Predicting Facebook-Users' Personality based on Status and Linguistic Features via Flexible Regression Analysis Techniques

Prantik Howlader*
Cisco Systems, Bangalore
prhowlad@cisco.com

Kuntal Kumar Pal*
Cavium Networks, Bangalore
kpal@caviumnetworks.com

Alfredo Cuzzocrea
University of Trieste, Italy
cuzzocrea@si.dimes.unical.it

S. D. Madhu Kumar
NIT Calicut
madhu@nitc.ac.in

ABSTRACT

Social Media is ubiquitous in everyday life. Which has, in turn, led to social media being the psychological mirror of an individual. The psychological constructs of a social media user are clearly visible from their posts, messages and other activities. But predicting this is a challenging task. This paper explores the use of Support Vector Regression (SVR) and Decision Trees for predicting the Big Five Personality scores, which provides a quantitative measure of the personality traits of users. We attempted to answer three main questions in this work: Is it possible to predict personality of Facebook users based on their statuses? Can the prediction of personality be further improved by adding linguistic features? Which of the popular machine learning techniques like Support Vector Regression (SVR) with linear, polynomial, RBF kernels and decision trees provides better prediction? According to the results of the experiments done to answer the first question, it was found that it is indeed possible to predict personality of users based on their Facebook statuses. The prediction is better than the conventional questionnaire-based methods. For the second question, we have confirmed that the error in prediction decreased when we used some of the features extracted using Linguistic Inquiry and Word Count (LIWC) tool. As for the third question, it has been found that SVR with Radial Basis Function kernel provides better results than Decision tree techniques and Support Vector Regression (SVR) with linear and polynomial kernel in predicting Big Five personality traits. The application of Latent Dirichlet Allocation (LDA) on Facebook status using open dictionary approach is also explored in the paper.

CCS CONCEPTS

• **Computer systems organization → Computing methodologies**; • **Massively parallel and high-performance simulations**;

KEYWORDS

Personality Prediction; Support Vector Regression (SVR); Radial Basis Function (RBF) kernel; Polynomial kernel; Decision Tree; Linguistic Inquiry and WordCount; Big Five Personality Model, Latent Dirichlet Allocation (LDA)

ACM Reference format:
Prantik Howlader, Kuntal Kumar Pal, Alfredo Cuzzocrea, and S. D. Madhu Kumar. 2018. Predicting Facebook-Users' Personality based on Status and Linguistic Features via Flexible Regression Analysis Techniques. In *Proceedings of ACM SAC Conference, Pau,France, April 9-13, 2018 (SAC'18),* 8 pages.
https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Social Media is becoming part and parcel of our life. Most people use at least one social media platform to communicate with others, learn new information or express their views. Social media giants like Facebook and Twitter, today have a quarter of the world's population using it every month. Such pervasive nature of social media has been harnessed to see how mood varies across seasons[9], prediction of stock markets[3]. Emotional words used in social media has been used to find happiness[14]. Activities of users expose their personality traits leaving an enormous record of digital footprint over Social Media. Researchers have shown that the psychological setup of an individual can be traced from their preferences and behavior on the social media[18]. Knowledge of personality helps in predicting an individual's preferences and is used to improve the recommendations by recommendation systems[15]. Personality has been shown to influence the preferences of websites, products and services[13]. Personality is now being extensively used for real-time marketing where knowledge of personality can help in product recommendations[11], in the choice of movies, television shows and books[4] and even in listening to music[22]. There have been studies on automatic personality assessment based on languages used on social media[19]. Researchers have also tried to predict personality in Twitter[20] or even automatically predicted crime based on events extracted from Twitter posts[29]. The Five Factor Model of personality is the most commonly accepted model which claimed to represent the basic structure of human personality traits[8]. It consists of five personality traits called The Big Five (BIG5)[6] personality traits. They are :

- Openness: People who score high in openness are creative, imaginative and politically liberal. They appreciate changes and new ideas.

*Both the authors contributed equally

SAC'18, April 9-13, 2018, Pau,France

ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/xx.xxx/xxx_x

- Conscientiousness: People scoring high on Conscientiousness are well organized, reliable and consistent. They are well-planned and pursue long-term goals.
- Extroversion: People with extroversion traits like to express their emotions and be in the company of the others. They are socially active, friendly and outgoing. They like to be the center of attraction and make friends very easily. They can be characterized as energetic and talkative.
- Agreeableness: People having high agreeableness try to maintain strong social relations. They are compassionate and co-operative. They usually trust others.
- Neuroticism: People having high Neuroticism are usually moody and experience emotions like anxiety, anger and guilt. They are likely to suffer from stress, depression and nervousness. On the contrary, people scoring low on Neuroticism are usually self-confident and calm.

Various analysts are continuing their studies using these widely accepted BIG5 model. We are also using this model as the basis of our experiments and attempted to predict these traits for each user.

