IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. III (July – Aug. 2015), PP 39-47
www.iosrjournals.org

Personality prediction of Twitter users with Logistic Regression Classifier learned using Stochastic Gradient Descent

Kanupriya Sharma[1], Amanpreet Kaur[2]
[1](Computer Science Department, Chandigarh University, India)
[2](Computer Science Department, Chandigarh University, India)

Abstract: *Twitter is a popular social media platform with millions of users. The tweets shared by these users have recently attracted the attention of researchers from diverse fields. In this research, we focus primarily on predicting user's personality from the analysis of tweets shared by the user. An associative study of different research works done in personality prediction shows that different techniques have been used to predict a user's personality from tweets but there are certain shortcomings which still need to be addressed. In the introduction section we have given the gist of personality prediction with Twitter research. In the next section we provide the importance and framework of predicting a user's personality with Twitter. In the consecutive section, literature survey, we have discussed briefly about the different research ideas in context to personality prediction with a side by side overview of various resources, tools and machine learning algorithms that were used alongside their potential and limitations, followed by research gap, which provides different fields of possibilities to which can be rectified to improve the efficiency of predicting personality with machine learning algorithms. Finally, a new approach is proposed to predict personality with new insights to predict personality on crucial factors such as scalability and counter measures to improve the research based on previous work by using a Logistic Regression Classifier with parameter regularization using stochastic gradient descent.*

Keywords: *Classification, Logistic Regression, Machine Learning Algorithms, Personality Prediction, Stochastic Gradient Descent, Twitter*

I. Introduction

The advent of the internet into the world gave an endless variety of ways to its users to indulge and savor themselves with the rich pool of knowledge and entertainment. The way social platforms have burgeoned in the recent years has provided an opportunity to study and harvest enormous data which is being produced rapidly at a continuous rate per second of time. Millions of users create profiles about themselves on social media platforms such as Twitter and use the services to connect with their friends and relatives all around the world. At Twitter, 288 million active users per month express themselves with short informal text messages called tweets which amounts to 500 million tweets in a single day. [1] It can be easily figured the amount of data generated by such volume of users at a daily basis. Thus, such factors have triggered research in opinion mining, sentiment analysis and predicting various aspects of real world using social media. Since, knowing what the world actually thinks provides solutions and insights in shaping the fate of economies, business ventures of enterprises and debatable issues of the world. [2] Moreover, the likelihood of a movie to perform and even the future of a commercial product to succeed in market have been quantified by predictive analysis of tweets. [3] Sentiment analysis using social media has been establishing itself as an explicit field of study under Natural Language Processing and it's relation with psychological sciences has been said to bound firmly with insights from the outcome of various researches done till date. [4] The relationship between personality and the factors on which the personality of an individual affects such as job satisfaction, success in personal and professional aspects of life is shown to be beneficial in making predictions using machine learning techniques for an automated process of integrating different fields together. [5][6][7][8] However, the tweets are basically informal and unstructured text therefore, they bring along certain constraints associated with text analysis. Significant use of slang, emoticons, abbreviations and short term words pose a greater difficulty for standardizing the entire process. All these issues have been widely faced and accepted with grace in research work done till date. Not only these issues have held back in generalizing the entire process of personality prediction but have also adulterated with the results and insights of different research attempts made to achieve adequate efficiency. For instance, various approaches have been made to predict personality from social media platforms using different machine learning techniques such as Zero R, Gaussian, Naïve Bayes and a few more. These approaches have helped to push forward the research of predicting personality from tweets but still lack in certain areas due to earlier mentioned constraints. [9][10][11][12] However, in the next sections we will discuss in depth about the various approaches that have been followed for predicting personality from Twitter and their different levels of efficiency as per the potential of resources used and the factors undertaken to perform analysis on tweets text.

II. Anatomy Of The Research

The above section provides a general overview of the entire text however, the aim of this section is to provide an elaborated view of the entire framework of personality prediction which has been divided into the following subsections.

