Personality Classification from Online Text using Machine Learning Approach

Alam Sher Khan[1], Hussain Ahmad[2], Muhammad Zubair Asghar[3]*
Furqan Khan Saddozai[4], Areeba Arif[5], Hassan Ali Khalid[6]
Institute of Computing and Information Technology
Gomal University, D.I. Khan, Pakistan

***Abstract*—Personality refer to the distinctive set of characteristics of a person that effect their habits, behaviour's, attitude and pattern of thoughts. Text available on Social Networking sites provide an opportunity to recognize individual's personality traits automatically. In this proposed work, Machine Learning Technique, XGBoost classifier is used to predict four personality traits based on Myers- Briggs Type Indicator (MBTI) model, namely Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P) from input text. Publically available benchmark dataset from Kaggle is used in experiments. The skewness of the dataset is the main issue associated with the prior work, which is minimized by applying Re-sampling technique namely random over-sampling, resulting in better performance. For more exploration of the personality from text, pre-processing techniques including tokenization, word stemming, stop words elimination and feature selection using TF IDF are also exploited. This work provides the basis for developing a personality identification system which could assist organization for recruiting and selecting appropriate personnel and to improve their business by knowing the personality and preferences of their customers. The results obtained by all classifiers across all personality traits is good enough, however, the performance of XGBoost classifier is outstanding by achieving more than 99% precision and accuracy for different traits.**

***Keywords*—Personality recognition; re-sampling; machine learning; XGBoost; class imbalanced; MBTI; social networks**

I. INTRODUCTION

Personality of a person encircles every aspect of life. It describes the pattern of thinking, feeling and characteristics that predict and describe an individual's behaviour and also influences daily life activities including emotions, preference, motives and health [1].

The increasing use of Social Networking Sites, such as Twitter and Facebook have propelled the online community to share ideas, sentiments, opinions, and emotions with each other; reflecting their attitude, behaviour and personality. Obviously, a solid connection exists between individual's temperament and the behaviour they show on social networks in the form of comments or tweets [2].

Nowadays personality recognition from social networking sites has attracted the attention of researchers for developing automatic personality recognition systems. The core philosophy of such applications is based on the different personality models, like Big Five Factor Personality Model [3], Myers- Briggs Type Indicator (MBTI) [4], and DiSC Assessment [5].

The existing works on personality recognition from social media text is based on supervised machine learning techniques applied on benchmarks dataset [6], [7], [8]. However, the major issue associated with the aforementioned studies is the skewness of the datasets, i.e. presence of imbalanced classes with respect to different personality traits. This issue mainly contributes to the performance degradation of personality recognition system.

To address the aforementioned issue different techniques are available for minimizing the skewness of the dataset, like Over-sampling, Under-sampling and hybrid-sampling [9]. Such techniques, when applied on the imbalanced datasets in different domain, have shown promising performance in terms of improved accuracy, recall, precision, and F1-score [10].

In this work, a machine learning technique, namely, XGBoost is applied on the benchmark personality recognition dataset to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P). Furthermore, to improve the performance of the system, resampling technique [11] is also utilized for minimizing the skewness of the dataset.

A. Problem Statement

Predicting personality from online text is a growing trend for researchers. Sufficient work has already been carried out on predicting personality from the input text [6, 7, 8].

However, more work is required to be carried out for the performance improvement of the existing personality recognition system, which in most of the cases arises due to presence of imbalanced classes of personality traits. In the proposed work. A dataset balancing technique, called re-sampling is used for balancing the personality recognition dataset, which may result in improved performance.

B. Research Questions

RQ.1: How to apply supervised machine learning technique, namely XGBoost classifier for classifying personality traits from the input text?

RQ.2: How to apply a class balancing technique on the imbalanced classes of personality traits for performance improvement and what is the efficiency of the proposed technique w.r.t other machine learning techniques?

*Corresponding Author

RQ.3: What is the efficiency of the proposed technique with respect to other baseline methods?

C. Aims and Objective

1) Aim: The aim of this work is to classify the personality traits of a user from the input text by applying supervised machine learning technique namely XGBoost classifier on the benchmark dataset of MBTI personality. This work is the enhancement of the prior work performed by [6].

2) Objectives

a) Applying machine learning technique namely XGBoost classifier for personality traits recognition from the input text.

b) Applying re-sampling technique on the imbalanced classes of personality traits for improving the performance of proposed system.

c) Evaluating the performance of proposed model with respect to other machine learning techniques and base line methods.

D. Significance of Study

Personality is distinctive way of thinking, behaving and feeling. Personality plays a key role in someone's orientation in various things like books, social media sites, music and movies [12].

The proposed work on personality recognition is an enhancement of the work performed by [6]. Proposed work is significant due to the following reasons: (i) performance of the existing study is not efficient due to skewness, which will be addressed in this proposed work by applying re-sampling technique on the imbalanced dataset, (ii) proposed work also provide a basis for developing state of the art applications for personality recognition, which could assist organization for recruiting and selecting appropriate personnel and to improve their business by taking into account the personality and preferences of their customers.

II. Related Work

A review of literature pertaining to personality recognition from text is presented here in this section. The literature studies of this work is categorized into four sub groups, namely, i) Supervised learning techniques, ii) Un-supervised machine learning techniques, iii) Semi-supervised machine learning techniques and, iv) Deep learning techniques.

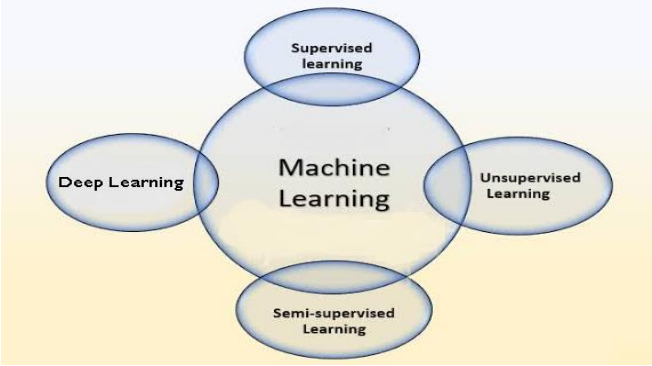

Fig. 1. Categorization Sketch of Literature Review.

Fig. 1 depicts the classification sketch of the literature review on personality recognition from text.

A. Supervised Learning Technique

These supervised learning algorithms are comprised of unlabeled data/ variables which is to be determined from labelled data, also called independent variables. The studies given below are based on supervised learning methodologies.

A system is proposed by [6] for analysing social media posts/ tweets of a person and produce personality profile accordingly. The work mainly emphasizes on data collection, pre-processing methods and machine learning algorithm for prediction. The feature vectors are constructed using different feature selection techniques such as Emolex, LIWC and TF/IDF, etc. The obtained feature vectors are used during training and testing of different kinds of machine learning algorithms, like Neural Net, Naïve Bayes and SVM. However, SVM with all feature vectors achieved best accuracy across all dimensions of Myers-Briggs Type Indicator (MBTI) types. Further enhancement can be made by incorporating more state of the art techniques.

MBTI dataset, introduced in [7] for personality prediction, which is derived from Reddit social media network. A rich set of features are extracted, and benchmark models are evaluated for personality prediction. The classification is performed using SVM, Logistic Regression, and (MLP). The classifier using all linguistic features together outperformed across all MBTI dimensions. However, further experimentation is required on more models for achieving more robust results. The major limitation is that the number of words in the posts are very large, which sometimes don't predict the personality accurately.

To predict personality from tweets, [8] proposed a model using 1.2 Million tweets, which are annotated with MBTI type for personality and gender prediction. Logistic regression model is used to predict four dimensions of MBTI. Binary word n-gram is used as a feature selection. This work showed improvement in I-E and T-F dimensions but no improvements in S-N and even slightly drop for P-J. In terms of personality prediction, linguistic features produce far better results. Incorporating enhanced dataset may improve performance.

A system was developed to recognize user personality using Big Five Factor personality model from tweets posted in English and Indonesian language [13]. Different classifiers are applied on the MyPersonality dataset. The accuracy achieved by Naive Bayes(NB) is 60%, which is better than the accuracy of KNN (58%) and SVM (59%).Although this work did not improve the accuracy of previous research (61%) yet achieved the goal of predicting the personality from Twitter-based messages. Using extended dataset and implementing semantic approach, may improve the results.

Personality assessment/ classification system based on Big5 Model was proposed for Bahasa Indonesian tweets [14]. Assessment is made on user's words choice. The machine learning classifiers, namely, SVM and XGBoost, are implemented on different parameters like existence of (n_gram minimum and n_gram weighted), removal of stop words and using LDA. XGBoost performed better than the SVM under

the same data and same parameter setting. Limited dataset of only 359 instances for training and testing is the main drawback of their work.

