Personality Mining from Biographical Data with the "Adjectival Marker " Technique

Shivani Poddar, VenuMadhav Kattagoni and Navjyoti Singh

Center for Exact Humanities, IIIT Hyderabad
shivani.poddar92@gmail.com, venumadhav.kattagoni@gmail.com, singh.navjyoti@gmail.com

Abstract

The last decade has witnessed significant work in personality mining from lexical cues in social media data. Not much work has yet been undertaken in extracting these lexical cues from biographical data populating social media. Most of this work involves a large crowd of researchers leveraging dictionary-based approaches such as LIWC (which primarily focus on function words). By means of this paper we intend to introduce a novel method of personality mining from social media data called "Adjectival-marker Technique". This method involves extracting lexical features from descriptive texts (e.g. biographical data) to train a learning model, so as to predict the respective personality traits of the subject. Conceptually, it draws heavily from the last 78 years of work in lexical psychology and the Big Five personality test. However, it is not only a computational variant of the primordial theories of lexical psychology, but is also competent in conferring a substantial accuracy of personality prediction, matching that obtained by psychometric tests. In this study, we propose a variant of the Lexical Hypothesis from psychology. This modified hypothesis is validated by the computational results of personality prediction achieved by the Adjectival Marker Technique discussed below. The paper also discusses some insights illustrating the coherence of people's judgments about the subject's personality (virtual personality). The average accuracy (i.e. matching that achieved by psychometric tests for Big 5) for prediction approximated to Extraversion - 82.82% Agreeableness - 89.62%, Conscientiousness - 92.48% and Imaginativeness/Intellect - 81.67%.

Keywords: Social Computing, Psychology, User Personality Determination, Natural Language Processing, Machine Learning

1. Introduction

1.1. Motivation

Social Media has become the most abundantly used means of communicating and propagating information online. Most information here is extensively descriptive of the users who channel themselves through it. It is not only the user who gives away information about himself (Goldbeck et al, 2011), but also his peers (Staiano et al, 2012). This paper mainly unravels how the latter approach is nearly an absolutely accurate predictor of certain personality traits. The judgements of not only peers but of people who know us remotely over time can be an important window into solving the labyrinth of our personalities. The future of social media will witness individuals choosing workplaces, friends, books, movies, products etc, in synchrony with their own personalities. The tomorrow of the advertising industry will witness a transformation from "spammers" to "personalized suggestors". This has also been cited in various discussions wherein advertisers are advised to study personalities instead of demographics (documented in the paper personalized persuasion, (Jacob, 2012)). The aforementioned applications are just a tip of the iceberg. Relationships have been discovered between personality and psychological disorders, job performance (Digman et al, 1990) and satisfaction (John et al, 1990), and even romantic success. An extremely dynamic field of study which also benefits from the research in the area of Human Computer Interaction (HCI) is interface design. Many interface designing projects revolve around modelling interfaces based on people's personality oriented preferences. This study, thus, aims to contribute to bridge this gap between biographical data and personality research. We also attempt to expedite the process of personality prediction, making it more automated instead of relying heavily on psychometric tests written by the subject.

1.2. The Big Five Personality Model

There have been several personality models (The Big Three, The Big Five, The Alternative Five, etc.) that claim to encapsulate the traits that need to be summoned so as to effectively predict user personalities from social media data. However, out of all these models, the most robust and tested model, which has been consistent for the last few decades, is the Big Five (Big5) model (Goldberg et al, 1992). This personality model, being one of the most supported in lexical psychology research, stood out as being most resilient to carry out research of biographical social media resources (Saucier et al, 1996). Another one of the instrumental personality theories that has spawned the landscape of personality models is the set proposed by Carl Jung (Myers-Briggs Type Indicator (MBTI), Socionics, Kiersey et al, 1921). Following the paucity of data (for evaluating our model) available for personality determination via reliable psychometric tests for the Big 5 model, we decided to refer to a publicly published research dataset,[1] that abundantly provided us with the MBTI personalities for people. So as to bridge this gap between the MBTI (for personalities which needed to be used for evaluation) and Big 5 (the personalities which were being predicted by our model), (Capraro et al (2002), Furnham et al (1996), McCrae et al (1989)) we used correlations shown in Tables 1 and 2. Thus, one of the major motivations of this paper is also to draw the most effective traits (namely: Extraversion, Agreeableness, Conscientiousness and Imaginativeness) from the intersection of these two instrumental

[1]The dataset can be found at http://www.celebritytypes.com.

paradigms of personality qualifiers. Hence, in scope of this study, the traits we predict are Extraversion, Agreeableness, Conscientiousness and Imaginativeness/Intellect.

1.3. Motivation for using Biographical Data

This research builds on the confluence of two major domains, the primordial theories of the lexical hypothesis and the recent computational techniques of data modeling. Allport's personality trait names (Allport et al, 1936) lead to the creation of Goldberg's adjective marker (Goldberg et al, 1992) and have ignited various studies. Goldberg et al (1990), Digman et al (1990), John et al (1990), Ostendoff et al (1990) built on the same foundation. All of these converge at a single point that cites a "descriptive", "adjectival" lexicon to be the key into a person's personality. Social media today is littered with biographical or descriptive content of its over 1.4 billion users. Tapping this reservoir of content by the principles and techniques discussed below, the paper aims at unveiling a substantial part of this personality descriptive content.