The rest of the paper is outlined as follows. In section 2, we present the related works. section 3 deals with the methodology adopted for our work. In section 4 we give an outline of our approach, the dataset used, machine learning techniques applied, data cleaning and normalization used. Experimental results have been provided in section 5 leading to the conclusion in section 6.

2 RELATED WORKS

There are research works that have investigated the use of user personality to augment recommender systems and Human-Computer Interface. Hu et al.[10] and Fernandez et al.[7] used personality of users along with collaborative filtering frameworks to improve recommender systems. This is useful in the scenario where a new user is added to a system and thus does not have much historical data for the recommender systems to provide proper recommendations. Oliveira et al.[17] showed that relationship exists between the personality of users and their satisfaction using mobile phone services. User personality can be leveraged to make adaptive and personalized systems that improves user experience. Lee et al.[16] showed that when synthesized voice personality matches the personality of the user, it affects the user in a positive way. These help to improve the virtual reality systems and human-computer interface.

Researchers have shown that there is a correlation between users' personality and each of the features of their Facebook profiles[1]. Bachrach et al. have used multivariate regression techniques to predict personality traits of an individual as per Big Five model based on Facebook profile properties like density and size of their friendship network, number of events they attended, number of photos they uploaded, number of times a user has been tagged in the photos and others. These researchers achieved the best accuracy for Extroversion and Neuroticism. Agreeableness recorded the lowest accuracy with Openness and Conscientiousness in between.

R. Wald et al.[28] successfully applied machine learning and data mining techniques like LinR, REPTree and DTable on Facebook profiles based on extracted 31 demographic and 80 text-based attributes. These researchers could achieve nearly 75% accuracy in predicting top 10% most Open individuals as per Big Five personality traits. Across all traits, they predicted top 10% users with 34.5% accuracy.

With the growing popularity of Twitter, researchers have started analyzing Twitter users as well. A group analyzed the relationship between different types of Twitter users and their personalities[20]. They found that both the popular and influential users achieved a low value of Neuroticism. The users who are popular are high in Openness while influential users are high in Conscientiousness. They have also accurately predicted an active user's personality with a mean squared error of below 0.88 based on following, followers and listed counts.

Sumner et al.[26] demonstrates links between dark triad personality traits namely narcissism, Machiavellians and psychopathy and Twitter usage. They used a number of machine learning methods for prediction of these dark triads in users.

Schwartz et al.[23] showed correlations between the words used by users in social media and personality traits like extroverts have a high propensity to use social words, unlike introverts who use more words associated with solitary activities.

Social media languages have been used to predict personality across various domains like Facebook, Twitter and Youtube[6]. The results suggest that demographic features such as age and gender have a high correlation with personality scores in all domains. The correlation between gender and agreeableness is positive on Facebook but negative on Twitter and Youtube. There is a positive correlation with word count for agreeableness personality score in Facebook and Twitter, but negative in Youtube.

Farnadi et al.[6] has used closed vocabulary approach, which extracts linguistic features from a text using a dictionary. Schwartz et al.[23] on the other hand, has used open-vocabulary approach, which extracts linguistic features as present in the text without any pre-defined dictionary. For open vocabulary, the approach provides more features than can be given by a priori lexicon used in closed vocabulary approach[23].

So, BIG5 personality prediction has become a significant topic for the researchers over the years. But we didn't find much research on how the performance of prediction of personality by using various regression methods varies with change in the parameter α of LDA (open vocabulary approach). The research that has worked closest to this aspect is by Farnadi et al.[6] but he focuses on the performance of regression algorithms in closed vocabulary approach which extracts linguistic features from a text using a dictionary. Schwartz et al.[23] on the other hand, has used open-vocabulary approach, which extracts linguistic features as present in the text without any pre-defined dictionary. But has fixed $\alpha = 0.30$. Moreover, they have applied only SVM with linear kernel and linear regression for predicting the personality. They have pointed out that open vocabulary approach gives much more features than closed vocabulary approach. In this paper, we have applied various regression techniques and have done a comparative study of these techniques over a large number of users (around fifty thousand), and also shown how the prediction performance of these techniques vary with change in α.

The traditional methods to predict the personality traits of human beings is to go through a set of questionnaires, diligently and honestly answered by the individuals. Based on the scores received

by those people, the personality traits are calculated. This method is dependent totally on the individual's awareness, honesty and willingness to answer the specific questions. It would be easier and helpful if we can predict their personality traits without going through the laborious questionnaire process. This led us to set the following goals for our research. The goals of our study are mainly driven by following three questions :

- Can we predict the personality traits of an individual using their Facebook statuses without manual intervention?
- Can we improve the predictions using some other factors along with their Facebook statuses?
- Which of the machine learning techniques is best for getting the predictions from Facebook status and other features?

3 METHODOLOGY ADOPTED

To achieve our goals mentioned in section 2, we have used machine learning techniques: Support Vector Regression with Linear, Polynomial, Radial Basis Function (RBF) kernels and Decision Trees.

