Predict Personality with Social Networks

By

Hanzi Li
Yuan Su
Zhaowei Zheng

03/19/2016
Santa Clara University

Table of Contents

1 Abstract

In this project, we have developed quantitative prediction models for people's Big 5 personality: Extraversion, Conscientiousness, Openness, Agreeableness and Neuroticism. We extracted language features and emoticon feature from Facebook users' status, and we also used data regarding the users' network information, such as network size and network betweenness, which total accounts to 101 features. We have implemented Linear Regression (LR), Artificial Neural Network (ANN) and M5P (M5') algorithms. Using root mean square error (RMSE) as evaluation standard, we found that our M5P provides best performance, and our model is better than the model reported by literature for all 5 personalities. We further filtered the features based on statistical hypothesis test. After feature filtering, LR shows comparable performance to M5P. Our optimized model demonstrates much better performance than the model reported by literature.

2 Acknowledgement

Special thanks to Professor Wang. We learned a lot from your class and we really appreciate it.

3 Introduction

3.1 Objective

Propose a method by which a user's personality can be accurately predicted through the publicly available information on their profile.

3.2 What Is the Problem

Personality prediction is more and more important in social network nowadays, as there's

significant correlation between personality and real-world behavior. When we focus on computational recognition of personality, this technique is more and more promising.

In the business field, personalized marketing and application design can help companies make great success on gaining more customers. It can be also tied to romantic relationships success with personality prediction.

3.3 Why This is a Project Related to This Class

In this project, we will use text analysis and mining, such as swear words counts. For data mining algorithms, we choose SVM, M5 to build the model. Other data mining topics, including neural networks, as well as decision tree are also applied.

3.4 Drawbacks of other approaches

Basically, there are already several different tries to make the prediction convincing. However, personality recognition is a challenge task, due to the fact that there are no obvious predictive features, and it correlated and strongly depends on datasets. How to make computational personality recognition more accurate and efficient is more attractive.

3.5 Advantages of our approaches

Our expectation is to be more accurate and efficient. Firstly, we based on linguistic features (LIWC), then add more features, such as emotion-icons to get more accurate result, sine we

suppose the emotion represent more sentimental meaning than texts. Secondly, we try to use top-down approaches, based on lexical resources (including sentiment analysis) ensemble to get more efficient when facing large datasets. Finally, we hope to make our model stronger and more robust by using cross-domain learning features and algorithms.

3.6 Statement of the Problem

In ubiquitous daily used social network, social media is a place where users present themselves to the world, revealing personal details and insights into their lives. Personality has been shown to be relevant to many types of interactions. User's personality can be accurately predicted through the publicly available information on their social network profiles.

3.7 Area or Scope of Investigation

Data mining / machine learning

4 Theoretical Bases and Literature Review

4.1 Definition of The Problem

Personality has been shown to be relevant to many types of interactions; it is useful in predicting job satisfaction, professional and romantic relationship success, and even preference for different interfaces. Until now, to accurately gauge users' personalities, they needed to take a personality test. This made it impractical to use personality analysis in many social media domains. We attempt to propose a method by which a user's personality can be accurately predicted through the publicly available information on their Twitter profile.

4.2 Related Research to Solve the Problem

In recent years the interest of the scientific community in personality recognition has grown very fast.

The first pioneering applied personality recognition to long texts, such as short essays or blog posts. The current challenges are instead related to the extraction of personality from mobile social networks from social network sites and from languages different from English. There are also many other applications that can take advantage of personality recognition, including social network analysis, recommendation systems, deception detection, sentiment analysis/opinion mining, and others.

4.3 Our solution

In project, we attempt to predict an author's personality features from his or her social network account, briefly described by following steps.

First, gather gold standard labelled dataset with personality scores and Facebook statuses in raw text.

Second, select features from LIWC, add with our new features.

Then, predict social media users' personalities with different algorithms, such as SVM, M5', etc.

Next, tune the results using cross validation, and compare the difference between predicted

values and observed values.

4.4 Where Your Solution Different from Others

We attempt to figure out some specific features have stronger relationship with personality than other features, by adding emotion features. From this proposal, we hope to get more accurate output.

More important, we suppose when facing large and various datasets, we need more efficient algorithms and more robust model.