1.4. Proposed modification in the "Lexical Hypothesis of Psychology"

The theories of psychology were influenced by various revolutionary concepts, for instance, "trait" - a theoretical construct which describes a basic dimension of a person's personality (Allport, 1937). The idea of trait gave birth to the "Lexical Hypothesis of Psychology". The initial direction of this paper was solely governed by this exact hypothesis (worked upon by Klages, 1926/1932; Cattell, 1943; Norman, 1963; Goldberg, 1982) -

> *"Those individual differences that are most salient and socially relevant in people's lives will eventually become encoded into their language; the more important such a difference, the more likely is it to become expressed as a single word."*

The Lexical Hypothesis has been used in its entirety in author's personality prediction systems, like the one for Greek Language described by Kermanidis et al, (2012). Motivated by the same inspiration, we too expected to extract author's personality traits from the text they wrote. This involved mobilizing huge datasets of web blogs and essays and extracting "names" from them to determine the author's personalities. However, by the course of our study, we found out that this was not as effective as the initial hypothesis proposed (Goldberg et al, 1982). The average accuracy of the initial experimentation was less than 50%, which was as good as a randomly predicted personality set.

Thus, we propose a modification of the Lexical Hypothesis in psychology which suggests that the personality of a person is predicted based on cumulative judgements of various authors about him/her. These judgements are indicative of the respective traits of the person described along the lines of the Big5 personality Model. The "Adjectival Marker " Technique helps us unravel these judgements, and is derived from the adjectival markers of Big5 personality traits as discussed by Goldberg & Saucier (1996). Thus, the modified Lexical Hypothesis of Psychology proposed and verified in this paper is as follows:

> *"Those individual differences that are most salient and socially relevant in people's lives will eventually (over time) become encoded into their language as well as that of people who describe them (via the knowledge they have of them, these people could be peers, associates, friends, family members, followers etc.); the more important such a difference, the more likely is it to become expressed as a single word".*

The "Adjectival Marker Technique" introduced in this paper is most accurate when it is used to analyze the personality of the subject who the social media resource is descriptive of and not the author himself. We also inferred an interesting observation that suggested that the views of different people describing the subject are coherent amongst themselves and also with the results of the psychometric tests. The average accuracy of the traits, based on the proposed hypothesis, for a series of data spread temporally and spatially (as compared to the results obtained by psychometric tests) in social media came out as discussed below.

1.5. Structure of the Paper

We begin by presenting a brief background on the Lexical Psychology theories of personality determination and related work on personality in conjuction with social media in Section 2. We then present our dataset in Section 3 & 4 and methodology for analyzing, quantifying and modelling biographical data content for 574 personalities in Section 5. The study proceeds on to describe the adjectival features used along with the machine learning techniques for classification and demonstrate significant improvements that the model was able to achieve over baseline classification on each personality factor. In subsequent sections, the paper presents the results in Section 6 and analysis of the study, and discusses the methods we incorporated which were instrumental in escalating the accuracy of the model for each of the traits discussed earlier in Section 7. We finally wrap up the paper with brief discussions about the future work, sparked by this study in Section 8.

2. Related Work

The last few years have witnessed a considerable escalation in studies which are directed at mining user personalities from social media data. Those which are related to this work can be mined in mainly 2 sections. (i) Studies which are based on lexical cues to mine author's personality, (ii) Studies which have used social media based features to study the personality of the user.

The former section includes work by Tausczik and Pennebaker (2010) wherein they mined author personality via LIWC (Linguistic Inquiry and word count) approaches. Another such study used linguistic features such as function words, deictics, appraisal expressions and modal verbs to classify 2 of the Big Five traits namely neuroticism and extraversion (Argamon et al, 2005). Oberlander & Nowson (2006) classified extraversion, stability, agreeableness and conscientiousness of blog authors' using n-grams as features and Naive Bayes algorithms. Mairesse et al, (2007) reported a long list of correlations between Big5 personality

traits. They obtained those correlations from psychological factor analysis on a corpus of Essays and audio cues (Pennebaker & King 1999) to develop a supervised system for personality recognition. Luyckx et al, (2008) built a corpus for stylometry and personality prediction from text in Dutch using n-grams of Part-Of-Speech (POS) and chunks as features. They used the MBTI schema in place of the Big5 (it includes 4 binary personality traits, see Briggs & Myers (1980)). Along the same lines, Iacobelli et al, (2011) used as features, word n-grams extracted from a large corpus of blogs, testing different extraction settings, such as the presence/absence of stop words or inverse document frequency. They found that bi-grams, treated as Boolean features and keeping stop words, gave substantial results using Support Vector Machines (SVM) as learning algorithm. Kermanidis (2012) followed Mairesse et al, (2007) and developed a supervised system for POS tagging in Modern Greek, based on low level linguistic features, such as Part-of-Speech tags, and psychological features, like words associated to psychological states like in LIWC. Kermanidis (2012) also somewhat operated along the lines of Lexical Hypothesis by mining author personalities via KMeans clustering algorithms.