For our first goal: predicting personality traits from individual's Facebook statuses, we have used Latent Dirichlet Allocation (LDA) on the status to get the set of words forming topics. Here, these topics have been used as features for each of the machine learning techniques mentioned above. Then we have compared the results achieved with that of a traditional questionnaire-based method to see, whether it is even possible to predict human personality without any manual interventions.

Next, we explored the possibility of improvement of the prediction by adding new features along with Facebook statuses. We have considered a number of features generated by Linguistic Inquiry and Word Count (LIWC) tool like word-count, values of analytic, authentic, clout characteristics, number of pronouns, adverbs, verbs and others. Based on the results obtained, we have analyzed whether we can succeed in our second goal: improving the accuracy of prediction of personality.

Finally, from the results achieved by the two experiments, we have chosen which of the machine learning techniques provides the best result for personality prediction from Facebook statuses and other linguistic features.

4 EXPERIMENT

In this section, the steps involved in our experiment have been described. Along with it, there is a brief description of the dataset used, a background about some of the machine learning approaches applied and pre-processing used to produce the final dataset.

4.1 MyPersonality: Facebook Dataset

David Stillwell and Michal Kosinski have collected data from over 4 million users who participated in taking a real psychometric test including a standard Five-Factor Model questionnaire. They allowed recording their profile through a popular Facebook application MyPersonality[12]. This dataset provides a huge variety of data including psychological profile, demographic data, Facebook likes, status updates, photos, activities, social networks and even last FM music listening data. For our work, we have used 3 sub-datasets namely BIG5 Personality Scores, Facebook status updates and LIWC tags for status updates aggregated on user level.

4.2 Machine Learning Techniques

4.2.1 Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation is a generative model that uses a huge collection of structured text (corpus)[2] and generates underlying (hidden) topics from the text along with the importance of each topic as well. Each of the topics is further characterized by words. The topic selection through LDA is controlled by a parameter α which represents document-topic density. For our experiments, we have varied this parameter to incorporate randomness into the experiments.

4.2.2 Support Vector Regression (SVR) . Support Vector Regression is a widely used regression technique which is based on Support Vector Machine[5, 24]. The goal is to find the separating boundary which best fits the training data. An error of no greater than ϵ is allowed and hence all the training points lie beyond ϵ distance from the separating boundary. Various kernel functions can be combined with SVR to incorporate non-linearity. We have used linear kernel where the separating boundary takes the form of a line. The polynomial kernel represents a decision boundary which incorporates the non-linearity based on degree n. We have also used radial basis function (RBF) kernel where the non-linearity is controlled by the degree n and kernel coefficient γ.

4.2.3 Decision Tree. Decision Tree[25] is a widely applied machine learning approach that can be used both for classification and regression problems. Decision tree for regression is formed with decision nodes and leaf nodes by breaking down the dataset into smaller subsets and taking a decision at the decision nodes. Here we have applied decision tree for regression guided by a parameter h, which is the depth of the decision allowed.

4.2.4 Mean Squared Error (MSE) . The Mean Squared Error (MSE) technique is a traditional method for evaluating the performance of predictions done by regression models. In our experiments we have used this as the evaluator for each of the predictions. Mathematically it can be represented by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2 \tag{1}$$

where N is the sample size, y_i is the actual value and $\bar{y}_i$ represents the predicted value. The value of the Mean Squared Error (MSE) ranges from 0 to ∞, where 0 specifies no error. The lower the value, the better is the performance of the evaluator.

4.2.5 Linguistic Inquiry and Word Count tool (LIWC) . Linguistic Inquiry and Word Count tool (LIWC) is a text analysis tool that is widely used in psychological studies[27]. This tool can be used to learn how different people use words in everyday language. It can reveal their thoughts, personality, feelings and motivations. Extraction of words that reflect different emotions is also possible through LIWC by calculating their percentages in any given document. It extracts emotions like thinking styles, parts of speech, various social concerns and along with it, a number of features related to psychological processes analyzing hate, swear, anger words and personal concerns.

4.2.6 Term Frequency-Inverse Document Frequency (TF-IDF). This is a numerical, statistical measure (weight) that reflect the

importance of a word in any given document. The technique is very often used in text mining and information retrieval[21].

Term Frequency (TF) measures how frequently the word occurs in the document and is represented as

$$TF = \frac{N(t)_d}{n_d} \tag{2}$$

where $N(t)_d$ is the number of times term t appears in document d and n_d is the total number of terms in document d.

Inverse document Frequency (IDF) measures the importance of a term and is represented by

$$IDF = log_e(\frac{d}{n(d)_t}) \tag{3}$$

where d is the total number of documents and $n(d)_t$ is the number of documents with term t in it.

4.3 Our Approach

To achieve all the three goals mentioned earlier, we have aggregated the Facebook statuses for each user into a single status for that user. Then we have cleaned and cut down the dataset as a part of our pre-processing step. To clean the content of each of the statuses, we performed stop-words removal and finally, the bag of words and dictionary is created as well.