4.5 Advantages of Proposed Solution

More accurate: based on linguistic features (LIWC), add more features, such as emotion-icons.

More effective: using top-down approaches, based on lexical resources (including sentiment analysis).

More robust: ensemble features/algorithms: try cross-domain learning.

5 Hypothesis and Goals

5.1 Goals

This project attempts to predict an author's personality features from his or her social network account.

We use the Big Five Personality Traits to measure the personalities of a person. The score for

each of the five factors is from 1 to 5.

We analyze additional linguistic features and apply three different algorithms, trying to figure out which one is the best for this specific purpose.

5.2 Positive Hypothesis

Position Hypothesis: Since we add more features to analyze and try different algorithms, the accuracy of the personality score is higher than the previous research.

6 Methodology

6.1 How to Generate / Collect Input Data

We will get a massive dataset from mypersonality.org. In this dataset, it contains 10,000 records of Facebook status updates from 250 users. The attributes in it includes the user identity, status update, personality scores of each of the big five traits, flag for each of traits, and some network structural data.

We use the LIWC tool to quantize and capture linguistic features of the status updates. The feature categories that we collect are: standard count, psychological processes, relativity, personal concerns, and other dimensions such as swear words and emoticons.

For all the features, we calculate the Pearson correlation between every single feature and a personality factor. And from there, we do the student's t hypothesis test against the correlations. If a feature is significantly correlated to a personality factor (with p-value < 0.05), then we are going to use that feature in the later algorithm.

6.2 How We Solve the Problem

We are going to describe how we solve this problem in the following four sections including the algorithm design, programming languages used, tool used, and the output generation.

6.2.1 Algorithm Design

We plan to implement three algorithms to develop the predictive model for the Big Five personality traits.

6.2.1.1 M5' (M5P) Algorithm

M5' algorithm is developed by Wang et al [1], which is an improvement of Quinlan et al's M5 algorithm. It is an advanced decision tree algorithm, and nowadays it is used more and more by data mining projects.

Essentially, M5' algorithm combines a conventional decision tree with the possibility of linear regression functions at the nodes. The pseudo-code for M5' is shown in Wang et al's [1] paper. There are two main parts. The first one is to create a tree by successively splitting nodes, as named by "split". The other is to prune it from the leaves upwards, as named by "prune".

At each inner node, compared with conventional decision tree where the information gain is maximized when splitting, M5' minimizes the intra-subset variation in the class. What is worthy mentioning is, M5' takes into account the missing value by using modified standard deviation. When the values of all instances that reach a node vary very slightly, or only a few instance

remains, the splitting procedure would stop in M5' thus stop the whole procedure.

Additionally, in M5', When pruning an inner node is turned into a leaf with a regression plane, the tree is pruned back from each leaf. Third, to avoid sharp discontinuities between the subtrees, a smoothing procedure is applied that combines the leaf model prediction with each node along the path back to the root, smoothing it at each of these nodes by combining it with the value predicted by the linear model for that node.

Previous work on Big 5 Personality has implemented this algorithm and reported decent result. In this project, we plan to implement this algorithm on the dataset. With the additional features, we expect to obtain better result.

6.2.1.2 Neural Network

Neural Network Algorithm mimic people's brain. It is said to be the "second better" algorithm for all data mining problem. So far, we have not read any publication where neural network algorithm is used for people's personality prediction. We think it is worthy trying this algorithm and compare with M5' which is so far the best one for the personality dataset.

Generally speaking, a neural network is typically defined by three types of parameters [3]:

The interconnection pattern between the different layers of neurons

The learning process for updating the weights of the interconnections

The activation function that converts a neuron's weighted input to its output activation.

Based on our dataset, since we do not have too many features (the total number of features is estimated to around 20), for now, we think a two-layer neural network might be enough. However, we will try more layers and see if we could get a better result.

6.2.2 Language Used

We used Java to combine data, and R for data process and modeling.

6.2.3 Tools Used

We will Weka Data Mining Software for modeling and LIWC Tool for textual analysis.

6.3 Output Generation

6.3.1 Model Development

When we use training data for model development, weka will allow us to tune the setting to get best result for each algorithm. All the settings and final parameters for each algorithm will be output.

6.3.2 Model Testing

We will test our model towards test sets. For model evaluation, root mean square error which compares the difference between predicted values and observed values will be output and used as standard for model comparison.

