IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Emotion Recognition by Textual Tweets Classification Using Voting Classifier(LR-SGD)

Anam Yousaf[1], Muhammad Umer[1,5], Saima Sadiq[1], Saleem Ullah[1], Seyedali Mirjalili[2,3,4], Vaibhav Rupapara[6,*], and Michele Nappi[7,*]

[1]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan
[2]Center for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia
[3]YFL (Yonsei Frontier Lab), Yonsei University, Seoul, Korea
[4]King Abdulaziz University, Jeddah, Saudi Arabia
[5]Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[6]School of Computing and Information Sciences, Florida International University, USA
[7]Department of Computer Science,University of Salerno,Fisciano, Italy

Corresponding author: Vaibhav Rupapara (vaibhav.rupapara.sept@gmail.com) and Michele Nappi (mnappi@unisa.it)

This work is supported by Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan.

ABSTRACT The proliferation of user-generated content on social media has made opinion mining an arduous job. As a microblogging platform, Twitter is being used to collect views about products, trends, and politics. Sentiment analysis is a technique used to analyze the attitude, emotions and opinions of different people towards anything, and it can be carried out on tweets to analyze public opinion on news, policies, social movements, and personalities. By employing Machine Learning models, opinion mining can be performed without reading tweets manually. Their results could assist governments and businesses in rolling out policies, products, and events. Seven Machine Learning models are implemented for emotion recognition by classifying tweets as happy or unhappy. With an in-depth comparative performance analysis, it was observed that proposed voting classifier(LR-SGD) with TF-IDF produces the most optimal result with 79% accuracy and 81% F1 score. To further validate stability of the proposed approach on two more datasets, one binary and other multi-class dataset and achieved robust results.

INDEX TERMS Sentiment Analysis, Text Classification, Machine Learning, Opinion Mining, Emotion Recognition, Artificial Intelligence

I. INTRODUCTION

AUTOMATIC emotion recognition, pattern recognition and computer vision have become significantly important in Artificial Intelligence lately with applications is a wide range of areas. Recently, social media platforms such as Twitter have generated enormous amounts of structured, unstructured and semi-structured data. One of the most recent example is COVID-19 infodemic that shows misinformation in social media can be far more important and devastating than a disaster such as a pendemic.

There is a need to analyze to accurately assign sentiment classes on a large scale. To perform such tasks, accurate NLP techniques and machine learning (ML) models for text classification are required. Twitter provides an opportunity to its users to analyze its data on a large and broader point of view. Efficient methods are important to automatically label text data due to its noisy nature. In the past many studies have been performed on Twitter sentiment classification [1]. As Twitter is very fast and an efficient micro-blogging examination that facilitates the end users to transmit small posts are said to be tweets. Twitter is a highly demanding app in the world and is a successful platform in social media.

Free account can be created by using Twitter that can provide an enormous audience potential. With the purpose of business and marketing, Twitter can be proved as the best platform, through which one can get in touch with very rich and famous personalities like stars and celebrities, so their purchasing can be very charming for them as well as for advertisers. Using Twitter, every celebrity is linked with fans as well as to grant a communication to followers. Such a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

platform is one of the superlative approaches for lovers as well. But, it has a short note range; only 140 letters for each post and it can type a post or link on the website since it has no cost and also open as the advertisements as well. There is no problem with clusters of personal ads which are similar to other social networking sites. It is quick because as a tweet is posted on Twitter, the public who is subsequent to respective business will get it without delay.

Companies and advertisers can compose utilization of this source to check the diverse operational point of views which are very considerable. With help of this, they will obtain an immediate response from their followers. Remarkably, a lot of businesses with the intention of purchase, Twitter followers increase their deals. Twitter facilitates the followers by making them identify regarding fresh business, products, services, websites, blogs, eBooks etc. Consequently, Twitter clients might tick lying on link and also optimistically endow in a manufactured goods or examine the products presented and to get share in profit. It is extremely effortless to utilize as people can follow to get the news and updates, as organizations can tweet or re-tweet, they can mark favorite or selected people to send the tweets, also know how to propel the posts plus to be able to endow their money and instance through it. Academy, Industry, super bowls and Grammy Awards of such major Sports and Entertainment events generate a lot of buzz in the global world by using it.

Competition is rising among different products on Twitter. People love to express their feelings about a particular product on social networks like twitter. Product owners are ready to spend more money on social media platforms to better advertise their products and to generate more revenue. When a person shares experience about a product, it helps the owner to change their market strategy, selling schemes, and improving the quality. Customer reviews serve as a feedback to the owners or manufacturers too.The data generated in such a way is of large amount and requires an analysis expert team to classify the customer sentiment from the reviews. Experts can make a human error in sentiment analysis, therefore it requires machine learning and ensemble learning classifiers to accurately classify the sentiment of the customers.

This study compares various machine learning models for emotion recognition by tweet classification using Tf and TF-IDF. This research presents a voting classifier (LR-SGD) and aims to estimate the performance of famous ML classifiers on twitter datasets. The key contributions are as follows:

- Machine learning-based classifiers including support vector machine (SVM), Decision Tree Classifier (DTC), Naive Bayes (NB), Random Forest (RF), Gradient Boosting Machine (GBM) and Logistic Regression (LR) trained on Twitter dataset are compared for emotion recognition.
- A voting classifier (VC) designed to classify tweets which combines LR and SGD and outperformed using TF-IDF.
- The proposed model stability is further validated by applying it on two different datasets, one binary dataset (containing hatred or non-hatred classes) and other multi-class dataset (containing product reviews having 1 to 5 ratings) .

The rest of the paper is organized as follows. Section II discusses literature related to the current research work. Section III presents the proposed methodology as well as as detailed description of the tweet dataset used in the experiment. Results are presented in Section IV and the stability of proposed model is given in Section V. Section VI finally conclude the research work and also suggest future work.

II. RELATED WORK

Sentiment analysis inspires corporations to define clients' preferences about products, services, and brands. Further, it plays an important role in interpreting information about industries and corporations to reserve them in making entity review. Sarlan et al [2] established a sentiment analysis through extracting number of tweets with the help of prototyping and the results organized customers' views via tweets into positive and negative. Their research divided into two phrases. The first part is based on literature study which involves the Sentiment analysis techniques and methods that nowadays are used. In the second part, the application necessities and operations are described preceding to its development.