In our first experiment, that is meant to achieve our first goal, we applied LDA per user on their aggregated status. This gave us the latent (hidden) topics that have been used as features to predict the BIG5 scores using the common machine learning techniques like SVR (Linear, Polynomial and RBF kernels) and Decision Tree.

Our second experiment directly follows from the first experiment. Here we added some more features in pursuit of achieving better prediction performance. A reduced set of all the LIWC features present in the dataset have been used.

The third and final experiment is derived from the results achieved from both of our first and second experiment where we look for the technique which provides best prediction performance.

For each of the experiments, while finding topics from Facebook statuses, a parameter of LDA (α), have been varied over a range of 0.1 to 0.9 to incorporate randomness. This parameter affects the status-topic and topic-word distribution in LDA. Similarly, the internal parameter of SVR with the polynomial kernel, degree, has been varied over a considerable range to incorporate its effect on the performance. The kernel coefficient γ for SVR with Radial Basis Function has also been varied. This is done to check the effect of γ on regression performance. Likewise in the decision tree, different heights h have been checked. For both the experiment the topics extracted from LDA have been converted to TF-IDF matrix which has been used to predict the BIG5 personality scores.

Finally, we have compared the predicted scores with the scores acquired through Mypersonality questionnaires using MSE.

4.4 Final Dataset Preparation

Initially, we had 153726 users along with their statuses in the dataset. A few statuses have non-English characters and were not possible to apply language-specific manipulations on them. So we decided to exclude those statuses which do not have any English words. We rejected 303 such userid and their statuses as a part of the cleansing process. The remaining 153423 users along with their statuses are moved on to the second phase of cleansing.

We have seen that some of the statuses are short or have too much repetition of same words. It affected the meaningfully different topics creation through LDA and the words under each topic resulted in repetition. Also for the LDA to extract various topics efficiently we needed the status to be of considerable length. So we created the dictionary of the status of each user. Then we have taken only those where the length of the dictionary is more than and equals to 500. Another reason beyond our choice of the value is to get a dataset which has sufficient text data in the statuses for LDA to work effectively.

Not all the users selected, have their corresponding questionnaire based BIG5 scores without which it is not possible to evaluate the prediction. So after all cleansing process, in the final dataset, we have 48701 users with valid and complete statuses and BIG5 scores.

Table 1 gives more insight on the statistics of the reduced dataset. From the table, we can say that we have a standard data where we can see that the mean of questionnaire-based BIG5 scores across all traits is nearly same, except for Neuroticism which indicates that selected users have less amount of the trait. The standard deviation is provided as well for all the traits, with Openness showing less variation.

Table 1: Statistics of BIG5 scores in Reduced dataset

	Mean	Standard Deviation
Openness	3.912669	0.6592319
Aggreableness	3.562032	0.7080322
Extrovertion	3.585833	0.8069222
Neuroticism	2.772607	0.8098645
Conscientiousness	3.453712	0.7301083

4.5 LIWC feature choices and Normalization

Along with the dataset, there are 93 pre-extracted features present. These have been already extracted using the LIWC tool. To reduce the number of LIWC features we have removed those features which have a high mean absolute correlation. To do this, we normalized each of the features calculating the z-score[30]. Then the correlation among the features have been calculated and after which the features have been selected keeping a threshold of 0.6. This reduced our LIWC feature set to 54. The threshold value is chosen such that it retains a good amount of data and have a low overall correlation.

5 EXPERIMENTAL RESULTS

In this section we present the results of the three experiments sequentially:

5.1 Experiment 1

To achieve our first goal we were to predict the five personality traits from merely their Facebook statuses. The experimental setup has been illustrated below.

After the cleaning of data, LDA is applied on the Facebook statuses. Since LDA produces latent topics and the words constituting each topic mainly based on a parameter α, we have experimented

Table 2: MSE for Openness in Expr #1

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.6628	0.6632	0.6620	0.6614	0.6599
SVR-Poly	0.6597	0.6598	0.6598	0.6601	0.6600
SVR-RBF	0.6403	0.6405	0.6401	0.6399	0.6396
Decision Tree	0.6520	0.6521	0.6524	0.6521	0.6523

Table 3: MSE for Aggreableness in Expr #1

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.7196	0.7192	0.7194	0.7202	0.7187
SVR-Poly	0.7089	0.7089	0.7089	0.7090	0.7090
SVR-RBF	0.6944	0.6943	0.6943	0.6941	0.6940
Decision Tree	0.6992	0.6992	0.6999	0.7000	0.7002

Table 4: MSE for Extrovertion in Expr #1

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.8066	0.8067	0.8062	0.8076	0.8053
SVR-Poly	0.8092	0.8092	0.8093	0.8093	0.8093
SVR-RBF	0.7815	0.7819	0.7815	0.7811	0.7810
Decision Tree	0.7967	0.7966	0.7965	0.7971	0.7970

with a number of values of the parameter. This has been mainly done so that the parameter does not become the driving factor in our experiments. Then the TF-IDF matrix is created from the topics extracted from the statuses. Finally, the MSE is calculated for each machine learning techniques using a 5-fold cross-validation.

