What we post, reveals who we are: Relating Facebook posts to personalities

Mohammed Aldhoayan
University of Pittsburgh
mda34@pitt.edu

Akash Parvatikar
University of Pittsburgh
akp47@pitt.edu

Yuqian Zhanfu
University of Pittsburgh
yuz98@pitt.edu

ABSTRACT

People's personality detection can be used for personalized marketing, recommendations to match individuals with similar personality and academic research. Social media data reflects people's traits through their posts at different times. Text mining can be used to unravel hidden topics in the individuals' post which may reflect their state of mind.
General information such as age, sex, and demographic details of individuals was used to predict personalities. In addition, other models included word- document matrix and topics analysis to predict a person's trait. However, all the data is not always available for analysis. We propose an approach to use different models based on emotions extracted from individuals' posts to predict different personalities. The outcome of the study shows that this approach is promising and can be used for different applications. A possible improvement can be tested by using a larger feature space and more data for each individual.

Keywords

Social media, emotion, personality, data mining, time

1. INTRODUCTION

Social media platform has become a main data generation resource nowadays. Facebook and Twitter social media platforms along with other social media such as Instagram, Pinterest, and Whatsapp bring multiple types of social media data. Lots of data is being generated by such platforms which offer large research opportunities in different areas. The marketing industry adopts social media data to personalize customer's preference and invest in advertising [8]. Psychologists use the prediction of individual's traits from their posts to reveal some insights into human personality [7]. Scientists compared the automatic personality assessment and human assessment and found that computational personality recognition is more accurate and accessible than human-made measures [2]. Based on the previous works, our project proposes to extract emotions from Facebook posts based on "emolex" corpus and then treat the processed emotions as predictors to predict BigFive personality through various algorithms. In this paper, we will discuss the dataset used in this project, the data processing, exploration part, the data mining methodologies and the evaluation followed by discussing the challenges faced.

2. RELATED WORK

Social media data have largely exploded in the era of mobile technology. The social media data can be used both in business and academic research. From the research conducted by Gregory Park, H. Andrew Schwartz[1] , social media data can extend access to many more people quickly, cheaply, and with low participants burden. To further understand the dataset, a comprehensive data analysis conducted by Golnoosh Farnadi [4] visualized the personality characteristics by gender, age, time and personalities, which shows some essential insight of the Facebook posts dataset. Furthermore, some researchers did a deep analysis regarding the data mining and classification methods to the personality recognition. For example, Dejan Markoviky and Sonja Gievaska , etc. [5] provided several classification techniques to mine Facebook Data, and they found Simple Minimal Optimization (SMO) and Boost algorithms (MultiBoost AB and AdaBoost M1) that showed a significant precision advantage. Laura Parks-Leduc, Gilad Feldman [6] provided a Meta-Analysis technique to identify valuable personalities regarding the My-personality project data. The research from Michal Kosinski, David Stillwell [7] used experience result to verify that digital records of behavior can be used to automatically and accurately predict a range of highly sensitive personal attributes, which can arrive 88% accuracy.

3. DATASET

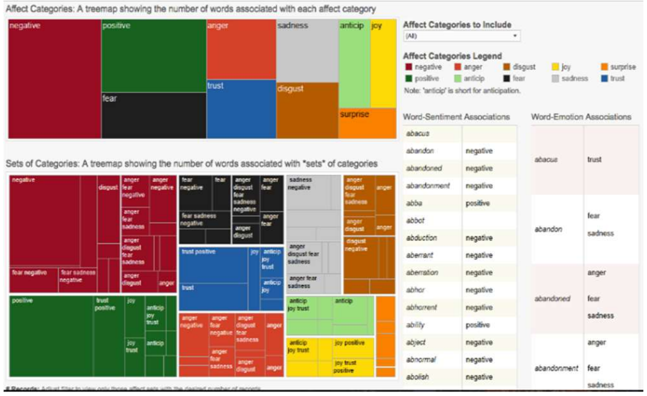
The original dataset is obtained from myPersonality project page. My-Personality was a popular Facebook application that recorded users' psychological and Facebook profiles. The database contains more than 6,000,000 test results and more than 4,000,000 individual Facebook Profiles which includes sex, gender, likes etc. The dataset includes registered dataset and non-registered dataset. In our project, we used the non-registered personality dataset. The features in this dataset include:
STATUS: Facebook posts from each user;
#AUTHID: Anonymous user ID for each user;
BETWEENNESS: Network betweenness score;
NETWORKSIZE: The Facebook network size number between different users;
DENSITY: The degree of compactness between different users
BigFive Personality predicted labels: The personality score predicted from users' posts
BigFive Personality self-labels: Users' self-identification of personalities