In another research Alsaeedi et al [3] analyzed various kinds of sentiment analysis that is applied on to Twitter dataset and its conclusions. The distinct approaches and conclusions of algorithm performance were compared. Methods were used which were supervised ML based, , lexicon-based, ensemble methods. Authors used four methods that were Twitter sentiment Analysis using Supervised ML Approaches; Twitter sentiment Analysis using Ensemble Approaches. Twitter sentiment Analysis is using lexicon based Approaches.

Lexicon based approaches have been explored by many researchers for emotion classification. Bandhakavi et al. [4] performed emotion-based feature extraction using domain specific lexicon generation. They captured association of words and emotions using a unigram mixture model. They used tweets that are weakly labelled to classify emotions. Their proposed architecture outperformed other state-of-the-art approaches such as Latent Dirichlet Allocation and Point wise Mutual Information. Event related tweets are identified by researchers on geo related tweets [5]. They used specific tweets of local festivities in one year. They also identified different parameters that helped in event discovery. Alsinet et al. [6] analyzed tweets from political domains. They claimed accepted tweets are stronger as compared to the rejected tweets. Rumor detection in tweets is performed by using an encoder to analyze human behavior in comments [7].

Hakh et al [8] used SMOTE method to remove excessive challenges of Twitter dataset. In addition, they applied different feature selections for rapidity of sentiment analysis method. Authors projected methodology that was estimated beside the dataset application decision, squashy favorable results on all operated evaluation metrics. Pre-processing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

steps were applied on their dataset after that they used TF-IDF features that were used to measure important weight of terms. Then classification methods were used (i.e. AdaBoost, Linear SVM, Kernel SVM, Random Forest, Decision Tree, Naïve Bayes and K-NN) and at last to relate classification's effectiveness: Accuracy and F1-score measures were used.

In [9], Xia et al. created the proportional training of the efficiency about collaborative method on behalf of Sentiment's arrangement. They set two types of feature in the context of sentiment analysis. Firstly, the feature set was totally depend on the part of speech and word relation was depending on the feature set. Secondly, the following familiar text classification algorithms that were maximum entropy,support vector machines and naive Bayes. Thirdly, the following ensemble strategies, that was the fixed combination, meta-classifier combination and weighted combination. They used 5 document-level datasets broadly utilized along with arena of Sentiment's arrangement. Experiments shown in this research the ensemble techniques are more effective than rest of the classifier which is also shown in our search that ensemble of two classifiers that are Logistics regression and stochastic gradient decent classifiers ensemble and give better result than other classifiers.

Deep learning has been utilized by many researchers for image classification [10] and tweet classification [11]. Rustam et al. [12] presented a Tweets Classification for US Airline Companies Sentiments. The researcher applied pre-processing on the dataset. The influence about feature extraction methods, together with TF, TF-IDF, along with word2vec, proceeding the classification accuracy has been examined. In addition, execution about the long short-term memory (LSTM) was studied in certain dataset. Paper of researcher proposes a Voting Classifier (VC) who helps to process similar administrations. Voting Classifier must dependent the Spatial Estimation (SE), Stochastic Gradient Descent classifier (SGDC) along with simple ensemble method for concluding results. Various types of ML classifiers tested with the use of precision, accuracy, recall and F1-score by way of working metrics. Results indicate that proposed VC is more efficient than one of the phase actors. The experiment also demonstrated the efficiency of machine learning students improved while TF-IDF utilizes a feature input.

Santos et al [13] examined a sentiment analysis of short texts. In the experiment, researchers suggest a first-hand profound convolution neural network that achieve from character to sentence level material to accomplish sentiment analysis of little texts. Mohamed et al. [14] evaluated a sentiment analysis of mining halal food consumers. This examination fills this gap through the investigation of an irregular example of 100,000 tweets managing halal food. To lead the examination, a specialist predefined dictionary of seed descriptors was utilized. By investigating halal food feelings communicated via web-based networking media, this examination adds expansiveness and profundity to the discussion over such an underrepresented region. Distinct investigation recognized for the most part positive estimation toward halal food, while geo-found Twitter maps indicated that "strict diaspora" broadly utilizes computerized presents on impart about halal food.

Parveen et al. [15] studied sentiment analysis on Twitter dataset that uses NB algorithm. Analyst use Hadoop Framework for preparing film informational collection which is reachable on Twitter site as reviews, input and opinions. Sentiment analysis on Twitter data is explored in three classes that are positive, negative and neutral. Alomari et al. [16] analyzed SVM utilizing TF-IDF. The study presented the Arabic Jordanian Twitter corpus where Tweets are explained seeing that any positive or negative. It researched distinctive directed machine learning opinion examination classifiers when applied to Arabic client's online life of general subjects that are found in either Modern Standard Arabic (MSA) or Jordanian tongue. Analyses were conducted to assess the utilization of various weight plans, stemming and N-grams terms strategies and situations.

Gamal et al [17] built Twitter benchmark dataset for Arabic Sentiment Analysis. A benchmark Arabic dataset suggested in experiment for estimation investigation demonstrating social event strategy about the latest tweets in various Arabic vernaculars. The experiment dataset incorporates in excess of 151,000 unique assessments which marked into two classes, negative and positive. ML algorithms are functioned in SC; ML algorithm attached through learning arrangements. Sentiment analysis ordinarily executed using one fundamental methodology from a ML(lexicon-based approach) based approach. The calculations functioned via SC on the dataset accomplished 99.90% precision utilizing TF-IDF.

Kumar et al [18] explored the sentiment analysis of multimodal Twitter data. The experiment utilized a multi-method feeling examination approach to decide slant extremity mark for approaching tweet that is printed picture information realistic. Picture estimation marking was accompanied by utilizing SentiBank along with SentiStrength marking for Regions with convolution neural network (R-CNN). For a picture posted in Twitter, the picture module is executed which utilizes a current module of SentiBank along with R-CNN that decide the feeling estimation mark of the picture. After pre-processing, the content module utilizes an AI-based troupe strategy gradient boosting to characterize tweets into extremity classifications, to be specific, positive, negative or neutral High execution exactness of 91.32% is watched on behalf of arbitrary multi method tweet dataset utilize assess the planned model. Sailunaz et al. [19] investigated the feeling through the dataset that analyzed by a sentiment analysis from Twitter texts. The objective this work was to recognize and investigate assessment and feeling communicated by individuals from content in their Twitter posts and to use them for creating suggestions.