Table 2 to Table 6 show the MSE between the personality predicted using only topics and that predicted using questionnaire-based approach. They provide the MSE for each of the BIG5 personality traits Openness, Agreeableness, Extroversion, Neuroticism and Conscientiousness.

From the data, it can be seen that the MSE achieved for openness ranges from 0.63-0.67 in all the four regression techniques. The results of Agreeableness is relatively higher in the range 0.69-0.72. Extroversion and Neuroticism record MSE in almost the same range 0.78-0.81 and 0.79-0.83 respectively. For Conscientiousness, we got the MSE in a very narrow range 0.71-0.73.

The parameter α of LDA is kept in range from 0.1 to 0.9. It can also be seen that the parameter α does not have much of an influence on the MSE achieved in this approach. We have experimented with values above 0.9 and below 0.1 but it didn't affect the MSE much and hence have not been shown in the results.

5.2 Experiment 2

To achieve our second goal we needed additional features apart from topics extracted from Facebook statuses using LDA. So we added a reduced set of LIWC features as mentioned earlier. The experimental setup is same as that of experiment 2. Only addition is the linguistic features added as inputs to regressors. Here also we performed a 5-fold cross-validation.

Table 5: MSE for Neuroticism in Expr #1

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.8238	0.8234	0.8226	0.8224	0.8221
SVR-Poly	0.8100	0.8101	0.8101	0.8101	0.8101
SVR-RBF	0.7948	0.7945	0.7944	0.7942	0.7937
Decision Tree	0.8063	0.8058	0.8058	0.8069	0.8072

Table 6: MSE for Conscientiousness in Expr #1

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.7311	0.7302	0.7295	0.7298	0.7291
SVR-Poly	0.7316	0.7316	0.7317	0.7318	0.7319
SVR-RBF	0.7037	0.7033	0.7033	0.7032	0.7030
Decision Tree	0.7163	0.7160	0.7173	0.7167	0.7169

Table 7: MSE Change for Openness in Expr #2

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.0144	0.0147	0.0137	0.0113	0.0100
SVR-Poly	0.0233	0.0224	0.0210	0.0189	0.0163
SVR-RBF	0.0146	0.0211	0.0073	0.0142	0.0138
Decision Tree	0.0093	0.0097	0.0098	0.0095	0.0098

Table 8: MSE Change for Aggreableness in Expr #2

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.0213	0.0137	0.0189	0.0189	0.0198
SVR-Poly	0.0194	0.0185	0.0171	0.0148	0.0120
SVR-RBF	0.0078	0.0093	0.0091	0.0086	0.0084
Decision Tree	0.0120	0.0121	0.0128	0.0128	0.0128

Table 7 to Table 11 record the change in the MSE achieved when we have only topics as features (case 1) and when we have topics along with LIWC features (case 2). The positive value indicates that MSE in case 1 is higher than that in case 2. This, in turn, specifies that the MSE has been reduced by adding more features and thus improving the performance of the regressor.

Based on the result, it can be inferred that there is a minute improvement for decision tree regressor on all 5 personality traits. Also, SVR with RBF kernel shows less improvements for Conscientiousness and Extroversion with the addition of LIWC features. The improvements have been maximum for SVR-Poly for all five personality traits across all values of α for LDA.

We have represented this in the bar graph for better visualization of the changes. Each bar represents the average MSE across all α values for each personality trait for each regressor. In each of the figures 1-5, the first blue bar represents the MSE achieved without the LIWC features and the next green bar represents the MSE for Facebook status with LIWC features. The change is visible distinctly from the graphs. In each of the graphs, the first bar graph is higher than the second bar indicating the improvement of the performance.

Table 9: MSE Change for Extrovertion in Expr #2

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.0154	0.0145	0.0136	0.0130	0.0096
SVR-Poly	0.0260	0.0246	0.0224	0.0191	0.0155
SVR-RBF	0.0137	0.0026	0.0036	0.0133	0.0132
Decision Tree	0.0099	0.0093	0.0091	0.0097	0.0097

Table 10: MSE Change for Neuroticism in Expr #2

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.0104	0.0178	0.0129	0.0102	0.0110
SVR-Poly	0.0194	0.0185	0.0170	0.0147	0.0123
SVR-RBF	0.0105	0.0040	0.0037	0.0101	0.0096
Decision Tree	0.0074	0.0068	0.0072	0.0078	0.0082

Table 11: MSE Change for Conscientiousness in Expr #2

α	0.1	0.3	0.5	0.7	0.9
SVR-Linear	0.0113	0.0121	0.0073	0.0115	0.0166
SVR-Poly	0.0310	0.0282	0.0287	0.0263	0.0233
SVR-RBF	0.0042	0.0062	0.0052	0.0018	0.0017
Decision Tree	0.0087	0.0089	0.0093	0.0084	0.0095

