Pertanika J. Sci. & Technol. 25 (S): 237 - 248 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Personality Prediction Based on Social Media Using Decision Tree Algorithm

Tan Lee Chee Yoong, Nor Rahayu Ngatirin* and Zurinahni Zainol

School of Computer Sciences, Universiti Sains Malaysia (USM), 11800 Minden, Pulau Pinang, Malaysia

ABSTRACT

Personality represents the mixture of features and qualities that built an individual's distinctive characters including thinking, feeling and behaving. Traditionally, self-assessment method via questionnaire is the most common means to identify personality. Since recommender systems and advertisement campaigns have evolved rapidly, personality computing has become a popular research field to provide personalisation to users. Currently, researchers have utilised social media data for automatically predicting personality. However, it is complex to mine the social media data as they are noisy, free-format, and of varying length and multimedia. This paper proposes a decision tree C4.5 algorithm to automatically predict personality based on Big Five model. The Big Five Inventory and ZeroR algorithm were included to be served as the baseline for performance evaluation. Experimental evaluation demonstrated that C4.5 performs better than ZeroR in terms of accuracy.

Keywords: Big Five, decision tree, personality, social media

INTRODUCTION

Personality is a word derived from Latin, persona, which means the theatrical mask used by the actors (Ahmad, 2015). It can be defined as a set of attributes that characterise a unique individual's behaviour, temperament, emotions and mental (Mairesse et al., 2007). Personality differentiates an individual from others in characteristic patterns of thinking, feeling and behaving. There are many different personality models used to characterise personality such as the Big Five model (Five-factor model) (John et al., 2008) and the Myers-Briggs Type Indicator (MBTI) (2016). The Big Five model conceived by Tupes and Christal (1961) consists of five traits which are the openness to experience, conscientiousness, extraversion,

ARTICLE INFO

Article history:
Received: 25 October 2016
Accepted: 17 March 2017

E-mail addresses:
tlcyoong.ucom12@student.usm.my (Tan Lee Chee Yoong),
nrn14_com057@student.usm.my (Nor Rahayu Ngatirin),
zuri@usm.my (Zurinahni Zainol)
*Corresponding Author

ISSN: 0128-7680

agreeableness and neuroticism. Each of these dimensions has their own set of poles (low/high) and will be discussed further in the next section. Meanwhile, the MBTI developed by Briggs and Myers in 1950s measures the preferences on four dichotomies which are energising (introversion/extraversion), attending (sensing/intuitive), deciding (feeling/thinking) and living (judging/perceptive) (Bishop-Clark, & Wheeler, 1994). In this study, the Big Five model was selected as it is one of the most well-researched and well-regarded measures of personality structure (Golbeck, Robles, & Turner, 2011). Many psychologists have come to a consensus of it is the current definitive personality model as it is essentially correct in its representation of the structure of traits (McCrae & John, 1992).

Many studies have related personality to various real-life behaviours including Internet usage (Tan, & Yang, 2012), personality and privacy concerns (Sumner, Byers, & Shearing, 2011), movie preferences (Golbeck, & Norris, 2013) and a correlation between personality and job performance (Barrick, & Mount, 1991). These indicate that the personality model is beneficial in capturing the important aspects of an individual and its life dimension, thus triggering the interest of the computing community.

Since the last decades, social media has become a major communication tool of human beings. It consists of a group of internet-based applications which serves multiple purposes and is categorised into collaboration, social commerce, blog platforms, social networks and wikis (Kaplan, & Haenlein, 2010). It has been utilised as the platforms for creating and sharing of the users' generated contents. With hundreds of millions of people spending countless hours on social media to share, communicate, interact, and create data at an unprecedented rate, social media have become one unique source of big data (Zafarani et al., 2014). The useful knowledge extracted from big data may lead to a more confident decision making. This makes the social media mining become a popular field for research. Social media mining is the process of representing, analysing and extracting actionable patterns from social media data (Zafarani et al., 2014). However, it is complex to mine these social media data as they are noisy, free-format, of varying length and multimedia (Zafarani et al., 2014). Many different mining techniques have been proposed to mine these semi-structured data from social media including Naïve Bayes, classification trees, and association rules. Performance of each technique needs to be determined to ensure only the most accurate and meaningful data are extracted.