6.4 Test Against Hypothesis

We will use LIWC Tool to get more features from dataset. Then use Weka Data Mining tool to implement data mining algorithms: M5' and Neural Network. After developing predictive models based on the three algorithms. We then use quantitative statistic values, eg: Root Mean Square Error to tell which model works best for personality prediction.

7 Implementation

7.1 Code

Please refer to our code submission.

7.2 Flowchart

Please see figures below.

8 Data analysis and discussion

8.1 Output Generation

8.1.1 Group Data by Users

The data we got from mypersonality.org contains 10,000 rows of data from 250 users, so we need to group the data by users, by concatenating all the status updates that are from the same user. This is done by P3.java.

8.1.2 Capture Linguistic Features

We use LIWC to capture the linguistic features. The feature categories that we are going to collect were mentioned previously in the data collections section.

The build-in LIWC2015 dictionary satisfies most of our need, but it doesn't capture the emoticon feature. So we created a custom dictionary to include all the emoticons on Facebook and load it

to LIWC. In this way, LIWC is able to capture all the features we need in this project.

8.1.3 Personality and Feature Correlations

We utilized R to calculate the Pearson Correlations for every pairs of feature and big five factor. And did t-test for each correlations we found. This is done by correlation-test.R. The output of this file is a csv file, indicating the correlations and what are the big five factors that are significantly correlated to a specific feature.

Below are some of our observations.

The numbers of features that are strongly correlated to different personality factors are different. Extraversion has the most features (16). And then Neuroticism and Conscientiousness have 8 features. Openness to Experience has 4 features, and the Agreeableness has the least features (2).

From the figure, we can see that network size is strongly and positively correlated to Extraversion with rho 0.362. This makes sense because extroverts tend to have more friends. And the network density is strongly and negatively correlated to Extraversion with rho -0.313. This is intuitive because extroverts' friends are crossing different friend circles, so their network density is sparser. Introverts usually stay in some certain friend circles, so their network is den ser. This explains why network density and extraversion is negatively correlated.

We have found other intuitive correlations with weaker correlations. For example, anger is positively correlated to neuroticism with rho 0.208 and negatively correlated to conscientiousness with rho -0.182.

An interesting phenomenon we found is that, some big five factors are correlated to each other.

For example, Agreeableness is strongly and negatively correlated to Neuroticism with rho -0.421. This indicates that high agreeableness is less neurotic.

8.1.4 Generate Input Files for Weka

From our correlations, we prepared 5 files for Weka. Each file is for one big five factors, and its related features. These files are with ARFF extension.

8.1.5 Use Weka for Model Output

We input the files from last step into Weka and train the model. And then get the output.

8.2 Output Analysis

M5'(M5P) is used to build a model tree for personality prediction. Essentially, M5P is a decision tree with linear regression functions at the leaves. It predicts a numeric target (class) attribute and produces a piecewise linear fit to the target. Figure Extraversion Score Using M5P with all 101 features is an example of our prediction model using M5P. This tree builds 6 rules, as labeled LM in the figure. For example, when the brokerage feature is less or equal to 27379.5 and reward feature is less or equal to 1.355. The extraversion score is predicted using LM1. In (37/74.024%) of LM 1, 37 means the number of covered instances while 74.024% means root mean square error (RMSE) divided by the global absolute deviation.

In M5P model tree development, the important parameters which determine model performance include minNumInstances which is the minimum number of instances to allow at a leaf node, unpruned which represents whether unpruned tree/rules are to be generated and useUnsmoothed which represents whether to use unsmoothed predictions. In our model development, we

choose unpuned false and useUnsmoothed false since pruned tree would avoid over-fitting problem and smooth tree would lead to better performance. When varying minNumInstance as shown in Figure RMSE Varies with minNumInstance for Extraversio Score in M5P and Figure RMSE Varies with minNumInstance for Conscientiousness Score in M5P, we found that there is always an optimal value for minNumInstance for each dataset. Using that optimal minNumInstance, RMSE value is recorded and used for future evaluation.