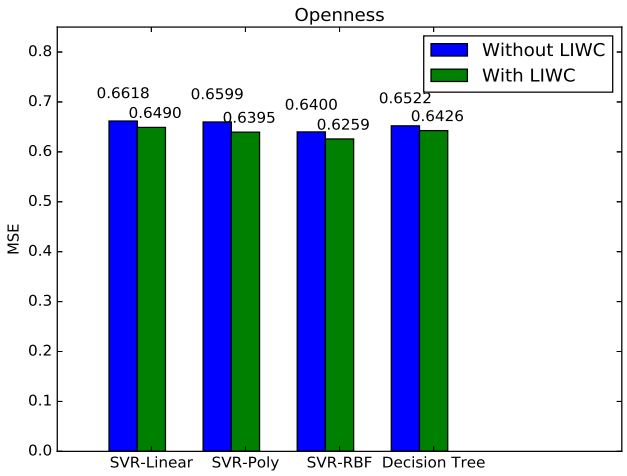

Figure 1: Openness with and without LIWC

5.3 Experiment 3

In our final experiment, our goal was to find which of the techniques applied, provides the best performance. This can be directly shown from the results which we have achieved in experiment 1 and 2.

From the experiment 1, we have the MSEs achieved for each of the four regression techniques - SVR with linear kernel, SVR with the polynomial kernel, SVR with RBF kernel and Decision Tree, for each of the five personality traits with only topics from status as feature. From Table 2 to Table 6 it can be seen that the minimum

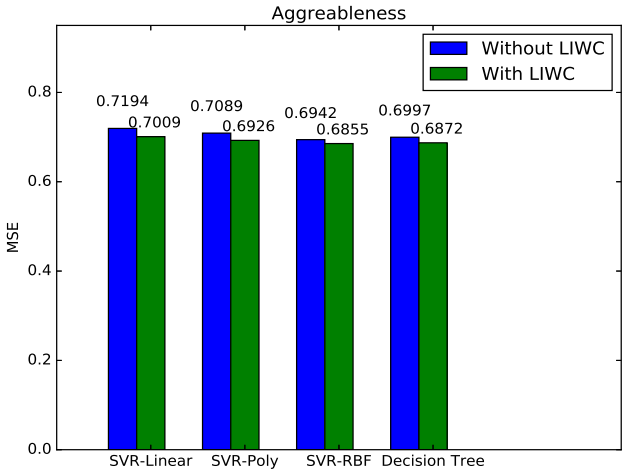

Figure 2: Aggreableness with and without LIWC

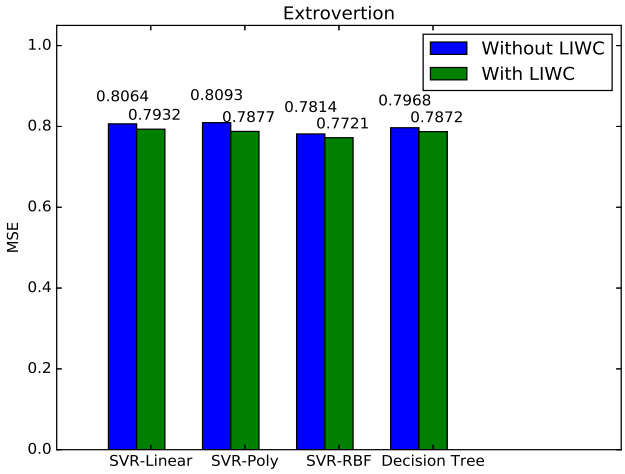

Figure 3: Extrovertion with and without LIWC

MSE is achieved for SVR with RBF kernel for each of values of α, across all the five personality traits. Hence the best performance is achieved using SVR with RBF kernel.

Since we have already shown in the second experiment that adding more LIWC features improves the performance, we concentrated more on finding which technique works best for the case. This we have illustrated in graphical representation for each of the personality traits separately.

In the figures 6-10, the black dashed dot line, red continuous line, the blue dashed lines and green dotted lines represents the SVR with linear kernel, SVR with the polynomial kernel, SVR with RBF kernel and decision tree techniques respectively. In each of the figures, the SVR with RBF kernel shows minimum MSE. The difference is distinct in Openness and Extroversion traits. For the rest of the traits as well, SVR with RBF kernel turns out to be the best technique for personality prediction. SVR with linear kernel shows the worst performance among the four techniques.

