[TITLE]

Abstract—

Keywords —

I. INTRODUCTION

Personality basically refers to a person's behavior and characteristic in managing a variety of situations. It tells a lot about an individual's choice towards things like their entertainment interests, reading, music, nature, and more. NLP includes a large number of techniques for identifying user's personality, the way people think, how people behave and communicate and much more.

In simple terms, we can say that NLP helps in detecting patterns in an individual's behavior [1]. Personality classification is a process of detecting and classifying a personality of users through different classification measures. This approach helps in identifying of user's feelings, likes, interests, and thoughts that demonstrates people's course of actions. It tells how an individual behave around other people and environment. All this information can help in career counseling, health analysis, employee recruitment, and relationship counseling.

Machine learning is a broad area. Researches nowadays utilize different machine learning algorithms for personality prediction. By applying different machine learning techniques on historical data, people can predict future data based on the results. Similarly, people also use these techniques for learning personality patterns [3, 4]. Meta programmes are considered an important factor in the area of NLP. According to Brian [2], people can sort information in their brains by using different approaches. Personality trait referred to habitual patterns of thoughts, behavior, and emotions [3]. There is an increased usage of different social networking sites by people of all age groups. It includes Facebook, Snapchat, Instagram, and Facebook. People use these social media platforms to make a tweet, and post regarding all the current topics. All this represents their actions, personalities, and behaviors. The personality of the user and the posts which they made on different social media platforms such as tweets are highly correlated with each other. The attitude behind each post and tweet can be measured subjectively, i.e., positive and negative. Nowadays a large number of researchers and experts are working on developing the automatic personality recognition systems.

The existing works is done by means of diverse supervised machine learning approaches. However, the major issue that is not handled completely is the skewness of the datasets. All these datasets contains imbalanced classes for different personality traits. This can badly affect the recognition system. There are different techniques that can help in reducing the skewness of the dataset. Some of the popular ones includes hybrid-sampling, under-sampling, and over-sampling [8].

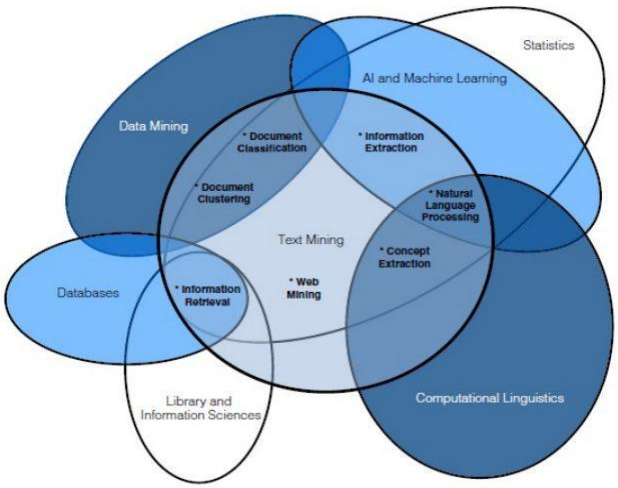

Figure. 1. NLP and Classification

Twitter is considered to be the popular social micro blogging platform among users these days. It allows people to make a tweet of up to 140 characters and share it with others. This social networking site keeps a real time track of user's views and thoughts related to different events [5]. It includes elections, social events, entertainment, sports, trending topics and more. The main thought behind this research is to demonstrate how community commotion can be used to infer personality traits. As mentioned earlier, there are different social media platforms. In this research, I am focusing on Twitter data for predicting user's personality type as per their tweets. People use twitter platform to tweet daily. Each tweet made by a user has some emotion associated to it. That emotion from the text helps in predicting the correct personality trait of a person[5,6]. A user often adds some personal touch in almost all the tweets they made.

All this usually reflect the undergoing thought process of that particular individual in his tweets which result in identifying a Conscientiousness personality type. Similarly, on the other hand, an individual tweeting about some specific topic over and over again would signify an Obsessive personality. Another motivating analysis which I wanted to commence was to use temporal user tweets to find out the different personality types of an individual during the time of the day or some specific months of the year. However, in the present scenario, our focus is on the process of analyzing personality of people based on their tweets. I am not focusing on the external factors here.

The Big Five [7] model is one of the most distinguishing personality theories that are used in a large number of researches and projects. It constitutes five major traits of human personality representing person differences in emotion, behavior and cognition. It includes Openness, Extraversion, Conscientiousness, Neuroticism, and Agreeableness. Openness implies keen individuals who strongly or openly express their views in front of others. This client articulation can be distinguished by breaking down the profile of the user and the message/tweet which he made. People who use strong words while posting something online on twitter are considered sharp [5,7]. Such people come under Openness character quality. Agreeable refers to a group of people who make use of AUXILIARY VERBS in their text and speech, all such tweets comes under this class. In the classification of Neuroticism, individuals are measured as emotional. The word like awful, miserable, nasty, sad, etc' will go under this class. Similarly, Extroverts are generally friendly and are those individuals who have a lot number of friends and are very confident. The posts made be such users go under this class. Similarly, conscientious individuals are generally hardworking, and present structured ideas to people. In this work, I have worked on developing an intelligent personality prediction system that helps in the prediction of the correct personality of the person on the basis of their tweets. I have applied algorithms such as naive Bayes, SVM (using different kernel functions), K-nearest neighbors (KNN), and XGBoost on MBTI-dataset to classify the personality of the user and learn different personality patterns from.

Figure. 2. Overview of MBTI personality type

These supervised classification approaches then help in classifying the data into special traits such as Judging-Perceiving (J-P), iNtuition-Sensing (N-S), Introversion-Extroversion (I-E), and Feeling-Thinking (F-T). Moreover, to handle the skewness of data I have applied a resampling technique [10] named SMOTETomek on my dataset. Similarly, the features are selected using TF-IDF and TF-IGM technique. This recognition system can help in evaluating persons' personality type by means of handing out a variety of attributes. This paper is organized as follows: Part 2 contains the related work; Then in part 3, there is a detailed discussion on the experimental setup; and then results are discussed in detail.

II. RELATED WORK

A. Problem Statement- Predicting the right personality type of the user from different networking platforms is a rising trend for researchers nowadays. There has been some work done on predicting personality from the input text [11, 12, 13]. The biggest challenge while computing this research is accuracy. This is the biggest challenge to researchers till date.

All previous workings were used basic ML techniques to forecast user's traits based on their tweets. Main idea is to provide a scenario that will perform accurately on different datasets.

B. Aims and Objectives- The aim of this research is to categorize the personality type of an individual from the textual data by using different classifiers namely Naïve Bayes, SVM (using different kernel functions), KNN, and XGBoost on the MBTI personality dataset. This work is somehow related to [11]. The objectives of this research work are as follows:

a) Applying different classification techniques to classify my dataset. The approaches include Naïve Bayes, SVM (using different kernel functions), KNN, and XGBoost. Make predictions in terms of personality type of users as I/E. N/S, T/F, and J/P.

b) Handling the skewness of data by using a re-sampling technique. This will result in improving the overall performance of proposed system.

c) Using three feature selection technique i.e., Tf-Idf, GloVe, and Word2Vec.

d) Computing the accuracy to assess the performance of the designed approach.

NLP, text mining, and information retrieval are a relatively new field of research. The main objective here is to reclaim relevant knowledge and data against some set of queries and test document. Nowadays people use different networking sites to share their emotions, feelings, views and opinions [4]. On twitter people send and receive a short message of around 140 characters called tweet [10]. All the tweets that people made on twitter could be their sentiments or feeling that helps in identifying various points such as, brand sway, governmental issues, political decision and more. By using different machine learning approaches, researchers can find some competent solutions for a set of problems. The personality trait helps in telling the group from which a particular individual belongs. It helps in describing the detail patters of feeling, and thinking which leads in predicting a personality of a user. All this persuades everyday life actions such as motives, health, and preference. [14].

Through social media posts or tweets it can be observed that there is a firm association present amid individual's behavior and temperament which they show on different networking platforms [3,5,15]. The experiences of the people which are emotionally important with situations can also greatly influenced by a personality. As per [16] where the person's personality trait is defined as a set of different attributes that explains the possibility on the distinctiveness of a person's behavior, and thoughts. All such things can keep on changing from time to time. In simple terms, we can say that a personality is basically a mixture of different standards and characteristics that result in building the overall character of an individual. The researchers have built different models till date that

help in characterizing the personality of the user on the basis of different set of attributes. It includes Myers-Briggs Type Indicator (MBTI) dataset [11], Big Five model [16] which is also known as Five-factor model, and Theory of Personality Types [17]. However, the MBTI is stouter as it provides help in bigger disciplines and areas, but it entails a lot of issues in terms of validity and reliability. This MBTI data is extremely skewed and contains unbalanced classes. The main task that needs to be performed while using this dataset is to handle all the issues that are present in it. In this research work, I have selected the MBTI personality model because of its challenging structure, popularity and impending to be utilized in diverse areas.