Personality Analysis in Social Media Analysis is a recently observed phenomenon. Herein, some substantial work was done by Goldbeck et al, (2011) wherein the authors predicted the personality of 279 users from Facebook, using either linguistic or social network specific features. Quercia et al, (2011) used network features to predict the personality of 335 Twitter users, using M5 rules as learning algorithm. Various means of evaluation have been used by the above researchers, ranging from accuracy to AUC (Area Under the Curve) values so as to establish relative accuracies of models against each other. The above have been discussed and captured very effectively by Celli et al, (2013). One important observation which comes to surface while analyzing relevant literature is that, none of the studies so far have exploited the primordial lexical hypothesis and 'adjectival traits' suggested by Saucier et al, (1996). Our work presented in this paper carves a very different niche for itself by computing this very approach of personality adjectives, compressing the last 80 years of psychological research in the lexical front and merging it with the latest computational techniques. This confluence has yielded encouraging results, predicting traits matching those predicted by a psychometric test.

3. Datasets

3.1. Biographical Data Mobilization

The data collected as a part of this study was by means of a Python-based crawler. We first used a simple web crawler to get a list of web-pages with the name of the respective "person" as the argument keyword to the crawler. These web pages were then filtered based on their meta-tags. To boost true positives, we only considered the pages which specified their content as "biographical" in the meta-tag descriptors. This resulted in mobilization of few Wikipedia resources, blog mentions and majorly some very descriptive biographical websites. We then manually cleaned the noisy data to assure entity disambiguation and irrelevant mentions. The same has been illustrated by means of Figure 1.

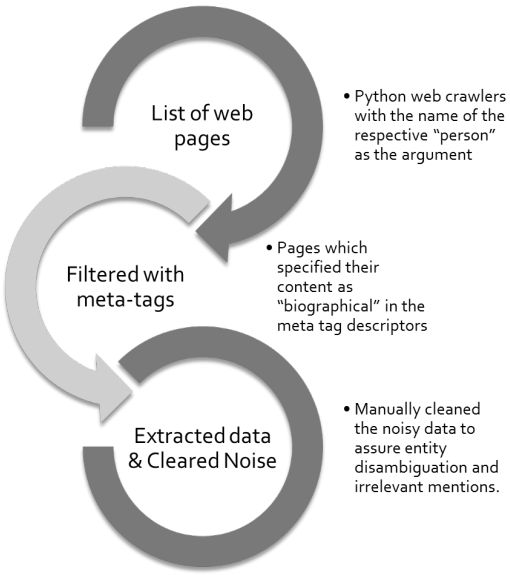


Figure 1: Data Mobilization

3.2. Personality Traits Data

The Jungian Personality functions of 574 personalities were extracted from the resource for eventual evaluation.[2] Since this was one of the most authentic reserves we found consisting of personality listings (so as to evaluate the ones our model predicts) we found it the most effective to be used for evaluating our own model. The "Adjectival Markers" that the paper is based on (as described below) are a proven indicator to reflect the Big5 traits of personality. Thus, to evaluate our computed predictive model via personalities for the respective subjects by an exclusively listed source, we scaled the Jungian Typology type to the closest traits of the Big5 using correlation factors as shown in Table 2 (Hall et al. 2009, Capraro et al. 2002, Furnham et al. 1996, McCrae et al. 1989). Table 1 shows the supporting notations of the personality systems.

Big5/ Global5	Jung/MBTI/Kiersey	Strength of Correlation
Extraversion	Introvert/Extrovert	High
Emotional Stability	Feeling/Thinking	Very Low
Conscientiousness	Judging/Percieving	High
Accommodation / Agreeableness	Feeling/Thinking	Medium
Intellect	Sensing/Intuition	Medium-High

Table 1: Notations for Personality Models

As illustrated, 4 final personality traits were scaled (each of which had medium to high correlation with the MBTI

[2]http://www.celebritytypes.com, wherein extensive cognitive functions have been used to derive the psychology of the given personalities.

Semi-Correlating Descriptions	
Jung/MBTI/Kiersey	Global 5
INFP	RCUAI, RLUAI
INTP	RCUEI, RLUEI
INFJ	RCOAI, RLOAI
INTJ	RCOEI, RLOEI
ISTJ	RCOEN, RLOEN
ISFJ	RCOAN, RLOAN
ISTP	RCUEN, RLUEN
ISFP	RCUAN, RLUAN
ENFP	SCUAI, SLUAI
ENTP	SCUEI, SLUEI
ENFJ	SCOAI, SLOAI
ENTJ	SCOEI, SLOEI
ESTJ	SCOEN, SLOEN
ESFJ	SCOAN, SLOAN
ESTP	SCUEN, SLUEN
ESFP	SCUAN, SLUAN

Table 2: Correlations between Personality traits

types) namely - Agreeableness (Accommodation - A/E), Extraversion (R/S), Conscientiousness (Orderliness - O/U) and Intellect (N/I).

3.3. Adjectival Marker Training Set

The adjectives mined from the biographical data were refined to extract the adjectival markers i.e. specific adjectives descriptive of the subject of the biographical data. These adjectival markers were used as features in the final LASSO logistic regression model. The adjectival markers extracted are based on the work of Saucier & Goldberg, (1996). Table 3 provides the factor loadings of few of the 435 adjectives (Saucier et al, 1996) on each of the five factors as discussed in their work. The order reflects the relative size (variance) of the factors (e.g. Factor II is the highest), and the sign reflects the relative size of the item subsets at each pole of the factor (e.g. the negative pole of Factor IV has more items). We have, as a part of our study, condensed this table to solely indicate whether or not the trait is descriptive of a particular trait, so as to achieve a binary matrix for them (for the respective 4 of the Big 5 traits mentioned above). The binary equivalent for Table 3 is shown in Table 4.