We used two types of datasets, one type include other personalities as features (as called feature 105) and the other type only include the features we generated and network features (as called feature 101). As shown in Figure M5P vs. ANN(neural network) vs. LR (linear regression) for All Features, we have developed Big 5 personality prediction model using the three algorithms for the two types of features (101 and 105). Generally speaking, M5P shows best performance with RMSE as evaluation standard. The **GREAT** news is that it is better than the data reported by literature where twitter data was used for personality prediction. Since the data we use here are actually from the same user as in the twitter data, the fact that our model shows better performance indicate that the features we extracted from Facebook data is more relates to Big 5 personality and our model predicts people's personality better.

Figure M5P after feature filtering shows that, after feature filtering, RMSE improve, which indicates that the filtering algorithm based on statistics works well in our dataset. Feature filtering filtered noisy data for linear regression and keeps the most important features.

8.3 Compare Output Against Hypothesis

First of all, our model shows smaller root mean square error than the data reported by literature, which indicate that our model has better performance. And this output is consistent with our hypothesis.

Additionally, when using three different algorithms, M5P works best when using all features. Both ANN and M5P beat LR significantly, which is consistent with our hypothesis and expectation.

Besides, using our proposed feature filtering method, the model performance is further improved, which demonstrates that in our case, statistical hypothesis test works well.

8.4 Abnormal Case Explanation

At first, we did not expect that after feature filtering, the linear regression would provide decent result. However, after our statistical feature filtering, LR provides better result than M5P (101) and very close to M5P (filtered). As shown in Figure M5P after feature filtering, model built by LR is actually very similar to M5P model. This indicate that our feature filtering filtered noisy data for LR and keeps the most important features. Such function is comparable to M5P where variance was evaluated to rank the most significant features.

8.5 Statistical Regression

We use root mean squared error (RMSE) to evaluate the error between our estimated data and real data. Please see figures below.

8.6 Discussion

Figure comparison of our model with literature data shows our final results compared with literature data. Our model beats literature model significantly. And our feature filtering method which is based on hypothesis test works for our personality prediction case.

9 Conclusions and recommendations

In our model prediction for Big 5 personality, we have used Facebook data, while the literature uses Twitter data. Since they are the same users for both Facebook and Twitter data, it might be interesting to perform model development based on the combination of the two datasets.

Additionally, since we know that, nowadays social network like Facebook and Twitter have many users, while the user data we use is very small part, it might be worthy testing our model for more users or implementing our model for a dataset with more user datasets in it.

Besides, in our project, the algorithms we use are LR, M5P and ANN. We think they are the most promising algorithms for this kind of dataset. They also demonstrate very good performance as what we expected. However, they are a lot of data mining algorithms. If time allows, some other algorithms might also be worthy trying.

10 Bibliography

Celli, Fabio, Fabio Painesi, David Stillwell, and Michal Kosinski. 2014. "Workshop on Computational Personality Recognition: Shared Task." 4.

Conference, 12th Austrialian Joint. 1999. *Advanced Topics in Artificial Intelligence.* Sydney: Springer.

Golbeck, Jennifer, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. "Predicting Personality from Twitter." *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* 8.

Golbeck, Jennifer, Cristina Robles, and Karen Turner. 2011. "Predicting Personality with Social Media." *ACM.*

Golveck, Jennifer, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. "Predicting Personality from Twitter." *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing.*

Luo, Zhunchen, Miles Osborne, Sasa Petrovic, and Ting Wang. 2013. "Improving Twitter Retrieval by Exploiting Structural Information." *Preceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence.*