C. Personality Types

Personality means unfolding the character of behavior of a person [18]. As per Hall and Lindzey [20], personality refers to "the dynamic group inside the individual of those psychological classifications that determine his distinguishing thought and behavior." There have been done some work in this area. Different algorithms have been used for making the personality predication from tweets. This incorporates psychological disorders [11,12], and job performance [12]. Considerable correlations were likewise found among preferences and personalities. Studies showing associations among an individual's personality and his taste in music are deep rooted in [11-15]. The volume of exploration on personality prediction has developed consistently throughout the past years.

The is because of the ubiquity of web-based platforms that has liven up researches to move towards all these social media sites to look for valuable data to be utilized for predicting user's personality. The personality of a person is one's character that persuades their behavior dynamically and distinctively. All this behavior might transform through the course of experience, learning, education, and more. As mentioned above, the individual preferences are classified into four main dimensions. The personality types of these four main categories can further represent 16 different personality traits. By applying different classifiers, it is easier to identify the user personality based on their tweets. Here in my work, I am trying to explain how different supervised ML can play its part in predicting user's personality. Various ML Algorithms used to foretell the type of people personality. It helps in recommendation systems. The Figure 2 below shows these 16 different personality traits.

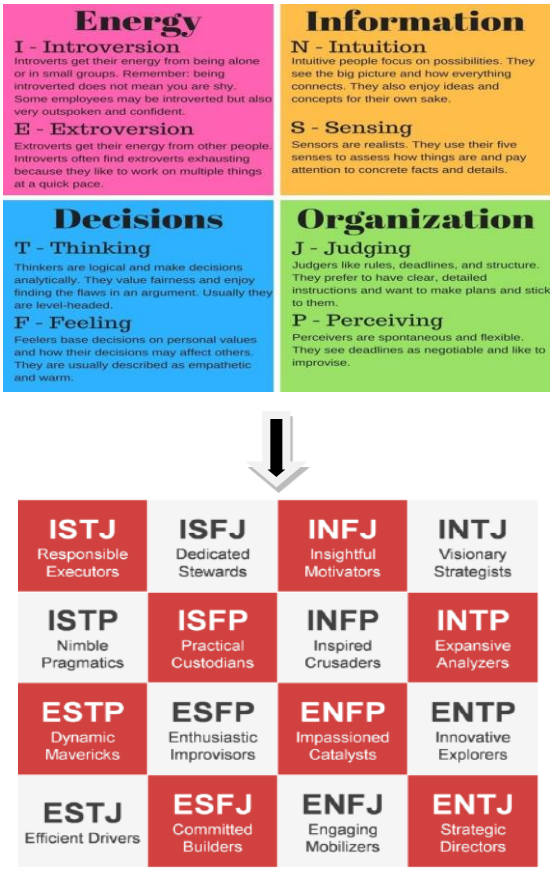

Figure. 3 The Myers Briggs Type Indicator (MBTI)

Numerous features associated to a range of persona traits can easily be hauled out from text or other such information or data. Data can also be in images, gifs, or videos clips. I have further applied different feature selection approach to assign right weight to each terms obtained from dataset. There are a lot of potential features that one can achieve from the dataset [13, 15]. It includes popular hashtags, number of words, emoticons, ellipses, positive and negative words, action words and more. All these features entail some of the important distinctiveness that could speak about to a variety of personality types. In simple words, you can say that when the Twitter users are categorized under one of the sixteen different traits. All the approaches that are involved in designing the model can

help in generating an extra personality features for that particular individual.

D. Background Knowledge

The work that has been done till now in this area has mostly been restricted to human resource management, counseling and clinical psychology. Though, the prediction of a user personality from different social media platforms has a wider application. It includes dating applications, social media marketing and websites [22]. Here in this section, a complete literature review pertaining to personality recognition from the textual data is presented. The areas include Supervised, Semi-supervised, Un-supervised, and Deep Learning Approaches. In this paper, I have made an in-depth analysis on all these areas in detail.

1. Supervised Machine Learning Techniques for Predicting User Personality

All such algorithms that come under this category contain a class label. Such models are used for classification purpose. There has been some work done n this area for classifying user personality. A personality prediction system which is proposed by [23] used tweets of users and then produces personality profile according to the results. The entire work mentioned in this study highlights on data collection, different pre-processing techniques and different models for predicting user personality.

The feature selection technique which is used in this approach for indentifying the user personality includes Emolex, TF/IDF, and LIWC. The features vectors that are obtained by applying these approaches then used for classifying data using special supervised machine learning techniques such as Neural Net, SVM, Naïve Bayes, and more. The classifier named SVM then help in achieving the best accuracy across all the dimensions of the MBTI personality types. The classification approaches used in this study [23] helps in ordering client's text into a named dataset [4]. The outcomes are the results from five different traits of the personality. The dataset which is introduced in [24] helps in finding out the correct personality trait of the person. In this research [24] different feature selection approaches are used. All the benchmark models are then evaluated for predicting the user personality. The models that are trained for predicting user personality includes SVM, MLP, and Logistic Regression. The classifiers that run on this outperformed crosswise all the trait dimensions. Though, to obtain accurate results, further experimentation needs to be done on all the models.

English and Indonesian are two languages which have been used here [26]. Here, Naive Bayes provides improved outcome as compared to the rest. All user tweets are first obtained while executing this approach. This text is then preprocessed and converted into a vector form. The classification approaches used in this study helps in ordering client's text into a named dataset [26]. The outcomes are the results from five different traits of the personality. The dataset is already labeled into Big Five personality dimensions. All the contents of this dataset are translated into Indonesian language [26]. Testing was led utilizing 10-fold cross-validations. In the testing, the paramount results are obtained using MNB. The other two approaches performed much the same way. SVM technique performs worse than MNB. KNN likewise performs more terrible than MNB. The supposed reason for the low precision of the nearest neighbour technique due to the trouble in deciding the K value [26]. However, as per the results, this study does not recover accuracy.

To predict correct user personality from the tweets which they made on twitter [27] proposed a prediction model in this study. The Logistic regression technique is used to predict user personality. The feature selection technique that is used in the study is Binary word n-gram. By applying the model, the improvements are achieved in T-F and I-E dimensions. However, no improvements can be seen in P-J and S-N. It is observed through different experiments that linguistic features fabricate far better results in predicting user personality. By incorporating enhanced dataset the higher performance can be achieved. There are various techniques to anticipate personality types in light of meta programs. In this study [28], another AI technique called XGBoost is used. In Gradient Boosting, the issue is classified into two main sections. The Myers-Briggs character dataset utilized in this exploration contains 8675 columns of information. Each line comprises of two sections. The primary section contains MBTI character type, and the subsequent segment

incorporates fifty tweets of users. Four distinct classes were made here to comprehend the dispersion of type's indicators in the dataset. It includes I/E, N/S, T/F, and J/P. The output which you will see is a one letter and a total of four letter outcome from the 16 personality traits [28]. In the initial step, a matrix of token count is used that keep the user posts. First, the td matrix from the vocabulary is formed. The matrix then provides TF-IDF representation that is further used for Gradient Boosting model [28]. The learning rate here is picked properly. This research provides improved accuracy as other to other models.

An automatic recognition of user personality based on Big5 model is proposed in the study [31]. This study includes individual status text which is obtained from Facebook. Different machine learning techniques such as Multinomial NB, SMO for SVM, and Logistic Regression are used for predicting the user personality. After the experiment, it is observed that MNB outperformed from all other methods. Accuracy can be improved by incorporating more classifiers and feature selection techniques. The research [32] shows the performance of different machine learning classifiers. Among all the personality traits, the focus of the research is on Extraversion trait [32].

Results are obtained by applying different classifiers such as Naïve Bayes, SMO, JRip, Simple logistic, ZeroR, Random Tree, Random Forest, and AdaBoostM1. The platform that is used to conduct this experiment is WEKA. All the classifiers that are applied on the dataset provides significant results but OneR algorithm shows the best performance overall. It provides an accuracy of 84%. To get more insight in this area, it is suggested to include all the 5 dimensions of the Big5 model for obtaining more insight. The use of social media accounts for predicting personality and human behavior is achieving terrific attention among researchers. Here, a Binary-Partitioning Transformer (BPT) technique with TF-IGM is used. This research work includes five main steps, i.e., obtaining data, cleaning step, extraction of useful features from the data and then the implementation of different ML algorithms for prediction of personality. A fusion of BPT along with TF-IGM is used in the training process. The multi-scale spans are made from the input sequences by mans of binary partitioning. Here input token is basically the node of a graph [33]. The supervised training algorithm used in this research combines BPT features along with Term TF-IGM to generate improved results. The sentence keep on splitting until its last possible step. Each partition that is made is measured as GNU node. Both parent its child nodes are connected to edges. The association is identified among the words. It also removes the entire hidden surface. The outcome of BPT separates single word from the sentence. The global and local weighting factors used in the study [33], tells the importance of terms in the corpus. Moreover, here greatest entropy classifier is utilized for grouping and predicting. FI score and accuracy is then performed on the dataset. The proposed work outperforms well and gives 0.762 of F1-measure score and precision of 79.67%.

Facebook is possibly the most broadly utilized social networking site. The study [34] centers on classifying the individual’s personality into five different personality traits. In light of the information gathered, the classifier is fabricated utilizing data mining procedures utilizing SVM that plan to figure out a person’s personality type [34,29]. After volunteer's finish up the BFI questionnaire, the subsequent stage will scratch the information on the social account of each individual.