Automatic identification of Big Five Factor Personality Model was proposed by [15] using individual status text from Facebook. Various techniques like Multinomial NB, Logestic Regression (LR) and SMO for SVM are used for personality classification. However, MNB outperformed other methods. Incorporating feature selection and more classifiers, may enhance the performance.

Personality profiling based on different social networks such as Twitter, Instagram and Foursquare performed by [16]. Multisource large dataset, namely NUS-MSS, is utilized for three different geographical regions. The data is evaluated for an average accuracy using different machine learning classifiers. When the different data sources are concatenated in one feature vector, the classification performance is improved by more than 17%. Available dataset may be enriched from multi (SNS) by user's cross posting for better performance.

The performance of different ML classifiers are analysed to assess the student's personality based on their Twitter profiles by considering only Extraversion trait of Big 5 [17]. Different machine learning algorithms like Naïve Bayes, Simple logistic, SMO, JRip, OneR, ZeroR, J48, Random Forest, Random Tree, and AdaBoostM1, are applied in WEKA platform. The efficiency of the classifiers is evaluated in terms of correctly classified instances, time taken, and F-Measures, etc. OneR algorithm of rules classifier show best performance among all, producing 84% classification accuracy. In future, all dimensions of Big5 can be considered for evaluation to get more insight.

The performance of different classifier is evaluated by [18] using MBTI model to predict user's personality from the online text. Various ML classifiers, namely Naïve Bayes, SVM, LR and Random Forest, are used for estimation. Logistic Regression received a 66.5% accuracy for all MBTI types, which is further improved by parameter tuning. Results may further be improved by using XGBoost algorithm, which remained winner of most Kaggle and other data science competitions.

The oversampling and undersampling techniques are compared by [11] for imbalance dataset. Classification perform poorly when applied on imbalanced classes of dataset. There are three approaches (data level, algorithmic level and hybrid) that are widely used for solving class imbalance problem. Data level method is experimented in this study and result of Over-sampling method (SMOTE) is better than under-sampling technique (RUS). More re-sampling techniques need to be evaluated in future.

Authors in [19] briefly discussed and explained the early research for the classification of personality from text, carried out on various social networking sites, such as Twitter, Blogger, Facebook and YouTube on the available datasets. The methods, features, tools and results are also evaluated. Unavailability of datasets, lack of identification of features in certain languages, and difficulty in identifying the requisite pre-processing methods, are the issues to be tackled. These issues can be addressed by developing methods for non-English language, introducing more accurate machine learning algorithms, implementing other personality models, and including more feature selection for pre-processing of data.

Twitter user's profiles are used for accurate classification of their personality traits using Big5 model [20]. Total 50 subjects with 2000 tweets per user are assessed for prediction. Users content are analysed using two psycholinguistic tools, namely LIWC and MRC. The performance evaluation is carried out using two regression models, namely ZeroR and GP. Results for "openness" and "agreeableness" traits are similar as that of previous work, but less efficient results are shown for other traits. Extended dataset may improve the results.

A connection has been established between the users of Twitter and their personality traits based on Big5 model [21]. Due to inaccessibility of original tweets, user's personality is predicted on three parameters that are publicly available in their profiles, namely (i) followers, (ii), following, and (iii) listed count. Regression analysis is performed using M5 rules with 10-fold cross validation. RMSE of predicted values against observed values is also measured. Results show that based on three counts, user's personality can be predicted accurately.

TwiSTy, a novel corpus of tweets for gender and personality prediction has been presented by [22] using MBTI type Indicator. It covers six languages, namely Dutch, German, French, Italian, Portuguese and Spanish. Linear SVM is used as classifier and results are also tested on Logistic Regression. Binary features for character and word (n-gram) are utilized. It outperformed for gender prediction. For personality prediction, it outperformed other techniques for two dimensions: I-E and T-F, but for S-N and J-P, this model did not show improvement. In future, the model can be trained enough to predict all four dimensions of MBTI efficiently.

The Table I represents the summaries of above cited studies for classification and prediction of user's personality using Supervised Machine Learning strategies.

B. Unsupervised Learning Approach

Unsupervised learning classifiers are using only unlabeled training data (Dependent Variables) without any equivalent output variables to be predicted or estimated.

The Twitter data was annotated by [23] for 12 different linguistic features and established a correlation between user's personality and writing style with different cross-region users and different devices. Users with more than one tweets are considered for evaluation. It was observed that Twitter users are secure, unbiased and introvert as compared to the users posting from iPhone, blackberry, ubersocial and Facebook platforms. More Twitter data for classification may enhance the efficiency of personality identification model.

TABLE I. PERSONALITY RECOGNITION BASED WORK USING SUPERVISED MACHINE LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Performance	Limitation and Future Work
1	Bharadwaj et al. (2018) [6]	Personality prediction from online text	SVM, Neural Net and Naïve Bayes TF-IDF, Emolex, LIWC and ConceptNet	SVM with all feature vectors achieved best accuracy across all dimensions of MBTI	Less weightage is given to the word's gravity. Incorporating more state-of-the-art techniques in future will yield better result.
2	Gjurković and Šnajder (2018) [7]	Personality classification of Reddit user's posts.	SVM, Logistic Regression and MLP with linguistic features	MLP using all linguistic features together outperform across all MBTI dimensions	Demographic data like age and gender is not considered Accuracy of T/F dichotomy may be improved in future.
3	Plank and Hovy (2015) [8]	Personality and gender prediction from tweets.	Logistic regression Model and Binary word n-gram is used as a feature selection.	Accuracy for personality prediction: I/E = 72.5% S/N = 77.5% T/F = 61.2 % J/P = 55.4%	A lot of Gap between general population personality types and this corpus personality types. Incorporating of enhanced dataset will improve the performance.
4	Pratama and Sarno (2015) [13]	To recognize user personality using Big-5 personality model from tweets posted in English and Indonesian language	Supervised • KNN • NB • SVM	Accuracy KNN = 58% NB = 60% SVM = 59%	Using extended dataset and implementing semantic approach, may improve the results.
5	Ong et al. (2017b) [14]	A personality assessment based on Big5 Model for Bahasa Indonesian tweets using user's words choice.	Supervised • XGBoost • SVM	Accuracy XGBoost = 97.99% SVM = 76.23%	Limited dataset of only 359 instances for training and testing is the main drawback of this work.
6	Alam et al. (2013) [15]	Automatic identification of Big Five Factor Personality Model using individual status text from Facebook	Multinomial NB, Logestic Regression (LR) and SMO for SVM are used for personality classification	MNB = 61.79% BLR = 58.34% SMO = 59.98% ›MNB outperformed other methods	Incorporating feature selection and more classifiers, may enhance the performance.
7	Buraya et al. (2017) [16]	Multisource large dataset, namely NUS-MSS, is utilized for personality profiling.	Supervised	By concatenating different data sources in one feature vector, the classification performance is improved by more than 17%.	In future the available dataset may be enriched from multi (SNS) by user's cross posting for better performance.
8	Ngatirin et al. (2016) [17]	Using different ML classifiers to assess the student's personality based on their Twitter profiles.	Naïve Bayes, Simple logistic, SMO, JRip, OneR, ZeroR, J48, Random Forest, Random Tree, and AdaBoostM1,	OneR with F1_Score = 0.837 outperform among all.	In future, all dimensions of Big5 can be considered for evaluation to get more insight.
9	Chaudhary et al. (2018) [18]	To predict user's personality from the online text using MBTI model.	Supervised learning methodology namely Naïve Bayes, SVM, LR and Random Forest, are used for estimation.	Accuracy NB = 55.89% LR = 66.59% SVM = 65.44%	Lower accuracy is due using traditional classifiers. Deep learning approach will definitely improve the performance.
10	Kaur and Gosain (2018) [11]	Comparing of oversampling and undersampling techniques for imbalance dataset.	Decision tree algorithm C4.5 is used.	Result of Over-sampling method (SMOTE) is better than under-sampling technique (RUS).	More re-sampling techniques need to be evaluated in future.
11	Ong et al. (2017a) [19]	Classification of personality from text, carried out using various social networking sites.	Survey paper using supervised learning approcah	Best result among all was attained by twitter with 91.9% accuracy using words frequency.	Unavailability of datasets, and lack of identification of features in certain languages, are the issues to be tackled. In future methods for non-English language may need to be developed.
12	Golbeck et al. (2011) [20]	User's Twitter profiles for accurate classification of their personality traits using Big5 model.	Two regression models, namely ZeroR and GP are used.	Accuracy Higher for Open = 75.5% Lower for Neuro =42.8%	Extended dataset may improve the results.

