International Journal of Mechanical Engineering

Liklihood Weighted Bagging Ensemble Approach to Analyze Public Sentiments About Covid-19 Pandemic

E. Prabhakar*
Assistant Professor,Department of Computer Science and Engineering, Nandha College of Technology, Erode – 638 052, Tamil Nadu, India.
Mail Id: prabhakaranptk@gmail.com

V.S.Suresh Kumar
Assistant Professor, Department of Computer Science and Engineering, Nandha College of Technology, Erode – 638 052, Tamil Nadu, India.

S. Nandagopal
Professor, Department of Computer Science and Engineering, Nandha College of Technology, Erode – 638 052, Tamil Nadu, India.

T. Kumutha
Assistant Professor, Department of Pharmacy Practice, Nandha College of Pharmacy, Erode – 638 052, Tamil Nadu, India.

Abstract

Social media emerges as a powerful tool in disseminating news trends in the world. The recent pandemic, COVID-19 is a catastrophic event that has changed the lifestyle and mindset of the people across the globe. This has created physical and mental trauma even in the lives of common man. Government and researchers are taking intensive efforts in analyzing and understanding the trends and patterns in COVID-19 related tweets to mine public opinion. These opinions could be further transformed into actionable plans. As the Governments are striving hard to combat new mutations of the virus, mining public opinion would be immensely helpful in formulating policies and decision making. Though the current day techniques can effectively analyze the sentiments, they are prone to suffer from overfitting. Hence, thisarticle proposes a new likelihood weighted bagging ensemble approach that mines the tones of people from tweets. The work is validated on Kaggle dataset and proves to be more efficient than the other state of art base learning models.

Keywords - Covid-19 Pandemic, Social Media, Likelihood Ensemble, Sentiment Analysis, Public Opinion Mining, COVID 19 Tweets.

I. INTRODUCTION

SARS-CoV-2 commonly known as COVID-19 [1] has emerged as a global pandemic that devastated the medical infrastructures and world's economy, which eventually led to a rampant surge in unemployment. Apart from being an infectious disease, COVID-19 has created a grim mood across the borders. It has created a mental trauma and has emerged as a new source of depression, stress, and anxiety. This is further aggravated by floods of misleading information posted on social media [2].

The genre of Sentiment analysis deploys human knowledge in identifying the polarity of documents into three different classes, such as negative,neutral, and positive. Recent times witnesses its usage in fields like opinion mining, business organizations,health informatics and recommendation systems. Leading business forums like Amazon, Flipkart etc. are devising their strategic plans by mining people's opinion to increase their profits. Similarly, assessing thesocial media posts under the canopy of sentiment analysis helps to understand the human psychology and their allied behaviors, which are directlyassociated to physical and mental well-being of an individual.

Sentiment analysis has emerged as most attractive research area in the field on Natural Language Processing (NLP) as it mines people's opinion in a particular topic of interest [4, 5].NLP can be envisioned as a major pillar under the broad umbrella of Machine Learning and Artificial Intelligence [12]. The process of sentiment analysis uses textual documents and posts from Facebook and Twitter for analyzing the underlying tone of the textual content. This study focus on most burning issue of analyzing COVID-19 tweets in the context of sentiment classification [3].

The present work concentrates on the following aspects:

(a) Labeling the real values into the following tones such as fear, anger, joy and sad sentiment scores.

(b) Classifying the sentiments based on a novel likelihood based weighted bagging ensemble approach.

(c) Assessing and comparing the performance of the proposed model with other state-of-the art machine learning techniques that handles sentiment analysis of COVID-19 related tweets.

(d) Deploying the proposed model in real world scenario to classify the sentiments of anger, fear, joy and sad.

The organization of the paper is as follows: Section 2 provides a brief but comprehensive background of selected works in sentiment analysis and emotional intelligence. Section 3 describes the detailed methodology of the proposed novel approach. Further, the experimental results of the proposed model have been described and assessed in Section 4. Section 5concludes the work augmenting the future directions of research.

