment and arriving investment decisions, healthcare industry to identify patient's sentiment and more. AI techniques such as Machine learning (ML) and deep learning (DL) has significantly advanced sentiment analysis, which enables organizations across diverse industries to extract and understand valuable information from social media posts shared by the users and this helps organizations to make data-driven decisions. As AI techniques continue to evolve, sentiment analysis remains a powerful analysis for making data driven decisions.

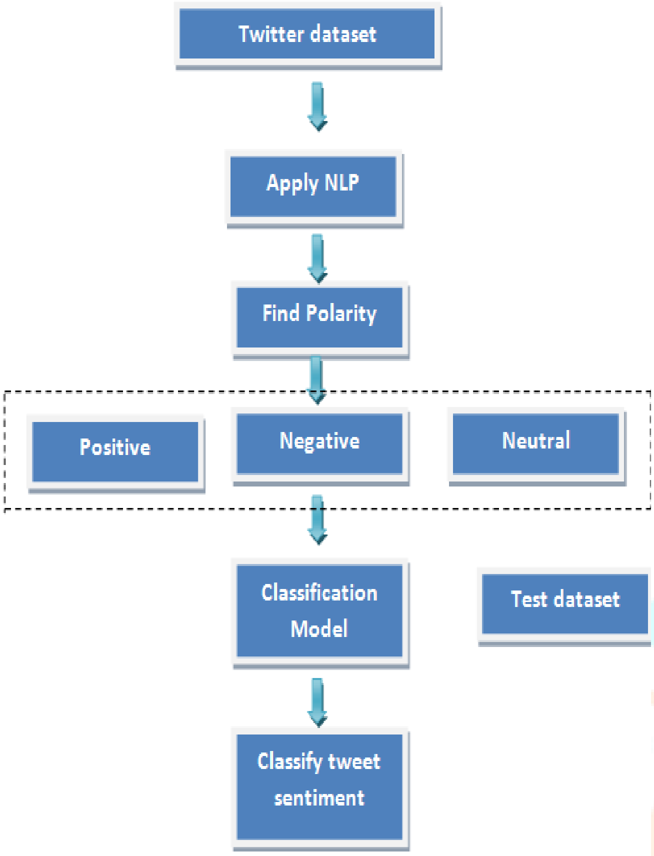

Figure: 1 Overview of Sentiment Classification

The above figure is the overview of Twitter sentiment Analysis. The figure shows that the dataset is extracted from twitter through API, which requires the consumer key and token for authentication and extracting real time tweets. The tweets are further analyzed through Natural Language Processing (NLP) and using polarity values of the tweet, they are categorized as positive and negative sentiments. The ML models are applied for classification as shown in the figure.

Many research on Twitter sentiment is available including fuzzy methods, some other traditional methods have not considered the features completely. Thus an effective and accurate model for sentiment analysis is highly required to fulfill the requirements of these problems. In this work, to overcome these problems, the classification of tweets sentiments are proposed with machine learning models. The contributions of this work includes the following

o Real time data from twitter is web scraped for live data analysis, this ensures the model the updated.

o Implementing machine learning models such as DecisionTree Classifier model (DT) and Randomforest Classifier (RF), Naive Bayes, Support Vector Machine (SVM) and XGBoost (XGB)

o Grid search cross validation (GSCV) technique is used for hyper tuning the parameter model to find the best parameter.

o Evaluating accuracy and error metrics for implemented algorithms

o Generate and optimized accuracy model using novel learning models.

In the upcoming chapters include literature review and work related to Sentiment analysis in Chapter 2. Under chapter 3, proposed algorithms and models are elaborated. In chapter 4, results and evaluations are discussed. In chapter 5, this work is Conclusion and further enhancements are discussed.

2. RELATED WORK

Sentiment analysis is currently gaining attraction due to sharing of content online has increased and thus sentiment analysis helps in extraction of useful information from data shared online. There are several researchers studied on Twitter sentiment analysis using various techniques including machine learning models. Some of these works related to sentiment analysis are discussed in this section.

The work [1] addressed the sentiment classification on Egyptian telecom related tweets with a labeled dataset. This work used three ML models SVM, Naive Bayes and maximum Entropy. The ensemble models were proposed considering these ML models as base classifiers and combining the other ML models. Experimental results showed that the ML classifier SVM has gained the highest accuracy of 78.45%. The ensemble model Random forest classifier has achieved the highest accuracy of 80.7%. It was observed from the experiments that the ensemble model has performed better than the normal ML classifier.