1. What is personality prediction from social media platforms such as Twitter?

Predicting ones personality is not a recent venture and different approaches have existed so far already in context to it. The famous Big Five Personality Inventory standardizes the personality of an individual into five categories i.e. Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. [100] The authors of [101] conducted a research on 279 subjects by using the Big Five Inventory and by collecting different statistics from their Facebook profiles and were able to construct a model which predicted personality of the subjects to within 11% of their actual values. Later, the authors of [9] based on insights of [101] conducted a research by collecting 2000 recent tweets of 279 subjects and then by linguistic analysis of the tweets text they were able to correlate the results with Big Five inventory and achieved 11% to 18% of their actual values. Both these researches laid the foundation of predicting personality traits from social media such as Facebook and Twitter. The authors of [102] further extended the relationship between personality prediction and Twitter by conducting a research with 142 participants which resulted into realization of new associations between linguistic cues and personality.

2. Why personality prediction is a significant area of research and its scope of applications in the real world?

There exists a vast array of real world implications that have been and can be made using insights from personality prediction of the global audience that is available with Twitter. Previous researches have suggested different areas which are directly affected with a user's personality. These include marketing, user interface designs and recommendation systems. Users can be viewed customized web pages, advertisements and products which suits their personalities. These applications may as well be scaled to customized search engines and user experience with different matrimonial and online dating websites. [9] [10] Also, in [11] authors have realized darker traits of a user's personality which can help in understanding criminal and psychopathic behaviors of different personalities. The authors of [12] have suggested that personality prediction can also aid in understanding collective behaviors which provides a qualitative view to social media text mining such as sentiment analysis and clustering of text. All these applications fit well into the present world and hence make personality prediction an interesting field to pursue.

3. How personality prediction of a Twitter user is done using machine learning?

The entire process of personality prediction of a Twitter user can be divided into following steps starting data collection to creating a classification model using an appropriate machine learning algorithm. Tweets can easily analyzed following the text mining and classification techniques as tweets are also mere textual data.

Data Collection – Twitter provides an API (Application Programming Interface) which facilitates the users to retrieve data, public info and user info. There are two API namely Search API and Stream API where the former provides access to limited recent tweets while the latter provide access of real time messages flow. Twitter4J (a java library) and certain Python scripts also provide easy access to data for tweets collection. After an adequate amount of tweets are collected which usually is a time consuming process, tweets are transferred into a single corpus of text.

Data Preprocessing – Tweets are unstructured, informal short messages and hence contain certain parts of text which are not necessary for analysis. These parts are nothing but emoticons, slang, short abbreviations, stop words and special characters. All these parts are removed from text after text tokenization. The remaining text is further processed by either stemming, dimension reduction or a few more techniques. This part is also called as feature extraction.

Model Training – If we consider supervised learning, the learning algorithm should be trained first with labeled examples so as to learn classification for new test unlabeled examples.

Classification/Regression Model – The personality prediction of tweets can be performed using two methods. In classification, the main aim is to classify the text based on certain set of rules into different labels or classes of personalities while in regression, the main aim is to find a value or score of an individual for each of the personality traits. Both these models can be used as per the nature of problem undertaken.

Model Evaluation – After the model has been built it needs to be evaluated using certain techniques such as Accuracy, Recall, Precision, and Root Mean Square Error based upon the nature of problem undertaken.

The above subsection intends to provide a basic framework of steps required to perform personality prediction using tweets. However, selection of machine learning algorithms, their evaluation metrics and textual features selection and linguistic analysis varies from the choice of the researcher to nature of problem undertaken. In the next section we will discuss about the different research literatures done in context with personality prediction using tweets. The next section aims to provide a brief image of the approaches followed by the researchers, different statistics that were used, shortcomings and future scope of their respective works.

III. Literature Survey

J. Golbeck et al. [9] stated in their research that they were the first ones to look at the relationship between personality traits and social profile statistics. They created a Twitter application through which they undertook recent 2000 tweets of fifty subjects. The subjects were presented with the 45 question version of the Big Five Personality Inventory. The collected tweets corpus was processed with the help of two tools, first LIWC (Linguistic Inquiry and Word Count) from which they were able to extract a total of 79 features. Second, MRC Database which yielded 14 language features. They also performed a word by word sentiment analysis with the help of General Inquirer dataset. Then the authors ran a Pearson Correlation analysis between features obtained and personality scores of each user. However, little weightage is given to correlation in this research and is left open for analysis over larger datasets. The authors used regression analysis to predict the score of specific personality features in WEKA. Two algorithms, Gaussian Process and ZeroR were used with 10 fold cross validation iterated 10 times. The authors were able to predict scores from 11% to 18% of their actual values. However, smaller sample size did affect the overall efficiency of the model as mentioned in the paper. Although, authors were able to open a new pathway to predict personality of Twitter users and provide insights about real world applications of their research in the field of marketing and interface design. The results are very likely to improve if larger datasets are taken for consideration.

