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

Prantik Howlader
Cisco Systems
Bangalore, India
prantikbubun@gmail.com

Kuntal Kumar Pal
Cavium Networks
Bangalore, India
kuntal.octo@gmail.com

Alfredo Cuzzocrea
University of Trieste and ICAR-CNR
Trieste, Italy
alfredo.cuzzocrea@dia.units.it

S.D. Madhu Kumar
National Institute of Technology
Calicut, India
madhu@nitc.ac.in

ABSTRACT

The psychological constructs of a user of social media are clearly visible from his/her posts and other activities. But predicting this is a challenging task. This paper explores the use of *Linear Regression* (LR) and *Support Vector Regression* (SVR) for predicting the *Big Five Personality* scores, which provide a quantitative measure of the personality traits of users. A performance comparison is made about the regression models on topics from Facebook users' statuses and topics from Facebook statuses along with features extracted via using *Linguistic Inquiry and Word Count* (LIWC) tool. Further, we have investigated the effect of number of topics found by *Latent Dirichlet Allocation* (LDA) on the performance of regression models. We found that SVR with Polynomial and *Radial Basis Function* kernel, respectively, provides better results in predicting big five personality traits. We found that the mean squared error of LR increases with the number of topics. But this increase is less in the case when we consider additional LIWC features for regression.

CCS CONCEPTS

• **Information systems → Social networks**;

KEYWORDS

Support Vector Regression; Linguistic Inquiry and WordCount; Big Five Personality Model; Latent Dirichlet Allocation

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 (SAC'18).* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/https://doi.org/10.1145/3167132.3167166

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

ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/https://doi.org/10.1145/3167132.3167166

1 INTRODUCTION

In today's world, social media has become a 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. These activities expose their *personality traits* leaving a pervasive record of digital footprints over Facebook, Twitter and other similar social media. The constant evolution of these social media and the big social data available led various researchers to convert their small-scale questionnaire-based experiments to a large-scale research on the human psychology over social networks and systems, like in Clouds (e.g., [6]). Researchers have shown that the psychological setup of an individual can be traced from their preferences and behavior on the social media [17]. With this knowledge about personality, they have attempted to predict an individual's preferences and tried to improve systems that provide recommendations [14]. People have attempted to enhance product brand services [13], for instance, which are now being extensively used for real-time marketing. Even decision making of an individual can be traced back to their personality which is shown in website preferences [11], as for the case of the choice of movies, TV-shows and books [4], and even in listening music [22]. There have also been studies on automatic personality assessment based on languages used on social media [18]. Researchers have even tried to predict personality in Twitter [20], or even automatically predicted crime based on events extracted from Twitter posts [28].

The *Five Factor Model* (FFM) of personality is the most commonly-accepted model that claimed to represent the basic structure of human personality traits[9]. It consists of five combined personality traits, called *The Big Five* (BIG5) [8]:

- *Openness*: people who score high in openness are imaginative, politically liberal, creative, appreciate changes and new ideas.
- *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 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 this widely-accepted BIG5 model.

On the other hand, the social media is so popular that, for an active user, the amount of data over a small period of time is really huge. These data mirror the personality traits of the user. The motivation of our work is to predict the BIG5 personality traits of users of social media through their Facebook statuses. The goal of our research is to experimentally prove how regression analysis can allow us to discover the BIG5 personality scores of Facebook users via analyzing their Facebook statuses. In more detail, in this paper we finally show the following experimental evidences:

- The comparison between the BIG5 personality scores evaluated using only the Facebook statuses of a user and the BIG5 personality scores computed via using separate questionnaires by *MyPersonality* [12] Facebook application. This would prove that social media personalities of users mirror their true personalities.
- The comparison between the BIG5 scores evaluated using the Facebook statuses of a user along with *Linguistic Inquiry and Word Count* (LIWC) [19] features and the BIG5 scores evaluated using only Facebook statuses of the user. This would prove whether the performance of personality prediction can be increased with the inclusion of more features from the digital footprints of a user.
- How the number of topics found by *Latent Dirichlet Allocation* (LDA) [2] on Facebook status affects the performance of the following regression models: *Linear Regression* (LR) [16], *Support Vector Regressions* (SVR) [23] with linear, polynomial, and *Radial Basis Function* kernel, respectively.