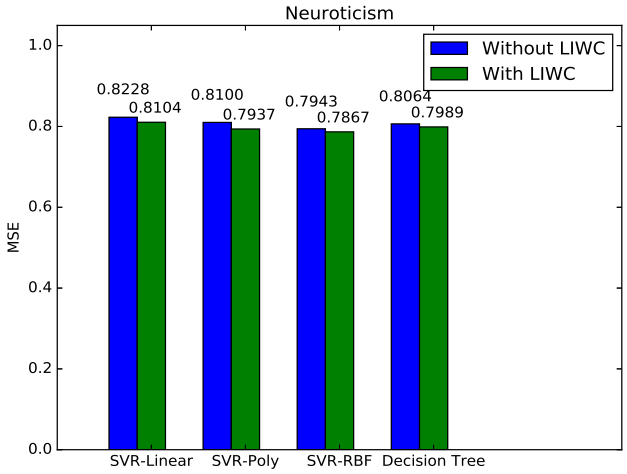

Figure 4: Neuroticism with and without LIWC

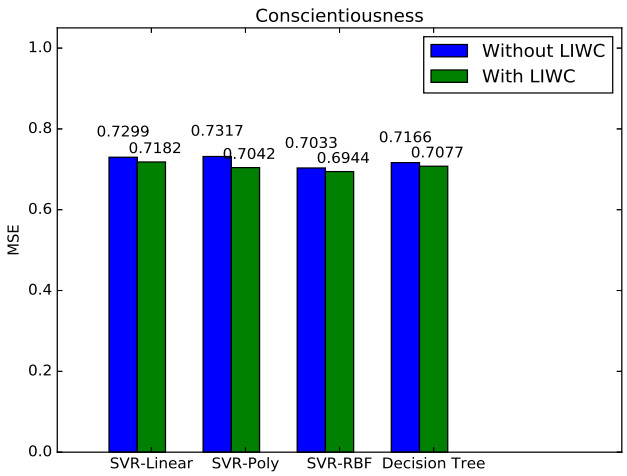

Figure 5: Conscientiousness with and without LIWC

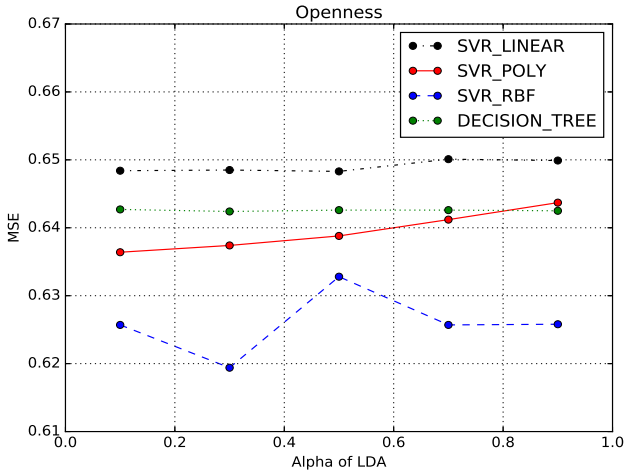

Figure 6: Openness across all 4 techniques

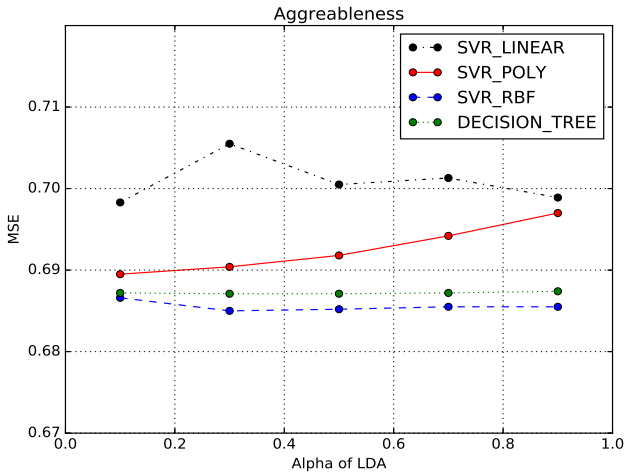

Figure 7: Aggreableness across all 4 techniques

6 CONCLUSIONS

This paper introduces BIG5 score-based prediction on Facebook status of users and their LIWC features using Regression techniques like SVR and Decision Trees. Based on our experimental results it is evident that SVR with RBF kernels has the best performance in predicting the personality of users. Moreover, we have observed that incorporating LIWC features in addition to status, significantly increases the performance of regression model.

In future, we would like to see how the addition of other features like demographic details, IQ scores etc. affect the performance of these regression models.

ACKNOWLEDGMENTS

We would like to thank David Stillwell and Michal Kosinski for providing the MyPersonality data that have been used for our work.

REFERENCES

[1] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. 2012. Personality and patterns of Facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 24–32.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR* abs/1010.3003 (2010).

[4] Iván Cantador, Ignacio Fernández-Tobías, and Alejandro Bellogín. 2013. Relating personality types with user preferences in multiple entertainment domains. In *CEUR Workshop Proceedings*. Shlomo Berkovsky.

[5] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. 1997. Support vector regression machines. *Advances in neural information processing systems* 9 (1997), 155–161.

[6] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. 2016. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction* 26, 2-3 (2016), 109–142.

[7] Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. 2016. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction* 26, 2-3 (2016), 221–255.

