Personality Prediction From Facebook Posts In Gurmukhi Script

Article · October 2017

CITATIONS
0

2 authors:

Gurpreet Josan
Punjabi University, Patiala
34 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Jagroop Kaur
Punjabi University, Patiala
6 PUBLICATIONS **13** CITATIONS
SEE PROFILE

CSI JOURNAL OF COMPUTING

Personality Prediction From Facebook Posts In Gurmukhi Script

Jagroop Kaur[1], Gurpreet Singh Josan[2]

[1,2]Assistant Professor
Department of Computer Engineering, Punjabi University Patiala
Department of Computer Science, Punjabi University Patiala
josangurpreet@pbi.ac.in

Abstract

Human personality is a complex trait. Lot of attempt has been found in literature to predict the personality of a person using different models. Big 5 model is one of them. Researchers attempt to develop supervised as well as unsupervised techniques to automate the prediction of personality from written text. This paper is focused on automatically predicting personality from facebook posts in Gurmukhi text by exploiting the correlations among linguistic cues and personality traits. Unsupervised technique for predicting personality has been presented. Scores representing personality traits are generated using correlations between personality trait and features. These scores are then converted into personality description of a person. Despite of the limited resources, evaluation proved that system's performance is quite good.

Keywords: Automatic *Personality Prediction,Big 5 Model, Social Media Text, Gurmukhi Scrip*

1. Introduction

Human personality is a complex trait. Personality is defined as the way of mind setting, feeling and acting in any situation. This is not about specific activities that are going through again but, it is regarding the pattern followed in those activities and the tendencies. Identity of any person is defined as a combination of different characteristics and qualities of a person. A Personality is made up of different characteristic patterns of thoughts, feelings, acting and behaviours that make a person unique. From time to time various models have been proposed by researchers to categorise the personality of a person like DISC assessment, Eysenck's personality model etc. butcurrently, the most widely used model of personality is the Big Five model. As per this model, personality is a combination of five factors viz. openness, conscientiousness, extraversion, agreeableness, and neuroticism also termed as OCEAN [9, 13]. Further, number of techniques has been devised by the researchers to measure the personality of a person. These include subjective, objective, projective and Psycho-analytic methods. But all these techniques require the interaction with subject in order to estimate its personality.Recently, researchers are attracted toward using vast amount of text data available on social media like facebook and twitter to analyse personality of a person. This way a personality of person can be predicted without the knowledge of subject under consideration.

Although looking promising, the task is not trivial. Considerable amount of research has been done in languages like English, Arabic, and other European languages. With the availability of input methods for regional languages, it is a matter of investigation to check how already devised methods work on regional languages. This paper is focused on testing this idea on text in Gurmukhi script. Next section discusses some of the issues in predicting personality from social media text, followed by literature review. Then linguistic features which can be used to correlate personality with words are discussed.Section VI discusses the experimentation followed by results of experimentation. Finally section VIII concludes the paper with discussion.

2. Issues

Although it is believed that there is a direct correlation between personality and written text, but estimating this correlation is a fuzzy task. Different traits of personality are reflected from the choice of words, punctuations, and style of sentence formation. Identification and extraction of these traits is not trivial task when the text is from social media. Various issues involved are: multilingual text (data is code mixed), unlabeled data, privacy (Few public data availability), annotation is dependent on annotators own perception. Due to fuzziness in definition of personality itself, there is no one single model that fits each user. Finally, the lack of available reference data, evaluation of personality models is also cumbersome.

3. Literature Survey

In literature, two types of approaches are used by researchers viz. machine learning and linguistic cues. [5]reports the comparison of Support Vector Machine, Bayesian Logistic Regression (BLR) and Multinomial Naïve Bayes (MNB). Their input data is the status of facebook users. Similarly, [6] also reported the use of support vectormachine,Nearestneighbour with k=1 (kNN) and Naïve Bayes for identifying personality of a person. Classification techniques has been reported by various researchers which uses linguistic features to build the model [2, 10, 12, 16]. [7] proposed a model which is based on linguistic features (such as number of words) as well as features from social network (like friends count) to predict personality from Facebook profile. Machine learning algorithms were used to develop the model. Similarly, [8] also uses LIWC tool, structural features and sentiment features to build a personality predictive model. In [19],M5 rules based learning model has been developed using network features (like followers, following, etc.). Authorsin [3] presented an analysis considering number of social network features which includes uploaded photos, number of persons infriendshipnetwork, eventswhich are attended by person,timesuserhasbeen tagged in photos etc.Theyperform the analysis on 180000 Facebookusers. In [20],authors employed sentiment feature of user text. For obtaining this information, sentNet was being used. Other features like affectiveandcommonsenseknowledge was also used for identifying personality. ConceptNet, EmoSenticNetandEmoSenticSpacehas been employed for this task. They developed a supervised classifier using SVM and report increased accuracy due to these features.

As per [16], there are many linguistic factors associated with Big Five personality traits model.The authors also reports correlations between the linguistic features and personality characteristics. [14 and 15] also study the relation of words with different personality traits. [17]proposes text analysis system named Linguistic Inquiry and Word Count (LIWC) which analyze emotional, cognitive, andstructuralcomponents present in individuals' written samples. Later [8, 12, and 19] report sets of correlations between linguistic cues and personality dimensions. [4]propose an unsupervised method to find personality of person from social media text by exploiting correlations between language cues and personality traits.