II. LITERATURE REVIEW

Some of the important milestones in sentiment analysis from social media posts are discussed here.

Authors in [6] developed a model to predict degree of awareness about the precautionary procedures in five major regions in Saudi Arabia. The feature extraction was done using N-gram technique. The prediction was done by powerful machine learning models like Support Vector Machine (SVM), Knearest neighbors (KNN), and Naïve Bayes (NB). The detailed analysis of the results shows that SVM classifier augmented with bigram that deployed Term Frequency–Inverse Document Frequency (TF-IDF) measure outperformed other peer models.

Rustam et al. [7] compared the performance of five significant machine learning algorithms namely random forest, extra tree classifier, XGBoost classifier, decision tree, and long short-term memory (LSTM) in the context of analyzing the opinion behind COVID-19 tweets. The profound feature extraction techniques such as Bag-of-Words (BOW) and Term-Frequency and Inverse Document Frequency (TF-IDF) were deployed to extract the predominant features from the texts. The models were trained to perform multi-label classification. The performance assessment of these models indicate that extra tree classifier is superior to its rivals.

Another notable work in analyzing sentiments behind COVID-19 tweets is contributed by Authors in [8]. This is a comprehensive work that gives insights on people's perception about various factors about the pandemic namely its mode of outbreak, awareness, common and uncommon symptoms, precautionary measures, adherence to Standard Operating Protocols (SOPs) etc. The ML algorithm could effectively mine the real sentiments from the tweets with a promising accuracy of 70%.

It is apparent that SVM is the most predominantly used algorithm for classification as well as regression problems. Since, sentiment analysis involves classifying data into multiple labels, SVM occurs as a natural choice. Rani and Singh [9] used SVM for mining the opinionfrom Twitter data. Features were extracted through TF-IDF and tones were classified using both linear SVM and kernel based SVM. The efficacy of the model was assessed using metrics such as F1-score, precision, recall, and accuracy. The results delineated that Linear SVM is more effective in extracting sentiments than the kernel based SVM.

NawNaw [10] administrated the sentiment analysis from tweets through SVM and KNN classifiers. The data collected from Twitter API was subjected to extensive preprocessing such as data transformation, tokenizing, negation handling, normalization, and filtering to enhance the classification efficacy. More suitable features are chosen by assessing the TF-IDF and classified using SVM and K-NN classifiers.

Al-Tamimi et al. [11] presented a sentiment analysis study that culminates several classifiers in analyzing 5986 Arabic Youtube comments. This extensive study includes Radial Basis Function (RBF) based SVM, K-nearest neighbors (KNN), and Bernoulli Naïve Bayes models to classify the sentiments from raw as well as normalized datasets. The SVM-RBF kernel model outperformed other models with highest accuracy score of 88.8% when supplied with normalized input data.

Authors in [14], [15], [17], [18], [21], [22], [23], [25], [29] proposed various novel methodologies to improve the performance of the data mining models. Researches in [20], [24], [26], [27], [28] discussed the applications and algorithms related to improvement of data mining models. Importance of social media and streaming data conversed in [16], [19], [30].

Thus, the literature in the field of sentiment analysis asserts the predominant of role of Artificial Intelligence and Machine Learning. Further the literature indicates that only very few models aim to classify the real tones such as joy, grief, anger, fear etc. Hence the proposed work aims to classify the COVID-19 based tweets into various emotion genres.

III. METHODOLOGY

Ensembling is a technique of integrating the prowess of each of the base learners to improve the overall prediction efficacy. The proposed methodology is a ensemble model of several classifiers like Naïve Bayes, Stochastic Gradient Descent, Random Forest, Extreme Gradient Boosting, Support Vector Machine and Logistic Regression. The dataset that is employed to build the model is collected from Kaggle. Since, the data is noisy it may be subjected to multitude of preprocessing and data cleaning activities. The data from the dataset infiltrate through three stages namely data collectionand pre-processing;data cleaning and selection; Modelling and Evaluation.