4. Biographical Data

Biographical data was mined for 574 personalities from online resources as discussed in the former Section 3.1. This data was divided into 2 categories. Testing data and Training data. Users with no substantial data (>100 words were discarded for the analysis as of now). The data mining undertaken for acquiring these datasets is spread across various social media resources including Wikipedia articles, blog posts, social Q & A sites and community media sites (sharing biographical book excerpts, for building datasets of word count >10,000)

4.1. Training Data

The training data set, used to mine adjectival markers, comprised of biographic data content of 283 personalities. The word count of the dataset ranged from 500 - 10,000 words. The ratio of the number of adjectives to the total number of words in the dataset ranged from 0 to 0.005.
This data content was mined by means of a Python-based web crawler, which parsed biographic websites, Wikipedia, and social media mentions.

4.2. Biographical Testing Data

The testing dataset comprised of biographic data content of a different set of 291 personalities than the ones used for training. These were mined from the social media reserves like Wikimedia articles, blog posts about the respective personalities, social Q& A sites etc. The word count and the number of adjectives to the total number of words ratio ranged from 100 10,000 words and 0.0001 to 0.003 respectively.

Adjectives*	II	I	IV	V	III
Sympathetic	0.62	0.02	0.07	0.03	-0.05
Kind	0.60	0.07	0.02	0.00	0.06
Sensitive	0.46	-0.10	0.35	0.23	0.00
Rude	-0.50	0.08	0.00	0.06	-0.15
Adventurous	0.00	0.38	-0.19	0.10	-0.04

Table 3: Factor Loadings of 5 of the 435 adjectives presented by Saucier et al (1996). (Factor I - Extraversion, Factor II - Agreeableness, Factor III - Conscientiousness, Factor IV - Emotional Stability, Factor V - Intellect/Imagination)

5. Methodology

The training data (283 users) was mined for adjectival markers according to Saucier's adjectival marker list (Saucier et al, 1996). Personality traits and their adjectival markers were represented as a sparse User-Trait Adjective Matrix for each of the 4 adjectival traits to be predicted. The entries of the respective Trait (say T) matrix were set to 1 if there existed an adjectival marker in the user's descriptive biographical data and 0 if the respective adjectival marker was not there. Thus, each personality trait was contained in a matrix wherein the Row of the matrix M, consisted of adjectival-features and the corresponding column entry consisted of the User-trait. The matrix entity M_{ij} was a binary number which was 1 if the adjectival marker in the i^{th} row indicated the presence of the trait T in the personality of the subject contained in the j^{th} column of the Matrix M. To predict the binary score of a given personality feature, we then performed a LASSO logistic regression (Tibshirani et al., 1996, Meier et al., 2008) analysis in Weka(Hall et al., 2009). A variety of regression algorithms were tested, each with a 10-fold cross-validation with 10 iterations. The best result out of all algorithms was using a binary classifier with Lasso regression (with 10 fold cross validation).
Using the LASSO Technique ensured that there was no overfitting because of extra adjectival features for certain

Adjectives	Agreeableness		Conscientiousness		Extraversion		Imaginative	
	Decimal*	Binary	Decimal*	Binary	Decimal*	Binary	Decimal*	Binary
Sympathetic	0.62	1	-0.05	0	0.02	1	0.03	1
Kind	0.60	1	0.06	1	0.07	1	0.00	0
Sensitive	0.46	1	0.00	0	-0.10	0	0.23	1
Rude	-0.50	0	-0.15	0	0.08	0	0.06	0
Adventurous	0.00	0	-0.04	0	0.38	1	0.10	1

Table 4: Adjectival Marker samples for various traits. Samples with values > 0 in the Saucier Goldberg table have been given a binary count of 1, while those lower than 0 have been given 0. (*Decimal Values taken from Saucier et al (1996)).

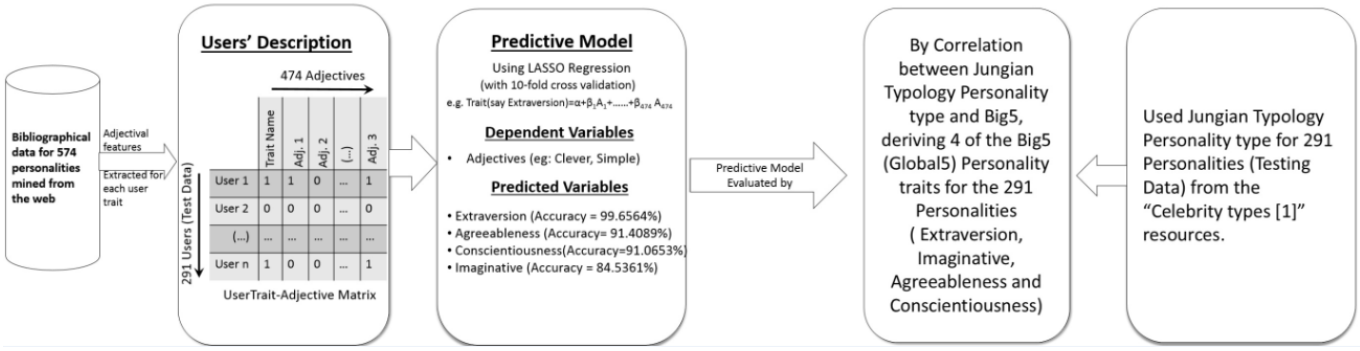


Figure 2: Descriptive of the methodology

traits.