The aim of this paper is to develop a personality prediction model based on the Big Five model. To achieve this, 107 undergraduate students completed the Big Five Inventory (BFI) (John et al., 1991) and their profile data were extracted from Twitter. However, only 100 participants' data were qualified to be used in this research. The other seven participants' data were eliminated due to some reasons including unintentional repeated entry, missing Twitter account username, and unresponsive friend requests. Using the profile data as a feature set, we were able to train the data using implemented decision tree C4.5 algorithm to predict the students' personality traits. At this moment, our proposed model is only considering extraversion dimension due to the limitations associated to the integration with education-related decision-making system that is currently being developed by our team. Thus, each node on the decision tree represents the Twitter profile attributes and the personality dimension of extraversion. Finally, the evaluation on the precision of the proposed algorithm against other prediction algorithm is carried out.

Personality Model

Personality indicates individual's preferences and may influence his/her decision making. Efforts were put in generating a descriptive personality model or taxonomy in which personality can be understood in a simpler way (John & Srivastava, 2008). Based on the lexical hypothesis, Allport and Odbert (1936) proposed that the most important individual differences are encoded in the language which represents the personality on a set of adjective terms. Then, Cattell (1943) used the Allport and Odbert's list and began the introduction of the Big Five model. Fiske (1949) constructed much simplified descriptions from Cattell's list of personality factors. To clarify these factors, Tupes and Christal (1961) found five relatively strong and recurrent factors, after doing an analysis of the correlation matrices from eight different samples. These factors eventually became known as the Big Five (Goldberg, 1981) which consists of five dimensions. After decades of intensive research, the psychologists reached the consensus on using the Big Five model with the five dimensions of personality trait to describe individual's personality. The Big Five personality traits are characterised by the following (John & Srivastava, 2008):

- Openness to experience: Intellectual, imaginative, and independent-minded.
- Conscientiousness: Orderly, responsible, and dependable.
- Extraversion: Talkative, assertive, and energetic.
- Agreeableness: Good-natured, cooperative, and trustful.
- Neuroticism: Moody, tense, neurotic, and, not confidence.

In the recent years, many previous works exist that infer users' personality from social media which studied the relations between individuals' personality and their interactions with the social media based on the Big Five model. Among social media sites, Twitter has become a popular social media for extracting data because it allows users to view anything about anybody even though they protect their tweets. The retweet feature in Twitter enables protected tweets to leak to the public by 'copy and paste' method of the text of protected tweets into other's own Twitter feed (Meeder et al., 2010). Due to this, even though one does not follow a specific protected tweet, he/she might somehow access the information later through the retweet feature. Literature shows that several studies extracted data from Twitter and used Twitter's feature set to study the relationship between personality and their interactions with this social site. For instance, Quercia et al. (2011) predicted the Big Five personality traits of Twitter users with the basic network properties such as followings, followers, and listed counts. Similarly, Golbeck et al. (2011a) also predicted personality using the profile data such as number of followers, number of followings, density of the social network, and number of hashtags. Lima and deCastro (2013) proposed a system to predict personality in tweets using linguistic information such as sentiment words, social processes, and family words. Celli and Rossi (2012) also correlated the Big Five personality trait of neuroticism and users' interactions in Twitter through the linguistic analysis of tweets. In addition, Sumner et al. (2012) predicted the dark triad personality from the Twitter profile attributes and the linguistic analysis of tweets. On the other hand, Gou, Zhou, and Yang (2014) automatically derived three types of personality traits from Twitter, including Big Five personality, basic human values, and fundamental needs based on individual's word choices

in written samples. Finally, Chen et al. (2015) demonstrated the Twitter's derived personality traits of openness and neuroticism for ad targeting using several features such as likelihood to click the link, follow the account, and a short written explanation for users' reported likelihoods.