The dataset is utilized to recognize slant and feeling from tweets and their answers and estimated the impact scores of clients dependent on different Tweet based and client based parameters. The strategy we utilized in this paper include several fresh approaches: (I) remembering answers to tweets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

for the dataset and estimations, (II) presenting understanding score, slant score and feeling score of answers in impact score computation, (III) producing customized and general proposal consisting rundown of clients who conceded to a similar subject and communicated comparable feelings.

III. PROPOSED METHODOLOGY

In this research, different techniques have been used for methodology in ML for its objectives. Versatile experiments were examined using different methods and techniques. Multiple classifiers applied on the dataset, but the Voting classifier is an ensemble of Logistic Regression and Stochastic Gradient Descent outperforms than all other ML models in terms of accuracy, recall, precision and F1-score.

Twitter dataset used in this experiment is scrapped from Kaggle repository. First the dataset is pre-processed by removing unwanted data. Then, the data was split into two sets: training set and testing set. The training set was given the percentage of 70% while the test set portion is 30%. After that feature engineering techniques are applied on the training set. Multiple machine learning classifiers are trained on the training set and tested using the test set. The evaluation parameters used in this experiment are: (a) Accuracy (b) Recall (c) Precision (d) F1-score.

A. DATASET

Dataset contains a lot of contrary tweets. The dataset is called “Sentiment Analysis on Twitter data" and contains 99989 records. Every record is labeled as happy and unhappy according to its sentimental polarity using symbol 1 and 0. Tweets which are in English are remembered for the finished dataset. The dataset contains different features. Table 1 contains features and description of each feature.

TABLE 1: Dataset Specifications.

Features	Description
Item ID	This is the index of record
Sentiment	This column contains Sentiment happy and unhappy corresponding to tweets
Sentiment Text	This column contains the textual tweets

B. DATA VISUALIZATION

Data Visualization helps to understand the hidden patterns lying inside the dataset. It helps to qualitatively get more details about the dataset by visualizing the characteristics of the attributes. Figure 1 shows the ratio of two target classes happy and unhappy. Figure 1 also illustrates that the happy class has more average than the unhappy class.

Figure 1 show the percentage of classes, percentage classes show that 56.5% tweets are happy tweets and 43.5% tweets are related to unhappy tweets.

1) Data pre-processing

Datasets contain unnecessary data in raw form that can be unstructured or semi-structured. Such unnecessary data

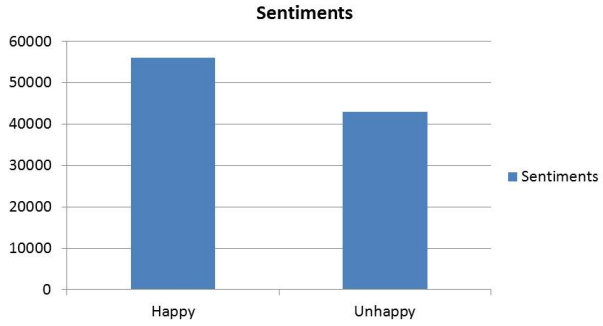

FIGURE 1: Countplot showing class-wise data distribution

increases training time of the model and might degrades its performance. Pre-processing plays a vital role in improving the efficiency of ML models and saving computational resources. Text pre-processing boosts the prediction accuracy of the model [20]. Following steps are performed in pre-processing; tokenization, case-conversion, stopwords removal and removal of numbers.

2) Feature extraction

After the data pre-processing step, the next essential step is the choice of features on a refined dataset. Supervised machine learning classifiers require textual data in vector form to get trained on it. The textual features are converted into vector form using TF and TF-IDF techniques [21]–[23] in this work. Features extraction techniques not only convert textual features into vector form but also helps to find significant features necessary to make predictions. For the most part all features do no contribute to the prediction of the target class. That is the reason feature extraction is the important part in the recognition of happy and unhappy related tweets.

What actually Term Frequency(TF) means that, according to what often the term arises within the document? It’s measured by TF. This will be achievable with the intention of a term would seem a lot further in lengthy documents than short documents because every document is variant in extent. Like the mode about standardization:

$$TF(t) = \frac{\text{No. of times term t shows in a document}}{\text{Total no. of terms inside document}} \quad (1)$$

The term frequency be frequently divided with the document length (the total number of terms in the document). IDF: Inverse documents frequency proceeds to find how much a term is significant within the text. Every term is measured equally when TF is computed. Nevertheless it is recognized that convinced terms, like "is", "of", and "that",

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

can show much more times except contain small prominence. Therefore frequent terms are needed to be weighed down as level up exceptional ones, through calculating following:

$$IDF(t) = log(e)\frac{\text{Total No. of documents}}{\text{No. of documents through term t in it}} \quad (2)$$

Term frequency (TF) is utilized regarding data recovery and shows how regularly an articulation (term, word) happens in a report.

C. PROPOSED MODELS FOR TWEETS SENTIMENT CLASSIFICATION

In this section classifiers utilized for tweet classification will be discussed. Figure 2 shows the proposed methodology of data and work flow of this research work. This work utilized five supervised machine learning algorithms: Support Vector Machines (SVM) , Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Gradient Boosting model (GBM), Logistic Regression (LR) and Voting Classifier(Logistic Regression + Stochastic Gradient Descent classifier).

1) Random forest

RF is a tree based classifier in which input vector generated trees randomly. RF uses random features, to create multiple decision trees, to make a forest. Then class labels of test data are predicted by aggregating voting of all trees. Higher weights are assigned to the decision trees with low value error. Overall prediction accuracy is improved by considering trees with low error rate.