A. Phase I: Data Collection and Pre-processing

As the data from Twitter is highly unstructured, it has to undergo data transformation and reduction of undesired elements like Hashtags, hyperlinks, whitespace, URLs, stop words, usernames, etc.

B. Phase II: Data cleaning and selection

Tweets in its original, raw form cannot be processed for sentiment analysis as they are prone to sarcasm and negations. Hence, by applying tokenization technique, the text can be partitioned into viable strings called tokens. The so formed tokens will be parts of speech like nouns, verbs, adverbs and adjectives etc. Similarly grouping of words that infer same meaning is also one of the most significant preprocessing technique in NLP. For example, the words ran, runs and running will be placed in same bucket. In addition to this, stemming and lemmatization are two popular techniques.The former eliminates undesirable suffixes or prefixes while the later analyses the word in the context of vocabulary. The cleaned and processed tweets are subjected to text processing.

Selection of central ideas from the tweets plays a major role in sentiment analysis. Hence, word density based data selection is determined by the metric Term Frequency Inverse Document Frequency (TF-IDF) method. The words are separated based on the frequency of its occurrence in the text under study. This is an excellent tool to gain insights about the polarity of the tweet into positive and negative genres. This TF-IDF is exercised on the entire terms in the dataset to rank the occurrence of each word. The word with highest rank contributes more to the polarity.

C. Phase III: Model Building& Evaluation

A learning model behaves like human cognition [13]. All the machine learning and deep learning algorithms are rooted on deriving insights from the data by observing the inherent patterns and trends. Nevertheless, sentiment analysis also focusses on delving knowledge from the texts in the view of understanding the tone as positive and negative. The models learn from the training data presented to them and their performance is tested by subjecting them to operate on test data.

The predictive power of the base learners like SVM, neural networks etc. can be further improved by integrating multiple models to form an ensemble model. In conventional ensemble algorithms like Adaboost, the weights were assigned to homogeneous base classifiers depending on the overall accuracy and error in each learning iteration. This leads to overfitting, where the model fails to meet its desired performance in testing phase. To mitigate the impact of overfitting, this article presents a novel weight allocation strategy confined to the multiclass ensemble approach. The weight assignment occurs based on both true positive and false positive outcomes of a specific emotion. For instance, in the class labelled as anger, the weight assignment will be done based on both true anger as well as false anger. Hence, the model evolves to be highly dynamic, thus avoiding overfitting. In addition to this, the model greatly reduces the occurrences of false positive and false negative cases.

TABLE 1: Pseudocode for Likelihood Weighted Bagging Ensemble Approach

```
Classifier generation:
Let N be the size of the training set.
for each of t iterations:
    sample N instances with replacement from the original training set.
    apply the learning algorithm to the sample.
    Assign weights based on previous classification
    store the resulting classifier.

Classification:
for each of the t classifiers:
    predict class of instance using classifier.
return class that was predicted most often.
```

TABLE I illustrates the step by step pseudocode for the proposed likelihood weighted bagging ensemble approach. It consists of training the model and model generation.

IV. RESULTS AND DISCUSSIONS

The dataset has been collected from Kaggle. Publicly available COVID 19 Tweets dataset has been considered. It consists of 4 labels namely sad, joy, anger and fear.

TABLE II: Confusion Matrix

	Anger	**Sad**	**Joy**	**Fear**
Anger	True Anger	False Sad	False Joy	False Fear
Sad	False Anger	True Sad	False Joy	False Fear
Joy	False Anger	False Sad	True Joy	False Fear
Fear	False Anger	False Sad	False Joy	True Fear

TABLEII presents the confusion matrix pertaining to four emotions namely anger, sad, joy and fear that were predicted by the proposed ensemble approach. Weight updates at every class happens by considering the false positives and true positives of the sentiments. The pseudocode for the proposed Likelihood Weighted Bagging Ensemble Approach is presented in TABLE II.