A novel sentiment majority voting classifier (SMVC) was proposed for identifying deepfake tweets. This work [2] combined the Long Short Term Memory (LSTM) and Decision tree model for embedding the features and gets outputs features which is used by the classification models SVM, K-nearest neighbor, Logistic regression and LSTM models. This work also performed cross validation for hyper parameter tuning. Experimental results showed that the Logistic regression model has achieved the highest classification accuracy of sentiment as binary class for the deepfake tweets. The accuracy of the LR model is 85% for hashing features and the accuracy of the LR model is 90% for Bag of Words (Bow) features.

The pandemic has created huge demand for online services, people around the world shared their thoughts during this pandemic Covid 19. The tweets related to Covid 19 pandemic was studied by many researchers. This work addressed the literature study [3] of machine learning models for sentiment classification of the pandemic tweet dataset. From this research it was observed that the Lexicon based study used Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analyzer. It was observed from this study that ensemble models have performed well for Covid 19 dataset.

Sentiment analysis is performed based on different feature extraction techniques in the work [4], this work used feature extraction techniques like TF-IDF and n-gram technique and combination of both techniques. This work also addressed machine learning as well as deep learning classification algorithms for sentiment classification. For ML models SVM, Naive Bayes and DT are used and ensemble models like Random forest and AdaBoost were used. The deep learning models used in this study include Convolutional Neural Network (CNN), Recurrent neural network (RNN) and Deep belief network (DBN). Experimental results showed that the absolute error for Adaboost model is less which is around 0.007

while using feature extraction of combination of TF-IDF and n-gram.

Sentiment analysis on contextual based approach was proposed in [5], this work addressed the context analysis which involves constructing a relationship between the words and sources. The dataset based on different contexts such as Electronic, Books, Electronics + kitchen were studied. The correlation of Tree similarity index and Tree differences index were studied. The classifiers used were Random forest and Naive Bayes. Experimental results showed that Tree similarity index along with machine learning models has given positive correlation and Recall value 0.90.

This study showed that sentiment analysis highly demanded research as it can provide useful information on various contexts. There are few observations made from this literature survey that various artificial intelligence techniques like machine learning and deep learning were used. The feature extraction plays a crucial role in this work, thus various researchers explored techniques like TF-IDF, n-gram, Word2Vec and more. It is also observed that machine learning classifiers are performed well when they are used as ensemble models. However, these existing studies have shown that the accuracy of classification on sentiment analysis is less and still need to be improved. This motivated us to bring a hyper tuned model based on Grid search cross validation.

3. PROPOSED WORK

Sentiment analysis is the study of analyzing twitter data by applying feature extraction techniques and applying machine learning models In this work, we proposed five machine learning models are evaluated their performance to identify the best model. Hyper parameter tuning is performed using Grid Search Cross Validation (GSCV) to find the best fit parameter for this model, thus improves the classification accuracy. The proposed work classifies tweets as positive, negative sentiments, which is useful for an organization or individual to find sentiment and analyzes a vast amount of tweets.

DATASET DETAILS

Dataset for this study is extracted from social media portal Twitter by using developer Application programming Interface (API). This requires authentication keys including access tokens and consumer keys to extract data from twitter website. Each extracted tweet contains various information, then required details are filtered are saved as comma separated values (CSV). The dataset variable names are described in the below table

Table1: Dataset extracted from twitter

Variable name	Attribute Description
Text	Tweet text information
ID	Twitter user ID
Screen_name	User's screen name
followers_count	Number of followers of users who tweeted
friends_count	Number of friends count for users who tweeted
Sentiment	Positive, negative
Label	0- positive and 1-negative

DATA PRE-PROCESSING

Data pre-processing involves cleaning the tweet and identifying polarity values to arrive at the sentiment. The proposed work requires the supervised learning dataset, thus the tweets are extracted and cleaned, which involves a few steps including removing numeric values and extract characters in text using regular expression module. Each tweet is real using a textblob and the polarity of the tweet is arrived and labelled as 'positive' and 'negative'. If the polarity value is positive or equal to zero then it is considered as positive sentiment and if the polarity values is negative, then it is considered as negative sentiment.