2) Support Vector Machine

The Support vector machine (SVM) is understood that executes properly as sentiment analysis [24]. SVM typifies preference, confines and makes usage of the mechanisms for the assessment and examines records, which are attained within the index area [25]. Arrangements of vectors for every magnitude embody crucial details. Information (shown in form of vector) has been arranged in type to achieve this target. Next, the border is categorized in two training sets by stratagem. This is a long way from any area in the training samples [26]. Support-vector machines in machine learning includes focused learning models connected to learning evaluations which inspect material that is exploited to categorize, also revert inspection [27].

3) Naive Bayes

Ordering approach, Naive Bayes(NB), with sturdy (naive) independent assumptions among stabilities, depends on Bayes' Theorem. NB classifier anticipates that the proximity of a specific element of class that is confined to the closeness of a couple of different variables. For instance, a natural organic product is presumably viewed as an apple, if its shading is dark red, if type of it is round and it is roughly 3 creeps in expansiveness. In machine learning, Naive Bayes classifiers are a gathering of essential "probabilistic classifiers" considering applying Bayes' speculation with gullible opportunity assumptions between the features. They are considered as the minimum problematic Bayesian network models.

D. DECISION TREE

DT algorithm is the category of supervised ML and is being widely used in regression and classification tasks. Selection of root node of a tree of each level is its main challenge which is called as attribute selection [28]. Gini index and information gain are most commonly used methods for attribute selection. In this study, gini index is used to find probability of root node by calculating sum of squares of attribute values and then subtracted by 1.

1) Gradient Boosting Machine

GBM is a ML based boosting model and is widely being used for regression and classification tasks, which works by a model formed by ensemble of weak prediction models, commonly decision trees [29], [30]. In boosting, weak learners are converted to strong learners. Every new generated tree is a modified form of previous one and use gradient as loss function. Loss calculate the efficiency of model coefficients fitting over underlying data. Logically loss function is used for model optimization.

2) Logistic Regression

In LR class probabilities are estimated on the basis of output such as they predict if the input is from class X with probability x and from class Y with probability y. If x is greater than y, then predicted output class is X, otherwise Y. Insight, a logistic approach used for demonstrating the probability of a precise group or else, occurrence is obtainable, e.g., top/bottom, white/black, up/down, positive/negative or happy/unhappy. This is able to stretch out and to show a small number of classes about events, for example, to make a decision if a image includes a snake, hound, deer, etc., every article being famous in the image would be appointed a probability wherever in the series of 0 and 1 with whole addition to one [31].

3) Stochastic Gradient Descent

Gradient Descent's types include Stochastic Gradient Descent (SGD). SDGD is an iterative strategy for advancing a target work with appropriate perfection properties (for example differentiable or sub differentiable) [32]. Degree of advancement is calculated by it in light of development of alternative variables. It is very well, may be viewed as a stochastic guess of inclination plummet advancement, since it replaces the genuine angle (determined from the whole informational index) by a gauge thereof (determined from an arbitrarily chosen subset of the information) [33].

4) Voting Classifier

Voting Classifier(VC) is a cooperative learning which engages multiple individual classifiers and combines their pre-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

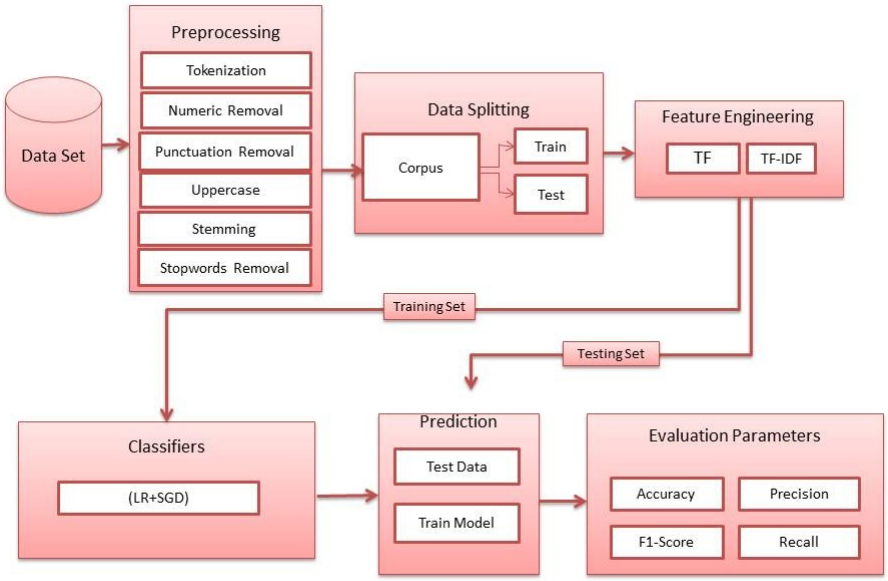

FIGURE 2: Proposed methodology architecture diagram.

dictions, which could attain better performance than a single classifier [34]. It has been exhibited that the mixture of multiple classifiers could be more operative compared to any distinct ones [35]. The VC is a meta-classifier for joining tantamount or hypothetically exceptional ML classifiers for order through greater part throwing a voting form. It executes "hard" and "soft" casting a ballot. Hard voting gives the researcher the chance to foresee the class name in place of the last class mark that has been anticipated often through models of characterization. Soft voting provides researchers the chance of anticipating the class names through averaging the class-probabilities [36].

Nowadays, progressively, researchers are concerned with cooperative learning because it gives better results [37]. This research contains voting classifiers by merging two classifiers that are VC(LR-SGD) and with the help of this voting classifier maximum results are achieved. SGD is an iterative strategy for advancing a target work with appropriate perfection properties (for example differentiable or sub differentiable). In this research, a voting classifier with multiple parameters is used, that has used two individual classifiers that are LR and SGD and also passes another parameter which is “voting” as “soft”. SGD is used to solve problems like redundancies in dataset and for big data. It performs classification by penalty and loss function [38]. It is similar to gradient decent and looks at one sample for each step [39]. On the other hand, LR calculates posterior probability $p(Ct|v)$ by applying sigmoid function on input for binary classification [40]. VC can be explained as:

$$\widehat{p} = argmax\{\sum_{i}^{n} LR_i, \sum_{i}^{n} SGD_i\}. \quad (3)$$

Here $\sum_{i}^{n} LR_i$ and $\sum_{i}^{n} SGD_i$ both will give prediction probabilities against each test example. After that, the probabilities for each test example by both LR and SGD passes through the soft voting criteria as shown in Figure 3.