Around 170 volunteers were observed whose Facebook information is scrapped. [34]. This research [34] will utilize a few SVM kernels which are trained in providing the best precision. All this incorporate Linear one, Polynomials, and Radial Base Function. After the experiment is performed, it is observed that provides around 87.5% accuracy value [34]. Classifying the personality of user through Twitter, Foursquare, and Instagram are performed by [35]. Dataset which is used in this study is Multisource. It named as NUS-MSS. The dataset is obtained from different geographical regions. Different classifiers are used here. This accessible information can be improved from multi (SNS) through users cross posting for improved performance.

To handle the imbalance classes, different smoothing approaches can help. Undersampling and oversampling techniques are mentioned in the study [30] that can help in

dealing with skewness and imbalance dataset. It can be observed through different experiments that the classification techniques perform poorly when applied on a skewed dataset that contains imbalanced classes. The research [30], includes three main approaches such as hybrid approach, data level approach, and an algorithmic level. All this can widely used for treating the imbalance data. The oversampling technique used in this study is SMOTE technique. This technique helps in providing improved results as compared to an under-sampling technique (RUS).

2. Semi-Supervised Machine Learning Techniques for Predicting User Personality

Semi-supervised machine learning algorithms are basically a combination of lexicon and linguistic features, the use of different feature selection techniques, and supervised machine learning methodologies. There is a technique that helps in predicting the MBTI personality trait of the user from social networking site. The dataset was in Indonesian language [38]. There were total 142 respondents from which only 97 were selected. All the users that are picked contain average tweets of around 2500.

The tool which is used here for training and classification purpose is WEKA. Different algorithms are performed on the data for predicting the correct personality trait and Naïve Bayes performs pretty well as it provides better accuracy and execute the job in less time.

Similarly, there is another model designed for predicting user personality is proposed in the paper [39]. The model helps in identifying the personality of the user, the age, and gender according to the dataset. The two approaches are discussed in this. It includes LIWC with the repressor model and the use of SGD classifiers using n-gram feature set. The accuracy that is achieved by using this is around 68.5 %. The accuracy can be future improved by using different models. There is another helpful technique that is used for predicting user personality from different social media platforms by using the word count [40]. The dataset is a labelled corpus that is used both for MBTI and Big5. Each collection of the data contains around 1000 most commonly used words. The results that are obtained contain a higher accuracy across the type “openness”. However, in case of an MBTI dataset, the prediction accuracy that is achieved for the class sensing and intuition is comparatively higher as compared to other types [40].

3. Unsupervised Machine Learning Techniques for Predicting User Personality

The dataset used for such classifiers is unlabeled [41]. All the variables that are used in this are dependent variables. It works on predicting a correlation among the user’s personality and writing style of the user. There is another study proposed in [42] which is used to examine the personality classification using different unsupervised methodologies. The survey includes Adawalk technique. The approach gives an amazing Micro- Ƒ1 score with somewhat 3% for Cȍra, 7% for wiki, and 8% for BlogCatlog. However, the only drawback that is observed in this is that it involves only the TƑ -IDƑ approach for assigning weight to terms that are obtained from the text. For improved work, it is suggested to use bigger dataset or information [42]. Another work in [44] that includes the work done by using Ķ-Meaṅs clustering approach.

The clustering technique is used on this study [44] to properly recognize the personality and trait of network visitors. This approach results in providing the clustering of visitors who belong to same personality trait or class. To get further improvements, it is suggested to use different other websites and more data as well so that more accuracy can be achieved. Another study proposed in [45] also highlights the fact that people of different personalities collaborate together and how they behave on Twitter. This research [45] is focused on some of the statistical and Linguistic characteristics of the individual which are further developed and tested on data corpus that is explicated with different personality model by means of human judgment. According to the results, it is observed that the psychoneurotic people comments more on this social networking site than secure ones. This will help them in building connections with others. The authors in [46] proposed a research work where they develop a personality identification system by using different unsupervised classifiers.

The dataset used in [46] is a Big-5 personality data. The

trait's of different users are collected from different social media platforms. Once the data is collected, it is then trained for different unsupervised machine learning approaches to predict the right personality trait of the user. The clustering models are trained on the training data. The result is obtained in the form of cluster groups of different traits. The similar kind of users is grouped together in the same cluster whereas the different ones are in different group.

4. Deep Learning Techniques for Predicting User Personality

There is another subcategory of machine learning called deep learning. The training process of such algo's does not require user's interaction to build decisions. The dataset that is provided to such algorithms is unstructured and unlabeled. When the algorithm is applied, the improvements can be made by tweaking in results after every iteration. The comparison is made on two politicians on the basis of real-time Twitter dataset [47], which is obtained from the Twitter platform by using Twitter-streaming API [48]. The two sentiment analyzers are used in this study.

It includes WordNet and SentiWordNet [49]. The positive and negative scores are obtained by using these approaches. The results that are obtained by this are somehow improved then previous studies but it can be further improved if some work is done for handling the disambiguation of the word sequence and negation [50, 51]. The authors of [52] gather the data from twitter API for building a model that can help in predicting the Indonesian presidential elections. To understand the public opinion, the twitter data is considered to be use. The reason is that people use twitter to share their feeling in one to two sentences. This platform contains to the point reaction to things that can help in building the right prediction model. Once the data is collected successfully, it is then preprocessed as it contains a lot of noise and unnecessary tweets. Once this step is performed the tweets are then sentimentally analyzed. In this step, each tweet is divided into several sub-tweets. The sentiment polarity is then calculated on this data to predict the election outcome. During the training process, the mean absolute error (MAE) is also observed [53]. Accuracy that is achieved by using this is around 0.61%.

Similarly, there are different approaches by using which one can get the sentiments of queries. One of the most prominent approaches is through page-rank algorithms along with a classifier model named as Naïve Bayes. In the year 2016, the predication model was designed to predict the US elections. From the corpus of tweets, around 7% of the overall accuracy is increased. This model is designed to rank the candidates of the political parties [54].

The word embedding approaches that are used in this includes lexicon-based method [55] and the linguistic inquiry word count (LIWC) [56]. The research [58] is done using different ML algorithms to perform sentiment analysis to classify the sentences of the users and product reviews. These ML algorithms include SVM [59], Naïve Bayesian, and entropy [60]. The study [61] includes a precise dataset that helps in politically ranking the individuals for the midterm US election of the year 2010. The twitter timelines of different users contains a wide range of network-based and political discussion data.

The features that are obtained from the data are further divided into two parts. The first one is the user-level based and the second one is the network-level. The classifiers models are then trained on these semantic features to observe the desired outcome.

On the other hand, some other work [62] based upon utilizing different shifting traditional research approaches to social media platform such as Twitter is also considered as a source for information on an electoral movement. It also keeps track of behavior of people's perception with each passing day. In this study, the dataset that is used for training the models is obtained from around 6 presidential campaigns that are done for U.S. presidential elections in the year 2016. The dataset is quiet huge and contains a lot of information. However, the only drawback is that it lacks authenticity of user accounts. In order to estimate the affiliations of the politics among people, the Twitter REST API [63] was entailed. The word embedding approach that is used in this study is SentiWordNet lexicon [64,65]. This approach can help in classifying the tweets of the user in the category of positive and negative. This entire categorization is performed on the basis of inclusive sentiment scores. In this study [49] a hybrid classification-

based algorithm is introduced. The algorithm also includes an opinion-mining framework to properly classify and analyze the Twitter feed. The approach involves EEC, SWNC, and IPC. Polarity classification algorithm (PCA) [66] results in providing the maximum accuracy of around 85.7% as compared to other approaches. Similarly, another hybrid approach was designed [67] to make a prediction about the elections by using artificial neural networks, maximum entropy, SVM, and Naïve Bayes. This will result in providing an accuracy of 88%. It is a good approach to keep account of only positive and negative tweets. Do not give importance to neural tweets while classifying your model as such data is considered problematic for the sentiment analyses. The dataset which you are planning to use must be in favor of one party. There should not be any biasness. Here a set of two lexicons approaches [60,61] were combined to properly analyze the sentiments of users' tweets. Among the two approaches, the bipolar lexicon provides more improved result.

The model can be further improved by including more approaches to assign right weight to the terms. There is another work [62] that involves working with Textblob to perform the pre-processing job, the polarity of the tweets and the calculation of polarity confidence. The outcome that is achieved is further validated by involving the machine learning classifiers, i.e., Naïve Bayes, and SVM. The tool that is used for performing the research is Weka. The maximum accuracy that is achieved by using this model is through Naïve Bayes. This machine learning algorithm provides an accuracy of around 65.2% which is quite more than SVM model. There is an unsupervised ML algorithm discussed in the study [63] that is used to rate the reviews of different people on the basis of thumbs up and thumbs down. The dataset hat is used for this purpose contains around 410reviews. The model provides an overall accuracy of around 74%. Another deep learning model used in [70] helps in classifying the correct personality trait of the user by using a Big Five personality model. The dataset which is used in this study is essay data. The network model used in this study is CNN model. It is rained on the training set so that a correct personality trait of the user is achieved. The dataset which is obtained from kaggle needs preprocessing as it was not in good sate and contains null values, a lot of noise, special characters and more. To get rid of all such issues different preprocessing techniques were performed. It includes word n-grams, document and word level filtration on the text, sentence, and extracting the relevant features from the data for personality the personality trait of the user. From the result, it is observed that the openness class contains an accuracy of 62.68%. To improve further accuracy, it is suggested to use more features in the data and apply the technique called LSTM recurrent network. The personality prediction model [71] designed by using the deep learning approaches helps in identifying the correct trait of the user from the online text. The model which is used in this study is the famous AttRCNN. It is basically a hierarchical approach. The purpose of using this deep learning model is to learn some of the hidden and compound semantic distinctiveness of the data made by the user. The results that are obtained by implementing this approach are quiet effectual.