Table III shows the classification accuracy of various based learners and the proposed ensemble approach. A remarkable improvement in the accuracy can be observed in the Enhanced Likelihood Ensemble approach, as the weight allocation for every class happens in accordance with the previous prediction. Though logistic regression and stochastic gradient descent shows better performance, they are shadowed by the proposed methodology.

TABLE III: Classification accuracy of various base learners and proposed approach

Models	**Accuracy**
Naïve Bayes	67.8
Stochastic Gradient Descent	68.93
Random Forest	63.1
Extreme Gradient Boosting	67.15
Support Vector Machine	61.97
Logistic Regression	69.58
Enhanced Likelihood Ensemble	72.58

The assessment of the base learners is displayed from Table IV to Table IX. The performance analysis of Naïve bayes model is elucidated in Table 4. The model exhibits comparable accuracy of 67.8% but its precision, recall and F1-scores are low especially in labelling anger and fear. This decline can be attributed to the fact that anger and fear are non-mutually exclusive emotions, hence the model failed to delineate them into proper classes.

TABLE IV: Performance of Naïve Bayes algorithm

Naive Bayes	**Precision**	**Recall**	**F1-Score**
Anger	54	67	60
Fear	66	57	61
Joy	71	82	76
Sad	81	69	74

Table V presents the results of labelling the COVID-19 related tweets into four classes by Stochastic Gradient Descent algorithm. Like Naïve Bayes algorithm, this base learner is not very successful in predicting the classes of anger and fear sentiments. Further the accuracy of this model is 68.93% which is marginally higher than Naïve Bayes.

TABLE V:Performance of Stochastic Gradient Descent algorithm

Stochastic Gradient Descent	**Precision**	**Recall**	**F1-Score**
Anger	65	59	62
Fear	54	67	60
Joy	81	72	76
Sad	77	77	77

The classification performance of random forest algorithm is shown in Table VI. The model exhibits relatively low accuracy of 63% apart from showing lower precision, F1-score and recall. Lower values of precision and recall indicates that the purity of classes obtained through Random forest model is crude. This may be because of overfitting; as homogeneous decision trees are used to classify the tweets.

TABLE VI:Performance of Random forest algorithm

Random Forest	**Precision**	**Recall**	**F1-Score**
Anger	64	57	60
Fear	42	71	53
Joy	81	54	65
Sad	67	80	73

Prediction results of Extreme Gradient Boosting is portrayed in TABLE VII. This model's classification accuracy is 67.5% which is highly competitive. On the other hand, this model also suffers from declined precision, F1-score, and recall score. As discussed above, the extreme gradient boosting also results in impure classes.

TABLEVII: Performance of Extreme Gradient Boosting algorithm

Extreme Gradient Boosting	**Precision**	**Recall**	**F1-Score**
Anger	61	60	61
Fear	52	67	59
Joy	79	66	71
Sad	78	75	77

The performance analysis of multi class SVM in classifying tweets is displayed in TABLE VIII. The classification accuracy of SVM is around 62%. This model was able to predict joy and sad classes with better purity than the other two classes.

TABLE VIII: Performance of Support Vector Machine algorithm

Support Vector Machine	**Precision**	**Recall**	**F1-Score**
Anger	53	61	57
Fear	49	62	55
Joy	77	54	63
Sad	70	73	72

The performance assessment of logistic regression algorithm in predicting the classes of COVID-19 related tweets is exhibited in TABLE IX. The prediction accuracy is 69.58% while the scores of precision, recall and F1-score for anger and fear labels are low.