Other than Twitter, researchers also employed Facebook data to infer individuals' personality. According to Quercia et al. (2011), Facebook differs from Twitter as it generally connects people who already know each other (e.g., friends, family, and co-workers); i.e. they need to be mutual friends on Facebook to fully share what they have been up to. However, they also mentioned that it is also possible to accurately predict individual's personality from information on Facebook even though the access is generally restricted. A study by Golbeck et al. (2011) predicted the Big Five personality traits of Facebook users based on several features such as structural features, personal information, activities and preferences, language features, and internal Facebook statistics. Using linguistic analysis, Sumner et al. (2011) studied the relationship between Facebook activity and personality traits, and they affirmed that there is a correlation between the two. Additionally, Alam, Stepanov and Riccardi (2013) employed bag-of-words approach and used tokens such as internet-slangs, smiles, and emoticons as features in classifying Facebook user's personality. In turn, in Celli and Polonio (2013), a Personality Recognition from Text (PRT) was presented to predict personality using the linguistic features in Mairesse et al. (2007). This work also addressed how users' personality determines their interaction and communication in Facebook. Besides Twitter and Facebook, Nie et al. (2014) used Sina Microblog and explored the unlabelled data to improve the prediction accuracy. The extracted features in their work were personal profile, social circles, social activities, and social habit.

There are several common algorithms that have been used by researchers for personality prediction. For example, M5' Rules has been used by Golbeck et al. (2011), Sumner et al. (2011), and Quercia et al. (2011), while Lima and deCastro (2013) and Sumner et al. (2012) adopted Naïve Bayes to automatically predict users' personality. Studies such as those by Golbeck et al. (2011, 2011a) and Sumner et al. (2011) applied Gaussian Processes to investigate the relationship between personality and their interactions with the social media. Golbeck et al. (2011a) also applied ZeroR in predicting Twitter users' personality. A recent work by Celli and Rossi (2012), and Celli and Polonio (2013) proposed a personality recognition algorithm from the linguistic analysis of tweets and texts to predict personality using the linguistic features in Mairesse et al. (2007). Meanwhile, Nie et al. (2014) adopted local linear semi-supervised regression algorithm for personality prediction. Chen et al. (2015) also utilised linear regression for predicting two dimensions of Big Five. Several classification methods including Sequential Minimal Optimisation (SMO), Bayesian Logistic Regression (BLR), and Multinomial Naïve Bayes (MNB) have been used by Alam et al. (2013) for automatic recognition of Big Five personality traits. Meanwhile, Gou et al. (2014) adopted lexicon-based approach for personality prediction and sharing preference of Twitter users. Finally, Sumner et al. (2012) also employed Sequential Minimal Optimisation (SMO), Random Forest, and J48 to predict the dark triad personality.

The most similar works with our current study are by Quercia et al. (2011), Golbeck et al. (2011a) and Sumner et al. (2012), but our work is focusing on predicting personality traits of Big Five model from students' social media. Additionally, in our work, we developed and

implemented C4.5 algorithm, whereas they performed classification using classifiers in WEKA (Hall, 2009). To the best of our knowledge, no study has been implementing C4.5 algorithm for personality prediction. Next, we present the process of automatic prediction of personality in the following section.

METHOD

This paper focuses on an approach to automatically predict personality based on social media data. Decision tree C4.5 algorithm and Big Five model were used for personality classification. Data were collected from two main sources: i) Personality from BFI; and ii) Twitter. Figure 1 illustrates the overall processes involved in the students' personality prediction framework.