4. METHODOLOGY

4.1 DATA PREPROCESSING

The dataset obtained from ‘myPersonality’ is processed to make it suitable for the application of this project. We used Python v2.7 to perform data cleaning and to access each post for analysis. Text preprocessing techniques were used to reduce the size of the posts which would not affect the performance of the algorithm for detecting the emotions pattern. The techniques included removing punctuation marks and stop-words by making use of ‘NLTK’ package in python. The posts were further converted to all lower case and bag of words representation was incorporated. Stemming was not performed on this dataset as different tense of the words indicate different emotions.

We built the **emotions-words dictionary** in Python v2.7 using the ‘EmoLex’ lexicon to weigh each post with respect to the emotion it expresses and obtains the severity of each emotion.


Affect Categories: A treemap showing the number of words associated with each affect category
negative
positive
anger
sadness
anticip
joy
fear
trust
disgust
surprise
Affect Categories to Include
(All)
Affect Categories Legend
negative
anger
disgust
joy
surprise
positive
anticip
fear
sadness
trust
Note: 'anticip' is short for anticipation.
Sets of Categories: A treemap showing the number of words associated with "sets" of categories

Word-Sentiment Associations

abacus	
abandon	negative
abandoned	negative
abandonment	negative
abba	positive
abbot	
abduction	negative
aberrant	negative
aberration	negative
abhor	negative
abhorrent	negative
ability	positive
abject	negative
abnormal	negative
abolish	negative

Word-Emotion Associations

abacus	trust
abandon	fear sadness
abandoned	anger fear sadness
abandonment	anger fear sadness

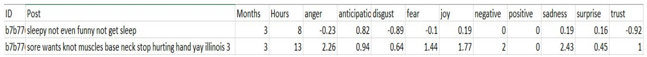
Figure 1- 'Emolex' Lexicon corpus

For this project, we extracted **10 emotions** from each user’s post. Each post is broken down into individual words and is compared to the block of 10 emotions, and a score is given to the word with respect to each emotion. If there is no association of the word with emotion, the score is measured to be 0. After the each ‘word-emotion’ analysis is done an overall score is obtained for the post and corresponding emotion expressed is obtained through the scores. For example, when analyzing a post with positive words (I like reading news), positive emotions, such as joy and trust will be high as shown in (Table 1). If the word “not” is encountered, then the emotion association score of the next non-neutral word will be reversed. For example, if the word love has an association score with the emotion joy as 0.8, then in a sentence like “I do not love ice-cream,” the association score between love and joy will be -0.8 because of the occurrence of the word “not.”

ang	antic	disgust	fear	joy	sadness	trust	surpr
0	1.87	0.06	0.01	0.95	0.13	1	0.51

Table 1- Processed Emotion features for “I like reading news”

The time of each post is rounded to the nearest lower approximation. That is, anytime between 9:00 – 10:00 is coded as 9. Further, the months are also coded to fit certain frame depicting seasonality. December- February is coded as ‘1’ referring to the **winter** season. March- May is coded as ‘2’ referring to the **spring** season. June- August is coded as ‘3’ referring to the **summer** season. September- November is coded as ‘4’ referring to the **fall** season.



ID	Post	Months	Hours	anger	anticipatic	disgust	fear	joy	negative	positive	sadness	surprise	trust
b7b77	sleepy not even funny not get sleep	3	8	-0.23	0.82	-0.89	-0.1	0.19	0	0	0.19	0.16	-0.92
b7b77	sore wants knot muscles base neck stop hurting hand yay illinois 3	3	13	2.26	0.94	0.64	1.44	1.77	2	0	2.43	0.45	1

Table 2 - Emotion extraction table

For instance, figure 2 considers 2 posts from the same user at different times, i.e. at 8 a.m. and 1 p.m. during the summer season. We can observe more anger, fear and sadness of the person while looking at her/his post when compared to the former which expresses more of disgust about not able to sleep.

4.2 DATA EXPLORATION

The first step of data mining is to get a brief data description from several angles. To have a basic knowledge of the dataset, we explore the emotion features, personality, and network size features from several dimensions.