Based on this rich experimental analysis, we finally select the regression technique that provides better results.

The remaining part of the paper is organized as follows. Section 3 illustrates the problem statement of our research. Section 3 describes related work in this area. In Section 5, we provide the whole approach of our experimental campaign, along with the dataset used and a brief background on the methodologies. We then present our experimental results in Section 6, and finally conclude in Section 7.

2 PROBLEM STATEMENT

In this paper, we experimentally address the problem of predicting the personality of Facebook users via computing their BIG5 scores via using SVR and LR analysis, respectively, on topics extracted from their Facebook statuses, by finally achieving a *flexible regression analysis methodology*. Basically, our experimental methodology consists in observing whether a performance increase in predicting the personality scores occurs when we consider topics from Facebook users' statuses as analysis features, by comparing with the features exploited by the popular LIWC tool. In addition to this, we also observe whether the number of topics found by *Latent Dirichlet Allocation* (LDA) [2] affects the performance of our regression models.

3 RELATED WORK

There has been steady research on social media over the years, but very few have specifically focused on personality traits. Researchers have shown that there is a correlation between users' personality and each of the features of their Facebook profiles [1]. They have used *multivariate regression techniques* [5] to predict personality traits of an individual as per BIG5 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 studies and proposals achieved best accuracies for *Extroversion* and *Neuroticism* traits, respectively. By the contrary, *Agreeableness* recorded lowest accuracy with *Openness* and *Conscientiousness* in between.

R. Wald *et al.* successfully apply machine learning and data mining techniques like *LinR*, *REPTree* and *DTable* on Facebook profiles based on extracted 31 demographic and 80 text-based attributes [27]. They achieve nearly 75% accuracy in predicting top 10% of most *Open* individuals as per the BIG5 personality traits. Across all the BIG5 traits, they predict top 10% users with 34.5% accuracy.

With the growing popularity of Twitter, researchers have started to focus on and analyze Twitter users as well. Quercia *et al.* focus on the relationship among different types of Twitter users and their personality [20]. They found that both popular and influential users achieve low values 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.

Social media languages have been used to predict personality across various domains, like Facebook, Twitter and YouTube [8]. 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* in Facebook and Twitter, but negative in YouTube.

Sumner *et al.* demonstrate links between dark-triad personality traits namely *Narcissism*, *Machiavellianism* and *Psychopathy*, and Twitter usage [25]. In particular, they make use of a number of machine learning methods for prediction of these dark-triad traits in Twitter users. Mario Cannataro *et al.* [3] present a probabilistic approach for the modelling of Adaptive Hypermedia Systems.

From the analysis of related work, it directly follows that, at a larger extent, researchers have applied machine learning and data mining methods and algorithms to support users' personality discovery over social networks and systems, with relevant real-life application scenario depicted by emerging social frameworks like Facebook, Twitter and YouTube.

4 EXPERIMENTAL METHODOLOGY AND SETTINGS

4.1 Using MyPersonality to Collect Experimental Facebook Datasets

Thanks to the popular Facebook application *MyPersonality* [12], researchers have collected Facebook data about over 4 million users who participated in taking real *psychometric tests*, including the standard FFM questionnaire, and allowed to record their profiles. The deriving baseline Facebook dataset provides a huge variety of data including psychological profiles, demographic data, likes, status updates, photos, activities, social networks, and even last FM music listening data, thus configuring itself as a "rich" dataset. From this universal dataset, we derived 3 datasets to be used in our experimental campaign, namely: (*i*) *BIG5 Personality Scores*, which has been obtained by applying the BIG5 analysis to the baseline Facebook dataset; (*ii*) *Facebook status updates*, which has been obtained by mining the status updates from the baseline Facebook dataset via inspecting Facebook statuses only; (*iii*) *LIWC status updates*, which has been obtained by mining the status updates aggregated on user level from the baseline Facebook dataset via applying LIWC tool, and considering derived LIWC tags.