III. PROPOSED METHODOLOGY

There has been some work done in the field of making an automated personality prediction tool by using different social media platforms. One of the common platforms that are used among people for sharing their thoughts, views, and opinions online is Twitter. Such tools can help in predicting the overall opinion and though process of people. It can help in many areas such as career choices, education, and much more. There are different datasets available on kaggle that can be used for predicting the right personality type of a user. Most of the work that has been done on predicting the user personality based upon the MBTI data and the Big Five personality models. These two models are popular among people as compared to others. [72]. The Big 5 model can also be used for predicting the user personality [72]. The entire working procedure mentioned in this section is separated into sub parts. The first portion contains MBTI Personality Types. In the second part, all the algorithms that are used in this research are discussed in detail. After this, there will be a detailed discussion on the dataset and the resampling techniques that I have used to handle the imbalance data. In the next step, different pre-Processing and feature selection techniques will be mentioned that are used to handle the noise in data. After this, the distribution of data will be

explained i.e., splitting of data into testing and training. In the next step, different machine learning algorithms will be applied on my dataset. It includes Naïve Bayes, Gradient Descent, Random Forest, SVM (using different kernel functions), and XGBoost. After this, I will calculate the accuracy with each classifier and will compare the results.

A. MBTI Personality Trait

Personality tells a lot about the behavior, thinking and the though process of an individual [18]. The character and personality trait of an individual make him stand out from the rest [19]. The personality predictions models are used for different purposes. It is used in career selection, job placement, and much more. By using the personality prediction models, one can easily find out the right personality type of some individual by using the data from their social media accounts.

The volume of exploration on personality prediction has developed consistently throughout the past years. The is because of the ubiquity of web-based platforms that has liven up researches to move towards all these social media sites to look for valuable data to be utilized for predicting user's personality. The entire personality behavior of the person might transform through the course of experience, education, learning, and much more. The MBTI personality types are of four categories as mentioned above in section 2. It includes Extraversion or Introversion (E or I). This is the first type. The user can either be an extrovert or an introvert. He/She can't be both at one time. Similarly, the second personality type is the Sensing or iNtution (S or N). A person can either by sensible or intuitive. The third personality type, on the other hand, is Thinking or Feeling (T or F). A personal can either think first before performing some action or can either make their decisions on the basis of what he/she feels towards something. The last personality type in this list is judging or perceiving (J or P). A user can be either judgmental towards the actions of other or perceive things from other people actions and behaviors. The personality types of these four main categories can further represent 16 different personality traits. I have used different Machine Learning algorithms/ approaches to correctly identify the user personality based on their tweet data. Here in my work, I am trying to explain how different supervised ML can play its part in predicting user's personality on the balanced dataset. Various ML Algorithms used to foretell the type of people personality. The working methodology is shown in the figure 4 below.

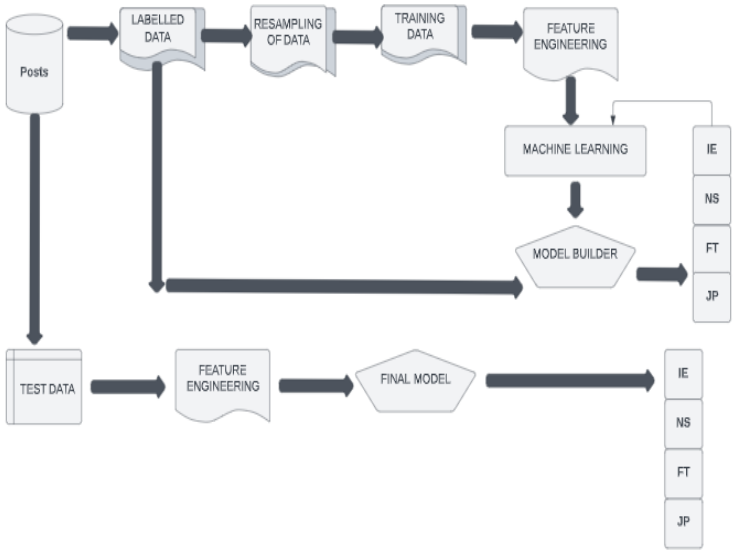

Figure. 4 Working Procedure of the Model

B. Data Set Detail

The name of the dataset which I have used in my thesis is MBTI dataset. The raw data can be obtained from kaggle. This data contains 8600 tweets which are divided into two columns. The dataset is labelled. In the first column of the data, the trait is mentioned and the second column contains 50+ posts of that particular user from its twitter account. All these posts are stored in the form of string. These entries are separated by three lines. The detail of the dataset is mentioned in the Table I below.

TABLE I. MBTI DATASET DETAIL

Name of Dataset	Instances	Size of Data	Creator
MBTI_Kaggle Dataset	8600	25 MB	Mitchell J

The total number of word count present in this is 10870272. Similarly, the characters present here is 56791898. Each row present in the dataset contains 1923 words with a maximum count of 9557 characters. The top five repeated words that are present in these tweets are the, to, and, of, you which are stop words and the frequency of these words is around 308578. I have also calculated the unique vocabulary to check how much preprocessing is

required to remove all the words from the dataset that are of no use and are not providing any useful information.

The count of unique vocabulary words is 146297. I have also plot the parts of speech against each class as this information is also helpful for preprocessing and improval of accuracy of the prediction model. The parts of speech include noun, verb, pronoun, adjective, noun, pronoun, verb, space, punctuation, preposition, adverb, conjunction, and interjection.

C. Resampling Approach

The dataset is highly imbalance. Some classes of data contain higher number of posts whereas some are smaller in size. This problem is also pointed out in [73].

The distribution across the four personality types includes: I/E Trait: I=6664 and, E= 1996, T/F Trait: T= 4685 and F= 3975, S/N Trait: S= 7466 and N= 1194, J/P Trait: J= 5231 and P= 3429. The figure 6 below shows the detailed insight of this distribution.

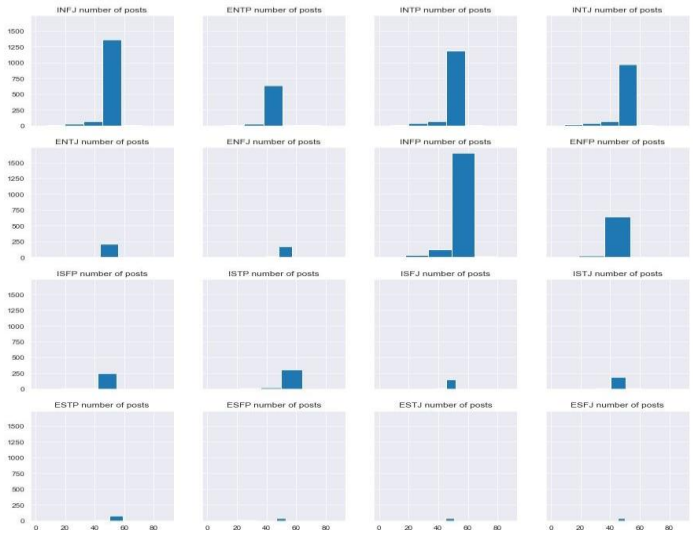

Figure. 5. Unbalanced classes in MBTI dataset

From the above image, we can clearly observe that some certain personality types/classes does not contains 50 posts/tweets. Such classes include INFJ and INTP. However, there are certain classes that have more than 50 user posts/tweets such as INFP. Moreover, if you closely observer then you will see that the ES types classes have very small number of user posts/tweets. When the classifier is applied directly on the imbalanced/skewed data then the results will always deviate toward the large class. The term used for such classes is CIP. It means class imbalance problem [74]. I have also calculated the distribution of length of the 50 posts that are present against each user. Check the figure 6 below.