D. Quercia et al. [10] conducted a research on 335 subjects having both Twitter and Facebook profiles. The data was gathered from a Facebook application called myPersonality and was mapped on the basis of the fact that the same person had profiles on both the platforms. The relationship between personality and different type of Twitters users was analyzed and the personality scores were predicted on the basis of input of three counts namely, following (count if people the user is following), followers (count of people following the user) and listed counts (number of times the user is listed). The authors performed regression analysis for each personality trait with 10 fold cross validation iterated 10 times using M5' rules. The authors also measured Root Mean Square Error between predicted and observed values with a maximum of 0.88. The authors were able to establish important personality relationships among different Twitter users and also provided an insight on accurately predicting personality based on simple profile attributes. This laid a foundation for using this research into good use at marketing, user interface designs and recommender systems. However, discussion were done regarding the extent of revelations that users make on such public profiles. This provides an insight of conducting research with verified users as can be seen now a days on Twitter so that exact predictions can be made efficiently without a doubt of adulterations in the resources.

C. Sumner et al. [11] extended the research of personality prediction beyond the Big Five Personality traits to the anti-social traits of narcissism, Machiavellianism and psychopathy, collectively known as the Dark Triads of personality. Language use and profile attributes were analyzed of 2927 Twitter users to predict dark triads of their personalities. The authors stated that they were the first to study the relationship between Twitter use and dark triads of personality. The study was conducted using custom made Twitter application which collected self-reported ratings on the Short Dark Triad. A maximum of 3200 tweets were collected and were analyzed using LIWC which resulted into a selection of 337 features for machine prediction usage. A comparative study of total six models was conducted by the authors namely, SVM using SMO and a polynomial kernel, Random Forest, J48 algorithm, Naïve Bayes Classifier and two Kaggle models, standard benchmark model and a competition winner model which was held by the authors respectively in context to predict psychopathy and other seven traits together in two different competitions. As stated by the authors, predictive models may not work well for predicting an individual's personality but may work well for predicting the trend of anti-social traits over a subset of population. However, the study resulted into new findings of strong relationships between anti-social traits and language use. The study also showed certain limitations such as selection bias of the subjects and the ever existing issues related to linguistic usage in social media. On the other hand, this study provides endless ports of opportunities for researchers to refine personality prediction from tweets and profile attributes. Rigorous work is required in linguistics used in social media and refinements on individual level predictions is open for study to build robust models of individual's personality prediction. The study also puts forward a greater need of better evaluation metrics for prediction models.

Ana C.E.S Lim et al. [12] have proposed personality traits prediction in text groups and extended the problem of personality prediction into a multi label classification problem. This is a novel approach as an individual may possess more than one personality traits. The authors used the Naïve Bayes Algorithm to analyze tweets and named their model as Bayesian Personality Predictor. Their approach was divided into three steps namely, preprocessing, transformation and classification, where in preprocessing certain attributes were extracted from the tweets and then in the second phase multi label sets were mapped into five single label training sets. Finally, in the third phase semi supervised classification takes place with the help of these training sets and meta-attributes as input to the classifier. The algorithm was evaluated with k-fold cross validation and metrics Accuracy, Recall and Precision. Brazilian TV shows were used as a benchmarking for personality analysis tool. The authors stated to achieve an average of 84% accuracy with their approach. However, more training sets are required to train classification models to achieve a high level of accuracy.