Since there was only single source where traits of major personalities are classified (i.e. celebritytypes.com) we used it to evaluate our model. We used the remaining 291 personalities for evaluation of the model. The testing biographical data was mined for adjectival trait markers and their respective traits were predicted. The results of this evaluation have been discussed elaborately in the next section. Figure 2, which can be found above, is also illustrative of the procedure define above.

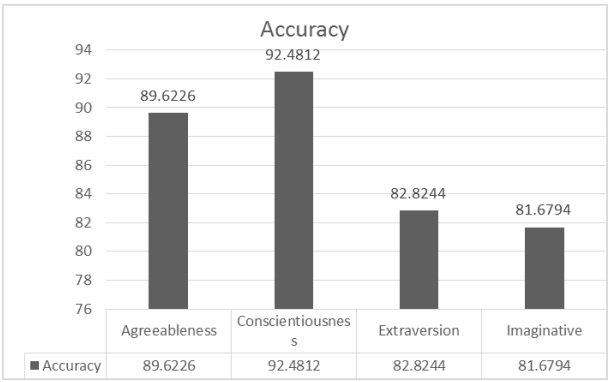


Figure 3: Average accuracy percentage of the personality traits by adjectival marker analysis

6. Results

The results by the above illustrated method are elaborated in this section. The average accuracies compared to the personalities obtained via psychometric tests (discussed in more detail in the following section) for considered four of the Big 5 traits were: Extraversion - 82.82% Agreeableness - 89.62%, Conscientiousness - 92.48% and Imaginativeness/Intellect - 81.67%. These readings do not necessarily demonstrate the prediction accuracy of the innate personality of a person but match that predicted by the psychometric tests with the given accuracies. They are also in league with few other techniques predicting the same for instance, the work of Iacobelli et al, (2011) attempted to decipher the personalities of bloggers has an average personality prediction accuracy of around 62.5%.

Thus, this paper proposes a technique which illustrates mainfold elevation in the overall accuracy of personality prediction (as indicated by psychometric tests) via social media.

7. Discussion

The results obtained illustrate that this method is competent for predicting the personalities of a person in coherence with other people's judgments about him/her. It gives substantial accuracies in the prediction of a person's personality matching with those obtained via psychometric tests. As an essential part of this study, we have also attempted to capture the variation in accuracy with the change in various factors, namely, word count of the corpus, and the ratio of the number of adjectives to the total number of words.[3] These are mainly intended to explore a threshold for word count and the adjective distribution (for the given

[3]Please note that the accuracies discussed here are the accuracy of the prediction as evaluated by the results via psychometric tests for Big 5 and should not be confused with accuracies used for predicting the baseline of the universal personality of a person.