4.2 Background on Comparative Approaches

In this Section, we provide a brief background on the comparative methodologies used in our experimental campaign.

4.2.1 Latent Dirichlet Allocation (LDA). LDA is a discrete generative model for a collection of structured text (i.e., corpuses) [2]. It is a three-level Bayesian model in which each document is represented by a finite random mixture of underlying (hidden) topics that, in turn, is characterized by words. In our experiments, we use LDA to extract n topics consisting of m words from the target text. Then, we vary the values of n and m to observe how they affect the performance of the proposed regression model.

4.2.2 Support Vector Regression(SVR). SVR is a widely used regression technique which is based on *Support Vector Machine* [7, 24]. Here, the goal is to find the separating boundary which best fits the training data. An error of no greater than ϵ is allowed and, therefore, all the training points lies beyond a ϵ distance from the separating boundary. Various kernel functions [10] can be combined with SVR to incorporate non-linearity. In our experiments, we use the following kernel functions: (*i*) *linear kernel*, where the separating boundary takes the form of a line; (*ii*) *polynomial kernel*, which represents a decision boundary that incorporates the non-linearity of degree n; (*iii*) *RBF kernel*, where the non-linearity is controlled by the degree n and the kernel coefficient γ.

4.2.3 Mean Squared Error (MSE). In order to perform our experimental analysis, we make use of the *Mean Squared Error* (MSE) [15] model for evaluating the performance of our comparative regression models, namely: LR, SVR with linear, polynomial and RBF kernels, respectively. According to a wide and well-established literature, this model has been extensively used over the years to evaluate performance of regression analysis. Formally, MSE is defined as follows:

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

wherein: (i) N represents the sample size; (*ii*) y_i represents the actual value; (*iii*) $\bar{y}_i$ represents the predicted value. The output value of the MSE ranges from 0 to ∞, where 0 specifies no error. The lower the value, the better is the MSE's performance.

4.2.4 Linguistic Inquiry and Word Count (LIWC) Tool. LIWC tool is a text analysis tool that is widely used in psychological studies [26]. This tool can be used to learn how the words people use in everyday language reveals their thoughts, personality, feelings and motivations. It takes as input a given document and extracts words that reflect different emotions by computing their percentages. It discovers emotions like thinking styles, parts of speech and various social concerns. Also, LIWC can extract a number of features related to psychological processes analyzing hate, swear, anger words and personal concerns, like finding words that refer to occupation (e.g., jobs and majors).

4.2.5 Term Frequency (TF) and Inverse Document Frequency (IDF). *Term Frequency* (TF) and *Inverse Document Frequency* (IDF) are numerical statistical measures (modeled as weights) that reflect the importance of a word in a given document. These techniques are very often used in text mining and information retrieval [21]. TF, which measures how frequently a word occurs in a document, is defined as follows:

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

wherein: (*i*) $N(t)_d$ represents the number of times the term t appears in the document d; (*ii*) n_d represents the total number of terms in document d.

IDF, which measures the importance of a term in a document, is defined as follows:

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

wherein: (*i*) d represents the total number of documents; (*ii*) $n(d)_t$ represents the number of documents that contain the term t.

4.3 Details of Proposed Experimental Methodology

In this Section, we provide the details on how experimental methodology. For all our experiments, we have used aggregated Facebook statuses and LIWC features of each user obtained from *MyPersonality* data [12]. As a pre-processing step, we have cleaned the data in order to eliminate statuses that have all non-ASCII data. Then, stop-words have been removed and bag of words have been created. After this cleaning phase, we have obtained a dataset of 115, 872 users. On this dataset, we have applied the following regression techniques: (*i*) LR, (*ii*) SVR with linear kernel (hereafter referred as L-SVR), (*iii*) SVR with polynomial kernel (hereafter referred as P-SVR), and (*iv*) SVR with RBF kernel (hereafter referred as RBF-SVR).