IV. Research Gap

This section aims to provide certain areas which need refinements in their respective manner since fewer approaches have been introduced to predict personality from Twitter. As the need to do so has been already made lucid in these different literatures, but still there exist specific areas which can be pursued and improved by finding new insights using an array of available resources such as

1. Tweets are nothing but textual data which can be analyzed for personality prediction based on linguistic analysis hence different approaches can be followed to improve recognition of linguistic constraints such as slang usage, communal bias, abbreviations and sentiment of tweets.
2. A rigorous research effort is required to make predictive models based on regression or classification algorithms and evaluation methods must be robust enough to complement these approaches.
3. Most of the work done is limited to English language and hence involvement of different language experts can open new pathways to make feature extraction and personality prediction language independent based upon semantic, lexical and grammatical rules.
4. The predictive models must be scalable and dynamic enough to meet the requirements of ever growing data and vast possibilities of considering different viewpoints based on certain groups of users belonging to different ethnicities, geographic regions and recognition of emerging training sets over time.

V. Proposed Methodology

In this section we propose our framework after crucial study of previous work done in this field and insights based on literature survey of respective research papers in context to personality prediction from Twitter, which comprises of a personality prediction model based on Logistic Regression Classifier. Since, personality prediction is a classification problem, our model puts forward a new approach of predicting one's personality using Logistic Regression algorithm with a minimized error function using stochastic gradient descent which addresses certain limitations that came across to researchers previously. The implementation of our model is described in a diagrammatic flowchart of various steps performed in this research. This approach carries the potential to overcome the drawbacks of previous work done. We have also evaluated our model with different evaluation metrics such as precision, recall and accuracy and also compared our results with Naïve Bayes Classifier used prior to predict user's personality by mapping linguistic features of tweets into different personality classes to which a user may or may not adhere.