13	Quercia et al. (2011) [21]	To establish a connection between the users of Twitter and their personality traits based on Big5 model.	Regression using M5 rules with 10-fold cross validation.	RMSE: O = 0.69 C = 0.76 E = 0.88 A = 0.79 N = 0.85	In future user personality classification may be utilized in marketing and recommender system.
14	Verhoeven et al. (2016) [22]	To predict gender and personality from a novel corpus of tweets, namely TwiSTy.	SVM and logistic Regression along words n_grams features.	Ƒ_score I/E =77.78 S/N =79.21 T/F = 52.13 J/P = 47.01 For italic lang:	In future, the model can be trained enough to predict all four dimensions of MBTI efficiently.

The purpose of the study carried out by [24], is to scrutinize the group-based personality identification by utilizing unsupervised trait learning methodology. Adawalk technique is utilized in this survey. The outcomes portray that while considering Micro- Ƒ1 score, the achievement of adawalk is exceptional with somewhat 7% for wiki, 3% for Cora, and 8% for BlogCatlog. While utilizing SoCE personality corpus, 97.74% Macro-Ƒ1 score was achieved by this approach. The drawback of this work is that it entirely depends on TƑ -IDƑ strategy, additionally the created content systems are not an impersonation of genuine social and interpersonal network like retweeting systems. Large and increased dataset will definitely enhance the performance of the proposed work in future.

An unsupervised personality classification strategy was accomplished by [25] to highlight the matter that to how extent different personalities collaborate and behave on social media site Twitter. Linguistic and statistical characteristics are utilized by this work and then tested on data corpus elucidated with personality model using human judgment. System investigation anticipate that psychoneurotic users comments more than secure ones and tend to develop longer chain of interaction.

An Unsupervised Machine learning methodology, namely, Ķ-Meaṅs was accomplished by [26] to recognize the network visitors' trait and personality. This proposed work is based on the quantifiable contents of the website. The obtained results portray that this strategy can be utilized to predict website and network visitors' personality traits, more accurately. Proposed system may be enhanced in future by adding more elements associated with websites and a greater number of websites for the better performance.

Author in [27] proposed a personality identification system using unsupervised approach based on Big-5 personality model. Different social media network sites are used for extraction and classification of user's traits. Linguistic features are exploited to build personality model. The system predict personality for an input text and achieved reasonable results. However, extended annotated corpus can boost the system's performance.

TABLE II. PERSONALITY RECOGNITION BASED WORK USING UN-SUPERVISED MACHINE LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Outcome	Limitation and Future Work
1	Celli (2011) [23]	Personality classification from individual's writing pattern	Un_supervised Score-based	Mean Accuracy =0.6651 and Mean validity= 0.6994	Additional Tweets for personality recognition may improve the accuracy of this proposed model.
2	Sun et al. (2019) [24]	group-based personality identification	Un_supervised Adȧwalk	97.74% (Macro-Ƒ1)	Large and increased dataset will definitely enhance the performance of the proposed work in future
3	Celli and Rossi, (2012) [25]	Impact of linguistic characteristics on personality traits.	Un-supervised Statistics-based	78.29% (Accurȧcy)	More tweets are needed for efficient investigation
4	Chishti and Sarrafzadeh (2015) [26]	To recognize the network visitors' trait and personality	Uṅ-supervised Ķ-Mean	Ķ=10 is accurate score	System may be enhanced in future by adding more elements associated with websites and a greater number of websites for better performance
5	Celli (2012) [27]	Impact of linguistic characteristics on personality traits using Big Five Model	Un_supervised Score-based	81.43% (Accuracy)	Extended annotated corpus can boost the system's performance
6	Arnoux et al. (2017) [28]	Developing personality model to predict individual's Big Five personality traits on much fewer data using twitter.	Word-Embedding	68.5% (Accuracy)	Findings of this method are based on English Twitter data, which may be extended to other languages

A model was proposed by [28] that requires eight times fewer data to predict individual's Big Five personality traits. GloVe Model is used as Word embedding to extract the words from user tweets. Firstly, the model is trained and then tested on given tweets. Further, the data is tested on three other combinations: (i) GloVe with RR, (ii) LIWC with GP, and (iii) 3-Gram with GP, and the proposed model performed better with an average correlation of 0.33 over the Big-5 traits, which is far better than the baseline method. Findings of this method are based on English Twitter data, which may be extended to other languages. Similarly, the performance of the model can be examined with small number of tweets.

The Table II illustrates the concise detail of above cited studies regarding user's personality and traits identification from textual data using un-supervised machine learning approach.

C. Semi-Supervised Learning Approach

The studies carried out by using the combination of linguistic and lexicon features, supervised machine learning methodologies and different feature selection algorithms are known as semi-supervised ML approaches. The following studies have utilized the semi-supervised and hybrid strategy.

Multilingual predictive model was proposed by [29], which identified user's personality traits, age and gender, based on their tweets. SGD classifier with n-gram features, is used for age and gender classification, while LIWC with regressor model (ERCC) is used for personality prediction. An average accuracy of 68.5% has been achieved for recognition of user's attributes in four different languages. However, author profiling can be enhanced by performing experiments in multiple languages.

A technique was devised to detect MBTI type personality traits from social media (Twitter) in Bahasa Indonesian language [30]. Among 142 respondents, 97 users are selected with an average 2500 tweets per user. WEKA is used for building classification and training set. Three approaches are used for prediction from training set. i) Machine Learning, ii) Lexicon-based, and iii) linguistic Rules driven. Among all, Naïve Bayes outperformed the comparing methods in terms of better accuracy and time. Its accuracy for I/E trait is 80% while for S/N, T/F and J/P, its accuracy is 60%. Lower accuracy on the part of linguistic rule-driven and lexicon-based, are due to limited corpus in Bhasha Indonesia. It is observed that by increasing the training data set, accuracy may get improved.

A technique proposed for personality prediction from social media-based text using word count [31]. It works for both MBTI and Big5 personality models using 8 different languages. Four kinds of labelled corpus both for Big5 and BMTI are used for conducting the experiments. In each corpus, 1000 most frequently used words are selected. Prediction accuracy for "openness" trait of Big5 is higher across all corpus, while for MBTI, prediction accuracy for S/N dimension is greater than other dichotomies. Using only word count for prediction is the main drawback of the proposed system, which may be covered by introducing different features selection and ML algorithms.

Detail of the above quoted studies regarding personality classification using Semi-supervised Machine Learning Approach are presented in Table III.

TABLE III. PERSONALITY RECOGNITION BASED WORK USING SEMI-SUPERVISED MACHINE LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Performance	Limitation and Future Work
1	Arroju et al. (2015) [29]	Multilingual predictive model is used to identify user's personality traits, age and gender, based on their tweets.	›SGD classifier with n-gram features. › LIWC with regressor model (ERCC)	Accuracy = 68.5%	Accuracy may be improved by using different personality model. Similarly, author profiling can be further enhanced by performing experiments in multiple languages.
2	Lukito et al. (2016) [30]	To recognize MBTI type personality traits from social media (Twitter) in Bahasa Indonesian language.	›Machine Learning, ›Lexicon-based, and ›linguistic Rules driven	I/E trait = 80% S/N, T/F and J/P accuracy is 60%	Lower accuracy is due to limited corpus in Bhasha Indonesia. By increasing the training data set, accuracy may get improved.
3	Alsadhan and Skillicorn (2017) [31]	Personality prediction from social media-based text using word count	Based on word count	Accuracy for "openness" trait of Big5 is higher, while for MBTI, accuracy for S/N dimension is greater than all other dichotomies.	Using only word count for prediction is the main drawback of the proposed system, which may be covered by introducing different features selection and ML algorithms.

D. Deep Learning Strategy

Deep learning is a subcategory of machine learning (ML) in artificial intelligence (AI), where machines may acquire knowledge and get experience by training without user's interaction to make decisions. Based on experiences and learning from unlabeled and unstructured corpus, deep learning performs tasks repeatedly and get improvement and tweaking in results after each iteration. The studies given below are in summarized form, showing the prior work performed in Deep learning.

A deep learning classifier was developed, which takes text/tweet as input and predict MBTI type of the author using MBTI dataset [32]. After applying different pre-processing techniques embedding layer is used, where all lemmatized words are mapped to form a dictionary. Different RNN layers are investigated, but LSTM performed better than GRU and simple RNN. While classifying user, its accuracy is 0.028 (.676 × .62 × .778 × .637), which is not good. The predictive efficiency of this work may be improved by increasing the number of posts per user. As the model is tested on real life example of Donald trump's 30,000 tweets, which correctly predict his actual MBTI type personality.