Feature extraction

Feature extraction is performed with TF-IDF (Term Frequency and Inverse Document Frequency), which converts the tweet text to vectors using the frequency of words. The dataset is prepared as supervised learning data and split into 80% training data and 20% test dataset. Supervised classification algorithm such as Decision Tree (DT), Random Forest (RF), Support vector machine (SVM), Naive Bayes and XGBoost models are used for sentiment classification

a. Decision Tree Classifier

Decision tree (DT) classifier is used for sentiment classification, this ML model is more effective for classifications, which can solve computation problems effectively. Decision tree can be applied for classification as well as regression problems. The process of generating a tree involves a root node from which branches are established. The leaf node from branches is the sentiment classification result, which is 0 or 1. Root node of DT is chosen as the best feature, this model has hyper parameters, which can be tuned using Grid search cross validation. The parameters tuned are given in the following table.

Table2: Hyper parameters for DT model

Parameter	Value
Criterion	Gini, entropy
Max_depth	None, 10, 20,30
Min_samples_split	2,5,10
Min_samples_leaf	1,2,4

Grid Search Cross Validation for 5 folds are performed and the Best Parameters chosen are : {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}

b. Random Forest Regression

Random forest is a supervised ML model, which can be used for classification as well as regression problems. RF model uses the voting process, in which it considered the maximum vote given by the trees will be considered as the classification result. The hyper parameters tuned for RF model are max_depth with values (10,20,30,40,50) and n_estimators with values (10,25, 50,100,200). Grid search cross validation of 5 folds is performed and the Best parameters chosen are : {'max_depth': 10, 'n_estimators': 25}.

c. Naive Bayes classifier

Naive Bayes is the classifier model based on probabilistic Bayes theorem, this model is very efficient for larger datasets. In this sentiment classification task, Multinomial Naive Bayes is used as it is more suitable for features which represents word frequency. Multinomial Naive bayes classifies documents effectively than Gaussian and Bernoulli models. The parameters tuned for this model are given in the below table.

Table3: Hyper parameters for Naive Bayes model

Parameter	Value
Alpha	0.1,0.5, 1.0, 1.5, 2.0
Fit_prior	True, False
Class_prior	None, [0.3,0.7],[0.5,0.5]

Grid search cross validation of 5 folds is applied on the above parameter values and the best fit parameter values chosen are {'alpha': 1.5, 'class_prior': [0.3, 0.7], 'fit_prior': True}.

d. Support Vector Classifier (SVM)

SVM model is a supervised machine learning model, which can be used for both classification as well as regression, however this model is more effective for classification task. This model performs the classification based on hyper plane that separates data points of different classes, in this case binary classes. There are four different kernels available in this model, which converts input data to higher dimensional space thus the linear separation is made possible. The parameter tuned for this model are given in the below table.

Table 4: Hyper parameters for SVM model

Parameter	Value
C	0.1, 1,10,100
Gamma	1,0.1,0.01,0.001
Kernel	Rbf, linear, poly, sigmoid

Grid search cross validation of 5 folds is applied on the above parameter values and the best fit parameter values chosen for SVM model are {'C': 1, 'gamma': 1, 'kernel': 'linear'}.

e. XGB classifier

Extreme Gradient Boosting model, which is an ensemble model and it combines the prediction of multiple weak learners to arrive a strong learner. The model avoid over-fitting problem with the help of regularization techniques and it is very effective in handling missing values. This model works iteratively to produce the results. This model can be used for classification as well as regression problems, the sentiment classification is performed with XGB model, the parameters are tuned using GSCV techniques. The list of parameters and values used are given in the below table.

Table 5: Hyper parameters for XGB model

Parameter	Value
Learning rate	0.1, 0.01, 0.001
Max_depth	3,5,7
Min_child_weight	1,3,5
Subsample	0.5,0.7,1.0
Colsample_bytree	0.5,0.7,1.0

Grid search cross validation of 5 folds is applied on the above parameter values and the best fit parameter values chosen are for XGBoost model are: {'colsample_bytree': 0.5, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'subsample': 0.7}.