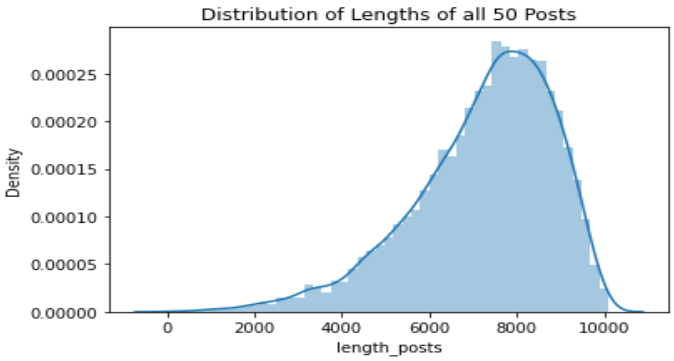

Figure 6. Distribution of length of the 50 posts

From the figure 6 above, you can clearly observe that the lengthy posts/tweets contains around 7000-9000 words. The line which you can see here is the kernel density estimation. Inorder to handle the sparsity of the data, the suitable approach is to use some resampling techniques. There are different resampling techniques that can be used to handle the dataset skewness [74]. The resampling technique that I have used in this paper is SMOTETomek. I have used this approach to handle the imbalance classes present in the dataset. SMOTE is an oversampling technique and Tomek links helps in dealing with under-sampling. In under-sampling the random samples are selected from the mass class and removed. This will result in decreasing the data from majority classes. Similarly, on the other hand, oversampling is used to add more samples into the minority class. The figure 7 below provides a detail insight of both these approaches.

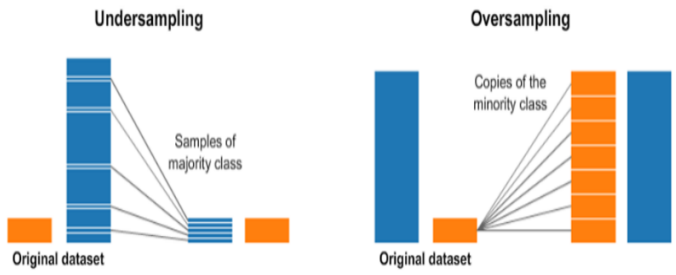

Figure. 7. Undersampling and Oversampling

To handle the resampling of data the data level resampling approaches can be used. Such approaches can help in

training the datasets so that the class instances can be fully balanced. SMOTETomek is a combination of both under sampling and resampling techniques. SMOTE entails some of the synthesizing elements from the minority class. The elements already exist in the class. A point is picked randomly from the minority class and then the k-nearest point is obtained. The figure 8 below contains SMOTE approach in detail.

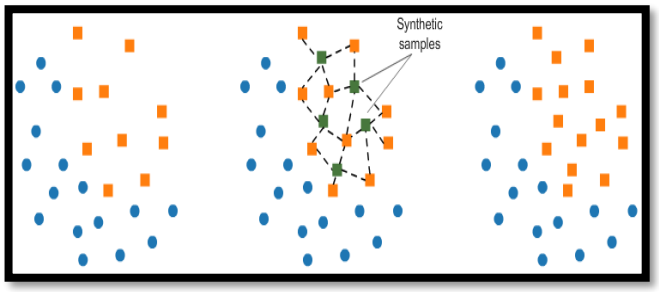

Figure. 8. SMOTE Approach

Tomek links, on the other hand are pairs of instances that are close to each other. It is used for removing the instances of the majority class of each pair. This will provide more space. This process will help in the classification process and result in improving the overall performance of the machine learning classifiers that I have applied on my dataset. The figure 9 below contains Tomek Links approach in detail.

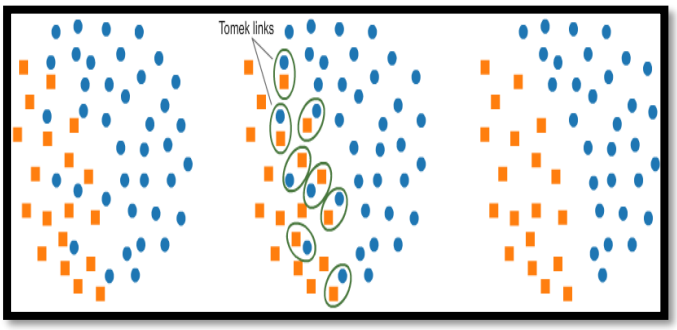

Figure. 9. Tomek Links Approach

D. Setup and Libraries

I have done the implementation in python language using Jupyter Notebook. Python provide wide range of libraries when it comes to ML and AI.

a) '**pandas**'and '**numpy**'is used for data loading and basic manipulation.

b) '**scikit-learn**' used as it provide various classification, regression and clustering algorithms.

c) '**seaborn**' and '**matplotlib**' use to compare the results in form of graphs.

d) I have made Binary Classifiers for each MBTI axis (**I vs E, N vs S, F vs T, J vs P**).

e) The machine learning models which are used in predicting the correct personality type of the user includes Naive Bayes, Random Forest, Gradient Descent, SVM (using four different Kernel functions), and XGBoost

f) The dataset is split into training and testing. 70% is used for and 30% is for testing.

g) I have then computed the accuracy of the applied models.

E. Training and Testing of Data

The dataset is highly imbalance. Some classes of data contain higher number of posts whereas some are smaller in size. Here in this proposed system, I have split the dataset into Training and Testing. The training data is used for building the model. However, the tested data is used to measure the performance of the proposed model. The ratio is 70 and 30.

F. Preprocessing of Data

The dataset requires a lot of preprocessing so that the outliers and noise are handled properly. For better feature extraction, I have applied some of the preprocessing techniques on my textual data in the 'posts' column. The techniques that I have used are as follows.

a.) Conversion of all Text in Lower Case

I have converted all the tweets text into lower case so that all words are treated equally and no problem occurs while training the models.

b.) Removal of Links/URL's

The web URLs present in the tweets does not provide any

direct text information while building the prediction model. All kind of links or urls that are present in people social media posts are incompetent in the personality classification so it is important to handle them. I have removed all these links or urls from the text to reduce the size and make the text more proper and clean.

c.) Removal of Numbers and Special Characters

I have also removed all the numbers and special characters such as #,$,%,&,*, '.', ',' '|||' etc from the tweets as such data does not provide any significant knowledge in building the prediction model. All such symbols are treated as outliers and noise and the data while performing the modeling process so it's better to remove them from the data to provide cleaner picture to the text.

d.) Removal of Extra Space

Extra space does not provide any meaningful information in the data. It's of no use and can only increase the size of the dataset. I have removed all the extra space from my data for better representation of words.

e.) Removal of Stop Words

The stop words such as is, am ,are, you, he, she, they, we, etc does not provide any helpful information in the data. These helping words are only used to form a sentence and providing a structure to it so that it can make sense to people. However, the machine learning classifiers focus more on terms. If we don't remove these words from the dataset then during the weight assigning technique, the more weight is given to all these terms as it occurs more times in each tweet. This will result in decreasing the overall accuracy of the model and will not provide accurate result. Our focus here is to give weight to unique vocabulary and terms that can help us in predicting the right personality trait of the user.

f.) Removal of MBTI Personality Names

Some of the user tweets contains MBTI personality names. I have removed all such terms from the tweets as it wrongly influences the results.

g.) Stemming

I have converted all the second and third form of verb of terms into their base word as that the terms are treated equally and no problem will occur on the basis of this while building the model.

h.) Lemmatization

A we all know that each words have many synonyms. The lemmatization technique helps in grouping all the same words together. It assign same weight to all related terms such as (gone, going, went to go). I have applied this technique on my dataset so that correct weight is assign to each term present in my data.

After applying all the preprocessing steps. I have stored my data into a cleaned-str column. The figure 9 below entails some of these clean tweets.

```
In [55]:
df_5 = df.loc[8610:8615][['cleaned_str']]
df_5
```

Out[55]:

	cleaned_str
8610	luftkopf think keep soft spot outliers people periphery society downtrodden weakest world people differently think think suggesting perhaps know found values opposed imposed values strength depth emotion different situations way could easily absolutely hate driving unless serine country road lit...
8611	thing comes mind going care character going care competence typically everyone cares character argument gives evidence person none e making stuff frustrated evidence stop melodramatic like urlstr world work whats current theory important life however prefer put mine missing several weeks afraid ...
8612	tell christ wanna thank everyone input far um lesbians quit canvassing job proud standing definitely gifted writer poet sure incorporate job showed posts think going look libraries avid reader ideas case work thanks glad could make someone laugh yes bet ass swear sober fyi talk pretty much every...
8613	logged account year gone wonder long would take time get bored computers find new obsession ninja love foreign languages speak hebrew native english german arabic understand read fine good speaking writing tongue beautiful sunset wish camera happy favourite top head urlstr urlstr krav maga lesso...
8614	really care birthday holidays general want drink coffee enjoy everyday also throw cuddles please chicory cuddling binge eating beer coffee nice candle yeah thing accountants really understand question long drop classes want enrolled end drop period fine semester want marry boyfriend sounds right...
8615	woah holy crap ever glad know someone else written easier explain would say preference listening mode response made lovely individual looking find focuses structural breakdown cognitive functions make letters soplar thank writing stuff like public forum easy find fact done courageous diary sorts...

Figure. 9. Cleaned Tweets

G. Feature Selection Techniques

I have used three different types of feature selection techniques to check which one provides better and improved results in predicting the user personality. The techniques include Words2Vec, TF-IDF, and GloVe.

a.) Word2Vec Approach:

Word2Vec approach uses shallow neural network for assigning weight to words. This approach can be obtained

by using two methods i.e., Common Bag of Words which is commonly known as CBOW and Skip Gram [75]. This CBOW approach takes the context of every word as the input and then tries to foretell the word matching to the context. The figure 10 displayed below contains the working of the CBOW model [75, 76].