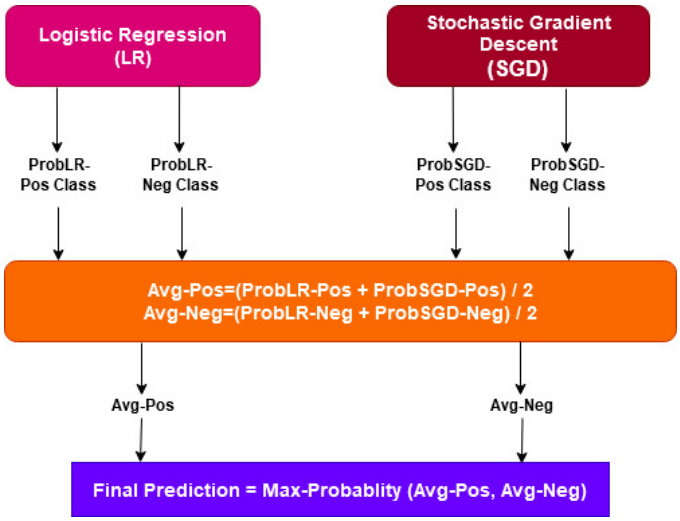

FIGURE 3: Proposed Voting Classifier Architecture (LR-SGD).

The functionality of the VC can be explained with an example. When a given sample passes through the LR and SGD, probability score is assigned to each class (that can be positive or negative). Let LR’s probability score be 0.966, 0.024, and for $ProbLR - Pos$ and $ProbLR - Neg$ classes and SGD’s probability score be 0.997, 0.002 for $ProbSGD - Pos$, and $ProbSGD - Neg$, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

Then the average probability for the two classes can be calculated as

Avg-Pos = (0.966+ 0.997)/2 = 0.9815
Avg-Neg = (0.024+ 0.002)/2 = 0.013

Final prediction is the $MaxProb(Avg-PosandAvg-Neg)$. In this example answer is the positive class. As predicted class is 'positive' and the actual class is also positive in the dataset. The proposed VC combines predicted probabilities of both classifiers to make the final decision. M_{LR} and M_{SGD} that are trained on the dataset and then predict the probability for both classes separately.An average probability is calculated for each class form the probability predicted by two classifiers. The decision function is then decides the final class of the review which is based on the maximum average probability for a class. The working mechanism of the LR-SGD is presented in Algorithm 1.

Algorithm 1 Ensembling of Logistic Regression and Stochastic Gradient Descent (LR-SGD).

Input: input data $(x,y)_{i=1}^{N}$
M_{LR} = Trained_ LR
M_{SGD} = Trained_ SGD

1: **for** $i = 1$ *to* M **do**
2: **if** $M_{LR} \neq 0$ & $M_{SGD} \neq 0$ & $training\_set \neq 0$ **then**
3: $ProbSGD - Pos = M_{SGD}.probibility(Pos - class)$
4: $ProbSGD - Neg = M_{SGD}.probibility(Neg - class)$
5: $ProbLR - Pos = M_{LR}.probibility(Pos - class)$
6: $ProbLR - Neg = M_{LR}.probibility(Neg - class)$
7: Decision function = $max(\frac{1}{N_{classifier}} \sum_{classifier}(Avg_{(ProbSGD-Pos,ProbLR-Pos)}, Avg_{(ProbSGD-Neg,ProbLR-Neg)}))$
8: **end if**
9: Return final label $\widehat{p}$
10: **end for**

E. EVALUATION METRICS

ML models are evaluated on many commonly used performance indicators such as accuracy, recall, precision and F1-score in classification tasks. Accuracy measures prediction correctness and is measured as:

$$Accuracy = \frac{\text{Number of correctly classified predictions}}{\text{Total predictions}} \quad (4)$$

while in case of binary classification, accuracy is measured as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

whereas TP is true positive , FP is false positive, TN is true negative, and FN is false negative and can be defined as [41].

TP: TP represents the positive predictions of a correctly predicted class.

FP: FP represents the negative predictions of a incorrectly predicted class.

TN: TN represents the negative predictions of a correctly predicted class.

FN: FN represents the positive predictions of a incorrectly predicted class

Precision measures the exactness of a classifier and determine percentage of positive labeled tuples that are actually positive. It can be measured as:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

While on the other hand recall measures completeness and it presents the percentage of correctly labelled true positive tuples. Recall can be measured as:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

For imbalance dataset, accuracy alone cannot be a good evaluation measure. F1 score, that is the harmonic mean of recall and precision, can help in such cases. It performs statistical analysis and computes score between 1 and 0 by considering both precision and recall of the model [42]. F1-score can be computed as:

$$F1score = 2\frac{precision.recall}{precision + recall} \quad (8)$$

IV. RESULTS AND DISCUSSION

This section provides the details of the experiment conducted in this research and the discussion of obtained results. Classification algorithms are tested using TF and TF-IDF features. Voting Classifier as an ensemble of Stochastic Gradient Descent and Logistic Regression gives highest accuracy. Table 2 presents the Accuracy, Recall, Precision and F1-score of classification with TF features.

TABLE 2: Classification result of all machine learning models using TF features.

Models	Accuracy	Precision	Recall	F1-Score
RF	74%	74%	79%	77%
SVM	76%	76%	80%	78%
NB	75%	75%	78%	75%
DT	74%	74%	77%	76%
GBM	74%	72%	79%	76%
LR	76%	79%	82%	80%
VC(LR-SGD)	78%	78%	84%	81%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

Figure 4 presents the results of all the classifiers and comparison between them. By using the TF feature. It can be seen that the Voting Classifier is best with accuracy 78% among all classifiers.