4. RESULTS AND DISCUSSIONS

Twitter sentiment classification on real time tweets is performed with hyper parameters tuned ML models. The tweets extracted are prepared as supervised learning data with the help of NLP technique. The polarity of the tweets is identified and labelled as positive and negative according to the polarity value computed for each tweet. The dataset is split into 80% training and 20% validation data for this experiment. The experimental results are discussed in the below table. The ML models are hyper tuned using Grid search cross validation of 5 folds. The experimental results showed that the Support Vector Machine (SVM) model has achieved the highest classification accuracy of 93.18%.

Table 6: Performance comparison of Hyper tuned ML models for sentiment classification

Algorithm	Accuracy (%)
Decision Tree algorithm	91.78
Random Forest	90.41
Naive Bayes	89.04
SVM	93.18
XGBoost	90.90

Accuracy of the model is arrived using formula 1, where TP is true positive, TN is true negative, FP is false positive and FN is false negative. Overall accuracy is computed by correctly classified instances to the total number of instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

---(1)

Decision Tree model classified tweets as positive and negative sentiments, the model is evaluated with accuracy, error values namely MAE (mean absolute error), MSE (Mean square error), RMSE (Root Mean square error) and R-squared error.

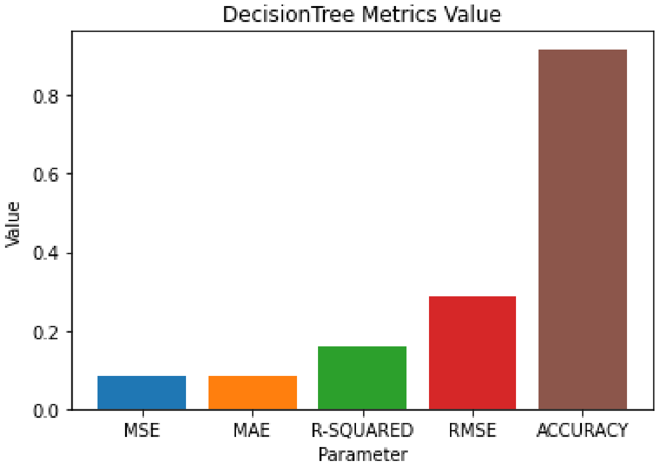

Figure 2: Performance of Hyper tuned DT model for sentiment analysis

Figure 8 shows the performance of Hyper tuned RF model for sentiment classification. RF model has achieved accuracy around 90.41% for twitter sentiment classification and MAE error is 0.0958.

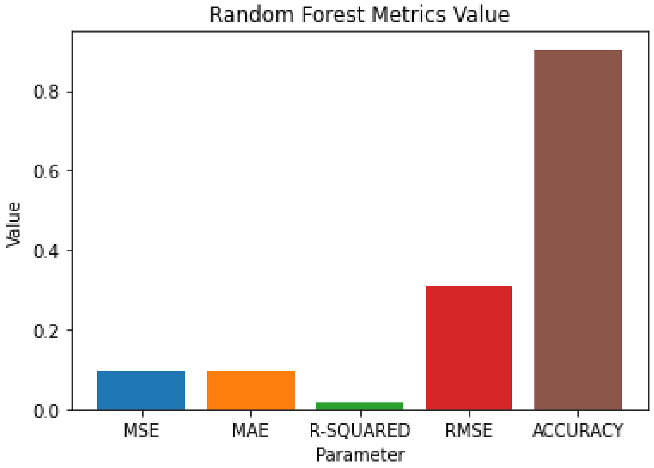

Figure 3: Performance of Hyper tuned RF model for sentiment analysis

Figure 9 shows the performance of hyper tuned SVM model, the classification accuracy of SVM model is the highest of all implemented ML models. The accuracy of SVM is 93.18% and MAE error is 0.0681.