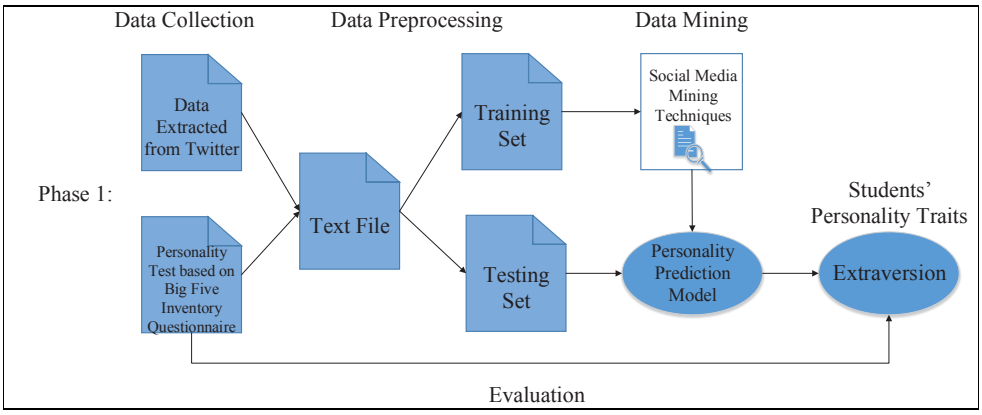

Figure 1. Framework of the student's personality prediction

A survey form was created using the Google Form and distributed online to undergraduate students through the Info Sharing Facebook Group. The survey form includes the BFI (John et al., 1991) questionnaire and their Twitter account username. As mentioned in the previous section, our research only focused on the trait of extraversion of Big Five model. Hence, only the eight questions regarding the trait of extraversion were adopted. The questionnaire begins with the statement, *"I see myself as someone who..."*, and is followed with eight phrases representing the behaviours of respondents. The respondents rated each question on Likert scales of 1 to 5 representing the respondents' level of agreement to each phrase, where 1 indicates "Disagree Strongly" and 5 indicates "Agree Strongly". The eight phrases related to the trait of extraversion include:

- Is talkative
- Is reserved (R)
- Is full of energy
- Generates a lot of enthusiasm with
- Tends to be quiet (R)
- Has an assertive personality
- Is sometimes shy, inhibited (R)
- Is outgoing, sociable

Data extracted from Twitter consisted of Twitter profile attributes such as the number of followers, number of followings, number of tweets, number of lists and number of likes of students. These data were extracted using the Twitter Rest API, Tweepy (a Python library for accessing the Twitter API).

Data file from the survey forms were processed by filtering out those responses that did not provide Twitter account name and repeated responses were also eliminated. All the data were checked to ensure there were no missing values. If there was any, the default value was used to replace the missing value. The total score of eight questions from the BFI was summed up based on the respondents' ratings after appropriately reversing the scores of the questions that denoted with "R" (John et al., 1991), as shown above. For example, when a respondent rated the phrase "Is reserved" on scale of 5, this rating was inverted to 1 and vice versa; a respondent rated the phrase "Tends to be quiet" on the scale of 4, this rating was inverted to 2, and vice versa. The final score indicating the subject's trait of extraversion/introversion would then be manually calculated using the formula provided in John et al. (1991). The scores calculated indicate the student is extrovert if the score is above 50, and vice versa.

The Pearson correlation analysis was used to calculate the significant score between subjects' personality scores and each of the features obtained from analysing their tweets and public account data. Among the five features, the number of followers contributes the highest correlation value to the trait of extraversion (Pearson $r = 0.422$, $p < 0.05$), followed by the number of followings (Pearson $r = 0.389$, $p < 0.05$), number of lists (Pearson $r = 0.326$, $p < 0.05$) and number of tweets (Pearson $r = 0.305$, $p < 0.05$). Both number of followers and number of followings showed significant positive correlation to extraversion trait. These indicate that extraverts tend to be outgoing, sociable, and like to make more friends compared to those low in extraversion (introverts). In contrast, the number of favourites (number of likes) shows no significant linear correlation to the extraversion trait (Pearson $r = 0.056$, $p > 0.05$). As a result, only four features with the significant correlation (number of followers, number of followings, number of tweets and number of lists) were selected to classify the students' personality. Table 1 shows the correlation values between the profiles attribute features and the personality trait of extraversion. Correlations that are statistically significant for $p < 0.05$ are bolded.

Out of 100 students' data, 70 of them were selected for the training set. In this training set, 37 subjects were with high level of extraversion, while 33 other subjects wee with low level of extraversion (introversion). As for the testing set, there were 16 subjects with high level of extraversion and 14 subjects with low level of extraversion. The students' data were manipulated such as getting more responses and eliminating some responses to ensure the data set is relatively balanced between each class. Besides, the data for the training set were also chosen to ensure it is balanced between classes to avoid the problem of training biased toward one class which can lead to inaccurate classification.