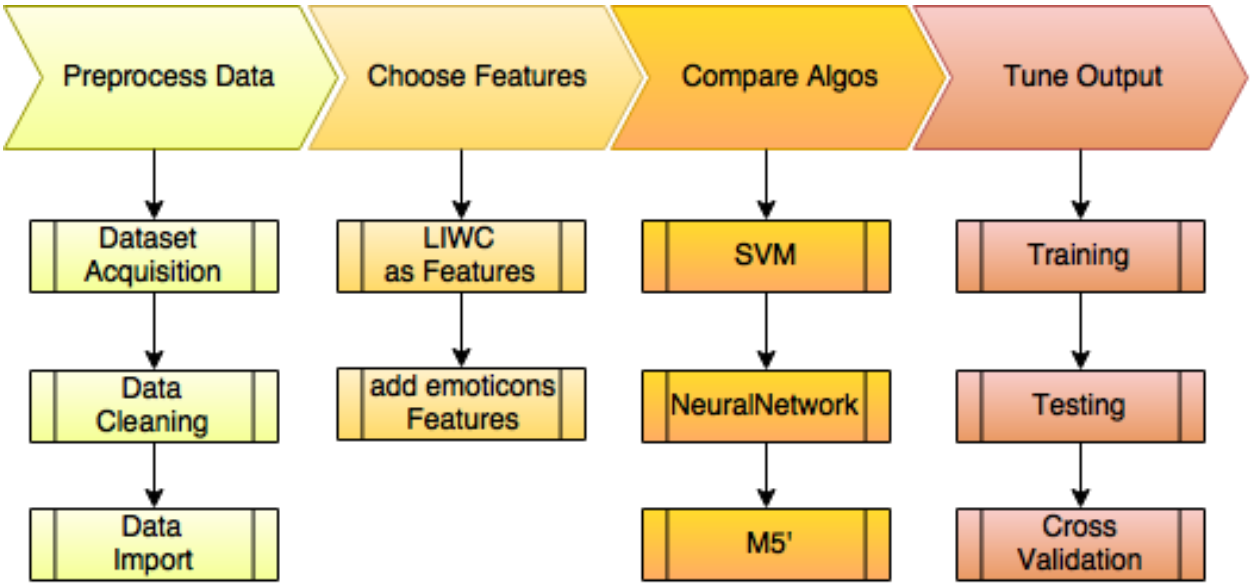
Quercia, Daniele, Michal Kosinski, David Stillwell, and Jon Crowcroft. n.d. "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter." *International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* 2011.

Quinlan, J. R. 2006. *Learning With Continuous Classes.* Sydney.

Wang, Yong, and Ian H Witten. n.d. *Iducing Model Trees for Continuous Classes.* University of Waikato.

11 Appendices

11.1 Program Flowchart



11.2 Figures

featureName	sExtCorrelation	sNeuCorrelation	sAgrCorrelation	sConCorrelation	sOpnCorrelation
NETWORKSIZE	**0.362139412**	**-0.180202417**	0.111331287	**0.156496408**	0.021652919
BETWEENNESS	**0.279393299**	**-0.136226906**	0.090130323	0.120864512	0.049911973
NBETWEENNESS	**0.250395646**	-0.051014033	0.115631858	0.111512119	-0.065374904
DENSITY	**-0.313391007**	0.111121629	-0.119949976	**-0.175826454**	0.042348396
BROKERAGE	**0.280104819**	**-0.136945562**	0.090449085	0.121301581	0.049551469
NBROKERAGE	**0.278453403**	-0.093519107	0.12208934	0.10199854	-0.029842516
TRANSITIVITY	**-0.298279662**	**0.18054047**	**-0.216499922**	-0.103805338	-0.058849298
Tone	0.096109401	**-0.178155924**	0.08102176	-0.045526526	0.032388085
Dic	**0.192334578**	-0.040054008	0.113777046	0.012306142	**0.131585323**
you	0.109684564	-0.079829542	0.01981891	-0.098091207	**0.181084175**
prep	0.055244997	**-0.12986321**	0.048425216	0.123329122	0.004771126
number	-0.115131704	0.022276307	**-0.157481129**	-0.069507523	**-0.131872093**
affect	**0.142116144**	-0.05664176	0.078888154	-0.02058511	-0.001694057
posemo	**0.142450272**	-0.101370894	0.077381937	0.049312079	-0.008115381
negemo	0.034380568	0.060098411	0.021343736	**-0.134775961**	0.020561052
anger	-0.115794461	**0.207849397**	-0.042721972	**-0.181889456**	0.045815762
sad	**0.126660315**	-0.030102596	-0.012464636	-0.045686346	0.006536196
social	0.094987273	-0.060658613	0.106779491	-0.02657665	**0.132283828**
female	**0.138809592**	-0.038396104	0.08923067	0.115788993	0.067036037
male	-0.02941115	0.059799625	-0.113741044	**-0.17834144**	-0.043543051
cogproc	**0.133918672**	0.038237564	0.067208633	**-0.125773263**	0.0658316
insight	0.013005093	0.091624421	-0.013812057	**-0.133654093**	-0.01834853
tentat	**0.125717747**	-0.025250378	0.098664015	-0.098399941	0.026640821
health	**0.130561732**	-0.064150581	-0.00783482	0.067394055	0.111651847
achieve	**0.127989645**	-0.112403492	0.091697426	**0.129490139**	0.005834973
Comma	-0.09292163	-0.023308452	-0.036427315	-0.009814168	**0.134924881**
Apostro	-0.043997865	**-0.132105586**	-0.010615943	0.012021057	0.098647827
sEXT	1	**-0.377247851**	**0.205419723**	**0.200926335**	**0.203026889**
sNEU	**-0.377247851**	1	**-0.420795931**	**-0.234329472**	**-0.135143751**
sAGR	**0.205419723**	**-0.420795931**	1	0.099657479	**0.179662488**
sCON	**0.200926335**	**-0.234329472**	0.099657479	1	0.082613401
sOPN	**0.203026889**	**-0.135143751**	**0.179662488**	0.082613401	1