In our first experiment, we used topics based on LDA as features extracted from aggregated Facebook statuses of users. We used

these latent topics as features to predict the BIG5 scores by running all mentioned regression techniques. In our second experiment, we used the topics based on LDA as features extracted from aggregated Facebook statuses along with the LIWC features of each user. This experiment was done in order to check if the performance of regression techniques increases with the inclusion of additional features that characterize users.

We ran both these experiments by varying the number of topics of LDA from 1 to 4. This would show the effect of a number of topics on the performance of each of the regression techniques (linear, polynomial and RBF). The degree of the polynomial kernel of SVR has been varied from 1 to 5 in order to check how degree affects the performance. The kernel coefficient γ for SVR with RBF have been varied with values 0.6, 0.8, 1.0 and 1.2, respectively. This was done in order to check the effect of γ on regression performance. For both the experiments, the topics extracted by LDA have been converted to suitable TF-IDF matrices that have been then used to predict the BIG5 personality scores.

Finally, we compared the predicted scores with the scores acquired through *MyPersonality* questionnaires using the MSE metrics.

5 EXPERIMENTAL ASSESSMENT AND ANALYSIS

From Table 1 to Table 8, we show the results of our experimental campaign. In each table, the values across each regression technique specify the minimum value of retrieved MSE, by varying their proper model parameters.

Table 1: Regression on 1 topic found by LDA.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.54	0.49	0.45	0.47
Conscientiousness	0.65	0.58	0.54	0.56
Extroversion	0.79	0.71	0.67	0.68
Agreeableness	0.61	0.54	0.50	0.52
Neuroticism	0.77	0.69	0.64	0.67

Table 2: Regression on 2 topics found by LDA.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.45	0.46	0.45	0.46
Conscientiousness	0.91	0.74	0.63	0.65
Extroversion	0.93	0.77	0.66	0.69
Agreeableness	0.82	0.62	0.53	0.55
Neuroticism	0.73	0.58	0.50	0.51

Then, we studied the behavior of MSE (still considering minimum retrieved values) with respect to the number of topics. Results are shown in Figure 1, Figure 2 and Figure 3, respectively, for each one of the BIG5 personality traits, for the different regressions analysis models considered, i.e. LR, L-SVR, P-SVR and RBF-SVR.

Table 3: Regression on 3 topics found by LDA.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.83	0.57	0.45	0.46
Conscientiousness	0.97	0.67	0.53	0.55
Extroversion	1.28	0.82	0.65	0.67
Agreeableness	0.97	0.63	0.49	0.50
Neuroticism	1.37	0.79	0.63	0.64

Table 4: Regression on 4 topics found by LDA.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	1.33	0.61	0.45	0.46
Conscientiousness	2.08	0.71	0.53	0.55
Extroversion	2.77	0.88	0.65	0.67
Agreeableness	1.49	0.67	0.49	0.50
Neuroticism	2.37	0.86	0.63	0.64

Table 5: Regression on 1 topic found by LDA and LIWC features.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.56	0.47	0.47	0.45
Conscientiousness	0.67	0.55	0.53	0.53
Extroversion	0.82	0.68	0.66	0.65
Agreeableness	0.64	0.54	0.51	0.52
Neuroticism	0.82	0.67	0.64	0.65

Table 6: Regression on 2 topics found by LDA and LIWC features.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.63	0.46	0.45	0.43
Conscientiousness	0.78	0.55	0.53	0.53
Extroversion	0.91	0.67	0.67	0.65
Agreeableness	0.70	0.52	0.50	0.51
Neuroticism	0.89	0.66	0.64	0.64

Table 7: Regression on 3 topics found by LDA and LIWC features.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	0.89	0.47	0.47	0.45
Conscientiousness	0.95	0.54	0.54	0.52
Extroversion	1.20	0.68	0.67	0.64
Agreeableness	1.09	0.53	0.51	0.51
Neuroticism	1.20	0.67	0.64	0.64

Table 8: Regression on 4 topics found by LDA and LIWC features.