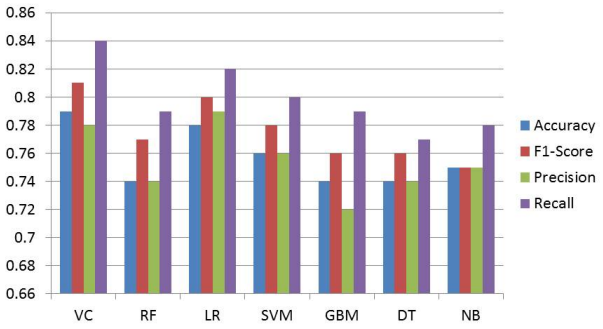

FIGURE 4: Classification result comparison of all machine learning models using TF features.

A Voting Classifier displays best outcome when it works with Stochastic Gradient Descent and Logistic Regression and provides maximum accuracy.

TABLE 3: Classification result of all machine learning models using TF-IDF features.

Models	Accuracy	Precision	Recall	F1-Score
RF	74%	74%	79%	77%
SVM	76%	76%	80%	78%
NB	75%	75%	75%	78%
DT	74%	74%	77%	76%
GBM	74%	72%	79%	76%
LR	78%	79%	82%	80%
VC(LR-SGD)	79%	78%	84%	81%

Table 3 shows the accuracy, recall, precision and F1-score of classification with TF-IDF technique. Voting classifier achieved the highest accuracy value with 79% and LR achieved 78%. LR achieved the highest precision value with 79% and the proposed model achieved 78%. Proposed model achieved the highest recall and f1 score with 84% and 81% values respectively. LR individually show reasonable results with 80% recall and 80% F1-score.

Figure 5 shows the results of all the classifiers and comparison between them Using TF-IDF feature. It can be seen clearly that the proposed voting classifier is performing best in terms of accuracy, recall and f1 score among all classifiers.

V. STABILITY OF THE PROPOSED MODEL

Different experiments are performed on the proposed approach to verify its stability under the different types of datasets. The second dataset used contains Dresses, Tweets of 20 garment products, Pants , Sweaters and KnitsBlouses.

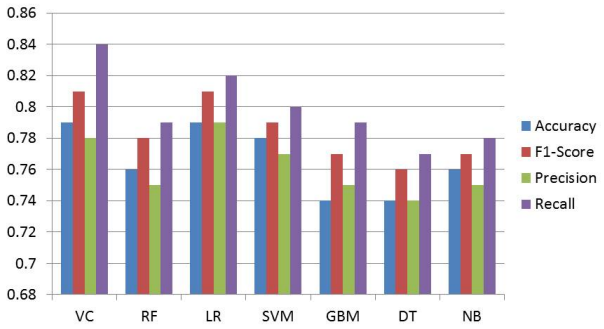

FIGURE 5: Classification result comparison of all machine learning models using TF-IDF features.

Rating range assigned by users are from 1 to 5, as shown in Figure 6. Dataset 3 consists of Tweets that contain supportive and hostile reviews, and that are to be classified as non hatred and hatred. The details of both datasets are presented in table 4 and table 5.

TABLE 4: Details of datasets used to check proposed model stability.

Dataset name	Records	Classes
Dataset 2: Women's ecommerce clothing reviews	23,486	5
Dataset 3: Sentiment Analysis of Hatred speech detection on Twitter	29,530	2

TABLE 5: List of features of dataset with their description.

Feature	Description
Women's e-commerce clothing reviews	
Clothing ID	Unique ID of the product.
Title	Title of the review.
Age	Age of the reviewer.
Review Text	Original text posted by the user.
Positive Feedback Count	Number of positive feed backs on the review.
Rating	Product rating by the reviewer.
Recommended IND	Whether product is recommended or not.
Department Name	Name of the product department.
Division Name	Product division name.
Class Name	Type of the product.
Twitter sentiment for hatred speech detection	
Label	Class label for the tweet.
Sentiment Text	Original text posted by the user

.

The proposed model which is ensemble of LR and SGD is applied on both dataset and the results are shown in 7. Results revealed that the proposed model outperformed other classifiers on both binary and multi-class classification dataset. The complete classification report of all classifiers is shown in table 6.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

TABLE 6: Classification report of both datasets.

Classifier	Dataset-2			Dataset-3		
	Precision	Recall	F1-score	Precision	Recall	F1-score
RF	0.49	0.59	0.47	0.85	0.84	0.84
SVM	0.60	0.62	0.62	0.88	0.88	0.88
NB	0.38	0.56	0.41	0.94	0.93	0.91
DT	0.49	0.50	0.49	0.93	0.94	0.93
GBM	0.58	0.63	0.58	0.89	0.89	0.89
LR	0.59	0.63	0.58	0.88	0.88	0.88
VC(LR-SGD)	0.62	0.63	0.62	0.94	0.94	0.92

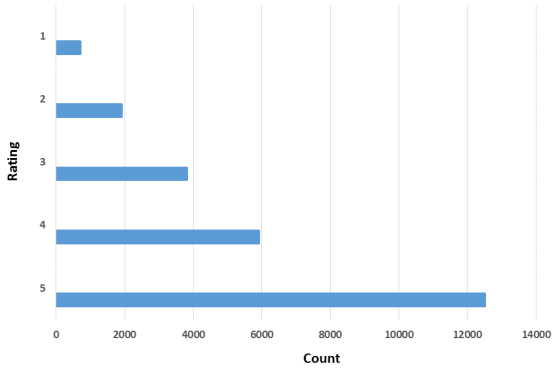

FIGURE 6: Ratings assigned by customers.

TABLE 7: Accuracy of classifiers with TF-IDF.

Classifier	Accuracy	
	Dataset-2	Dataset-3
RF	0.585	0.844
SVM	0.620	0.883
NB	0.559	0.934
DT	0.495	0.936
GBM	0.628	0.885
LR	0.629	0.880
VC(LR-SGD)	0.631	0.940

As it can be observed from the above results all traditional machine learning based models did not perform well on all three dataset. The proposed Voting Classifier ensemble outperforms all other traditional models. If the reason of poor performance of RF is explored specially on dataset 2 then it is concluded as RF is an ensemble technique which is composed of joining multiple trees which helps to deal with outliers and noise. But for the large size dataset it is difficult to grasp relationship in input data [43]. RF works on bootstrap samples and if samples are not fully representatives, prediction can be inaccurate.