Figure 1: Feature Personality Correlation

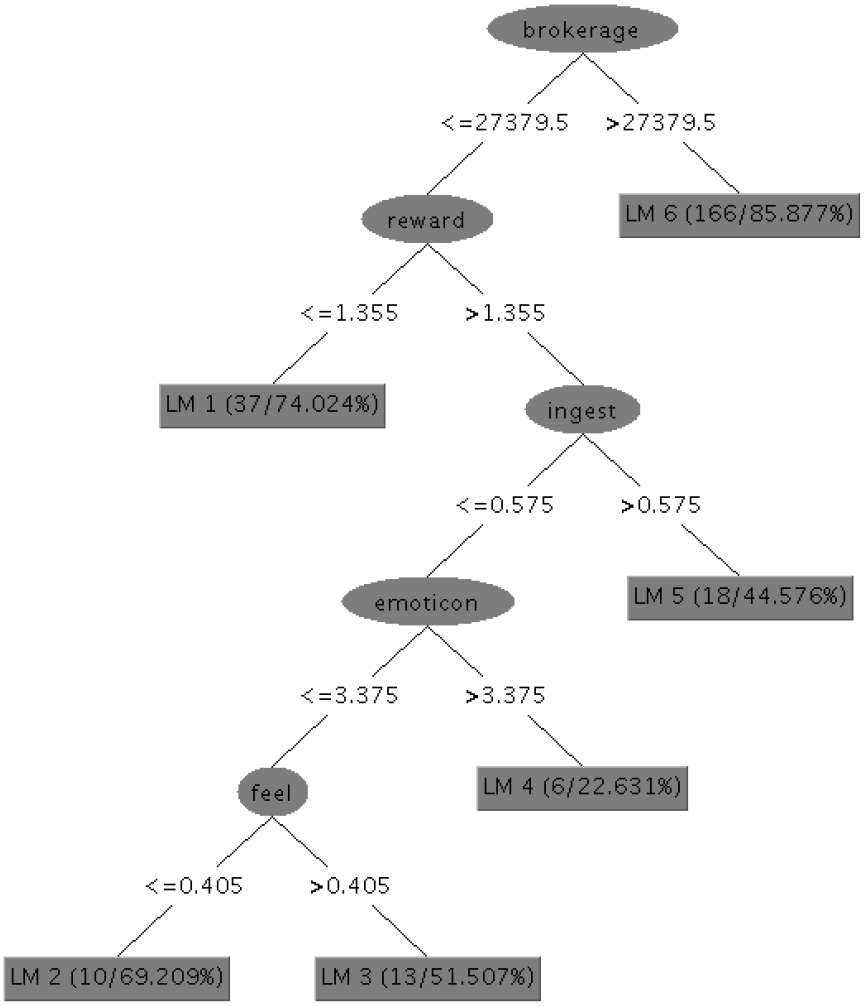


Figure 2: Extraversion Score Using M5P with all 101 features

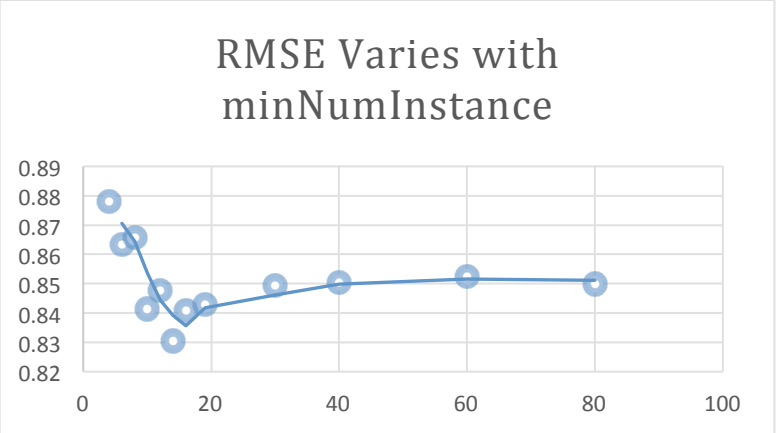


Figure 3: RMSE Varies with minNumInstance for Extraversion Score in M5P