TABLE IX: Performance of Logistic regression algorithm

Logistic Regression	**Precision**	**Recall**	**F1-Score**
Anger	59	64	61
Fear	58	63	60
Joy	85	71	77
Sad	77	80	79

TABLE X shows the results of the proposed Likelihood Weighted Bagging Ensemble model. The proposed model portrays classification accuracy of 72.58%, which is notably higher than other base learners discussed in this article. Also, the proposed model exhibits better precision, recall and F1-score, which is a direct implication that the classes are pure.

TABLE X: Likelihood Weighted Bagging Ensemble

Likelihood Weighted Bagging Ensemble	**Precision**	**Recall**	**F1-Score**
Anger	75	76	75
Fear	70	69	70
Joy	77	76	76
Sad	74	75	74

The classification results show that the proposed model is highly successful in mining the public opinion about COVID-19 related tweets. The likelihood based weight assignment has greatly reduced the formation of impure classes.

V. CONCLUSION AND FUTURE WORK

This article proposes a novel Likelihood Weighted Bagging Ensemble model that classifies multiple sentiments form the social media posts. The performance of the proposed model is validated by comparing with other peer learning algorithms like logistic regression, Decision tree, random forest, SVM, Extreme Gradient Boosting, stochastic gradient descent etc on classifying COVID-19 related tweets from Kaggle dataset. The analysis shows that the proposed algorithm outshines other base learners in terms of classification accuracy, recall, precision and F1-score.

Further, augmenting otherembeddings such as Word2vec and GloVe could improve the efficacy of the proposed model. In addition to this, development of a comprehensive framework that incorporates topic modelling with sentiment analysis that assesses the surge in COVID-19 cases in a locale, enforcement of SOPs and propagating information about vaccine plans is the need of the hour.

REFERENCES

[1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, “The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2”, Nature Microbiology, 5, pp.536–544, 2020.

[2] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, Aboul Ella Hassanien, “Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media”, Elsevier - Applied Soft Computing, 97, 2020.

[3] C. Sitaula, A. Basnet, A. Mainali, T. B. Shahi, "Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets", Computational Intelligence and Neuroscience, 2021.

[4] R.Liu, Y.Shi, “Survey of Sentiment Analysis Based on Transfer Learning”, IEEE Access 7, pp.85401–85412, 2019.

[5] P.Tyagi, R.C.Tripathi,“A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data”, SSRN Electronic Journal, 2019.

[6] S.SumayhAljameel , A. Dina Alabbad, A.NorahAlzahrani, M.ShouqAlqarni, A. Fatimah Alamoudi, M.LanaBabili, K.SomiahAljaafary and M.FatimaAlshamrani, “A Sentiment Analysis Approach to Predict an Individual’s Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia”, International Journal of Environmental Research and Public Health, 2021.

[7] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, “A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis,” PLoS One, 16, 2021.

[8] K B PriyaIyer, SakthiKumaresh, “Twitter Sentiment Analysis On Coronavirus Outbreak Using Machine Learning Algorithms”, European Journal of Molecular & Clinical Medicine, 7, 2020.

[9] S.Rani, J.Singh, “Sentiment Analysis of Tweets Using Support Vector Machine”, International Journal of Computer Science and Mobile Applications, 5(10), pp.83–91, 2017.

[10] N.Naw, “Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers”, International Journal of Scientific and Research Publication (IJSRP), 8, pp.407–411, 2018.

[11] A.K.Al-Tamimi, A.Shatnawi, E.Bani-Issa, “Arabic Sentiment Analysis of YouTube Comments”, In Proceedings of the 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, 2017.

[12] S.Sharanya, V.Revathi, G.Murali, “Estimation Of Remaining Useful Life Of Bearings Using Reduced Affinity Propagated Clustering”, Journal of Engineering Science and Technology, 16(5), pp. 3737 – 3756, 2021.

[13] S.Sharanya, V.Revathi, “An intelligent Context Based Multi-layered Bayesian Inferential predictive analytic framework for classifying machine states”, Journal of Ambient Intelligence and Humanized Computing, 12, pp.7353–7361, 2020.