BIG5	LR	L-SVR	P-SVR	RBF-SVR
Openness	1.16	0.47	0.47	0.45
Conscientiousness	0.47	0.54	0.54	0.53
Extroversion	1.57	0.68	0.67	0.65
Agreeableness	1.32	0.53	0.51	0.51
Neuroticism	1.51	0.67	0.64	0.64

In particular, Figure 1 reports on the case of using topics for the Facebook statuses of users only as features for regression analysis. Figure 2 reports instead the case of using even LIWC features in addition to users' Facebook statuses. Finally, Figure 3 shows the combination of both previous experiments, i.e. the case of using both settings: users' Facebook statuses only, and users' Facebook statuses and LIWC features, respectively.

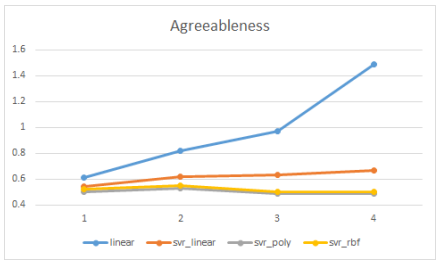
From Figure 1, it clearly follows that an increase of MSE for LR occur following an increase of number of topics retrieved by LDA. On the other hand, P-SVR and RBF-SVR hardly show any change in MSE with the increase of number of topics. Similar trends are also observed in Figure 2. Based on the analysis that derives from Figure 1 and Figure 2, we can infer that P-SVR and RBF-SVR perform better than L-SVR and LR, as they expose lowest MSE for all the BIG5 personality traits. This can also be observed from previous Tables 1 – 8, where MSE in almost all the personality traits exceeds 1 for LR whereas it reaches 0.64 at the maximum for P-SVR and RBF-SVR. In each of the reported cases, it can also be observed that the MSE for L-SVR is higher than MSE for P-SVR and RBF-SVR.

Figure 3 combines the experiments of Figure 1 and Figure 2, in order to compare the performance of different regression techniques with respect to using only Facebook status topics versus the combination of Facebook status topics and LIWC features, for different number of topics. We observe that, in both the cases, MSE for LR increases with number of topics, though the increase is less in the case of considering additional LIWC features for regression analysis. Contrary to this, P-SVR and RBF-SVR hardly show any change in MSE when considering additional LIWC features for regression analysis. From this evidence, we can infer that these two regression techniques are more robust when linguistic features for determining personality traits of Facebook users are considered.

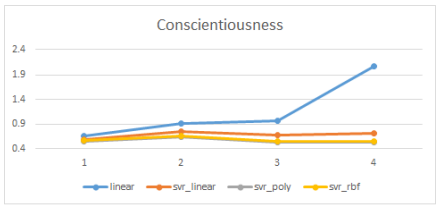
Overall, from analyzing all the results of our experimental campaign, we conclude that, for less number of topics, the performance of LR and SVR do not vary too much. Since LR is computationally cheap as compared to SVR, LR reveals itself to be the better approach to predict Facebook users' personality by using fewer topics and LIWC features.