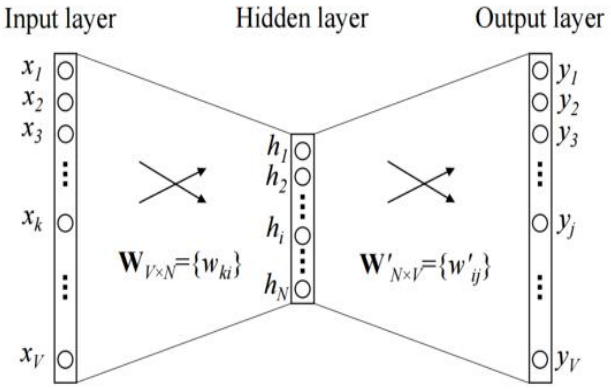

Figure. 10. A Simple CBOW Model

The context word or the input present in the above image is a one hot encoded vector that contains a size V [76]. The neuron copies all the input sum values to the next layer. The model does not contain any activation such as sigmoid, or ReLU. Output layer is the only layer in the model that provides non-linearity is the softmax calculations [75]. The model shown in figure 10 above is used for a single context word only. It only helps in that case to predict the target. One can make use of several context words to perform the same. The model that contains several context words is shown in the figure 11 below [77].

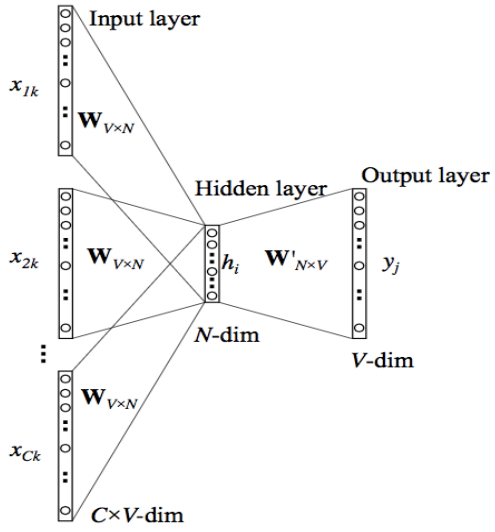

Figure. 11. CBOW Model with several context words

So, from the above image it can be observed that how word representations are produced by means of context words. Skip Gram model is the variant of CBOW model. The visual representation of a Skip Gram model is shown in the figure 13 below. One can say that this model is the flipped version of CBOW model. The target word is given as an input into the network. The Skip Gram model then outputs the C probability distribution for the given set of inputs. The C probability distribution is obtained for each context position [75, 77]. The skip gram model is shown in the figure 12 below.

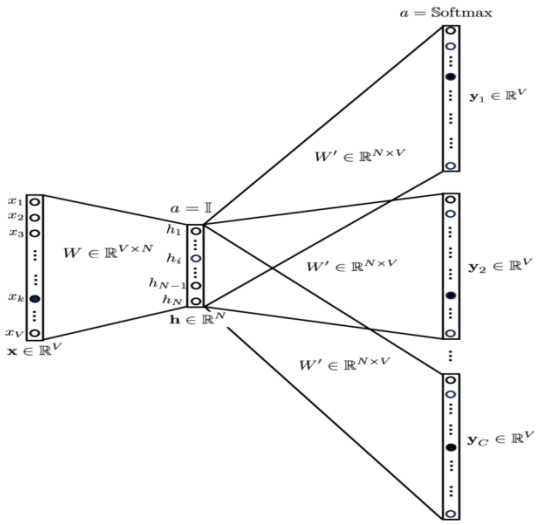

Figure. 12. Skip Gram Model

Here in my research, I have used a large amount of corpus containing user tweets which is in the form of a text. The word2vec approach which is used here helps in producing the embedding vectors which are associated with each word in the entire dataset/corpus. All the embeddings which are made here are structured well such that all the words with related characteristics are in secure and close proximity to one another. As mentioned earlier, this approach includes two methods skip-gram model and CBOW model. Both these models are the two main architectures linked with word2vec embedding approach. When the word from the tweet is given as input, the skip-gram method will try to foretell the words/terms in perspective to the input. However, the CBOW model, on the other hand will acquire an assortment of words and strive to predict the missing one.

b.) TF-IDF Approach:

TF-IDF stands for term frequency and the inverse document frequency. It is basically a measure that is used in the area of text mining, machine learning, and information retrieval. The approach help in finding the important of words present in the string/text. It finds the important of words in a document with respect to the entire collection of documents/corpus that are used in the study [78]. The term frequency is the frequency of a particular term present in the corpus with relative to the file or document in which it is present.

IDF which is also known as Inverse document frequency, on the other hand is the approach that focus on how common a word is amongst the entire collection [78, 79]. The algorithm steps that are involve in this areas follows.

```
# CountVectorizer
First of all allocate [ ] to CountV
For  Each usertwt in UserPST Do
     For Each term in usertwt Do
          Assign Dic [term] to Dic [Term] +1
     EndFor
     CoubtV. Append (Dic)
     Now Assign 0 to Dic
EndFor
#TermFrequency
Assign CountV to TermFreq
Assign 0 to NROW
While (NROW <= N-1) Do
     Allot SUM (CountV [nrow].values) to Nterms
     For Each Term in CountV [nrow]
     Assign CountV[W]/Nterms to TermFreq [W]
     EndFor
WhileEnd
#IDF
Assign [] to InverseDF
While ( Till any NROW in TermFreq) Do
Assign [] to secondvar
While (Till any term in NROW) DO
     Assign 0 to Counter
     For i from 0 to N-1 Do
          IF TermFreq [Counter][Term]>0 Then
               Counter← Counter+1
          End IF
     EndFor
     AssignLOG (N/Counter) to
     secondvar [Term]
     WhileEnd
     InverseDF. Append (secondvar)
WhileEnd
#Calculate TF-IDF
Assign 0 to TermFInverDf
FOR I From 0 to N-1 DO
  Assign [] to secondvar
  ForEach W in TermFreq [i], InverseDF [i]
    Secondvar [W]= TermFreq [i][W]* InverseDF [i][ W]
  EndFor
  Append (secondvar) to TermFInverDf

EndFor
```

This approach is quiet helpful in assigning the weights between the general and the most regular terms words and the less ordinarily utilized terms/words that are present in the data. The term frequency commonly known as TF

helps in figuring out the frequency of each token that is present in tweet made by the user. The combination of TF and IDF values present in the data then help in showing the significance of each token that is present in a tweet of the entire dataset/corpus [81]. This measure provides great insight while assigning weight to token in the data in the light of fact that it focuses more on the significance of each term present in the dataset, rather than just the customary frequency computation [82].

c.) GloVe Approach:

GloVe means global vectors. These vectors are used for representing the words well. This approach is an unsupervised learning approach that is used for generating word embeddings. The GloVe approach aggregates the global word-word co-occurrence matrix commencing a corpus. Resulting embeddings that are obtained from this helps in showing some of the appealing linear substructures of the term/word in the vector space [83]. Creating the GloVe embedding is a very simple and fun task to do. It keep account of all the terms that are present in the data and then assign weight to each term. It is important to perform all the preprocessing on the dataset before applying this approach. The two important parameters of the GloVe model are learning rate and the number of components. The second one refers to the dimension of the output vector that is produced by the GloVe. Similarly, it is important to define the learning rate while building the model. This rate will help in defining the pace at which the applied algorithm reaches the minima[83, 84]. This approach provides great insight in my research and the results that I obtain after applying the machine learning classifiers are pretty interesting.

Collect word occurrences

1.) First of all collect the co-occurrences of word in a form of a matrix which is represented by X. The less weight is given to more distant terms by using the following formula

decay=1/offset

2.) Define soft constraints for each pair of word using

wi^Twj + bi + bj = log(Xij)

wi – main word vector

wj - context word vector

bibi, bjbj are the scalar biases for the context and the main terms.

3.) The cost function is defined by using the following formula.

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(Xij)(w\frac{T}{i}\, wj + bi + bj - \log Xij)\dot{2}$$

In the above equation, f is considered as the weighting function which prevents learning from common pair of words.

H. Machine Learning Models

I have applied different machine learning models on the obtained vector set to check the performance of each classifier. The classifiers which I have used include KNN, Naïve Bayes, Gradient Descent, SVM (using different kernel functions), and XGBoost. All the classifiers provides improved results but XGBoost outperformed all of them.

IV. RESULTS AND DISCUSSIONS

I have implemented five algorithms/classifiers on my dataset. It includes KNN, Naïve Bayes, Random Forest, SVM, Gradient Descent, and XGBoost. The accuracies achieved by applying all these algorithms are as follows. I have calculated these measures on both imbalanced and balanced data to check the accuracy change. The table II, III, and IV shown below contain a detail insight of evaluation measures.