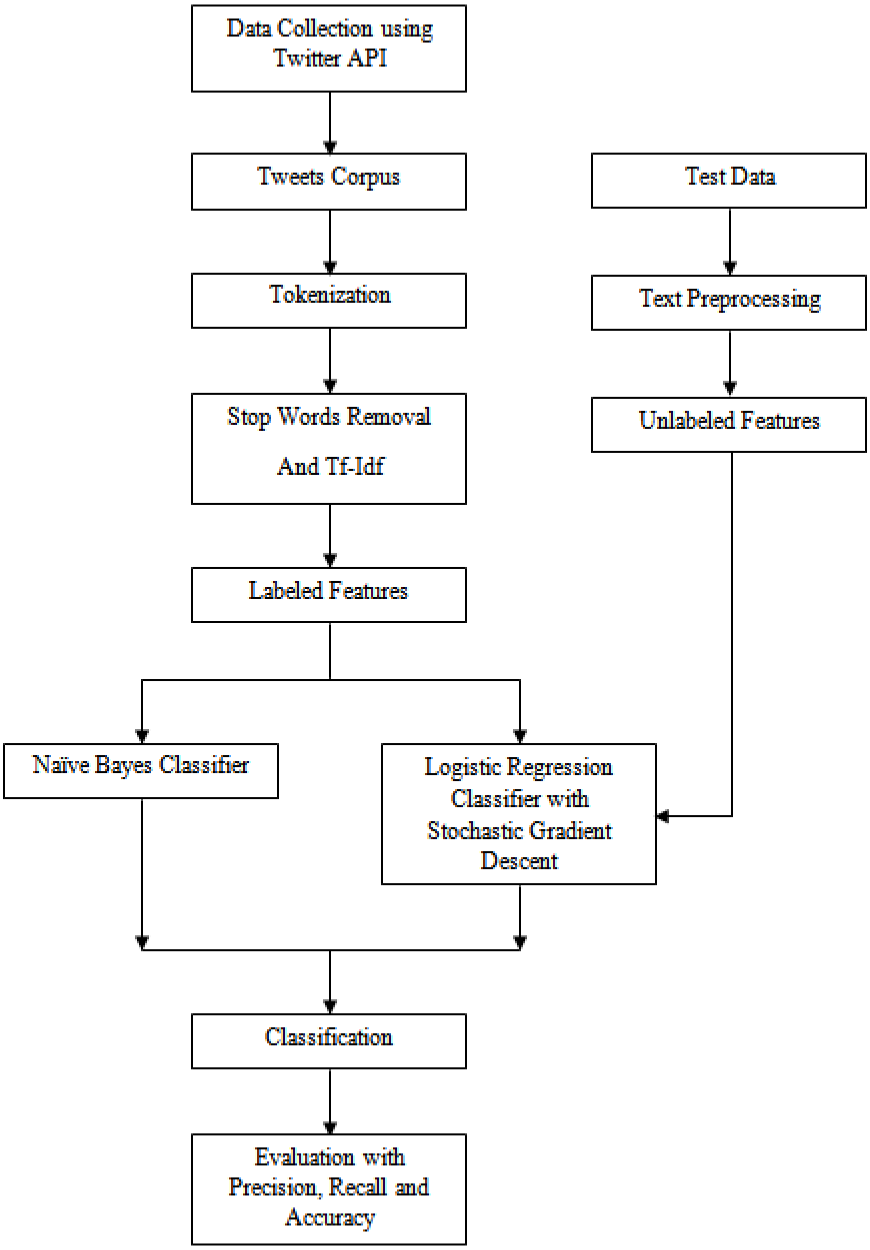

Fig. 1.0 Proposed Framework

The above diagram depicts the working framework of our proposed approach i.e. personality prediction with logistic regression regularized using stochastic gradient descent, which is agreeable to predicting user's personality withstanding variations related to data conditions such dependence between required variables and input features. Our model incorporates the working principle of parameter recognition in context to conditional posterior of data. The working model of our approach provides significant advantages over Naïve Bayes classifier due to dynamic flexibility of feature variation and improved prediction during the course of increase in features fed to the model for classification purposes. [13] The linguistic sentiment of tweets enabled our model to classify user's personalities into positive and negative intent based upon various features related to language use for expressing sentiments with tweets. Our model can be subdivided into four modules namely,

1. Feature Extraction
2. Training
3. Model parameters regularization using Stochastic Gradient Descent
4. Evaluation

Algorithm for Personality Prediction using Logistic Regression Model

Input: t = unstructured tweets corpus, t_test = test tweets corpus
Output: pos_p, neg_p = personality polarity

1. **for** I = 0 to I < length (t)
2. X ← tokenization (t)
3. Y ← stop words removal (X)
4.   **for each** word w in Y
5.   f ← calculate TF-IDF (w)
6.   **end**
7. **end**
8. **for each** labelled feature in f
9. f_label ← correlation (f)
10. Train Logistic Regression Model ← f_label
11. Parameter Regularization using Stochastic Gradient Descent
12. **for** I = 0 to I < length (t_test)
13. Run Logistic Regression Model ← t_test
14. Output ← pos_p or neg_p
15. **end**
16. Validate Logistic Regression Model ← Precision, Recall and Accuracy
17. Compare efficiency of Logistic Regression with Naïve Bayes Classifier

VI. Results and Discussions

Once again, to mention repetitively, that personality prediction from tweets still comes under linguistic analysis and is open to still a vast array of possibilities to improve its course of nature over time as more and more standardization will be introduced in extracting meaningful inferences from text. This approach complements research work related to not only scalable predictions using machine learning and Twitter [14] but also recommendation strategies for users and personalized user experience with advertisements and content suggestion. [15] However, our working implementation of the personality prediction model using NLTK toolkit, Scikit-learn and Python can be briefed with the following tabular data.

Number of features	Accuracy of Naïve Bayes	Accuracy of Proposed Model
100	68.91	70.25
500	71.48	72.24
1000	74.39	76.78
1500	79.21	81.21
10000	83.36	84.45
15000	85.19	87.84
20000	88.57	89.66