Table 1
Correlation coefficient values between feature scores and trait of extraversion

Personality Trait	Profile Attribute Sub-features	Correlation Coefficient
Extraversion	Number of followers	0.422
	Number of followings	0.389
	Number of tweets	0.305
	Number of lists	0.326
	Number of favorites	0.056

The calculated personality traits and data extracted from the Twitter were combined into a new CSV file. Next, the students' personality prediction model was constructed using the C4.5 algorithm, which was implemented in Java language using NetBeans IDE 8.1. The algorithm first read the input file and stored all the data into different variables. Table 2 shows the sample data file consisting of the training set which consists of the four aforementioned attributes and calculated poles (low/high) of extraversion dimension of each subject.

Table 2
Sample training set used to train students' personality prediction model

No. of followers	No. of followings	No. of tweets	No. of lists	Class
349	173	20775	0	High
130	299	840	4	High
241	92	3172	6	High
571	327	7978	9	High
440	185	13899	2	High
155	129	633	0	High
123	123	709	1	High
36	78	1204	0	Low
19	13	133	0	Low
133	166	705	0	High

After all the data had been stored, the algorithm calculated the overall entropy of the training data set and the information gain of each and every attribute. Since the training set involved numerical data, in order to identify the information gain of an attribute, the algorithm identified all the splitting points of the attribute and then calculated the information gain for every split on each of the splitting points. Next, the splitting point with the highest information gain was selected as the splitting point and information gain of the attribute. After the splitting point and information gain of all the attributes had been identified, an attribute with the highest information gain was selected as the first node of the decision tree. These processes were then repeated for the left sub-tree of the first node followed by the right sub-tree of the first node until the decision tree was completely built. Figure 2 shows the decision tree generated from the implementation of the prediction model which represents the students' personality prediction

model. The constructed personality prediction model was then tested by the algorithm to read another file which consists of the test data set. At the end of the algorithm implementation, personality traits of the test data were successfully predicted.

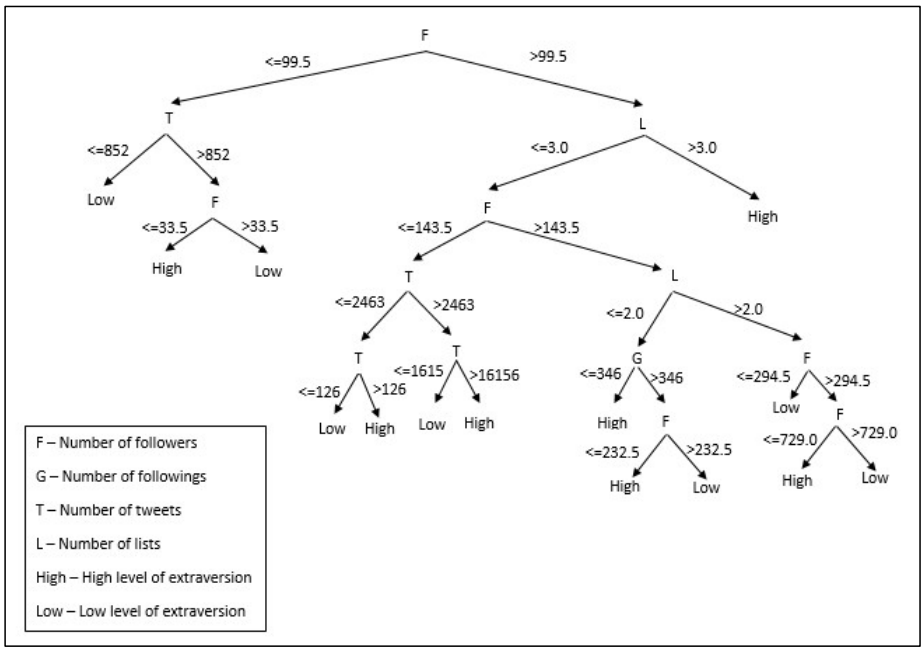

Figure 2. The decision tree generated from the students' personality prediction model

RESULTS AND DISCUSSION

The accuracy and weighted F-measure value of various classifiers in predicting the students' personality trait of extraversion are shown in Table 3. From the table, we see that the ZeroR algorithm as the baseline produced a moderate accuracy at 53.33% with a quite low weighted average F-measure value of 0.371. Both J48 classifier implementation in WEKA and C4.5 algorithm implemented in this research produced higher accuracy and weighted average F-measure value above the baseline. The J48 classifier has better performance with higher accuracy of 86.67% with weighted average F-measure value of 0.865 as compared to the C4.5 algorithm which has generated prediction result with 73.33% of accuracy and weighted average F-measure value of 0.733.