A model proposed by [33] that takes snippet of post or text as input and classify it into different personality traits, such as (INFP, ENTP, and ISJF, etc.). Different classification methods like Softmax as baseline, SVM, Naïve Bayes, and deep learning, are implemented for performance evaluation. SVM outperformed NB and softmax with 34% train 33% test accuracy, while Deep learning model shows more improvement with 40% train and 38% test accuracy. However, the accuracy is still low as it doesn't even achieve 50 percent.

Personality classification system is proposed by [34], to recognize the traits from online text using deep learning methodology. AttRCNN model was suggested for this study utilizing hierarchical approach, which is capable of learning complex and hidden semantic characteristics of user's textual contents. Results produced are very effective, proving that using deep and complex semantic features are far better than the baseline features.

A deep learning model was suggested by [1] to classify personality traits using Big Five personality model based on essay dataset. Convolutional Neural Network (CNN) is used for this work to detect personality traits from input essay. Different pre- processing techniques like word n-grams, sentence, word and document level filtration and extracting different features are performed for personality traits classification. "OPN" traits achieved higher accuracy of 62.68% by using different configuration of features and among all five traits. In future, more features need to be incorporated and LSTM recurrent network may be applied for better results.

Table IV represents the outline of the works regarding automatic personality recognition system using Deep learning methodology.

TABLE IV. PERSONALITY RECOGNITION BASED WORK USING DEEP LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Outcome	Limitation and Future Work
1	Hernandez and Scott (2017) [32]	To predict and classify people into their MBTI types using their online textual contents.	Deep Learning ›RNN ›LSTM ›GRU ›BiLSTM	Accuracy I/E= 67.6% S/N=62.0% T/F=77.8% J/P=63.7%	The predictive efficiency of this work may be improved by increasing the number of posts per user.
2	Cui, and Qi (2018) [33]	A model that takes snippet of post or text as input and classify it into different personality traits.	Deep Learning Multi-layer LSTM	Over all accuracy= 38% I/E= 89.51% S/N=89.84 % T/F=69.09 % J/P=69.37 %	In future more deep learning techniques with more word embedding features may be exploited. Using of unsupervised technique will also give better results.
3	Xue et al. (2018) [34]	To recognize the personality traits from online text using deep learning methodology.	Deep Learning using AttRCNN Approach	MAE OPN= 0.3577 CON= 0.4251 EXT= 0.4776 AGR= 0.3864 NEU= 0.4273	In future these deep and complex semantic features will be used as input of regression classifiers for more improvement in the performance.
4	Majumder et al. (2017) [1]	To classify personality traits using Big Five personality model based on essay dataset.	Deep Learning ›CNN	Accuracy OPN= 62.68% CON= 56.73% EXT= 58.09% AGR= 56.71% NEU= 59.38%	In future more features need to be incorporated and LSTM recurrent network may be applied for better results.

III. Methodology

The working procedure of this proposed system are as follows: (i) Data acquisition and re-sampling, (ii) Pre-Processing and feature selection, (iii) Text-based Personality classification using MBTI model, (iv)Applying XGBoost for personality classification, (v) Comparing the efficiency of XGBoost with other classifiers, (vi) Applying different evaluation metrics.

A. Dataset Collection and Re-sampling

The publically available benchmark dataset is acquired from Kaggle [6]. This data set is comprised of 8675 rows, where every row represents a unique user. Each user's last 50 social media posts are included along with that user's MBTI personality type (e.g. ENTP, ISJF). As a result, a labelled data set comprising of a total 422845 records, is obtained in the form of excerpt of text along with user's MBTI type. Table V describes the detail of acquired dataset.

1) Re-Sampling: As pointed out by [6], the original dataset is totally skewed and unevenly distributed among all four dichotomies, described as follows: **I/E Trait:** I=6664 and, E= 1996, **S/N Trait:** S= 7466 and N= 1194, **T/F Trait:** T= 4685 and F= 3975, **J/P Trait:** J= 5231 and P= 3429. Whenever, an algorithm is applied on skewed and unbalanced classified dataset, the outcome always diverge toward the sizeable class and the smaller classes are bypassed for prediction. This drawback of classification is known as class imbalance problem (CIP) [11].

Therefore, this sparsity is balancedby re-sampling technique [11]. As mentioned earlier, two traits are highly imbalanced, Data Level Re-sampling approach for class balancing is used [9]. This bridged the gap between each dichotomy traits and resulted in the efficient and predictable performance of the proposed system.

TABLE V. Detail of Dataset

Dataset Name	Description	Instances	Format	Default Task	Updated	Origin	Size	Creator
MBTI_kaggle	This dataset was acquired from Kaggle by using PersonalityCafe platform. The members of PerC have known MBTI personality type along their tweets or text. The dataset comprised of 8676 PerC members personality types.	8675	Text	Personality Prediction	2018	Kaggle	25 MB	Mitchell J

2) Data Level Re-Sampling Approach: Data manipulation sampling approaches focus on rescaling the training datasets for balancing all class instances. Two popular techniques of class resizing are over-sampling and under-sampling.

At the data level, the most famous methodologies are Oversampling and under sampling procedures. Oversampling is the way toward expanding the number of classes into the minority class. The least difficult oversampling is random oversampling, which basically duplicate minority instances to enhance the imbalance proportion.This duplication of minority class enhancement really improved the performance of machine learning classifier for efficient personality traits prediction [11].

Under samplingapproach is used to level class distribution by indiscriminately removing or deleting majority class instances. This process is continued till the majority and minority class occurrences are balanced out.

As illustrated in Fig. 2, the data level sampling-based methodologies including over-sampling and under-sampling have gotten exceptionalconsiderations to counter the impact of imbalanced datasets [35].

3) Training and Testing Data: In this proposed system, the data is divided into Training, Testing and Validation dataset. Mostly two datasets are required, one for building the model while the other dataset is needed to measure the performance of the model. Here training and validation are used for building the model, while Testing step is used to measure the performance of the proposed model [36]. Table VI shows the sample tweets from training dataset, while Table VII represents the sample of test data tweets.

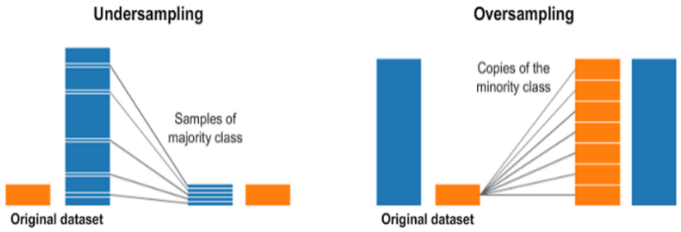

Fig. 2. Class Balancing using Undersampling and Oversampling.

TABLE VI. Sample Tweets from Training Dataset

Personality Type	Tweets
ISTP	'I'm only a mystery to myself.
INTP	Of course, to which I say I know; that's my blessing and my curse.
INFJ	Hello ENFJ7. Sorry to hear of your distress.
ENTP	'I'm finding the lack of me in these posts very alarming.
ENTJ	Lol. Its not like our views were unsolicited. What a victim.
INFP	That more or less finds myself in agreement, honey cookie.
ESTP	Most things hands on. For me, music. I'm very tactile. I like to write too.

TABLE VII. SAMPLE TWEETS FROM TEST DATASET

Personality Type Type	Tweets
ENFP	Patience is a virtue. So proud that you guys are still together.
ISFJ	We are always willing to help those in need
ENTJ	I'm scared of failure, but also throwing up...take that for what you will.
INFP	That would be the best description for what I usually am.
ENFJ	You're right. Not sure why I didn't think of that before hahah
ESTP	I have 0 friends. I don't trust anybody.

At the point when the dataset is divided into training data, validating data and testing data, it utilizes just a portion of dataset and it is clear that training on minor data instances the model won't behave better and overrate the testing error rate of algorithm to set on the whole dataset.

To address this problem a cross-validation technique will be used.

4) Cross-validation: It is a statistical methodology that perform splitting of data into subgroups, training on the subset of data and utilize the other subset of data to assess the model's authentication.

Cross validation comprises of the following steps:

- Split the dataset into two subsets.
- Reserve one subset data.
- Train the model on the other subset of data.
- Using the reserve subset of data for validation (test) purpose, if the model exhibits better on validation set, it shows the effectiveness of the proposed model.

Cross validation is utilized for the algorithm's predictive performance estimation.

a) K fold cross validation: This strategy includes haphazardly partitioning the data into k subsets of almost even size. The initial fold is reserved for testing and all the remaining k-1 subsets of data are used for training the model. This process is continued until each Cross-validation fold (of k iteration) have been used as the testing set.