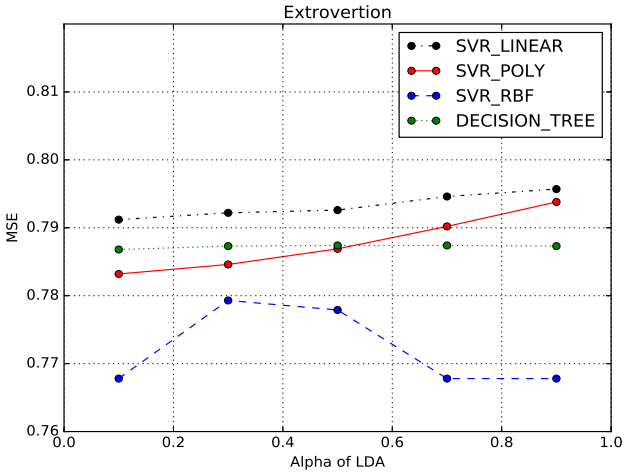

Figure 8: Extrovertion across all 4 techniques

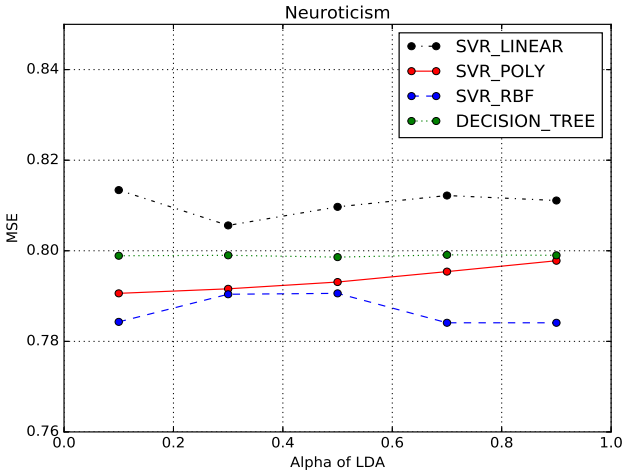

Figure 9: Neuroticism across all 4 techniques

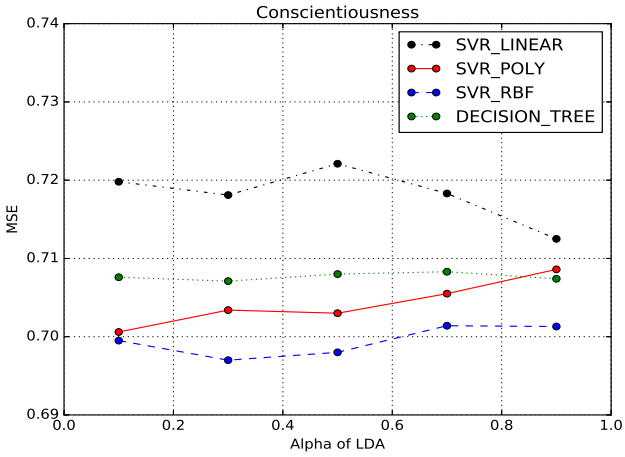

Figure 10: Conscientiousness across all 4 techniques

[8] Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48, 1 (1993), 26.

[9] Scott A. Golder and Michael W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051 (2011), 1878–1881. https://doi.org/10.1126/science.1202775

[10] Rong Hu and Pearl Pu. 2011. Enhancing Collaborative Filtering Systems with Personality Information. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 197–204. https://doi.org/10.1145/2043932.2043969

[11] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. 2014. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning* 95, 3 (2014), 357–380.

[12] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 6 (2015), 543.

[13] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[14] Adam D.I. Kramer. 2010. An Unobtrusive Behavioral Model of "Gross National Happiness". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 287–290. https://doi.org/10.1145/1753326.1753369

[15] Renaud Lambiotte and Michal Kosinski. 2014. Tracking the digital footprints of personality. *Proc. IEEE* 102, 12 (2014), 1934–1939.

[16] Kwan Min Lee and Clifford Nass. 2003. Designing Social Presence of Social Actors in Human Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 289–296. https://doi.org/10.1145/642611.642662

[17] Rodrigo De Oliveira, Mauro Cherubini, and Nuria Oliver. 2013. Influence of Personality on Satisfaction with Mobile Phone Services. *ACM Trans. Comput.-Hum. Interact.* 20, 2, Article 10 (May 2013), 23 pages. https://doi.org/10.1145/2463579.2463581

[18] Daniel J Ozer and Veronica Benet-Martinez. 2006. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.* 57 (2006), 401–421.

[19] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108, 6 (2015), 934.

[20] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 180–185.

[21] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

[22] Peter J Rentfrow and Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* 84, 6 (2003), 1236.

[23] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

[24] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[25] Dan Steinberg and Phillip Colla. 2009. CART: classification and regression trees. *The top ten algorithms in data mining* 9 (2009), 179.

[26] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *Machine learning and applications (icmla), 2012 11th international conference on*, Vol. 2. IEEE, 386–393.

[27] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[28] Randall Wald, Taghi Khoshgoftaar, and Chris Sumner. 2012. Machine prediction of personality from Facebook profiles. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*. IEEE, 109–115.

[29] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 231–238.

[30] Wikipedia. 2017. Standard score. (Sept. 2017). Retrieved September 27, 2017 from https://en.wikipedia.org/wiki/Standard_score