GBM converts weak learners to strong learners and it is sensitive to noise and outliers. If it gets trained on weak learners due to noisy data which can cause overfitting. GBM shows results similar to RF on the Twitter dataset but it perform better on Dataset-2 and Dataset-3.

NB works on the assumption that features are independent of one another, that is rarely correct. Features commonly depends upon each other and that is the major reason of low performance on NB on diverse featured dataset. NB performed better than tree based models (RF and GBM) on Twitter dataset but worse on dataset-2 and dataset-3.

SVM works on by separating classes with the help of hyperplane, and shows good results on binary classification problems. It separates class labels by constructing hyperplanes between classes but for multiclass problems mostly SVM is not able to separate the data. SVM performed better than most of the tradition Ml models like RF, GBM and NB on all datasets.

To overcome the deficiencies of ML models, this study utilized combination of ML models as voting classifiers. It can be seen clearly in table 3, 7 and 6, proposed VC(LR-SGD) outperformed on all datasets as compared to tradition ML based models.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a novel combination of LR and SGD as a voting classifier for emotion recognition by classifying tweets as happy or unhappy. Our experiments showed that one can improve the performance of models by recognizing patterns efficiently and through effective averaging combination of models. Experiments are conducted to test seven machine learning models that are; (1) SVM, (2) RF, (3) GBM, (4) LR, (5) DT, (6) NB and (7) VC(LR-SGD). This study also employed two feature representation techniques Tf and TF-IDF. The results showed that all models performed well on tweet dataset but our proposed voting classifier VC(LR-SGD) outperforms by using both TF and TF-IDF among all. Proposed model achieves the highest results using TF-IDF with 79% Accuracy, 84% Recall and 81% F1-score. The proposed model is further validated on two more dataset and achieved robust results. The future work will compare more feature engineering techniques and explore more combinations of ensemble models to improve the performance. In addition, new techniques will be investigated to deal with sarcastic comments.

REFERENCES

[1] Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66:170–179, 2014.

[2] A. Sarlan, Chayanit Nadam, and S. Basri. Twitter sentiment analysis. Proceedings of the 6th International Conference on Information Technology and Multimedia, pages 212–216, 2014.

[3] Abdullah Alsaeedi and Mohammad Zubair Khan. A study on sentiment analysis techniques of twitter data. International Journal of Advanced Computer Science and Applications, 10(2):361–374, 2019.

[4] Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. Lexicon based feature extraction for emotion text classification. Pattern Recognition Letters, 93:133–142, jul 2017.

[5] Joan Capdevila, Jesús Cerquides, Jordi Nin, and Jordi Torres. Tweet-SCAN: An event discovery technique for geo-located tweets. Pattern Recognition Letters, 93:58–68, jul 2017.

[6] Teresa Alsinet, Josep Argelich, Ramón Béjar, Cèsar Fernández, Carles Mateu, and Jordi Planes. An argumentative approach for discovering relevant opinions in twitter with probabilistic valued relationships. Pattern Recognition Letters, 105:191–199, apr 2018.

[7] Weiling Chen, Yan Zhang, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. Unsupervised rumor detection based on users' behaviors using neural networks. Pattern Recognition Letters, 105:226–233, apr 2018.

[8] Heba Hakh, Ibrahim Aljarah, and Bashar Al-Shboul. Online social media-based sentiment analysis for us airline companies. 04 2017.

[9] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci., 181:1138–1152, 03 2011.

[10] Muhammad Umer, Saima Sadiq, Muhammad Ahmad, Saleem Ullah, Gyu Sang Choi, and Arif Mehmood. A novel stacked cnn for malarial parasite detection in thin blood smear images. IEEE Access, 8:93782–93792, 2020.

[11] Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. Aggression detection through deep neural model on twitter. Future Generation Computer Systems, 2020.

[12] Furqan Rustam, Imran Ashraf, Arif Mehmood, Dr. Saleem Ullah, Gyu Sang Choi, and Khan Yar. Tweets classification on the base of sentiments for us airline companies. Entropy, 21, 11 2019.

[13] Cicero Dos Santos and Maira Gatti de Bayser. Deep convolutional neural networks for sentiment analysis of short texts. 08 2014.

[14] Mustafe Mohamed. Mining and mapping halal food consumers: A geo-located twitter opinion polarity analysis. Journal of Food Products Marketing, 24:1–22, 12 2017.

[15] Huma Parveen and Shikha Pandey. Sentiment analysis on twitter data-set using naive bayes algorithm. pages 416–419, 01 2016.

[16] Khaled Alomari, Hatem Elsherif, and Khaled Shaalan. Arabic tweets sentimental analysis using machine learning. pages 602–610, 06 2017.

[17] Donia Gamal, Marco Alfonse, El-Sayed El-Horbarty, and Abdel-Badeeh M.Salem. Twitter benchmark dataset for arabic sentiment analysis. International Journal of Modern Education and Computer Science, 11:33 – 38, 01 2019.

[18] Akshi Kumar and Geetanjali Garg. Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications, 78, 03 2019.

[19] Kashfia Sailunaz. Emotion and sentiment analysis from twitter text. Journal of Computational Science, 36, 07 2019.

[20] V. Kalra and R. Aggarwal. Importance of text data preprocessing & implementation in rapidminer. In Proceedings of the First International Conference on InformationTechnology and Knowledge Management, volume 14, page 71–75, 2018.

[21] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 841–842. ACM, 2010.

[22] Scikit learn. Scikit-learn feature extraction with countvectorizer. https://scikit- learn. org/stable /modules /generated/sklearn. feature_extraction. text.Count/. Online: accessed 5 April 2019.

[23] Scikit learn. Scikit-learn feature extraction with tf/idf. https://scikit-learn.org/stable/modules /generated/sklearn. feature_extraction. text.Tfidf/. Online: accessed 5 April 2019.

[24] Preeti Routray, C. K. Swain, and S. Mishra. A survey on sentiment analysis. International Journal of Computer Applications, 76:1–8, 2013.