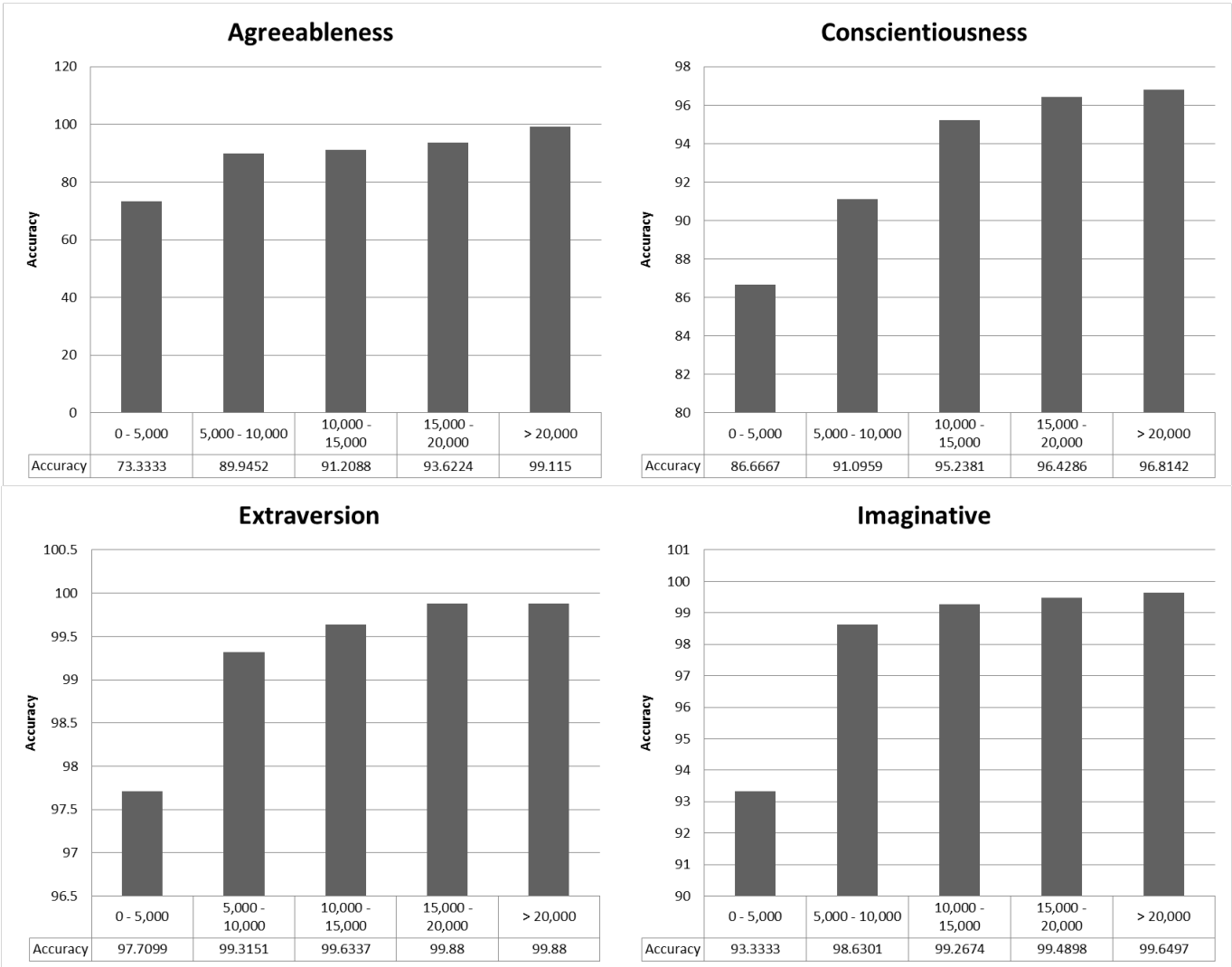


Figure 4: Accuracy variation over word count of testing data

technique) in the document set so as to get substantial results from the Adjectival Marker Technique. The following deductions can be made respective to each trait:

7.1. Collective Observations

Few collective observations can be drawn from the gathered data. As indicated in Figure 4, the accuracy in predicting the traits increases with an increase in the data word count. We also compared the accuracy results in predicting the respective traits on the basis of varying distribution of adjectives in the training dataset (Figure 5). The accuracy in predicting the traits is relatively low when the ratio of the AC/TWC is low and increases with a subsequent increase in the AC/TWC ratio.

7.2. Agreeableness

The accuracy in predicting Agreeableness is relatively low (73.33%) for data with word count $<$ 5000 words, and escalates up to 99.11% for big data reserves ($>$20,000 words). We also compared the accuracy results of predicting "Agreeableness" on the basis of varying distribution of adjectives in the training dataset.
The prediction of the "Agreeableness" trait is relatively low when the ratio of the adjectival count versus total word count is low. It illustrates an accuracy of 84.00% when the ratio is less than 0.001, improving to 94.18% when the ratio is between 0.001-0.002. Finally it escalates to 95.62% when increased to be greater than 0.003 (Figure 5). As expected there is a consistent increase in accuracy with increase in word count and the ratio AC/TWC.

7.3. Conscientiousness

The accuracy in predicting Conscientiousness varies from 86.66% when the word count of the data reserves is less than 5000 words, and subsequently increases with the increase in the number of words as shown in Figure 4.
We also varied the adjective distribution with the word count so as to obtain respective accuracies for the same model. It varies from an accuracy of 88.00% when the ratio is less than 0.001, improving to 93.60% when the ratio is between 0.001-0.002, and finally to 95.44% when increased to be greater than 0.003 (Figure 5). As expected there is a consistent increase in accuracy with increase in word count and the ratio AC/TWC.