4.2.1 Network size & Personality

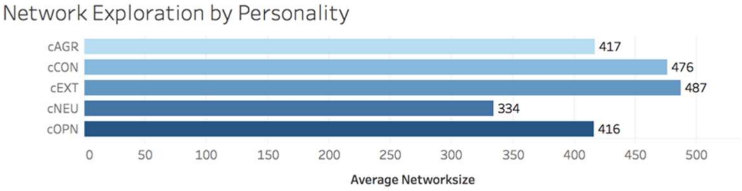


Figure 2 - Network Exploration By Personalities

We explored the average network size with different personalities and observed users who labeled themselves as consciousness and extravert have relatively larger network size.

4.2.2 Emotion Pattern

According to the psychological research, people’s emotion can change depend on the weather. We explored the emotion pattern by seasons and hours and expected to see some regular emotion change patterns. However, the emotion pattern from this dataset do not have a very regular emotion pattern. We checked the peak value in the season and hours and found data in season 2 to only have 8 records, which causes the peak value indicated in figure 3.

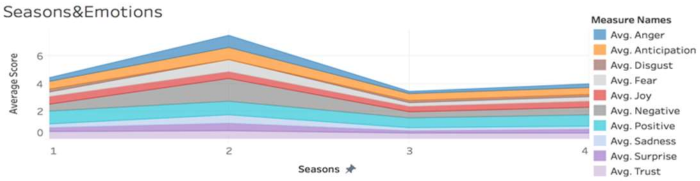


Figure 3 - Emotion Pattern by Seasons

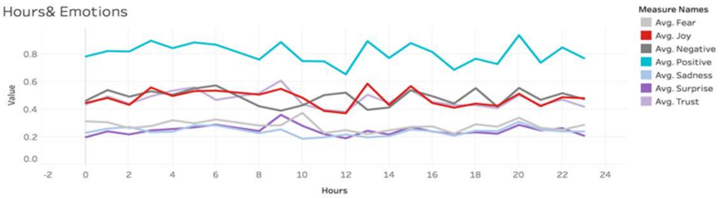


Figure 4 - Emotion Pattern by Hours

4.2.3 Emotion & Personality

From figure 5 we want to see what emotions are frequently shown in the posts from people labeled themselves with different personalities. Based on the plots, we found people labeled themselves with five personalities are all tend to show positive emotions like trust, joy, anticipation on Facebook post.

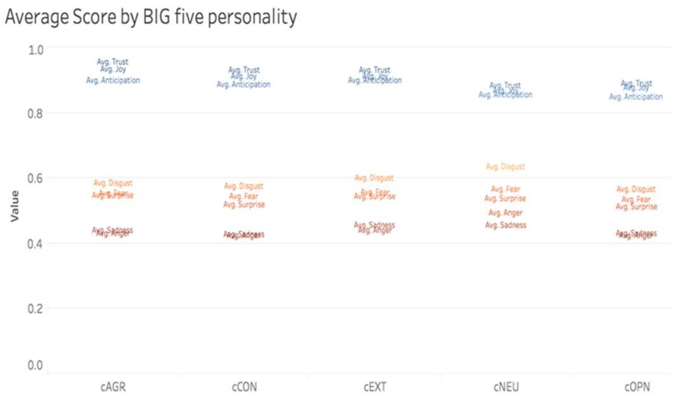


Figure 5 - Average Score by Big Five Personalities

4.3 APPROACH

4.3.1 Approach 1

The emotion scores obtained from the data preprocessing is used to train the classifier from each of the Facebook posts and this is done along with the personality of the author. Here, each post is considered separately and personality is trying to be predicted from such posts. During testing, the majority of the votes is considered of his/ her posts of each post's author. By doing this, we conclude the personality of the author. While doing this, more than 9000 posts were included and 10 fold- algorithm was applied by taking 10% of the posts for testing and remaining as the training data. Since we were considering each post individually here, we were able to include time factors in the prediction. To accomplish that, we included the hour, month, and season in which each post was posted as predictors of personalities in our models. However, this did not seem an optimal approach while detecting the personality using emotions. Thus, we considered another approach which shows some positive results and the limitation of this approach is discusses in the future section of this paper.