This procedure is repeated kth times; therefore, the Mean Square Error also obtained k times (from Mean Square Error-1 to kth Mean Square Error). So, k-fold Cross Validation error is calculated by taking mean of the Mean Square Error over Kfolds. Fig. 3, explain the working procedure of K-Fold cross validation.

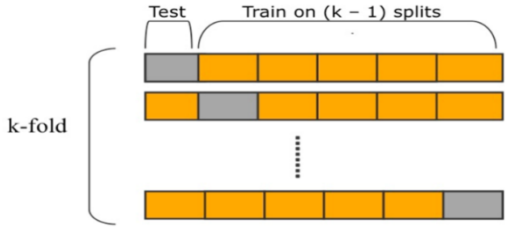

Fig. 3. K-Fold Cross Validation Working Procedure.

Algorithm 1. Dividing the Data set in Train and Test sets.

```
#Division of Data in training and testing sets:
Assign [] to X↔Train
Assign [] to Y↔Train
Assign [] to X↔Test
Assign [] to Y↔Test
Allocate Test→ Size to 20% of n
Assign RNDM (0, n -1, Test→Size) to TINDICES
For I = 0 to n-1
        Assign [] to Temp
        For each WORD in Tf-Idf [i]
                Append (If-Idf [i][WORD]) to temp
        END FOR
If I in TINDICES then
        Append (TEMP) to X↔Test
        Append (tweet [i][ I]) to Y↔Test
Else
        Append TEMP) to X↔Train
        Append TEMP) to Y↔Train
END IF
END FOR
```

B. Preprocessing and Feature Selection

Different pre-processing techniques and various feature selection are exploited, for more exploration of the personality from text. These techniques include tokenization, removal of URLs, User mentions and Hash tag, word stemming, stop words elimination and feature selection using TF IDF [28] and [32].

1) Preprocessing: The following preprocessing steps on mbti_kaggle dataset are applied before classification, acquired from the [37] work.

a) Tokenization: Tokenization is the procedure where words are divided into the small fractions of text. For this reason, Python-based NLTK tokenizer is utilized.

b) Dropping Stop Word: Stop words don't reveal any idea or information. A python code is executed to take out these words utilizing a pre-defined words inventory. For instance, "the", "is", "an" and so on are called stop words.

c) Word stemming: It is a text normalization technique. Word stemming is used to reduce the inflection in words to their root form. Stem words are produced by eliminating the pre-fix or suffix used with root words.

2) Feature Selection: The following feature selection steps are accomplished using different machine learning classifiers.

a) CountVectorizer: Using machine learning algorithms, it cannot execute text or document directly, rather it may firs be converted into matrix of numbers. This conversion of text document into numbers vector is called tokens.

The count vector is a well-known encoding technique to make word vector for a given document. CountVectorizer takes what's known as the Bag of Words approach. Each message or document is divided into tokens and the number of times every token happens in a message is counted.

CountVectorizer perform the following tasks:

- It tokenizes the whole text document.
- It constitutes a dictionary of defined words.

- It encodes the new document using known word vocabulary.

b) Term Frequency: It represents the weight of a word that how much a word or term occurs in a document.

c) Inverse document Frequency: It is also a weighting scheme that describe the common word representation in the whole document.

d) Term Frequency Inverse Document Frequency: The TF-IDF score is useful in adjusting the weight between most regular or general words and less ordinarily utilized words. Term frequency figures the frequency of every token in the tweet however, this frequency is balanced by frequency of that token in the entire dataset. TF-IDF value shows the significance of a token in a tweet of whole dataset [38].

This measure is significant in light of the fact that it describes the significance of a term, rather than the customary frequency sum [39].

Feature engineering module pseudocode is illustrated in the following Algorithm 2.

Algorithm2. Stepwise procedure for Feature Engineering

```
# CountVectorizer
Assign [] to CVectorizer
For        Each tweet in Post Do
           ForEach word in tweet Do
                      Assign Dict [word] to Dict [Word] +1
           EndFor
           CVectorizer. Append (Dict)
           Assign 0 to Dict
EndFor
TermFrequency
Assign CVectorizer to TF
Assign 0 to ROW
While (ROW <= N-1) Do
           Assign SUM (CVectorizer [row].values) to Nwords
           For Each Word in CVectorizer [row]
           Assign CVectorizer[W]/Nwords to TF [W]
           EndFor
WhileEnd
# TF/IDF
# IDF Calculation
Assign [] to IDF
While (Till the existence of ROW in TF) Do
Assign [] to temp
While (Till the existence of word in ROW) DO
           Assign 0 to Count
           For i from 0 to N-1 Do
                      IF TF [Count][Word]>0 Then
                        Count← Count+1
                      End IF
           EndFor
                      Assign LOG (N/Count) to Temp [Word]
           WhileEnd
           IDF. Append (TEMP)
WhileEnd
# TF/ IDF
Assign 0 to TF -IDF
FOR I From 0 to N-1 DO
           Assign []  to  TEMP
           ForEach W in TF [i], IDF [i]
                      TEMP [W]= TF [i][W]*IDF[i][ W]
           EndFor
           Append (TEMP) to TF -IDF
EndFor
```

C. Text-based Personality Classification Using MBTI Model

In this proposed work, supervised learning approach is used for personality prediction. The model will take snippet of post or text as an input and will predict and produce personality trait (I-E, N-S, T-F, J-P) according to the scanned text. Mayers-Briggs Type Indicator is used for classification and prediction [4]. This model categorize an individual into 16 different personality types based on four dimensions, namely, (i) *Attitude →Extroversion vs Introversion:* this dimension defines that how an individual focuses their energy and attention, whether get motivated externally from other people's judgement and perception, or motivated by their inner thoughts, (ii) *Information →Sensing vs iNtuition (S/N):* this aspect illustrates that how people perceive information and observant(S), relying on their five senses and solid observation, while intuitive type individuals prefer creativity over constancy and believe in their guts, (iii) *Decision →Thinking vs Feeling (T/F):* a person with Thinking aspect, always exhibit logical behaviour in their decisions, while feeling individuals are empathic and give priority to emotions over logic, (iv) *Tactics →Judging vs Perceiving (J/P):* this dichotomy describes an individual approach towards work, decision-making and planning. Judging ones are highly organized in their thoughts. They prefer planning over spontaneity. Perceiving individuals have spontaneous and instinctive nature. They keep all their options open and good at improvising opportunities [40].

D. Working Procedure of the System for Personality Traits Prediction

As depicted in Fig. 4, first, the proposed model is trained by giving both labelled data (MBTI type) and text (in the form of tweets). After training the model, it is evaluated for efficiency. For better prediction, the dataset will be split into three phases (training phase, validating phase and testing phase). The validating step will reduce overfitting of data.

The mbti_kaggle dataset is available in two columns, namely, (i) type and (ii) posts. By type it means 16 MBTI personality types, such as INTP, ENTJ and INFJ, etc. As we are interested in MBTI traits rather than types, therefore we through python coding added four new columns to the original dataset for the purpose of traits determination. As a result, the new modified dataset will look like as given bellow in Table VIII.

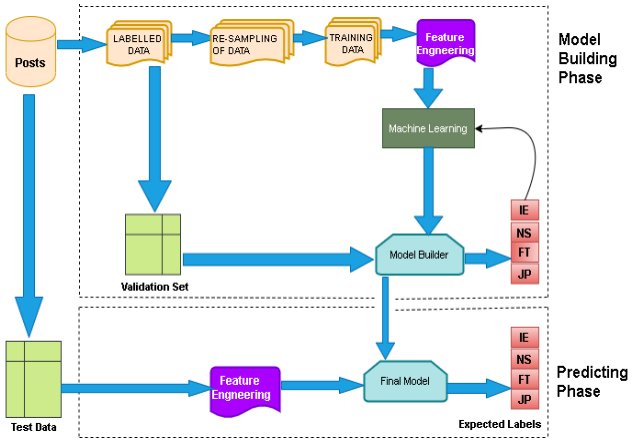

Fig. 4. Working Procedure of the System.