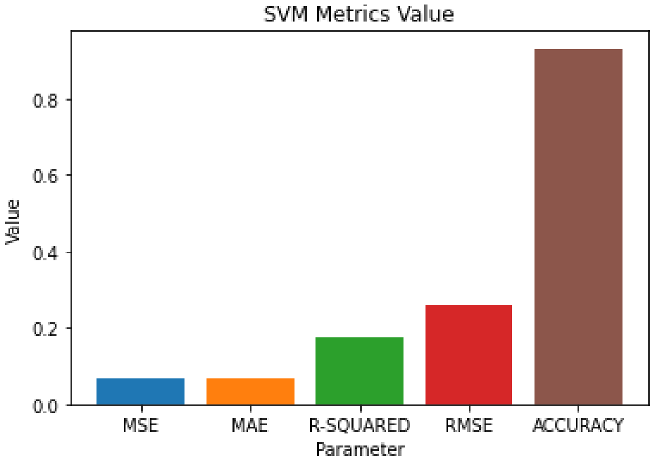

Figure 4: Performance of Hyper tuned SVM model for sentiment analysis

5. CONCLUSIONS

Twitter sentiment analysis using Hyper parameters tuned machine learning models is studied in this work. The tweets are extracted real time and the dataset is prepared and handled as a supervised learning model. This is achieved by finding the polarity of tweets using NLP. The dataset is pre-processed using TF-IDF technique. The hyper tuning of the model parameter is performed by Grid search cross validation. The ML models used for this project are Decision Tree, Random Forest, XGBoost, SVM and Naive Bayes and these models performances are compared for sentiment classification. The study shows that the ML models are more reliable for sentiment classification. The tweets are classified as binary classes with positive and negative sentiments. The experimental results showed that the hyper tuned SVM model has achieved the highest accuracy around 93.18%. As a future extension, this work can be extended to use more feature extraction techniques such as n-gram analysis, Glove vector models.

This work can also be extended with deep learning algorithms such as Deep Neural Network (DNN), LSTM model or Recurrent Neural Network model to achieve more efficiency.

This approach can also be extended with popular microblogging portals like instagram and Facebook datasets.

REFERENCES

[1] A. Shoukry and A. Rafea, "Machine Learning and Semantic Orientation Ensemble Methods for Egyptian Telecom Tweets Sentiment Analysis," in Journal of Web Engineering, vol. 19, no. 2, pp. 195-214, March 2020, doi: 10.13052/jwe1540-9589.1924.

[2] M. Khalid et al., "Novel Sentiment Majority Voting Classifier and Transfer Learning-Based Feature Engineering for Sentiment Analysis of Deepfake Tweets," in IEEE Access, vol. 12, pp. 67117-67129, 2024, doi: 10.1109/ACCESS.2024.3398582.

[3] N. Braig, A. Benz, S. Voth, J. Breitenbach and R. Buettner, "Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data," in IEEE Access, vol. 11, pp. 14778-14803, 2023, doi: 10.1109/ACCESS.2023.3242234.

[4] K. Ayyub, S. Iqbal, E. U. Munir, M. W. Nisar and M. Abbasi, "Exploring Diverse Features for Sentiment Quantification Using Machine Learning Algorithms," in

IEEE Access, vol. 8, pp. 142819-142831, 2020, doi: 10.1109/ACCESS.2020.3011202.

[5] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," in IEEE Access, vol. 8, pp. 17722-17733, 2020, doi: 10.1109/ACCESS.2019.2958702.

[6] H. He, G. Zhou and S. Zhao, "Exploring E-Commerce Product Experience Based on Fusion Sentiment Analysis Method," in IEEE Access, vol. 10, pp. 110248-110260, 2022, doi: 10.1109/ACCESS.2022.3214752.

[7] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," in IEEE Access, vol. 6, pp. 23253-23260, 2018, doi: 10.1109/ACCESS.2017.2776930.

[8] A. Mahmood, H. U. Khan and M. Ramzan, "On Modelling for Bias-Aware Sentiment Analysis and Its Impact in Twitter," in Journal of Web Engineering, vol. 19, no. 1, pp. 1-27, January 2020, doi: 10.13052/jwe1540-9589.1911.

[9] H. T. Phan, V. C. Tran, N. T. Nguyen and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," in IEEE Access, vol. 8, pp. 14630-14641, 2020, doi: 10.1109/ACCESS.2019.2963702.

[10] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem and N. Habib, "A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis," in IEEE Access, vol. 8, pp. 68580-68592, 2020, doi: 10.1109/ACCESS.2020.2983859.

[11] J. Samuel et al., "Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics," in IEEE Access, vol. 8, pp. 142173-142190, 2020, doi: 10.1109/ACCESS.2020.3013933.

[12] U. A. Siddiqua, T. Ahsan and A. N. Chy, "Combining a rule-based classifier with weakly supervised learning for twitter sentiment analysis," 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, Bangladesh, 2016, pp. 1-4, doi: 10.1109/ICISET.2016.7856499.