As shown in Table 3, the implemented C4.5 algorithm achieved a lower accuracy and weighted average F-measure value than the J48 classifier implemented in WEKA machine learning toolkit. The result is due to the use of unpruning method in the implementation of C4.5 algorithm. The tree generated from the C4.5 implementation is not pruned and it causes a problem called overfitting, which may led to a less accurate personality and learning style classification. This may affect the accuracy of results (Patil et al., 2010). In contrast, tree pruning used in J48 classifier, which converts a large tree into smaller tree and has eliminated those meaningless rules, generally results in faster and more accurate classification (Patil et al., 2010).

Table 3
Classification results comparison in predicting personality trait of extraversion

Algorithm	ZeroR* in WEKA	J48 in WEKA	C4.5 algorithm implemented in this research
Accuracy (%)	53.33	86.67 (v)	73.33
Weighted Average F-measure	0.371	0.865 (v)	0.733

Remark 1. *. Baseline
Remark 2. v. Victory (Significant above baseline at $p < 0.05$)

Based on the results, it was found that personality trait of extraversion is positively correlated with most of the Twitter profile attributes, except the number of favorites (number of likes) with the weakest linear relationship (refer Table 1). The findings are similar to the results generated from the work by Sumner et al. (2012), which showed that followers, friends (followings), number of tweets, and number of lists are significantly correlated to extraversion, while the number of favourites did not show significant relationship. Both our study and Sumner et al.'s (2012) showed that the number of followers has the highest correlation coefficient to the extraversion trait compared to other profile features. Additionally, Quercia et al. (2011) also found that the listeners (those who follow many users) and popular (those who are followed by many) users have the strongest and significant correlations with the personality trait of extraversion. Therefore, it supports the inference that extroverts have many people following them. In the same paper, they explained that the personality trait of extraversion is the predictor for number of friends in the real world and social network. Finally, the work by Golbeck et al. (2011a) also suggested that extroverts who are outgoing and sociable tend to have more friends. Hence, the results from this research are consistent with those observed from the previous studies.

In terms of evaluation, it is difficult to compare and critically evaluate the practical performance and precision of the aforementioned studies due to the use of different evaluation methods. While Quercia et al. (2011) applied the Root Mean Square Error (RMSE) to measure the performance of their work, Golbeck et al. (2011a) employed the Mean Absolute Error (MAE) in their study. In contrast, according to Sumner et al. (2012), evaluation methods such as MAE and RMSE can mask larger errors at the extremes of a unimodal population distribution by predicting the majority of instances around the mean value. This means that, for instance, the model may predict a high extraversion as a low introversion without significantly affecting the overall MAE. Due to this reason, they adopted several evaluation criteria such as Accuracy (Acc) for both the maximum Geometric and Arithmetic means and were presented in both median split and 90th percentile split classification. In conclusion, due to the fact that each error measure has weaknesses that can produce inaccurate evaluation of the predicting results, it is impossible for the researchers to choose only one measure (Mahmoud, 1984).

CONCLUSION AND FUTURE WORK

This study has resulted in one main insight, i.e. it is possible to predict students' personality trait from the public information they share on Twitter. The finding shows that J48 performs

better than C4.5 because of the use of unpruning method in the implementation of C4.5 algorithm. As discussed earlier, there ae many previous studies working on the algorithms to automatically predict personality using social media data. Regardless of this, we have found an opportunity to explore C4.5 algorithm in the automatic personality prediction. However, the current work does not consider all the dimensions of the Big Five model. The ongoing works aim at resolving the issues with the other dimensions of the Big Five model. Other algorithms are also ventured on for personality prediction to improve accuracy. In addition, although the proposed work has been tested for a maximum of 100 students, further work must consider more participants so as to provide a more accurate representation of the entire population. With the ability to infer a students' personality trait, currently we are integrating personality with education-related decision making such as predicting students' learning styles and suggesting teaching strategies that are tailored to suit their learning styles.