TABLE VIII. SAMPLE OF DATASET USED FOR EXPERIMENT

Type	Posts	I/E	S/N	F/T	J/P
ENTP	I'm scared of failure, but also throwing up...take that for what you will.	1	0	0	1
INFJ	Just a funny comment from my side. A bit serious maybe. If you don't care about the functions	0	0	1	0
INFP	I need a date with an INTJ! God dammit. Opps, wrong thread. lol	0	0	1	1

Algorithm 3: Pseudo code of the entire System

```
Inṗut:          Ṣet of tweets from mbti_kaggle dataset saved in CSV format
Output:         Classification of input text into personality traits
Personality Traits:     ["I_E", "S-N", "F-T", "J-P"]
ML-Classifier:          [ "XGboost"]
Stop-word List:         [There, it, on, into, under…….]
Start
//Inputting Snippet of Text
Assign Dataset text of post to Text
#Pre-processing steps.
#Tokenization/segmentation
        Assign Tokenize(text) to Token
# Dropping of stop words
        Set Post_text to Drop_stopwords(tokens)
#punctuation
# data set splitting into train/teṣt
        Set X↔Ŧráin, Ȳ↔Ŧráin, X↔Ŧest, Ȳ↔Ŧest to Ṣplit (post_text, test-ṣize=20%)
# counterVectorizer(Post_Text)
#Application of tf‣ idf
#Classifier implementation
Set Model to MLClassifier
ÄssignMödel: fit(X↔Ŧráin, Ȳ↔Ŧráin) to Classification
Set Classification to Mödel: fit(X↔Ŧráin, Ȳ↔Ŧráin)

#Traits Ṗrediction
Assign Classification: Ṗrediction (X↔Ŧest) to Ṗrediction
Set Trait_Prediction to Classification: Ṗrediction (X↔Ŧest)
#Accuracy
Set Accuracy to Accuracy (Trait_Ṗrediction, Ȳ↔Ŧest)
#Recall score
Set Recall to Recall (Trait_Prediction, Ȳ↔Ŧest)
#Precision score
Set Precision to Precsion (Trait_Prediction, Ȳ↔Ŧest)
#F1‣ score
Set F1‣ score to F1‣ score (Trait_Prediction, Ȳ↔Ŧest)
Assign (Accuracy, Re_call, Precession, F1‣ score) to Personality Traits
Return (Personality Traits)
```

E. Applying XGBoost for Personality Classification

XGBoost belongs to the family of Gradient Boosting. It is used to handle classification and regression issues that make a prediction/ forecast from a set of weak decision trees.

Although work has been performed on personality assessment using supervised machine learning approaches [13, 17]. Here state of the art Algorithm XGBoost with optimized parameters is used for MBTI personality assessment [41]. XGBoost classifier is good on producing better accuracy as compared to other machine learning algorithms [41, 42]. The proposed work is the first attempt to predict personality from text using XGBoost as classifier and MBTI as personality model.

Algorithm 4: XGBoost Working Procedure

```
Data: Dataset and Hyperparameters
Initialize
for k = 1,2, ……., M do
Calculate 𝑔𝑘 = ;
Calculate ℎ𝑘 = ;
Determine the structure by choosing splits with maximized gain
𝒜 =
Determine the leaf weights = ;
Determine the base learner 𝑏(𝑥) = ;
Add trees 𝑓_k(𝑥) = 𝑓_{k-1}(𝑥) + 𝑏(𝑥);
end
Result: 𝑓(𝑥) =
```

F. Comparing the Efficiency of XGBoost with other Classifiers

The overall prediction performance and efficiency of the proposed system has examined by applying other supervised machine learning classifiers. This comparison illustrates a true picture of the performance of this proposed classifier, namely XGBoost, as compared to the other machine learning algorithms and baseline methods regarding personality prediction capability from the input text [13].

G. Evaluation Metrics

The evaluation metrics, such as accuracy, precision, recall and f-measure, describe the performance of a model. Therefore, different evaluation metrics has been used to check the overall efficiency of predictive model.

Algorithm 5: Pseudo code of the Performance Evaluation

```
# Ṗerformance
ŦC
Äccuracy ↔ŦC/ɲ2
TṖ  CÖUNŦ (Ṗrediction = Ṗositive ÄNḊ Ẏ↔Ŧest =Ṗositive)
Tɲ ↔ CÖUNŦ (Ṗrediction =ɲegative ÄNḊ Ẏ↔Ŧest = ɲegative)
ƑṖ ↔ CÖUNŦ ((Ṗrediction = Ṗositive ÄNḊ Ẏ↔Ŧest = ɲegative)
ƑN ↔ CÖUNŦ (Ṗrediction = ɲegative ÄNḊ Ẏ↔Ŧest = Ṗositive)
Ṗrecisioń ↔TṖ / (TṖ + ƑṖ)
Recȧll ↔TṖ / (TṖ + ƑN)
CFM ↔ []
CƑM ['TṖ'] ↔TṖ
CƑM ['FN'] ↔ƑN
CFM['FP'] ↔ƑP
CƑM ['ŦN'] ↔ŦN
```

IV. RESULTS AND DISCUSSIONS

This chapter presents a set of results which are produced from the proposed system by systematically answering the raised research questions.

A. Answer to RQ.1

To answer to RQ1: "How to apply supervised machine learning technique, namely XGBoost classifier for classifying personality traits from the input text?", the supervised machine learning technique, XGBoost classifier is applied to predict MBTI personality traits from excerpt of text. Fine-tuned parameter setting for XGBoost is presented in Table IX.

Table X shows the results of XGBoost classifier with default parameter settings.

It is clear from Table XI that increasing or decreasing the values of different parameters for XGBoost classifier, has huge effect on the text classification results.

B. Answer to RQ.2

While addressing RQ2: "How to apply a class balancing technique on the imbalanced classes of personality traits for performance improvement and What is the efficiency of the proposed technique w.r.t other machine learning techniques?", An imbalanced dataset is considered first. Imbalanced dataset can be defined as a distribution problem arises in classification where the number of instances in each class is not equally divided.

Whenever, an algorithm is applied on skewed and unbalanced classified dataset, the outcome always diverge toward the sizeable class and the smaller classes are bypassed for prediction. This drawback of classification is known as class imbalance problem [11].

Therefore, it is attempted to balance this sparsity by re-sampling technique [11]. As two traits are highly imbalanced, therefore Data Level Re-sampling approach is used for class balancing [9].

TABLE IX. PARAMETER SETTING FOR XGBOOST

Parameters	Description
Learning_rate = 0.03	It describes the effect of weighting of adding more trees to the boosting model.
Colsample_bytree = 0.4	It corresponds to the fraction of features (columns) that will be used to train each tree.
Scale-pos_weight = 1	It controls the balance between negative and positive classes.
Subsample = 0.8	Subsample ratio of the training instance. Setting it to 0.5 means that XGBoost randomly collects half of the data instances to grow trees. This prevents overfitting.
Objective = 'binary:logistic',	It returns predicted probability for binary classification.
n_estimators = 1000	It represents the number of decision trees in XGBoost classifier.
Reg_alpha = 0.3	L1 regularization encourages sparsity (meaning pulling weights to 0).
Max-depth = 10	It represents the size (depth) of each decision tree in the model. Over fitting can be controlled using this parameter.
Gamma = 10	Its purpose is to control complexity. It represents that how much loss has to be reduced. It prevents overfittings.

TABLE X. RESULTS OF XGBOOST WITHOUT PARAMETER SETTINGS

	Metrics	I-E	S-N	F-T	J-P
No Parameter setting	**Accuracy**	87.04	92.32	89.00	85.85
	Recall	81.44	81.75	87.70	89.16
	Accuracy	91.59	68.98	91.65	87.80
	F1_Score	86.22	74.82	89.92	88.47

TABLE XI. RESULTS OF XGBOOST WITH DIFFERENT PARAMETERSETTINGS

	Metrics	I-E	S-N	F-T	J-P
learning_rate: 0.01 n_estimators: 1000 max_depth: 5 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	93.10	96.70	92.32	90.88
	Recall	89.56	96.24	92.07	94.24
	Precession	96.32	97.14	93.64	90.91
	F1_Score	92.82	96.68	92.85	92.55
learning_rate: 0.01 n_estimators: 1000 max_depth: 6 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	95.51	97.61	93.15	91.79
	Recall	93.39	97.21	92.91	94.77
	Precession	97.47	98.00	94.37	91.81
	F1_Score	95.39	97.60	93.64	93.27
learning_rate: 0.01 n_estimators: 500 max_depth: 6 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	90.95	94.51	91.20	89.84
	Recall	85.78	91.98	90.28	95.23
	Precession	95.48	96.88	93.28	88.69
	F1_Score	90.37	94.37	91.75	91.84
learning_rate: 0.01 n_estimators: 1000 max_depth: 10 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	99.37	99.92	94.55	95.53
	Recall	97.16	100	89.96	92.66
	Precession	100	99.50	100	100
	F1_Score	98.56	99.75	94.72	96.19

In this section the overall comparison of predicting personality traits is presented using all evaluation metrics to determine the performance of different classifiers. Results are reported in Table XII.