4.3.2 Approach 2

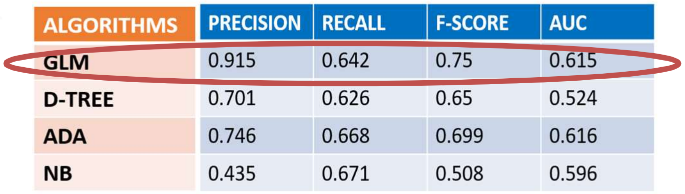
The emotion scores obtained from the posts is averaged for all the individuals with more than 10 posts. Here, we train the classifier using average emotions score from all the Facebook posts for every individual along with their personality. During testing, the personality is predicted from the average of the emotions in their posts. In each fold, 10% of the users are taken into account and not 10% of the posts for testing data, and the remaining 90% is used for training. It is worth mentioning here that we did not include any time factor in this approach since we are averaging all the posts, which were posted in different time segments, for each user. As a result, this approach has produced more accurate results than the first one.

5. EVALUATION AND RESULTS

The five personalities are considered separately, and the algorithms are run for each of them, and the best model for each personality is highlighted. The ROC plot is displayed for the best-suited model for that particular personality. Four different algorithms are run for each personality which includes, Generalized Linear Model(GLM), Decision Tree(D-Tree), ADA, Naive Bayes(NB). The outcome of these algorithms is dichotomous.

5.1 Personality 1: Extraversion

For this personality, different algorithms were run to infer from the emotions. **Table 3** illustrates the performance metrics for 'Extraversion'. As it can be observed from the table that, for logistic regression, the performance metrics such as Precision and F-score is the highest. Although the performance values for recall and AUC is comparable to that of ADA algorithm, GLM is chosen over it since it has a greater F-score and Precision by a larger margin.



ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.915	0.642	0.75	0.615
D-TREE	0.701	0.626	0.65	0.524
ADA	0.746	0.668	0.699	0.616
NB	0.435	0.671	0.508	0.596

Table 3 - Model Performance for Extraversion

The ROC plot for the best model chosen, that is, GLM is plotted in figure 6.

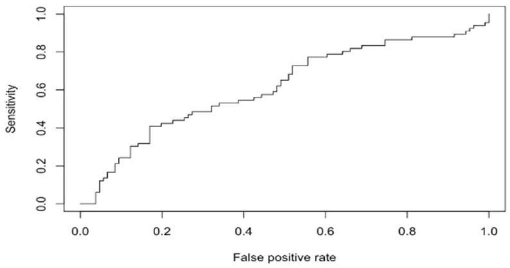
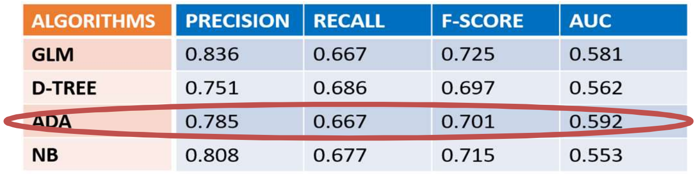


Figure 6 – ROC plot for Extraversion

5.2 Personality 2: Neuroticism

For neuroticism personality, ADA suits to be the best model by taking into account a larger value for AUC performance metric than GLM which although has a higher value for Precision is not preferred over ADA.



ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.836	0.667	0.725	0.581
D-TREE	0.751	0.686	0.697	0.562
ADA	0.785	0.667	0.701	0.592
NB	0.808	0.677	0.715	0.553

Table 4 - Model Performance for Neuroticism

The ROC plot for the best model chosen, that is, ADA is plotted in figure 7.

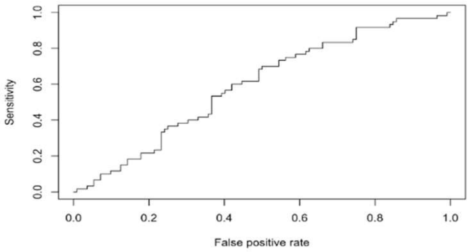
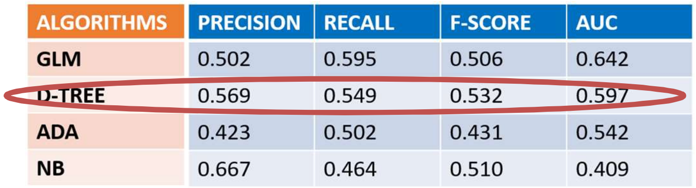


Figure 7 – ROC plot for Neuroticism

5.3 Personality 3: Agreeableness

Decision tree model seems to be the most suited model while predicting Agreeableness. The precision value for D-tree is comparable to that of NB method. Whereas, the recall and F-score value of the best-suited model is little low while compared to GLM. However, when the overall performance is observed and the confusion matrix D-tree seems to outperform others.



ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.502	0.595	0.506	0.642
D-TREE	0.569	0.549	0.532	0.597
ADA	0.423	0.502	0.431	0.542
NB	0.667	0.464	0.510	0.409

Table 5 - Model Peformance for Agreeableness

The ROC plot for the best model chosen, that is, D-tree is plotted in figure 8.

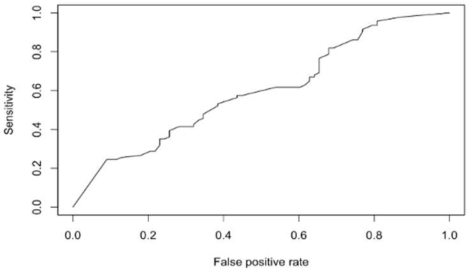
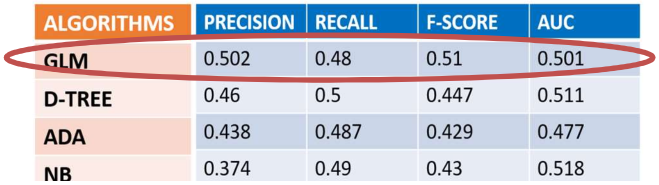


Figure 8 – ROC plot for Agreeableness

5.4 Personality 4: Conscientiousness

The fourth personality, Conscientiousness reflects a person's trait to be active. From the implementation of our algorithm for prediction of personality using emotions, among the four models discussed it's clear from table 6 that GLM is best suited. All the other algorithms have only few performance metric almost same or slightly greater than GLM but when taken all the metrics into account GLM is the preferred model.



ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.502	0.48	0.51	0.501
D-TREE	0.46	0.5	0.447	0.511
ADA	0.438	0.487	0.429	0.477
NB	0.374	0.49	0.43	0.518

Table 6 – Model Performance for Conscientiousness

The ROC plot for the best model chosen, that is, GLM is plotted in figure 9.

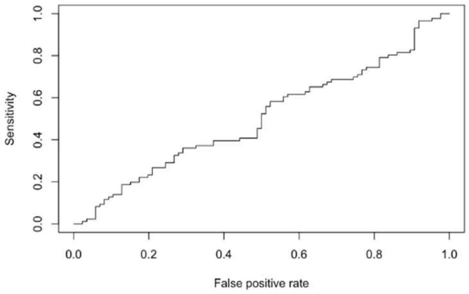
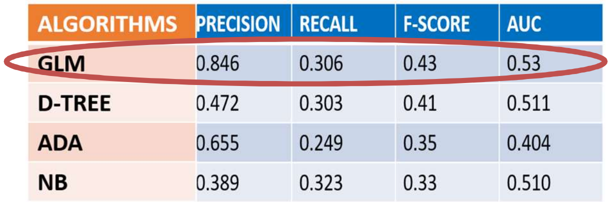


Figure 9 – ROC plot for Conscientiousness

5.5 Personality 5: Openness

For the prediction of last personality, Openness, GLM model clearly outperforms all the other models with respect to precision, F-score and AUC.



ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.846	0.306	0.43	0.53
D-TREE	0.472	0.303	0.41	0.511
ADA	0.655	0.249	0.35	0.404
NB	0.389	0.323	0.33	0.510

Table 7 - Model Performance for Openness

The ROC plot for the best model chosen, that is, GLM is plotted in figure 10.

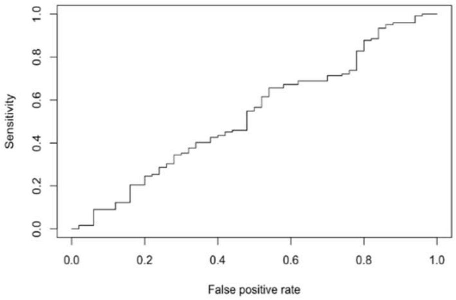


Figure 10 – ROC plot for Openness

6. DISCUSSION

From the results obtained after running the algorithm on the cleaned dataset, it is observed that, predicting personality from emotions although seems difficult but is achievable. Satisfactory results were obtained for few of the personalities, however, greater accuracy couldn't be achieved due to several factors. Lack of feature space makes it complicated to derive prediction models from the available data. The lesser number of feature space diminishes the predictive power for the algorithms to infer an individual's personality from the emotions.