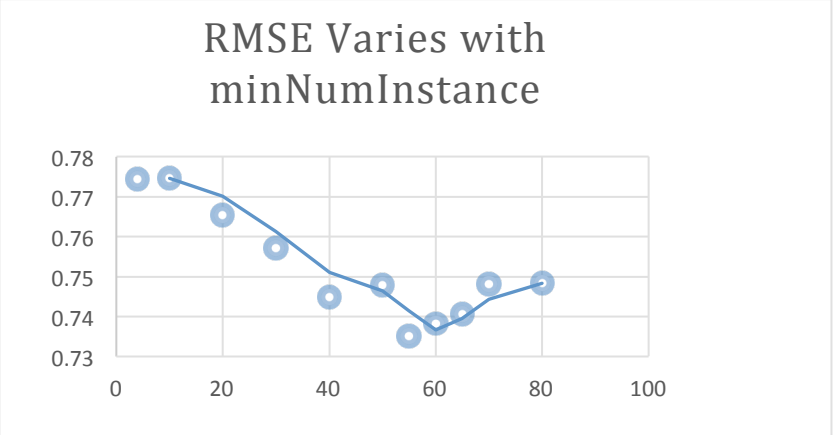


Figure 4: RMSE Varies with minNumInstance for Conscientiousness Score in M5P

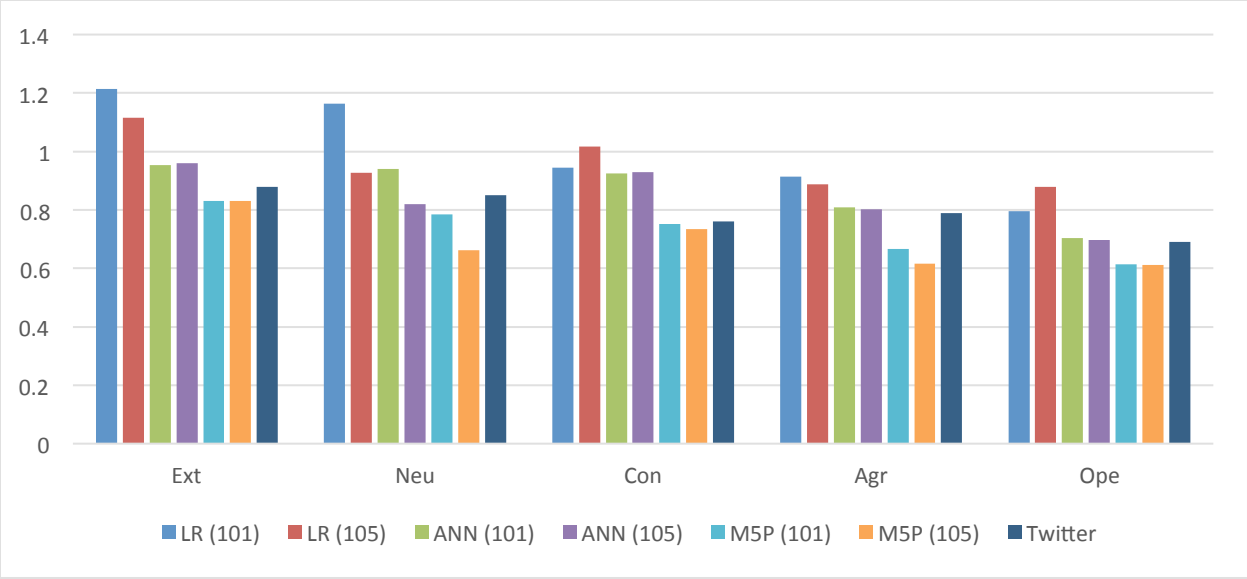
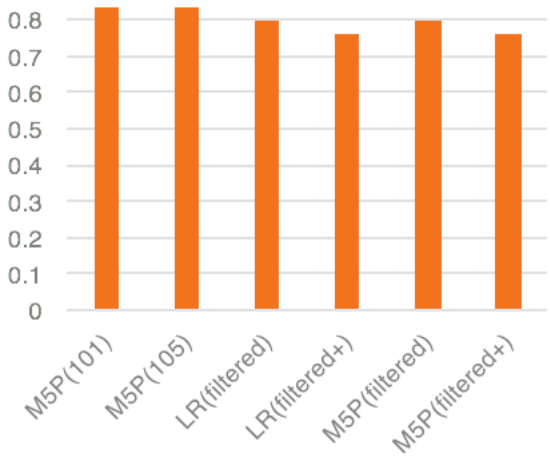