Different classifiers are applied over same mbti_kaggle dataset using Re-sampling technique and without Re-sampling technique. Results reported in Table XII depict that XGBoost obtained the highest score using all four-evaluation metrics and across all the MBTI personality dimensions, when imbalance dataset is experimented. However, Naïve Bayes and Random Forest on imbalance dataset, performed poorly. So, it is concluded from this experiment that applying classifiers on skewed data is not producing good results.

On the other hand, when different classifiers are tested over resampled dataset, an improved result is obtained for all dimensions over all classifiers.

The most accurate and precise algorithm for this proposed work is XGBoost. It got excellent results for all traits using all metrics. XGBoost obtained maximum accuracy (99.92%) for S/N trait. Its results are highest for all four dimensions and across all metrics.

1) Why our Class balancing technique is better: By applying class balancing technique results for all evaluation metrics and for all four personality traits are high and better than base line work. In this dataset two dimensions I/E and S/N are highly imbalanced, therefore a class balance technique is used for better prediction performance.

TABLE XII. COMPARISON OF DIFFERENT CLASSIFIERS PERFORMANCE USING RE-SAMPLE DATASET AND IMBALANCE DATASET

Classifier	Metrics	Without Re-sampling				With Re-Sampling			
		I-E	**S-N**	**F-T**	**J-P**	**I-E**	**S-N**	**F-T**	**J-P**
KNN	Accuracy	77.02	86.65	60.11	59.31	86.90	81.44	73.45	81.52
	Recall	20.34	15.5	89.89	77.17	86.44	98.00	93.82	89.19
	Precession	45.74	32.29	58.64	63.70	65.74	42.79	68.69	81.74
	F1_Score	28.16	20.94	70.98	69.79	74.51	59.57	79.31	85.30
Decision Tree	Accuracy	78.69	82.01	70.42	69.33	99.34	99.93	90.85	91.30
	Recall	53.31	38.25	71.94	75.11	97.00	99.50	83.14	85.72
	Precession	51.84	36.34	73.11	74.68	100	100	100	100
	F1_Score	52.56	37.27	72.52	74.89	98.48	99.75	90.79	92.31
Random Forest	Accuracy	77.93	86.03	74.,89	64.90	98.36	99.45	82.15	91.62
	Recall	00	0	84.49	97.7	92.59	98.94	74.07	86.24
	Precession	1	0	73.31	63.84	100	100	100	100
	F1_Score	00	0	78.50	77.22	96.15	99.44	85.10	92.61
MLP	Accuracy	83.83	88.40	83.41	75.86	99.27	99.93	94.52	92.18
	Recall	40.37	22.0	84.68	86.46	96.69	99.59	89.90	87.91
	Precession	83.83	88.40	83.41	75.86	100	100	100	81.906
	F1_Score	40.37	22.0	84.68	86.46	98.32	99.75	88.89	93.14
SVM	Accuracy	85.54	88.68	85.02	78.62	95.94	98.08	92.63	91.37
	Recall	43.69	22.75	85.64	90.36	91.32	97.00	89.45	91.11
	Precession	82.93	85.84	86.59	78.01	90.28	90.02	96.73	94.53
	F1_Score	57.23	35.96	86.12	83.74	90.69	93.38	92.95	92.79
MNB	Accuracy	77.86	86.03	54.63	60.92	79.32	88.82	84.04	60.11
	Recall	0	0	99.93	100	6.78	20.25	73.18	100
	Precession	0	0	54.47	60.91	97.73	98.78	96.68	60.11
	F1_Score	0	0	70.51	75.71	12.66	33.61	83.25	75.09
XGboost	Accuracy	86.52	89.21	83.16	80.82	99.37	99.92	94.55	95.53
	Recall	52.68	31.5	84.04	89.90	97.16	100	89.96	92.66
	Precession	79.52	78.26	84.80	80.78	100	99.50	100	100
	F1_Score	63.38	44.92	84.42	85.10	98.56	99.75	94.72	96.19
Logistic Reg	Accuracy	82.47	86.48	84.32	76.63	92.80	96.09	88.96	88.44
	Recall	25.86	4.5	86.35	93.52	85.33	90.25	85.28	92.14
	Precession	83.67	78.26	84.99	74.57	82.72	83.18	93.90	89.23
	F1_Score	39.51	8.5	85.66	82.98	84.01	86.57	89.34	90.66
SGD	Accuracy	85.26	90.29	85.19	79.36	94.31	97.42	91.86	90.99
	Recall	41.64	40.5	85.71	90.82	91.64	95.50	87.52	89.39
	Precession	83.54	80.19	86.83	78.61	84.08	87.21	97.21	95.53
	F1_Score	55.58	53.82	86.27	84.28	87.70	91.17	92.11	92.36

KNN classifier gives overall low performance, however its Recall for I/E and F/T is a little bit high.

The outcome of Decision Tree algorithm for I/E and S/N traits is better than F/T and J/P traits.

Random Forest gives highest for all traits. However, for J/P lowest Recall is obtained.

Logistic Regression classifier produced tremendous result for all traits, but again for J/P traits accuracy and Precision are not up to the mark.

The results obtained by applying Naïve Bays classifier is comparatively better for I/E and S/N traits.

Support Vector Machine when tested on the given dataset it gives better and balance results in respect to all traits. SGD Classifier showing remarkable performance for all four personality traits.

MLP classifier achieved outstanding results for all four traits using four metrics.

XGBoost classifier has proven to be very good for classification problems. The results obtained using XGBoost is very balance in respect to all personality traits

C. Answer to RQ.3

To answer RQ3**:** "What is the efficiency of the proposed technique with respect to other baseline methods." This proposed model is compared with two baseline methods [6, 7].

Classification performed by [6] for personality prediction using same mbti_kaggle dataset by applying three classifiers namely, (i) SVM, (ii) MLP and (iii) Naïve Bayes and got accuracy upto 88.4%. Due to imbalance data the result of [6] is not up to the mark. The results show that SVM in collaboration with LIWC and TF-IDF feature vectors gave accurate prediction score for all four traits, while MLP with all features Vectors got maximum accuracy score for S/N trait (90.45%) however its result for J/P trait is lower. Naïve bays also perform well for I/E and S/N traits but its performance for T/F and J/P is very poor. The reason behind better accuracy for I/E and S/N dimensions and least performance for T/F and J/P is due to class imbalance problem.

A very large dataset MBTI9k acquired from reddit is used for personality prediction [7]. The emphasis of this work is to extract features and linguistic properties of different words and then these features are used to train various machine leaning models such as Logistic Regression, SVM and MLP. Classifiers using integration of all features together (LR_all and MLP_all) obtained better results for all traits. The overall worst results using all classifiers obtained for the T/F dichotomy. The major limitation of this work is that the number of words in each post are very large, which lead to a little bit lower performance on the part of all classifiers.

1) Proposed Work: In this proposed system, the same dataset is used as experimented by [6], However re-sampling technique is applied over it, and hence obtained results in respect of all personality traits are very good, especially XGBoost achieved the best score across all dimensions and all traits as compared to previous work. It is observed that the mbti_kaggle dataset is very skewed, therefore when oversampling technique is applied the output is far better than all previous works. Up to 99% accuracy for I/E and S/N traits are achieved using XGBoost classifier, while Bharadwaj [6], got 88% maximum accuracy for S/N trait. Similarly, for T/F and J/P proposed work results are promising and obtained 94.55% accuracy for T/F and 95.53% accuracy for J/P dimension using XGBoost. While in previous work MLP classifier achieved accuracy of 54.1% for T/F and 61.8% for J/P dimension. Therefore, it is clear that by using resampling technique excellent and improved results are obtained for all four dimensions. The results reported in Table XIII, describe the comparison of proposed work with the baseline method.

2) XGBoost with Outstanding Performance: XGBoost belongs to the family of Gradient Boosting is a machine learning technique used for classification and regression problems that produces a prediction from an ensemble of weak decision trees.

The main reason of using this algorithm is its accuracy, speed, efficiency, and feasibility. It's a linear model and a tree learning algorithm that does parallel computations on a single machine. It also has extra features for doing cross validation and computing feature importance.