6 CONCLUSIONS AND FUTURE WORK

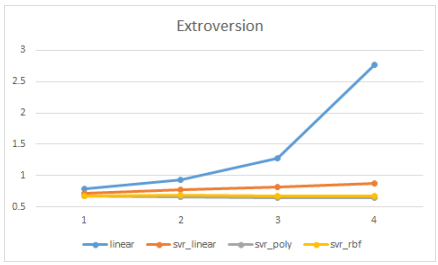
This paper has introduced a framework for supporting BIG5-based prediction of based on Facebook users' personality via exploiting Facebook statuses and their LIWC features on top of which flexible regression analysis techniques are performed (basically, LR and SVR). Our rich experimental campaign has demonstrated that P-SVR and RBF-SVR expose the best performance. Moreover, we



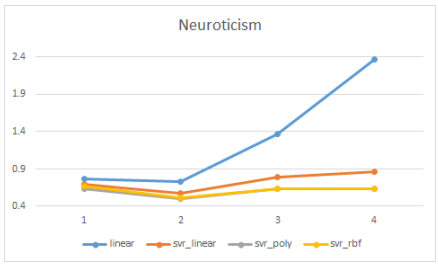
(a) MSE of regression for *Agreeableness*



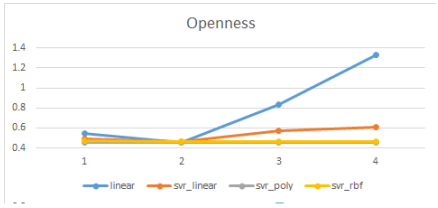
(b) MSE of regression for *Conscientiousness*



(c) MSE of regression for *Extroversion*



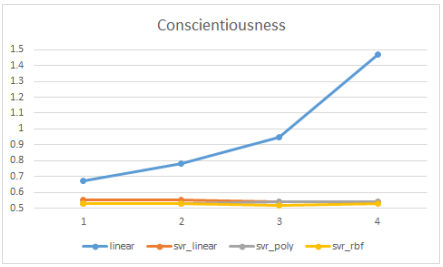
(d) MSE of regression for *Neuroticism*



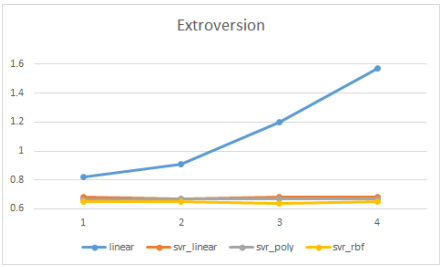
(e) MSE of regression for *Openness*

Figure 1: MSE for BIG5 scores with respect to number of topics when considering users' Facebook statuses only, for the regressions analysis models LR, L-SVR, P-SVR and RBF-SVR.

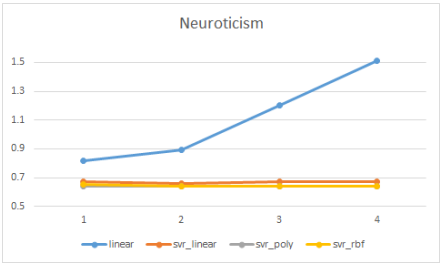
(a) MSE of regression for *Agreeableness*



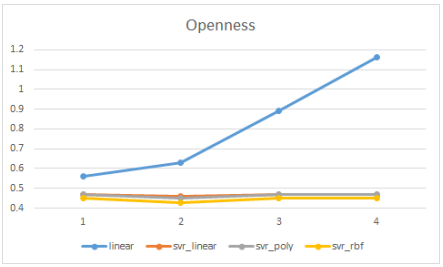
(b) MSE of regression for *Conscientiousness*



(c) MSE of regression for *Extroversion*



(d) MSE of regression for *Neuroticism*



(e) MSE of regression for *Openness*

Figure 2: MSE for BIG5 scores with respect to number of topics when considering users' Facebook statuses and LIWC features, for the regressions analysis models LR, L-SVR, P-SVR and RBF-SVR.

observe that, incorporating LIWC features in addition to Facebook statuses, significantly increases the performance of LR.

Future work is oriented to adding other features to our regression analysis, like demographic details, IQ scores, and so forth, and studying how they affect the performance of the various regression models.

7 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] M. Cannataro, A. Cuzzocrea, and A. Pugliese. 2001. A Probabilistic Approach to Model Adaptive Hypermedia Systems. In *1st International Workshop on Web Dynamics (WebDyn), in conjunction with the 8th International Conference on Database Theory (ICDT 2001)*.