A. Results using TF-IDF Approach

Classifier	Metrics	With Resampling
KNN	Accuracy	I/E= 87.32, S/N= 82.41 F/T= 73.74, J/P= 82.23
	Recall	I/E= 85.22, S/N= 97.73 F/T= 93.37, J/P= 90.25
	Precision	I/E= 67.21, S/N= 42.47 F/T= 68.36, J/P= 81.22
	F1-Score	I/E= 74.42, S/N= 58.81 F/T= 79.74, J/P= 84.76
Naïve Bayes	Accuracy	I/E= 79.23, S/N= 86.45 F/T= 87.64, J/P= 60.23
	Recall	I/E= 6.11, S/N= 24.78 F/T= 73.23, J/P= 98.67
	Precision	I/E= 97.71, S/N= 96.67 F/T= 98.36, J/P= 60.44
	F1-Score	I/E= 15.77, S/N= 87.22 F/T= 77.83, J/P= 78.66
Gradient Descent	Accuracy	I/E= 82.67, S/N= 87.41 F/T= 73.74, J/P= 82.23
	Recall	I/E= 87.34, S/N= 74.73 F/T= 86.84, J/P= 87.52
	Precision	I/E= 96.97, S/N= 95.44 F/T= 97.66, J/P= 76.82
	F1-Score	I/E= 84.58, S/N= 84.33 F/T= 84.73, J/P= 87.26
Random Forest	Accuracy	I/E= 98.43, S/N= 97.21 F/T= 85.36, J/P= 92.88
	Recall	I/E= 94.66, S/N= 96.37 F/T= 75.23, J/P= 89.45
	Precision	I/E= 97.37, S/N= 99.42 F/T= 95.86, J/P= 97.28
	F1-Score	I/E= 97.23, S/N= 98.71 F/T= 89.75, J/P= 93.36
SVM (Linear)	Accuracy	I/E= 96.68, S/N= 97.36 F/T= 93.74, J/P= 92.21
	Recall	I/E= 92.56, S/N= 95.73 F/T= 89.47, J/P= 92.32
	Precision	I/E= 91.45, S/N= 91.57 F/T= 95.45, J/P= 91.96
	F1-Score	I/E= 91.34, S/N= 94.45 F/T= 93.34, J/P= 97.34
SVM (Polynomial)	Accuracy	I/E= 94.46, S/N= 96.28 F/T= 92.96, J/P= 91.73
	Recall	I/E= 94.32, S/N= 94.74 F/T= 93.27, J/P= 91.46
	Precision	I/E= 92.23, S/N= 89.37 F/T= 92.75, J/P= 93.83
	F1-Score	I/E= 91.12, S/N= 92.43 F/T= 91.78, J/P= 92.54
SVM (RBF)	Accuracy	I/E= 94.36, S/N= 92.59 F/T= 94.46, J/P= 93.57
	Recall	I/E= 93.76, S/N= 96.43 F/T= 95.30, J/P= 92.54
	Precision	I/E= 98.77, S/N= 92.67 F/T= 97.74, J/P= 89.29
	F1-Score	I/E= 96.35, S/N= 94.77 F/T= 92.59, J/P= 98.38
XGBoost	Accuracy	I/E= 98.33, S/N= 97.31 F/T= 95.66, J/P= 98.73
	Recall	I/E= 97.64, S/N= 99.38 F/T= 89.46, J/P= 94.55
	Precision	I/E= 98.23, S/N= 99.75 F/T= 98.34, J/P= 98.76
	F1-Score	I/E= 96.74, S/N= 95.11 F/T= 98.36, J/P= 97.56

B. Results using Word2Vec Approach

Classifier	Metrics	With Resampling
KNN	Accuracy	I/E= 83.67, S/N= 81.93 F/T= 67.21, J/P= 79.69
	Recall	I/E= 82.79, S/N= 93.54 F/T= 89.43, J/P= 89.36
	Precision	I/E= 63.77, S/N= 41.29 F/T= 65.36, J/P= 79.68
	F1-Score	I/E= 72.11, S/N= 55.36 F/T= 76.57, J/P= 81.22
Naïve Bayes	Accuracy	I/E= 78.37, S/N= 84.21 F/T= 84.14, J/P= 59.65
	Recall	I/E= 5.47, S/N= 21.95 F/T= 70.46, J/P= 94.37
	Precision	I/E= 94.65, S/N= 93.85 F/T= 95.22, J/P= 57.36
	F1-Score	I/E= 13.92, S/N= 85.77 F/T= 74.61, J/P= 75.37
Gradient Descent	Accuracy	I/E= 80.45, S/N= 85.32 F/T= 68.29, J/P= 79.38
	Recall	I/E= 84.54, S/N= 71.88 F/T= 83.75, J/P= 84.36
	Precision	I/E= 95.63, S/N= 92.52 F/T= 94.38, J/P= 73.88
	F1-Score	I/E= 81.72, S/N= 82.47 F/T= 82.64, J/P= 85.69
Random Forest	Accuracy	I/E= 94.26, S/N= 94.76 F/T= 83.11, J/P= 89.51
	Recall	I/E= 91.35, S/N= 91.65 F/T= 72.88, J/P= 86.42
	Precision	I/E= 95.82, S/N= 96.36 F/T= 92.74, J/P= 95.30
	F1-Score	I/E= 95.73, S/N= 94.67 F/T= 86.29, J/P= 89.77
SVM (Linear)	Accuracy	I/E= 93.57, S/N= 94.85 F/T= 89.76, J/P= 90.44
	Recall	I/E= 89.31, S/N= 92.58 F/T= 86.66, J/P= 89.47
	Precision	I/E= 88.34, S/N= 89.42 F/T= 92.17, J/P= 89.37
	F1-Score	I/E= 88.16, S/N= 61.37 F/T= 89.72, J/P= 95.60
SVM (Polynomial)	Accuracy	I/E= 91.63, S/N= 92.34 F/T= 88.46, J/P= 88.27
	Recall	I/E= 91.69, S/N= 91.32 F/T= 89.59, J/P= 88.63
	Precision	I/E= 89.47, S/N= 86.43 F/T= 89.63, J/P= 90.58
	F1-Score	I/E= 89.39, S/N= 90.60 F/T= 91.26, J/P= 91.74
SVM (RBF)	Accuracy	I/E= 91.34, S/N= 89.67 F/T= 91.29, J/P= 90.83
	Recall	I/E= 89.53, S/N= 92.72 F/T= 91.45, J/P= 88.93
	Precision	I/E= 95.19, S/N= 89.36 F/T= 96.24, J/P= 86.44
	F1-Score	I/E= 93.25, S/N= 91.16 F/T= 89.35, J/P= 95.63
XGBoost	Accuracy	I/E= 94.56, S/N= 95.44 F/T= 92.93, J/P= 96.46
	Recall	I/E= 95.62, S/N= 96.11 F/T= 87.40, J/P= 91.23
	Precision	I/E= 95.83, S/N= 97.97 F/T= 96.72, J/P= 96.60
	F1-Score	I/E= 93.31, S/N= 92.75 F/T= 95.49, J/P= 94.65

C. Results using GloVe Approach

Classifier	Metrics	With Resampling
KNN	Accuracy	I/E= 70.16, S/N= 78.05 F/T= 64.09, J/P= 76. 30
	Recall	I/E= 79.24, S/N= 90.10 F/T= 86.18, J/P= 86.24
	Precision	I/E= 60.32, S/N= 38.15 F/T= 62.27, J/P= 76.21
	F1-Score	I/E= 69.40, S/N= 52.20 F/T= 73.36, J/P= 78.18
Naïve Bayes	Accuracy	I/E= 75.48, S/N= 81.25 F/T= 81.45, J/P= 56.15
	Recall	I/E= 3.56, S/N= 18.30 F/T= 67.54, J/P= 91.12
	Precision	I/E= 91.64, S/N= 90.35 F/T= 92.63, J/P= 55.89
	F1-Score	I/E= 10.72, S/N= 82.40 F/T= 71.72, J/P= 72.06
Gradient Descent	Accuracy	I/E= 77.80, S/N= 82.45 F/T= 65.10, J/P= 76.38
	Recall	I/E= 81.33, S/N= 68.50 F/T= 80.18, J/P= 81.03 //
	Precision	I/E= 92.07, S/N= 89.11 F/T= 91.20, J/P= 70.08
	F1-Score	I/E= 78.14, S/N= 79.23 F/T= 79.10, J/P= 82.16
Random Forest	Accuracy	I/E= 91.21, S/N= 91.36 F/T= 80.92, J/P= 86.24
	Recall	I/E= 88.28, S/N= 88.48 F/T= 69.84, J/P= 83.32
	Precision	I/E= 93.35, S/N= 93.59 F/T= 89.72, J/P= 92.40
	F1-Score	I/E= 92.42, S/N= 91.60 F/T= 83.60, J/P= 86.48
SVM (Linear)	Accuracy	I/E= 90.49, S/N= 91.77 F/T= 86.48, J/P= 87.56
	Recall	I/E= 86.56, S/N= 89.81 F/T= 83.36, J/P= 86.64
	Precision	I/E= 85.63, S/N= 86.93 F/T= 89.24, J/P= 86.72
	F1-Score	I/E= 85.70, S/N= 58.10 F/T= 86.12, J/P= 93.80
SVM (Polynomial)	Accuracy	I/E= 88.04, S/N= 89.19 F/T= 85.80, J/P= 85.91
	Recall	I/E= 88.08, S/N= 88.28 F/T= 86.39, J/P= 85.75
	Precision	I/E= 86.12, S/N= 83.37 F/T= 86.74, J/P= 87.39
	F1-Score	I/E= 86.16, S/N= 87.46 F/T= 88.72, J/P= 88.28
SVM (RBF)	Accuracy	I/E= 88.20, S/N= 86.55 F/T= 88.64, J/P= 87.55
	Recall	I/E= 86.24, S/N= 89.61 F/T= 88.56, J/P= 85.17
	Precision	I/E= 92.28, S/N= 86.72 F/T= 93.48, J/P= 83.38
	F1-Score	I/E= 90.32, S/N= 88.83 F/T= 86.40, J/P= 92.25
XGBoost	Accuracy	I/E= 91.34, S/N= 92.94 F/T= 89.32, J/P= 93.65
	Recall	I/E= 92.40, S/N= 93.05 F/T= 84.24, J/P= 88.44
	Precision	I/E= 92.44, S/N= 94.57 F/T= 93.16, J/P= 93.10
	F1-Score	I/E= 90.48, S/N= 89.25 F/T= 92.08, J/P= 91.95