Figure 5: M5P vs. ANN vs. LR for all Features




```
Linear Regression Model

sext =

      0.0022 * networksize +
      0      * brokerage +
     -0.9896 * transitivity +
      0.0129 * dic +
      0.0227 * posemo +
      0.0799 * health +
      1.6092


M5 pruned model tree:
(using smoothed linear models)
LM1 (250/87.596%)

LM num: 1
sext =
        0.0022 * networksize
        - 0 * brokerage
        - 0.9896 * transitivity
        + 0.0129 * dic
        + 0.0227 * posemo
        + 0.0799 * health
        + 1.6092

Number of Rules : 1
```

Figure 6: M5P after feature filtering

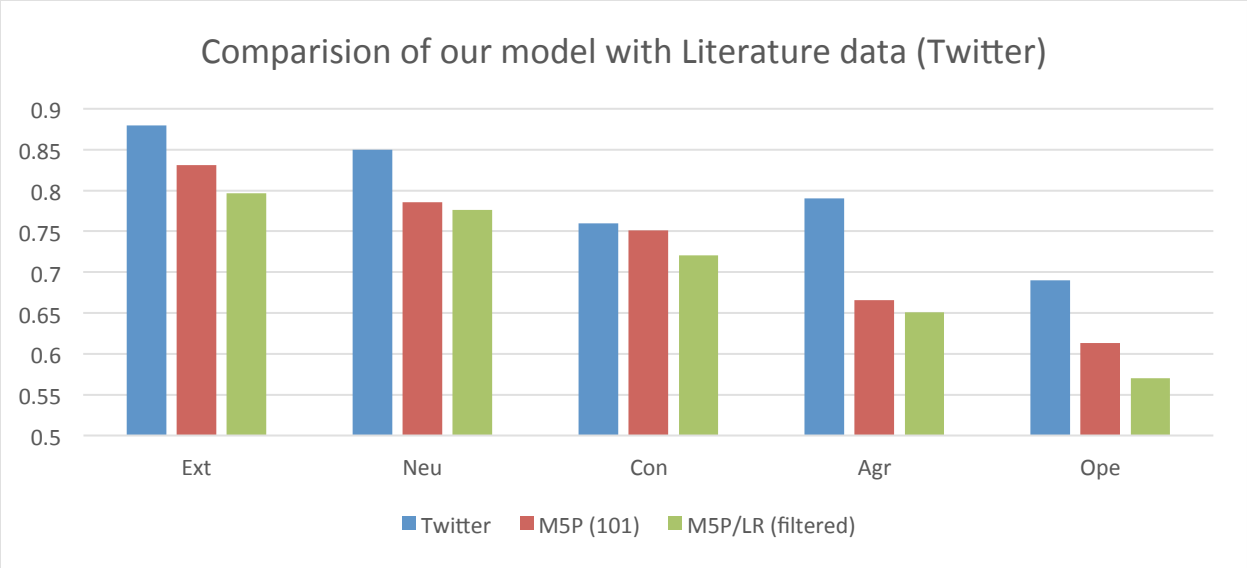


Figure 7: comparison of our model with literature data