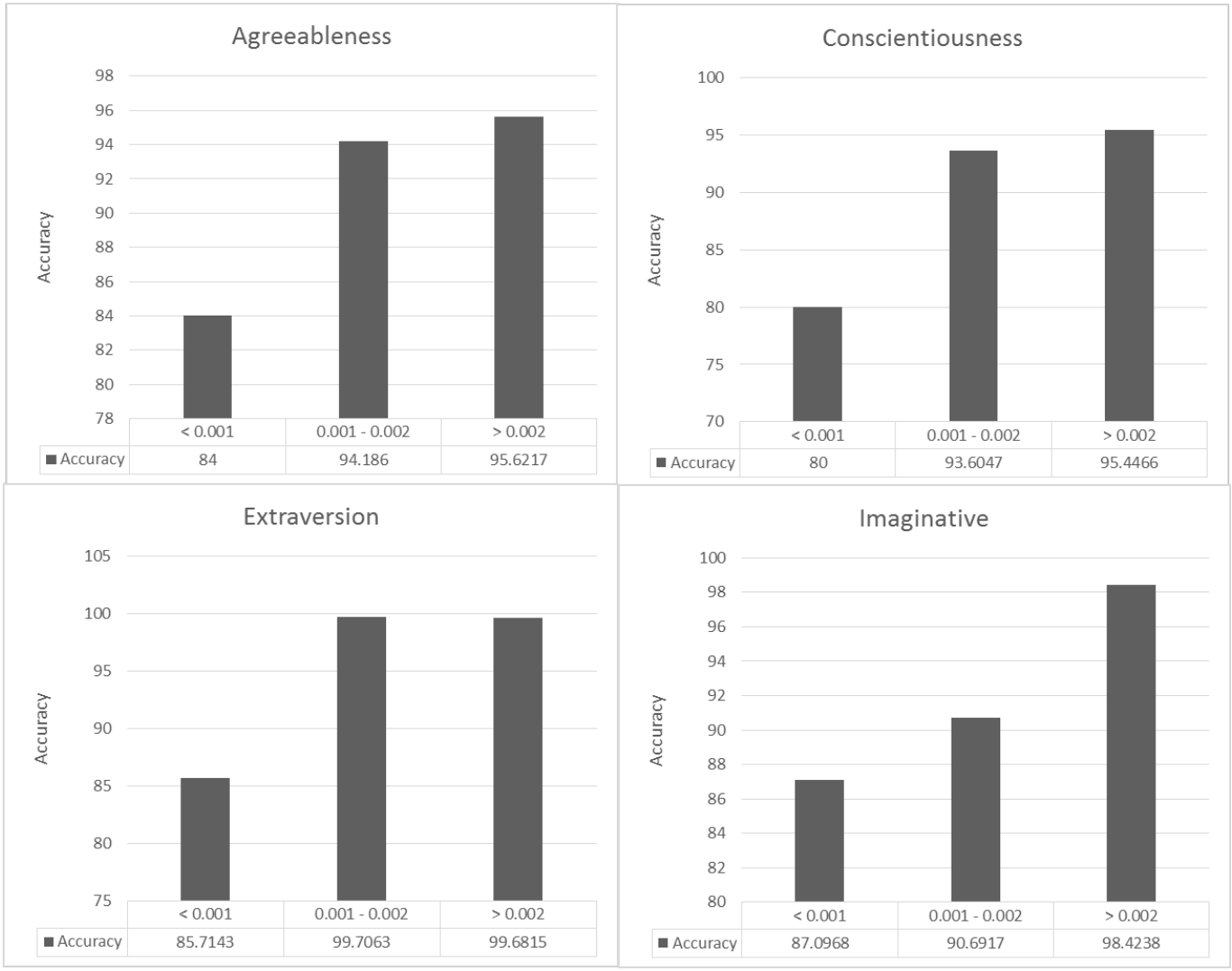


Figure 5: Accuracy variation over adjective distribution (AC/TWC) in testing dataset

7.4. Imaginative

The accuracy in predicting Imaginativeness varies from 93.33% at wordcount lower than 5000 words, and goes upto 99.88% for big data reserves (Figure 4).
The peaks observed in the variation of accuracy for "Imaginative" trait over the distribution of adjectives (AC/TWC) range from 85.71% accuracy for AC/TWC = 0.001, 90.69% accuracy for AC/TWC = 0.002 and finally 98.42% for AC/TWC >= 0.003 (Figure 5).

7.5. Extraversion

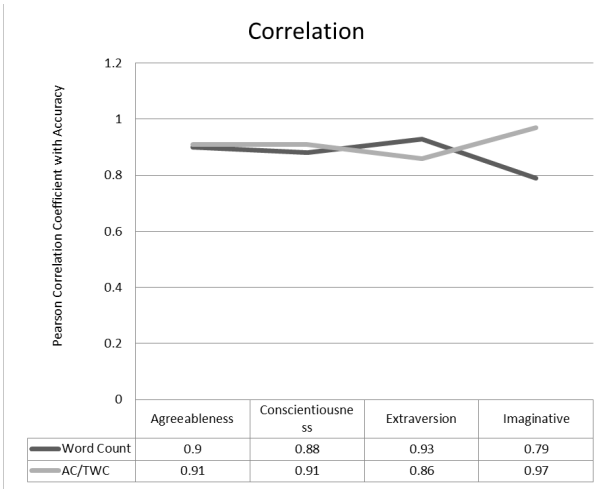
The accuracy varies from 97.70% for word count $<$ 5000 words and subsequently increases to 99.88% as shown in Figure 4.
The accuracy of this trait varied from 85.71% for AC/TWC = 0.001 and went on to increase upto 99.68% for AC/TWC = 0.002 and then 99.70% for AC/TWC >= 0.003.

The correlations for each of word count with accuracy and AC/TWC with accuracy for each of the above mentioned coefficient implies that for "Adjectival Markers" these are highly correlated to one another. This can also be validated by the graph in Figure 6.

8. Conclusion & Future Work

By means of this study we propose a simpler yet effective method to facilitate personality extraction of people in social media. In order to achieve this we have also reworked some perennial theories of Lexical Psychology and modified them with the newer concepts of machine learning models. This technique brings about a wave of novelty in the wide spread lexical concepts and techniques used to achieve user personality understanding in biographical data reserves. It is a significant contribution in the field of Computer Human interaction, since it is not just based on the modern model training techniques of artificial intelligence, but also finds solid ground in the foundational theories of human psychology. One major drawback of this study is that, it is (as of now) most optimized and accurate when tested on bigger data samples. This research is thus intended to pave way for extrapolating itself to smaller data reserves and microblogs. We intend to apply the same technique on not just adjectives but various other parts of speech (POS) in the near future. There are various studies which discuss the role of a person's personality in the development of diseases (Friedman et al, 1987). Thus, another goal that

this research aims to achieve is that in the very near future it would be able facilitate personality analysis for a wide range of people with varied handicaps which render them incapable of self-analysis in order to effectively predict their personalities. Statistics say that 11% of children 4-17 years of age (6.4 million)(Friedman et al, 1987) in the United States itself have been diagnosed with Attention-Deficit / Hyperactivity disorder (the number increasing by 3% this year). With valuable feedback from friends and family this model can help designing better technology for them and various other such people. Building upon this research and extending it to cover other POS would enable us to predict personalities from scanty as well as large datasets with good accuracy. The vision of this research is to train our next generation computers to not only understand people in terms of their choices, but the innate personalities which lead them to make those choices (leading to smart suggestive advertising systems etc). The future work of this research will also include combining this technique with pre-existing ones (e.g. LIWC, etc.) so as to increase the personality prediction accuracy to match that achieved by psychometric tests. We also intend to work on a lexical personality ontology, which analyzes the relationship of personality (both direct and indirect) with the various parts of speech (POS) i.e. extending it from being solely adjectival markers to various other POS. We would soon be graduating from solely Big5 trait prediction to evolving various mental states which can be predicted from the abundant lexical resources available online. Thus graduating the singly dimensioned Big5 model to a multi-dimensional graphical ontology tree of a person.