Table 1.0 Comparative accuracy of Naïve Bayes and proposed model

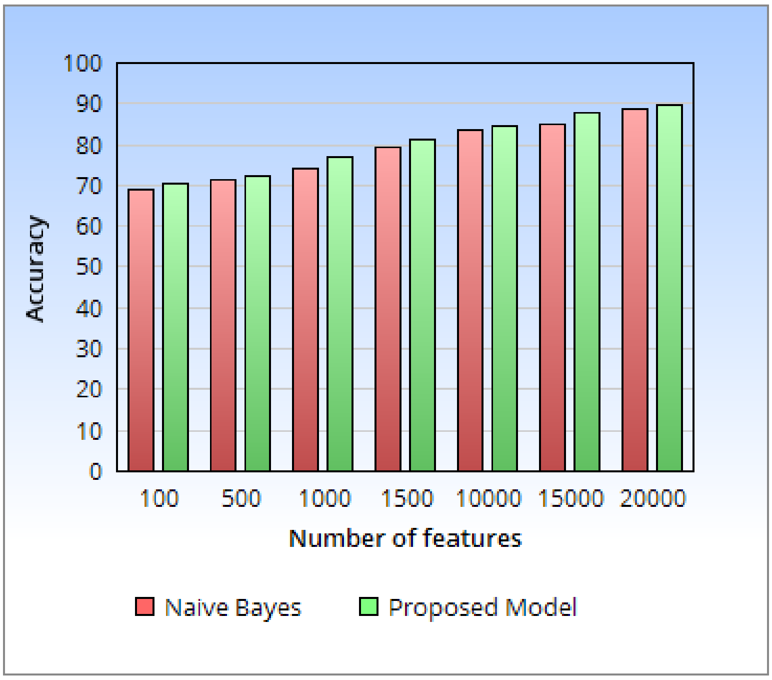

Fig. 2.0 Accuracy of Naïve Bayes and proposed model

Number of features	Precision of Naïve Bayes	Precision of Proposed Model
100	64.91	65.21
500	66.48	67.21
1000	68.39	69.45
1500	71.47	72.21
10000	76.36	77.23
15000	78.89	81.92
20000	82.57	84.34

Table 2.0 Comparative precision of Naïve Bayes and proposed model

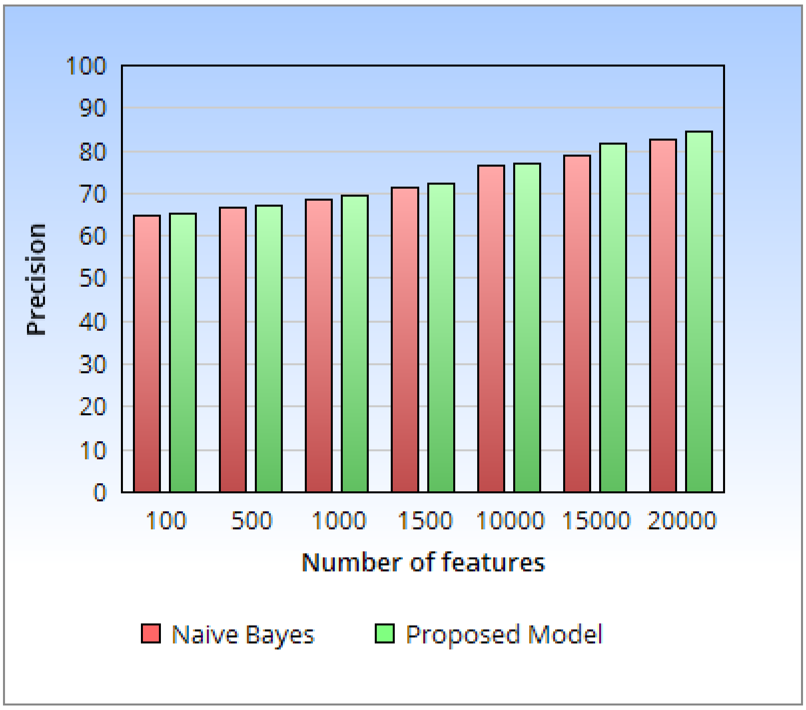

Fig. 3.0 Precision of Naïve Bayes and proposed model

Number of features	Recall of Naïve Bayes	Recall of Proposed Model
100	57.91	59.26
500	62.48	64.23
1000	68.39	70.65
1500	73.47	74.19
10000	77.36	79.67
15000	81.19	82.64
20000	83.57	84.23