[25] Ali Harb, Michel Plantié, Gerard Dray, Mathieu Roche, François Trousset, and Pascal Poncelet. Web opinion mining: How to extract opinions from blogs? In Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, CSTST '08, page 211–217, New York, NY, USA, 2008. Association for Computing Machinery.

[26] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. EMNLP, 10, 06 2002.

[27] Kristin P Bennett and Colin Campbell. Support vector machines: hype or hallelujah? Acm Sigkdd Explorations Newsletter, 2(2):1–13, 2000.

[28] David Hand. Data Mining. 09 2006.

[29] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7:21, 2013.

[30] Jerome Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 11 2000.

[31] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. Applied logistic regression, volume 398. John Wiley & Sons, 2013.

[32] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in neural information processing systems, pages 315–323, 2013.

[33] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, pages 177–186. Springer, 2010.

[34] Yong Zhang, Hongrui Zhang, Jing Cai, and Binbin Yang. A weighted voting classifier based on differential evolution. Abstract and Applied Analysis, 2014:1–6, 05 2014.

[35] Michael A. Arbib. The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA, USA, 2nd edition, 2002.

[36] Madiha Khalid, Imran Ashraf, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, and Gyu Sang Choi. Gbsvm: Sentiment classification from unstructured reviews using ensemble classifier. Applied Sciences, 10(8):2788, 2020.

[37] Stan Z. Li and Anil Jain. Encyclopedia of Biometrics. Springer Publishing Company, Incorporated, 2015.

[38] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[39] Jose Silva, Isabel Praça, Tiago Pinto, and Zita Vale. Energy consumption forecasting using ensemble learning algorithms. In International Symposium on Distributed Computing and Artificial Intelligence, pages 5–13. Springer, 2019.

[40] Marco Vicente, Fernando Batista, and Joao P Carvalho. Gender detection of twitter users based on multiple information sources. In Interactions Between Computational Intelligence and Mathematics Part 2, pages 39–54. Springer, 2019.

[41] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood. A novel stacked cnn for malarial parasite detection in thin blood smear images. IEEE Access, 8:93782–93792, 2020.

[42] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B. W. On. Fake news stance detection using deep learning architecture (cnn-lstm). IEEE Access, 8:156695–156706, 2020.

[43] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

ANAM YOUSAF received her BS degree in Department of Computer Science, Islamia University of Bahawalpur, Pakistan . Since Nov-2018, she got herself enrolled in Phd of Computer Science(KFUEIT). Her recent research interests are related to Data Mining, mainly working Natural Language Processing based problems.

MUHAMMAD UMER received his BS degree in Department of Computer Science, Khwaja Fareed University of Engineering & IT(KFUEIT), Pakistan (Oct-2014 to Oct-2018). Since Sep-2020, he got himself enrolled in Ph.D Computer Science(KFUEIT). He is also serving as Research Assistant at Fareed Computing & Research Center, KFUEIT, Pakistan. His recent research interests are related to Data Mining, mainly working machine learning & Deep Learning based IoT, Text Mining, and Computer Vision tasks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3047831, IEEE Access

SAIMA SADIQ working as Assistant Professor in Department of Computer Science at Government Degree College for Women. Since Sep-2020, she got herself enrolled in Ph.D Computer Science at Khwaja Fareed University of Engineering & IT (KFUEIT). Her recent research interests are related to Data Mining, Machine learning & Deep Learning based Text Mining.

SALEEM ULLAH was born in AhmedPur East, Pakistan in 1983. He received his B.Sc. and MIT degrees in Computer Science from Islamia University Bahawalpur and Bahauddin Zakariya University (Multan) in 2003 and 2005 respectively. From 2006 to 2009, he worked as a Network/IT Administrator in different companies. He completed his Doctorate degree from Chongqing University, China in 2012. From August 2012 to Feb 2016, he worked as an Assistant Professor in Islamia University Bahawalpur, Pakistan. Currently, he is working as an Associate Dean in Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan since February 2016. He has almost 14 years of Industry experience in field of IT. He is an active researcher in the field of Adhoc Networks, IoTs, Congestion Control, Data Science, and Network Security.

VAIBHAV RUPAPARA received Master of Science degree in Computer Science from Florida International University, Miami, FL, USA. He has worked on different domain including Finance and Healthcare. His expertise contributed towards achieving high quality, scalable deliverability with security. His recent research interests include Machine Learning, AI, Deep learning.

SEYEDALI MIRJALILI is an associate professor and the director of the Centre for Artificial Intelligence Research and Optimization at Torrens University Australia. He is internationally recognized for his advances in Swarm Intelligence and Optimization, including the first set of algorithms from a synthetic intelligence standpoint - a radical departure from how natural systems are typically understood - and a systematic design framework to reliably benchmark, evaluate, and propose computationally cheap robust optimization algorithms. Seyedali has published over 200 publications with over 25,000 citations and an H-index of over 55. As the most cited researcher in Robust Optimization, he is in the list of 1% highly-cited researchers and named as one of the most influential researchers in the world by Web of Science in Computer Science and Engineering. Seyedali is a senior member of IEEE and an associate editor of several journals including Neurocomputing, Applied Soft Computing, Advances in Engineering Software, Applied Intelligence, and IEEE Access.

MICHELE NAPPI received the laurea degree (cum laude) in Computer Science from the University of Salerno, Italy, in 1991, the M.SC. degree in Information and Communication Technology from I.I.A.S.S. "E.R. Caianiello," in 1997, and the Ph.D. degree in Applied Mathematics and Computer Science from the University of Padova, Italy, in 1997. He is currently a full professor of Computer Science at the University of Salerno. Author of more than 180 papers in peer review international journals, international conferences and book chapters, He is co-editor of several international books. His research interests include pattern recognition, image processing, image compression and indexing, multimedia databases and biometrics, human computer interaction, vr/ar. Dr. Nappi serves as associate editor and managing guest editor for several international journals. He is also member of tpc of international conferences. He is team leader of the Biometric and Image Processing Lab (BIPLAB) and received several international awards for scientific and research activities. IEEE Senior Member, GIRPR/IAPR Member, He has been the President of the Italian Chapter of the IEEE Biometrics Council. In 2014 He was one of the founders of the spin off BS3 (biometric system for security and safety).