	Agreeableness	Conscientiousness	Extraversion	Imaginative
Word Count	0.9	0.88	0.93	0.79
AC/TWC	0.91	0.91	0.86	0.97

Figure 6: Variation of correlation coefficient based on distribution of adjectives in testing dataset

9. References

Allport. 1936. Traitnames. *A psycho-lexical study, Psychological Monographs*, 47.

Argamon S, Dhawle S, Koppel M, Pennebaker J W. 2005. Lexical Predictors of Personality Type. In *Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America*.

Boele de Raad. 2000. *The Big Five personality factors: The psycholexical approach to personality*. Hogrefe & Huber.

Briggs, I Myers, P B Gifts differing. 1980. *Understanding personality type*. Davies-Black Publishing, Mountain View, CA.

Capraro RM. 2002. Myers-Briggs Type Indicator Score Reliability Across: Studies a Meta-Analytic Reliability Generalization Study. *Educational and Psychological Measurement*.

Cattell R B. 1943. The description of personality: basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, 38(4), 476–506.

Celli F, Polonio L. 2013. *Relationships between Personality and Interactions in Facebook*. Social Networking: Recent Trends, Emerging Issues and Future Outlook, pages 41–54 Nova Science Publishers.

Digman J. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 4(1):417–440.

Friedman, Howard S. 1987. The disease-prone personality: A meta-analytic view of the construct. *Booth-Kewley, Stephanie American Psychologist*, 42(6), 539–555.

Furnham A. 1996. The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences*.

Goldbeck J, Robles C, Turner K. 2011. Predicting Personality with Social Media. In *Proceedings of the annual conference extended abstracts on Human factors in computing systems*.

Goldberg L R. 1990. An alternative description of personality: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Goldberg L R. 1992. The Development of Markers for the Big Five Factor Structure. *Psychological Assessment*, 4(1):26–42.

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

Iacobelli F, Gill A J, Nowson S, Oberlander J. 2011. Large scale personality classification of bloggers. *Lecture Notes in Computer Science*, 6975.

Jacob B Hirsh, Sonia K Kang, Galen V. Bodenhausen 2012. *Personalized Persuasion : Tailoring Persuasive Appeals to Recipients Personality Traits*. Psychological Science.

Jacopo S, Bruno L, Nadav A, Fabio P, Nicu S, Alex P 2012. Friends dont Lie - Inferring Personality Traits from Social Network Structure. In *Proceedings of UbiComp*. 180–185, 5 Sep - 8 Sep, Pittsburgh, USA, ACM, 978-1-4503-1224-0/12/09.

Jill G, Midian K. 1982. Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, 33(3), 386–404.

John O E. 1990. *The Big Fivefactor taxonomy: Dimensions of personality in the natural language and in*

questionnaires. Handbook of personality theory and research, L.A. pervin, pages 66–100 Guilford Press, New York.

Karl Jung. 1921. *Psychological Types*.

Klages L. 1926. *Die Grundlagen der Chat-akterkunde [111e science of character]. Lcipzig: Barth*.

Lukas M, Sara van de Geer and Peter B. 2008. The group lasso for logistic regression.J. R. Statist Soc. *Eidgenssische Technische Hochschule, Zrich, Switzerland*, 70(1):53–71.

Kermanidis K L. 2012. Mining Authors Personality Traits from Modern Greek Spontaneous Text. *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals, in conjunction with LREC12*.

Luyckx K, Daelemans, W Personae 2008. A corpus for author and personality prediction from text. In *Proceedings of LREC- 2008, the Sixth International Language Resources and Evaluation Conference*.

Mairesse F, Walker, M Personage. 2007. Personality Generation for Dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics ACL*.

Mairesse F, Walker M A, Mehl M R, Moore R K. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial intelligence Research*, 30.

McCrae R, Costa P. 1989. Reinterpreting the Myers-Briggs Type Indicator From the Perspective of the Five-Factor Model of Personality. *Journal of Personality*, National Center on Birth Defects and Developmental Disabilities Division of Human Development and Disabilities, USA.

Norman, Warren T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6), 574–583.

Oberlander J, Nowson S. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics ACL*.

Ostendoff E. 1990. *prache und Personlichkeitsstruktur: Zur Validitat des Funf-Faktoren-Modeiis der Personlichkeit [Language and personality structure: On the validity of the five-factor model of personality]*. Regensburg, Federal Republic of Germany : Roderer Verlag.

Pennebaker J W, King L A. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77.

Quercia D, Kosinski M, Stillwell D, Crowcroft J. 2011. Our Twitter Proles, Our Selves: Predicting Personality with Twitter. In *Proceedings of SocialCom*. 180–185.

Robert T. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Saucier G, Goldberg L R. 1996. Evidence for the Big Five in analysis of familiar English Personality adjectives *European Journal of Personality*, 10:61–77.

Yla R Tausczik, James W Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

http://www.celebritytypes.com/.