REFERENCES

Ahmad, Z. I. (2015). *Prediction Models of Extraversion and Neuroticism of Malaysian Facebook Users*. Masters Thesis, School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia.

Alam, F., Stepanov, E. A., & Riccardi, G. (2013). Personality Traits Recognition on Social Network – Facebook. *Computational Personality Recognition*, 6-9.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*, 1-178.

Barrick, M. R., & Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology, 44*, 1-26.

Bishop-Clark, C., & Wheeler, D. D. (1994). The Myers-Briggs Personality Type and Its Relationship to Computer Programming. *Journal of Research on Computing in Education, 26*(3), 358-369.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476-506.

Celli F., & Rossi, L. (2012). The Role of Emotional Stability in Twitter Conversation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, 10-17.

Celli, F., & Polonio, L. (2013). Relationships between Personality and Interactions in Facebook. S*ocial Networking: Recent Trends, Emerging Issues and Future Outlook*, 41-53.

Chen, J., Haber, E., Kang, R., Hsieh, G., & Mahmud, J. (2015). Making Use of Derived Personality: The Case of Social Media Ad Targeting. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 1-10.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44*, 329-344.

Golbeck, J., & Norris, E. (2013). Personality, Movie Preferences, and Recommendations. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* Niagara, Ontario, Canada, 1414-1415.

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing System*. Vancouver, BC, Canada, 253-262.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011a). Predicting Personality from Twitter. *Proceedings of the IEEE International Conference of Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 149-156.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology 2* (pp. 141-165). Beverly Hills, CA: Sage.

Gou, L., Zhou, M., & Yang, H. (2014). KnowMe and ShareMe: Understanding automatically discovered personality traits from social media and user sharing preferences. *CHI 2014*. Toronto, ON, Canada, 955-964.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

John, O. P., & Srivastava, S. (2008). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research 2* (pp. 102–138). Guilford Press, New York.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory--Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research 2* (pp. 114-158). Guilford Press, New York.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*, 59−68.

Lima, A. C. E. S., & deCastro, L. N. (2013). Multi-Label Semi-Supervised Classification Applied to Personality Prediction in Tweets. *Proceedings of the BRICS Congress on Computational Intelligence & 11th Brazilian Congress on Computational Intelligence.* Recife, Brazil, 195-203.

Mahmoud, E. (1984). Accuracy in forecasting: A survey. *Journal of Forecasting, 3*(2), 139-159.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automation Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research, 30*, 457-500.

McCrae, R. R., & John, O. P. (1992). An Introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

Meeder, B., Tam, J., Kelley, P., & Cranor, L. F. (2010). RT@ IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network. *Proceedings of the Web 2.0 Privacy and Security Workshop co-located with IEEE Symposium on Security and Privacy*, 285-299.

Nie, D., Guan, Z., Hao, B., Bai, S., & Zhu, T. (2014). Predicting Personality on Social Media with Semi-supervised Learning. *Proceedings of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Warsaw, Poland, 158-165.

Patil, D. D., Wadhai, V. M., & Gokhale, J. A. (2010). Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy. *International Journal of Computer Applications, 11*(2), 23-30.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *Proceedings of the IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 180-185, IEEE Press, Boston.

Sumner, C., Byers, A., & Shearing, M. (2011). Determining personality traits & privacy concerns from Facebook activity. *In Black Hat Briefings*, Abu Dhabi, United Arab Emirates, 1-29.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets. *Proceedings of the 11th IEEE International Conference on Machine Learning and Applications*, 386-393.

Tan, W. -K., & Yang, C. -Y. (2012). Personality Trait Predictors of Usage of Internet Services. *Proceedings of the 2012 International Conference on Economics, Business Innovation*. Kuala Lumpur, Malaysia, 185-190.

The Myers & Briggs Foundation. (2016). MBTI® Basics. Retrieved from http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/

Tupes, E. C., & Christal, R. C. (1961). Recurrent personality factors based on trait ratings. Technical Report, USAF, Lackland Air Force Base, TX.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining: An Introduction (Draft version). Cambridge University Press. Retrieved from http://dmml.asu.edu/smm