[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] Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. 2002. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition* (third ed.). Routledge. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0805822232

[6] Alfredo Cuzzocrea, Giancarlo Fortino, and Omer F. Rana. 2013. Managing Data and Processes in Cloud-Enabled Large-Scale Sensor Networks: State-of-the-Art and Future Research Directions. In *13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013, Delft, Netherlands, May 13-16, 2013*. 583–588. https://doi.org/10.1109/CCGrid.2013.116

[7] 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.

[8] 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.

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

[10] Thomas Hofmann, Bernhard SchÃűlkopf, and Alexander J. Smola. 2008. Kernel methods in machine learning. *Ann. Statist.* 36, 3 (06 2008), 1171–1220. https://doi.org/10.1214/009053607000000677

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

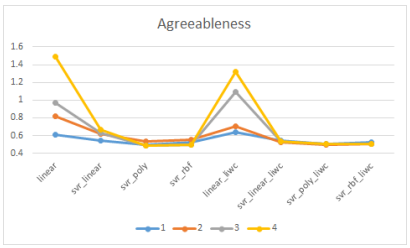
[15] E. L. Lehmann and George Casella. 1998. *Theory of Point Estimation (Springer Texts in Statistics)* (2nd ed.). Springer. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387985026

[16] John Neter, Michael H. Kutner, William Wasserman, and Christopher J. Nachtsheim. 1996. *Applied Linear Regression Models*. McGraw-Hill/Irwin. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/025608601X

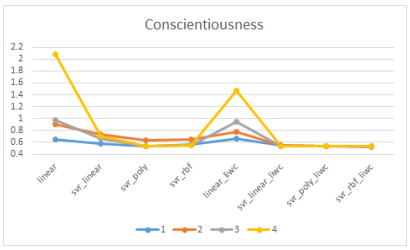
[17] Daniel J Ozer and Veronica Benet-Martinez. 2006. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.* 57 (2006), 401–421.

[18] 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.

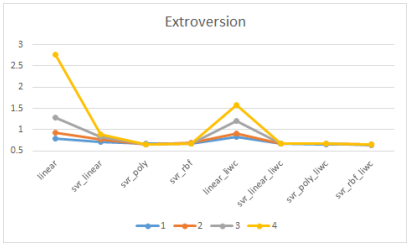
[19] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. *Linguistic Inquiry and Word Count: LIWC2007, Operator's manual*. LIWC.net, Austin, TX.



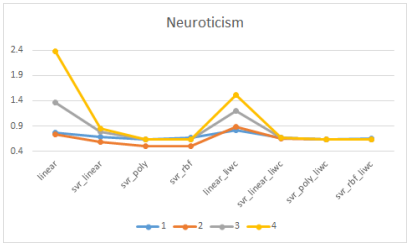
(a) Comparative MSE of regression for *Agreeableness*



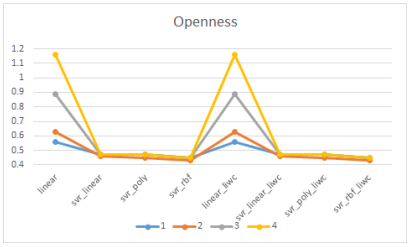
(b) Comparative MSE of regression for *Conscientiousness*



(c) Comparative MSE of regression for *Extroversion*



(d) Comparative MSE of regression for *Neuroticism*



(e) Comparative MSE of regression for *Openness*

Figure 3: MSE for BIG5 scores with respect to number of topics when considering both cases: users' Facebook statuses only, and users' Facebook statuses and LIWC features, respectively, for the regressions analysis models LR, L-SVR, P-SVR and RBF-SVR.

[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] AlexJ Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3 (1 Aug. 2004), 199–222. https://doi.org/10.1023/b:stco.0000035301.49549.88

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.