[14] K.Nithya, PC D Kalaivaani, R.Thangarajan, “An enhanced data mining model for text classification”, IEEE International conference on computing, communication and applications, pp.1-4, 2012.

[15] S.Nandagopal, VP.Arunachalam, S.Karthik, “A Novel Approach for Mining Inter-Transaction Itemsets”, European Scientific Journal, 8, pp.14-22, 2012.

[16] P.Devayani, D.Mohanapriya, A.Kiruthika, A.Megavardhini, K.Nandhini, “Social Distancing Detection using Deep Learning Model”, International Journal of Scientific Development and Research (IJSDR), 6(4), 2021.

[17] M.Vijayakumar, S.Prakash, “An Improved Sensitive Association Rule Mining using Fuzzy Partition Algorithm”, Asian Journal of Research in Social Sciences and Humanities, 6(6), pp.969-981, 2016.

[18] T.Dinesh Kumar, E.Prabhakar, K.Nandhagopal, “An Enhanced Ensemble Classification Algorithm (EECA) for Airline Services Big Data Sentiment Analysis”, Journal of Advanced Research in Dynamical & Control Systems, 11 (7), pp.1461-1467, 2019.

[19] M.Vijayakumar, RMS.Parvathi, “Concept mining of high volume data streams in network traffic using hierarchical clustering”, European Journal of Scientific Research, 39 (2), pp.234-242, 2010.

[20] S.Nandagopal, S.Karthik, VP.Arunachalam, “Mining of meteorological data using modified apriori algorithm”, European Journal of Scientific Research, 47 (2), pp.295-308, 2010.

[21] T.Malathi, S.Nandagopal, “Enhanced Slicing Technique for Improving Accuracy in Crowd Sourcing Database”, International Journal of Innovative Research in Science, Engineering and Technology, 3 (1), pp.278-284, 2014.

[22] S.Nandagopal, V.P.Arunachalam, S.Karthik, “Mining of Datasets with Enhanced Apriori Algorithm”, Journal of Computer Science, 8 (4), pp.599-605, 2012.

[23] K.Gopalakrishnan, S.Thiruvenkatasamy, E.Prabhakar, R.Aarthi, “Night Vision Patrolling Rover Navigation System for Women’s Safety Using Machine Learning”, International Journal of Psychosocial Rehabilitation, 23 (4), pp.1136-1148, 2019.

[24] M.Vaijayanthi, M.Karthick, “Method Level Bug Prediction Using Information Gain”, International Journal of Research in Computer Science, 4 (1), pp.9-13,2017.

[25] M.Karthick, R.Senthikumar, “Hybrid Approach for Image Restoration” ,International Journal of New Innovations in Engineering and Technology, 3 (3), 2015.

[26] M.Vijayakumar, S.Prakash, RMS.Parvathi, “Inter cluster distance management model with optimal centroid estimation for k-means clustering algorithm”, WSEAS transactions on communications, 10 (6), pp.182-191, 2011.

[27] M.Vijayakumar, RMS.Parvathi, “Performance of Distributed Hierarchical Cluster in Peer to Peer Network Traffic”, Journal of Computational Information Systems, 7(6), pp.1901-1909, 2011.

[28] P.Saveetha, Arumugam.S, “Privacy Preservation On Stream Ciphers Using Genetic Algorithm With Pseudo-

Random Series" Australian Journal of Basic and Applied Sciences 7 (12), pp.1-8,2013.

[29] V.Kavitha, C.Palanisamy, "SWT - SPIHT - NVF Based Blind Medical Image Watermarking", International Journal of Advanced Science and Technology,2020.

[30] T.Krishnakaarthik,K.DeepaK.Gowthamapriya,B.Keerthana, "Summarized Automated Hash-Tag Tweet Segmentation",Journal of Applied Science and Engineering Methodologies , 1 (1), pp.109- 112, 2015.