V. CONCLUSION

VI. REFERENCES

[1] Stevenson, M. Introduction to Neuro-Linguistic Programming. Availableonline:http://www.reenlphomestudy.com/membersonly/iNLP/iNLPManual.pdf (accessed on 8 September 2018).

[2] Brian, C. Metaprograms as a Tool for Critical Thinking in Reading and Writing. In Proceedings of the Second JALT Critical Thinking SIG Forum, Kobe Convention Center, Portopia Kobe, Tokyo, 25–28 October 2013; Available online: http://www.standinginspirit.com/wpcontent/uploads/2013/10/JALT2013-Critical-Thinking-Forum-Handout-Metaprograms-Explanation-Handout.pdf (accessed on 12 September 2018).

[3] Barrick, M.; Mount, M. The Big Five personality dimensions and job performance: A meta-analysis. Pers.Psychol. 1991, 44, 1–26.

[4] X. Teng and Y. Gong, "Research on Application of Machine Learning in Data Mining," IOP Conf. Ser. Mater. Sci. Eng., vol. 392, no. 6, 2018.

[5] Schonfeld, E.: Mining the Thought Stream. Tech Crunch Weblog Article (2009). http://techcrunch.com/2009/02/15/mining-the-thought-stream/

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

[7] 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

[8] 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," NeFurocomputing, 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] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] B. de Raad and B. Mlačić, "Big Five Factor Model, Theory and Structure," Int. Encycl. Soc. Behav. Sci. Second Ed., no. December, pp. 559–566, 2015.

[17] T. L. C. Yoong, N. R. Ngatirin, and Z. Zainol, "Personality prediction based on social media using decision tree algorithm," Pertanika J. Sci. Technol., vol. 25, no. S4, pp. 237–248, 2017.

[18] Darsana, M. The influence of personality and organisational culture on employee performance through organisational citizenship behaviour. Int. J. Manag. 2013, 2, 35–42.

[19] Alwi, H.; Sugono, D.; Adiwirmata, S. Kamus Besar Bahasa Indonesia; Balai Pustaka: Jakarta, Indonesia, 2003.

[20] Hall, C.; Lindzey, G. Theories of Personality, 2nd ed.;Wiley: New York, NY, USA, 1970.

[21] Nguyen, D.; Do˘gruöz, A.S.; Rosé, C.P.; Jong, F.D. Computational sociolinguistics: A survey. Comput.
Linguist. 2016, 42, 537–593. [CrossRef]

[22] Gjurkovic, M.; Snajder, J. Reddit: A gold mine for personality prediction. In Proceedings of the Second Workshop on Computational Modelling of People's Opinions, Personality and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 87–97. Available online: https://peopleswksh.github.io/pdf/PEOPLES12.pdf (accessed on 21 September 2018).

[23] 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.

[24] 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.

[25] Medium, Medium Articles, [Online]. Available: https://medium.com/@purnasaigudikandula/recurrent-neural-networks-and-lstm-explained-7f51c7f6bbb9. [Accessed 1 October 2020].

[26] builtin, "builtin.com," [Online]. Available: https://builtin.com/data-science/recurrent-neural-networks-and-lstm. [Accessed 1 September 2020].

[27] 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.

[28] M. f. w. Crawling. [Online]. Available: https://medium.com/@allisonmorgan/short-essay-on-web-crawling-scraping-8abf1b232b65. [Accessed 10 September 2020].

[29] Hansen, C.; Hansen, R. Constructing personality and social reality through music: Individual differences among fans of punk and heavy metal music. J. Broadcast. Electron. Media 1991, 35, 335–350. [CrossRef]

[30] 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.

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

[32] 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.

[33] Shaver, P.; Brennan, K. Attachment styles and the “Big Five” personality traits: Their connections with each other and with romantic relationship outcomes. Personal. Soc. Psychol. Bull. 1992, 18, 536–545. [CrossRef]

[34] Rentfrow, P.; Gosling, S. The do re mi’s of everyday life: The structure and personality correlates of music preferences. J. Personal. Soc. Psychol. 2003, 84, 1236–1256. [CrossRef]

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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

[44] 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.

[45] 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.

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

[47] Jose, R.; Chooralil, V.S. Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation. In Proceedings of the 2015 International Conference on Control Communication & Computing India (ICCC), Trivandrum, India, 19–21 November 2015; pp. 638–641.

[48] Twitter Apps. Available online: http://www.tweepy.org/ (accessed on 26 February 2018).

[49] Esuli, A.; Sebastiani, F. Sentiwordnet: A High-Coverage Lexical Resource for Opinion Mining; Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR): Pisa, Italy, 2006.

[50] Hogenboom, A.; Van Iterson, P.; Heerschop, B.; Frasincar, F.; Kaymak, U. Determining negation scope and strength in sentiment analysis. In Proceedings of the 2011 IEEE International Conference on

Systems, Man, and Cybernetics, Anchorage, AK, USA, 9–12 October 2011; pp. 2589–2594.

[51] Navigli, R.Word sense disambiguation: A survey. ACM Comput. Surv. 2009, 41, 10. [CrossRef]

[52] Ibrahim, M.; Abdillah, O.; Wicaksono, A.F.; Adriani, M. Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 1348–1353.

[53] Kaggle. Mean Absolute Error. Available online: https://www.kaggle.com/wiki/MeanAbsoluteError (accessed on 26 February 2018).

[54] Rezapour, R.; Wang, L.; Abdar, O.; Diesner, J. Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis. In Proceedings of the 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 30 January–1 February 2017; pp. 93–96.

[55] Ding, X.; Liu, B.; Yu, P.S. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 International Conference onWeb Search and Data Mining, Palo Alto, CA, USA, 11–12 February 2008; pp. 231–240.

[56] Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Linguistic Inquiry and Word Count: Liwc 2001; Mahway: Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2001; Volume 71.

[57] Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6–7 July 2002; Volume 10, pp. 79–86.

[58] Gautam, G.; Yadav, D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 7–9 August 2014; pp. 437–442. 43.

[59] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; pp. 137–142.

[60] Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A maximum entropy approach to natural language processing. Comput. Linguist. 1996, 22, 39–71.

[61]Conover, M.D.; Gonçalves, B.; Ratkiewicz, J.; Flammini, A.; Menczer, F. Predicting the political alignment of twitter users. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 192–199.

[62] Le, H.; Boynton, G.; Mejova, Y.; Shafiq, Z.; Srinivasan, P. Bumps and bruises: Mining presidential campaign announcements on twitter. In Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4–7 July 2017

[63] Developers. Twitter Rest API. Available online: https://dev.twitter.com/rest/public (accessed on 26 February 2018).

[64] Esuli, A.; Sebastiani, F. Sentiwordnet: A High-Coverage Lexical Resource for Opinion Mining; Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR): Pisa, Italy, 2006.

[65] Baccianella, S.; Esuli, A.; Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the International Conference on Language Resources and Evaluation, Valletta, Malta, 17–23 May 2010; pp. 2200–2204.

[66] Khan, F.H.; Bashir, S.; Qamar, U. Tom: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst. 2014, 57, 245–257. [CrossRef]

[67] MTRanslator. Available online: https://github.com/mouuff/mtranslate (accessed on 26 February 2018).

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

[69] 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.

[70] 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.

[71] 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.

[72] Soto, C.J. Big Five personality traits. In The SAGE Encyclopedia of Lifespan Human Development; Borstein, M.H., Arterberry, M.E., Fingerman, K.L., Lansford, J.E., Eds.; SAGE Publications: Thousand Oaks, CA, USA, 2018; pp. 240–241.

[73] 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.

[74] 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.

[75] https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

[76] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

[77] https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

[78]https://www.capitalone.com/tech/machine learning/understanding-tf-idf/

[79] https://monkeylearn.com/blog/what-is-tf-idf/

[80] https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3

[81] 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.

[82] 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.

[83] https://nlp.stanford.edu/projects/glove/

[84] https://medium.com/analytics-vidhya/word-vectorization-using-glove-76919685ee0b