TABLE XIII. COMPARISON OF XGBOOST WITH BASELINE TECHNIQUE

Study	Technique	Dataset	Classifier	Obtained Results				
				Metrics	I/E	S/N	F/T	J/P
Bharadwaj, et al. (2018)	SVM, MLP and Naïve Bayes	MBTI_Kaggle	NB	Accuracy	77%	86.2%	77.9%	62.3%
				Recall				
				Precession				
			SVM	Accuracy	84.9%	88.4%	87.0%	78.8%
				Recall				
				Precession				
			MLP	Accuracy	77.0%	86.3%	54.1%	61.8%
				Recall				
				Precession				
Gjurković et al. (2018)	SVM, MLP and Logistic Regression	MBTI9k	SVM	Accuracy				
				F1-Score	79.6%	75.6	64.8	72.6
				Precession				
			LR	Accuracy				
				F1-Score	81.6	77.0	67.2	74.8
				Precession				
			MLP	Accuracy				
				F1-Score	82.8	79.2	64.8	72.6
				Precession				
Proposed (our work)	XGBoost	MBTI_Kaggle						
			XGBoost	Accuracy	99.37	99.92	94.55	95.53
				Recall	97.16	100	89.96	92.66
				Precession	100	99.50	100	100
				F1-Score	98.56	99.75	94.72	96.19

V. CONCLUSION AND FUTURE WORK

The central theme of this study is the application of different machine learning techniques on the benchmark, MBTI personality dataset namely mbti_kaggle to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P).

The Mayers-Briggs Type Indicator (MBTI) model is used for text classification and personality traits recognition [4]. After applying class balancing techniques on the imbalanced classes, different machine learning classifiers, namely, KNN, Decision Tree, Random Forest, MLP, Logistic Regression (LR), SVM, XGBoost, MNB and Stochastic Gradient Descent (SGD) are experimented to identify the personality traits. Evaluation metrics, such as accuracy, precision, recall and F-score, are used to analyze and examine the overall efficiency of the predictive model. The obtained results show that score achieved by all classifiers across all personality traits is good enough, however, the performance of XGBoost classifier is outstanding. We got more than 99% precision and accuracy forI/E and S/N traits and obtained all about 95% accuracy for T/F and J/P dimensions. However, KNN classifier resulted in overall lower performance.

A. Constraints or Limitations

1) MBTI model is examined for personality traits classification, however, others personality models such as Big Five Factor (BFF) and DiSC personality Assessment models, are not experimented and investigated.

2) The textual data used in the proposed work for personality assessment is comprised of only English language, and the contents of other languages are not experimented.

3) Simple over-sampling and under sampling techniques are used to balance and level the skewness of dataset.

4) The dataset comes from only one platform namely personalitycafe forum, which may lead to biased results.

5) All the experiments conducted in this proposed work are based on the classical or traditional machine learning algorithms.

6) The textual contents which are classified for personality traits identification belong to only one site Twitter, however other social networking sites are ignored.

7) Only textual data is analysed and investigated for user's personality traits recognition in his proposed work.

8) Less weightage is given to feature extraction in classification of text, only TF-IDF technique is utilized.

B. Future Proposal

1) The predictive performance of MBTI personality model needs to be compared with the Big Five Factor (BFF) model for better assessment of the traits.

2) Multilingual textual content, especially Urdu and Pashto language textual data can be examined for personality classification.

3) SMOTE (Synthetic Minority Over-sampling Technique) can be utilized as class balancing method for more robust and reliable performance.

4) Labelled data may need to be collected from other platforms like "Reddit" using multiple benchmark datasets.

5) More experiments on personality recognition may be conducted using Deep learning algorithms.

6) Other social networking sites like FACEBOOK posts and comments are required to be examined for automated personality traits inference.

7) Data available in the format of images and videos on social networking sites can be experimented for the task of personality traits identification.

8) More advanced features selection approaches are required to be exploited for enhancement of the proposed work.

REFERENCES

[1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017.

[2] D. Xue et al., "Personality Recognition on Social Media With Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017.

[3] L. R. Goldberg, L. R. ,"An alternative" description of personality": the big-five factor structure," Journal of personality and social psychology, vol. 59, no. 6, p.1216, 1990

[4] I. B. Myers, "The Myers-Briggs Type Indicator: Manual" ,1962

[5] D. Shaffer, M. Schwab-Stone and P. Fisher, "Preparation, field testing, interrater reliability and acceptability of the DIS-C," J Am Acad Child Adolesc Psychiatry, vol. 32, pp. 643-648, 1993.

[6] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.

[7] M. Gjurković and J. Šnajder, "Reddit: A Gold Mine for Personality Prediction," In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.

[8] B. Plank, and D. Hovy, "Personality traits on twitter—or—how to get 1,500 personality tests in a week." In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

[9] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases," Neurocomputing, 175, pp. 935-947, 2016.

[10] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," 2016, arXiv preprint arXiv:1608.06048

[11] P. Kaur and A. Gosain, "Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," In ICT Based Innovations , pp. 23-30, Springer, Singapore, 2018.

[12] I. Cantador, I. Fernández-Tobías and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," In CEUR workshop proceedings, ShlomoBerkovsky, 2013.

[13] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.

[14] V. Ong et al., "Personality prediction based on Twitter information in Bahasa Indonesia," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 367-372.

[15] F. Alam, E. A. Stepanov and G. Riccardi, "Personality traits recognition on social network-facebook," WCPR (ICWSM-13), Cambridge, MA, USA, 2013.

[16] K. Buraya, A. Farseev, A. Filchenkov and T. S. Chua, "Towards User Personality Profiling from Multiple Social Networks," In AAAI, pp. 4909-4910, 2017.

[17] N. R. Ngatirin, Z. Zainol and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 435-440.

[18] S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, "A comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," IRJET, vol.05, pp.1410-1413, 2018.

[19] V. Ong, A. D. Rahmanto, Williem and D. Suhartono, "Exploring Personality Prediction from Text on Social Media: A Literature Review," INTERNETWORKING INDONESIA, vol. 9, no. 1, pp. 65-70, 2017a.

[20] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 149-156.

[21] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 180-185.

[22] B., Verhoeven, W. Daelemans and B. Plank, "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling," In Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. pp. 1-6, 2016.

[23] F. Celli, "Mining user personality in twitter, " Language, Interaction and Computation CLIC, 2011.

[24] X. Sun, B. Liu, Q. Meng, J. Cao, J. Luo and H. Yin, "Group-level personality detection based on text generated networks," World Wide Web, pp. 1-20, 2019.

[25] F. Celli and L. Rossi, "The role of emotional stability in Twitter conversations," In Proceedings of the workshop on semantic analysis in social media, Association for Computational Linguistics, pp. 10-17, 2012.

[26] S. Chishti, X. Li and A. Sarrafzadeh, "Identify Website Personality by Using Unsupervised Learning Based on Quantitative Website Elements, " In International Conference on Neural Information Processing, Springer, Cham. pp. 522-530, 2015.

[27] F. Celli, "Unsupervised personality recognition for social network sites," In Proc. of Sixth International Conference on Digital Society, 2012.

[28] P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju and V. Sinha, "25 Tweets to Know You: A New Model to Predict Personality with Social Media," 2017, arXiv preprint arXiv:1704.05513.

[29] M. Arroju, A. Hassan, and G. Farnadi, "Age, gender and personality recognition using tweets in a multilingual setting," In 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction, pp. 23-31, 2015.

[30] L. C. Lukito, A. Erwin, J. Purnama and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-6.

[31] N. Alsadhan and D. Skillicorn, "Estimating Personality from Social Media Posts," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 350-356.

[32] R. K. Hernandez and L. Scott, "Predicting Myers-Briggs type indicator with text," In *31st Conference on Neural Information Processing Systems* (*NIPS 2017)*, 2017.

[33] B. Cui and C. Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction".

[34] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao et al, "Deep learning-based personality recognition from text posts of online social networks," *Applied Intelligence*, vol. 48, no. 11, pp. 4232-4246, 2018.

[35] Y. Yan, Y. Liu, M. Shyu and M. Chen, "Utilizing concept correlations for effective imbalanced data classification," Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, 2014, pp. 561-568.

[36] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," in IEEE Communications Magazine, vol. 57, no. 5, pp. 76-81, May 2019.

[37] M. Z. Asghar, A. Khan, F. Khan and F. M. Kundi, "RIFT: A Rule Induction Framework for Twitter Sentiment Analysis," Arabian Journal for Science and Engineering, vol. 43, no. 2, pp.857-877, 2018.

[38] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, 57, pp. 117-126, 2016.

[39] L. H. Patil and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," 2013 3rd IEEE International Advance Computing Conference (IACC), Ghaziabad, 2013, pp. 858-862.

[40] M. C. Komisin and C. I. Guinn, "Identifying personality types using document classification methods," In Twenty-Fifth International FLAIRS Conference, 2012.

[41] D. Nielsen, "Tree Boosting With XGBoost-Why Does XGBoost Win Every Machine Learning Competition? (Master's thesis, NTNU)," 2016.

[42] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959-61969, 2018.