Table 3.0 Comparative recall of Naïve Bayes and proposed model

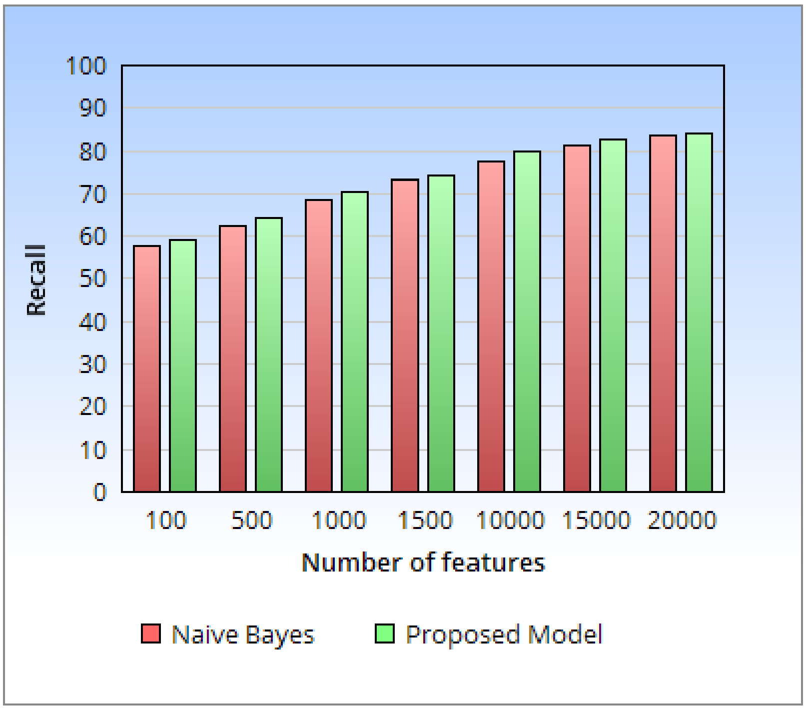

Fig. 4.0 Recall of Naïve Bayes and proposed model

VII. Conclusion

As the results seem promising enough to conclude that our proposed model surely has performance measures which are amenable to the continuing needs for this research to flourish over a span of time. However, as far as the limitations related to the tweets are concerned, the port is open wide enough for the research communities related to text mining and linguistic analysis. Since, tweets are unstructured and informal forms of text, which need rigorous efforts in standardization of a framework to dig deep into harnessing meaningful insights from mere text. The applications related to this field are endless and can only be exploited, unless certain measures are taken to fully understand and document literature related to linguistic analysis and degree of inter rater agreements. At some level researchers are still at the mercy of users for revealing adequate information that can be considered valid for research. There is ample scope for researchers to study the meaningful relationship between language use and intent of the user, from which valid predictions can only be made unless sufficient training data is available to train machine learning algorithms. The aim of this research is to serve as a foundation for reforming personality prediction models by considering several factors based entirely on the scope of literature done till date. Factors like false revelation of information, biased subjects and a variety of possible parameters to choose from are user dependent. However, efforts can be made to refine machine learning algorithms to possibly produce efficient results and adapt to the ever changing face of available data.

References

[1]. Twitter Stats 2015, Twitter Inc.
[2]. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008):1-135
[3]. Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Vol 1 IEEE, 2001
[4]. Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis. "IEEE Intelligent Systems 28.2(2013): 15-21
[5]. Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." CHI'11 extended abstracts on human factors in computing systems. ACM, 2011.
[6]. Barrick, Murray R., and Michael K. Mount. "The Big Five personality dimensions and job performance: A meta-analysis." (1991).
[7]. Judge, Timothy A., et al. "The big five personality traits, general mental ability, and career success across the life span." Personnel psychology 52.3 (1999): 621-652.
[8]. Shaver, Phillip R., and Kelly A. Brennan. "Attachment styles and the" Big Five" personality traits: Their connections with each other and with romantic relationship outcomes." Personality and Social Psychology Bulletin 18.5 (1992): 536-545.
[9]. Golbeck, Jennifer, et al. "Predicting personality from twitter." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.
[10]. Quercia, Daniele, et al. "Our Twitter profiles, our selves: Predicting personality with Twitter." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[11]. Sumner, Chris, et al. "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets." Machine Learning and Applications (ICMLA), 2012 11th International Conference on. Vol. 2. IEEE, 2012.
[12]. Lima, Ana CES, and Leandro N. De Castro. "Multi-label Semi-supervised Classification Applied to Personality Prediction in Tweets." Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI & CBIC), 2013 BRICS Congress on. IEEE, 2013.
[13]. Jordan, A. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems 14 (2002): 841.
[14]. Lin, Jimmy, and Alek Kolcz. "Large-scale machine learning at twitter."Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.
[15]. Goel, Ashish, et al. "Discovering Similar Users on Twitter." 11th Workshop on Mining and Learning with Graphs. 2013.