Furthermore, the sparse distribution of the predicted emotions was another such factor which made prediction difficult. Some of the posts were too short to extract any emotion out of it. Furthermore, emoji for emotions was not taken into account which sometimes is a direct representation of the emotion for a post.

These limitations posed a challenge for careful selection of parameters and the number of users for certain number of posts. High precision in the performance of the algorithms for some of the Big Five personalities indicates low False positive which results in high confidence in our predictions especially when the prediction is positive. For example, since our algorithm's precision is 0.846 for openness, when it predicts that some person is open then that person is high likely to have this personality. On the contrary, for some personalities with low precision the prediction accuracy is going to be affected by the distribution of that personality in the training dataset.

Models with high precision can be beneficial for different applications. For instance, it can be used for research studies which requires individuals possessing certain kind of personality as an inclusion criteria. Furthermore, it can also be used by the advertising companies who seek to only target a group of people with a certain kind of personality who can be benefited by the product being marketed or are potential customers for the company.

As mentioned in the methods section, we used two approaches to predict personalities. The first approach, which used each post individually, has failed to produce highly accurate results compared to the second approach, which uses the average of all posts for each individual. This outcome is logical since it reflects reality where we cannot judge someone's personality based on a single post or a single moment. Instead, to reveal someone's personality we would want to include the person's action and behavior over a period of time. Therefore, our second approach seems more promising since it mimics the real life situation.

7. CONCLUSION

Despite the challenges, we could test different algorithms to predict each one of the Big Five personalities. Each algorithm used different set of predictors and finally the algorithm with the best performance measure was chosen. In addition, the network size or the number of friends improved the personality prediction performance by a margin up to 20% for some personalities.

Data preprocessing played a vital role in this project, since we had to extract emotions from each of the individuals' posts and then progressed to predict personalities.

Individuals with higher number of posts have a higher chance of being predicted correctly which indicates more activity on social media reflects more about the individual's characteristics.

An improvement in the approach was presented despite of the challenges faced. This challenge was due to the sparseness in the dataset and no strong predictive features. Future work would include making use of the emoji for extracting emotions from the posts apart from words and trying out the algorithms on other datasets where the feature space is large. Furthermore, depending on the dataset presented, a possible emotion pattern can be generated based on temporal scale.

8. ACKNOWLEDGEMENT

We would like to appreciate the help from Professor Lin during our project process. She provided useful insights and suggestions to process the project. We also appreciate every member in our team who contributes to this project. Mohammed Aldhoayan provided project insights and was responsible for most of the data processing and mining work. Yuqian Zhanfu contributed to the data exploration and cleaning part. Akash Parvatikar worked on some of the data processing and data mining implementation parts. Our group met several times to discuss and come out ideas together for project presentation and paper. Everyone tried their best to work on this project.

9. REFERENCES

1. Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

2. Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, *26*(2-3), 109-142.

3. Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036-1040.

4. Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., Stillwell, D., Moens, M. F., ... & De Cock, M. (2014, January). How are you doing? Emotions and personality in Facebook. In *Proceedings of the EMPIRE Workshop of the 22nd International Conference on User Modeling, Adaptation and Personalization (UMAP 2014)*.

5. Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. J. (2013, June). Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*.

6. Parks-Leduc, L., Feldman, G., & Bardi, A. (2015). Personality traits and personal values: A meta-analysis. *Personality and Social Psychology Review*, *19*(1), 3-29.

7. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

8. Cantador, I., Fernández-Tobías, I., & Bellogín, A. (2013). Relating personality types with user preferences in multiple entertainment domains. In *CEUR Workshop Proceedings*. Shlomo Berkovsky.

9. Wang, Y., & Pal, A. (2015, July). Detecting Emotions in Social Media: A Constrained Optimization Approach. In *IJCAI* (pp. 996-1002).

10. Kozareva, Z., Navarro, B., Vázquez, S., & Montoyo, A. (2007, June). UA-ZBSA: a headline emotion classification through web information. In *Proceedings of the 4th international workshop on semantic evaluations* (pp. 334-337). Association for Computational Linguistics.

11. Lorincz, A., Jeni, L., Szabo, Z., Cohn, J., & Kanade, T. (2013). Emotional expression classification using time-series kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 889-895).