4. Proposed Work

This paper focus on identification of personality of a person from his/her posts on facebook. As described in [4], unsupervised approach has been followed due to lack of annotated and reference data. As per this approach, any person's personality can be defined by five variables based on Big 5 model. The value of these variables can be represented numerically. Given the correlation between Big 5 trait and linguistic feature, the numerical value of trait can be calculated. These values vary from negative to positive. The negative values can be converted into n (no) and postive value can be converted to y(yes). If value is zero then it can be converted into o (No information). Thus a code will be produced by the model which represents personality of person. For example a person can get the code ***yynoy*** interpreted as open-minded, Conscientious, Introvert and emotionally instable.

5. Features

Implementing above discussed model needs correlation between linguistic factors and personality traits. As our domain is Punjabi language, no such study exists which provide required correlations. As per [1], the linguistic features are universal and equally acceptable in all societies of world, we decided to use the list of correlation coefficients between linguistic factors and personality traits provided by [12]because it is the one of the largest and also it has more focus on language or literature related factors and cues. Further, as this list is developed from an essay corpus, some of the factors are irrelevant to social media text and some others are missing. We picked 30 factors some of which are language independent and others are language dependent. These are:

5.1 Language independent features:

1. All Punctuation (ap): Count of all punctuation marks in posts (Including marks of gurmukhi script)
2. Commas(cm): Count of commas in the user post
3. Exclamation mark(em): Number of exclamation mark (!) in the user post
4. Negative emotions (ne): the count of emoticons expressing negative feelings in the post. A list of negative emoticons as described by [11] has been used.
5. Numbers (nb): the count of numeric values in the post e.g. (2017, 11 etc)
6. Parenthesis (pa): the total number of phrases written in parenthesis '()'in the user post.
7. Positive emotions (pe): the total number of emoticons expressing positive feelings in the post as described by [11].
8. Question marks (qm): the total number of symbol '?' in the user post.
9. Long words (sl): The choice of long or short words to convey message is also a one of the important trait. Thus, this feature represents the total number of words which are longer than 6 letters in the user post.
10. Type/token ratio (tt): the ratio obtained by dividing the types (the total number of different words) occurring in a text or utterance by its tokens (the total number of words).
11. Word count (wc): words in the user post.
12. Mean word frequency (mf): It is a simple mean of the frequency of words in the user post calculated by summation of all word frequencies divided by total unique words.

5.2 Language dependent features

1. First person singular pronoun (im): The number of first person singular pronoun in the user post e.g. "ਮੈ","ਮੈਨੂੰ","ਮੈਥੋ", "ਮੇਰਾ"etc.
2. Negative particles (np): This provides the total number of negative particles in the user post e.g. "ਨਹੀ", "ਬਨਿਾਂ", "ਬਗੈਰ" etc.
3. Prepositions (pp): This provides the total number of prepositions in the user post e.g. "ਉੱਤੇ", "ਪਹਲਿਾਂ", "ਬਾਅਦ", etc.
4. Pronouns (pr): the count of pronouns in the user post.
5. Self reference (sr): The number describing how many first person (singular and plural) pronouns are present in the user post e.g. "ਮੇਰਾ", "ਸਾਡੀ", "ਅਸੀ","ਆਪਾਂ" etc.
6. Swears (sw): total number of vulgar expressions present in the user post e.g. "ਹਰਾਮਖੋਰ", "ਹਰਾਮੀ", "ਹਰਾਮਜ਼ਾਦਾ", "ਫੁੱਦੂ", "ਸਾਲੇ" etc.
7. First person plural pronouns (we): The number of first person plural pronouns present in the user post e.g. "ਮੇਰੇ","ਸਾਡੀਆਂ" etc.
8. Second person singular pronouns (yu): The number of second person singular pronouns present in

the user post e.g. "ਤੂੰ","ਤੈਨੂੰ","ਤੈਥੋ","ਤੇਰਾ" etc.

9. Quantifiers count (qntfc) : count of Quantifiers in the post e.g. "ਕੁੱਝ", "ਬਹੁਤ", "ਥੋੜ੍ਹਾ", "ਕਾਫੀ" etc.
10. Ordinal Count (ordnalc): count of ordinal words in the post e.g. "ਇੱਕੀਵਾਂ", "ਚੌਵੀਵਾਂ", "ਅਠਾਈਵਾਂ", "ਬੱਤੀਵਾਂ", "ਸੱਤਰਵਾਂ" etc.
11. Family word count (fmc): count of family word in the post e.g. "ਮਾਂ", "ਪਤਿ", "ਮਾਤਾ", "ਨਾਨਾ", "ਦਾਦਾ" etc.
12. Friend word count (frc): count of Friend word in the post e.g. "ਸਾਜਨ", "ਦੋਸਤ", "ਮਤਿਰ", "ਸਾਥੀ", "ਯਾਰ" etc.
13. Positive word count (posfeel): count of positive words in the post
14. Tentative word count (tentat): count of tentative words in the post e.g. "ਸ਼ਾਇਦ", "ਅੰਦਾਜ਼ਾ", "ਖਬਰੇ", "ਖੈਰ੍ਹ", "ਭਾਵੇ", "ਕਧਿਰੇ" etc.
15. Certainity word count (certain): count of certainity words e.g. "ਹਮੇਸ਼ਾ", "ਸਦਾ", "ਨਤਿ" etc.
16. Money word count (money): count of money words e.g. "ਪੈਸਾ", "ਧਨ", "ਚਾਂਦੀ", "ਮਾਇਆ", "ਮਾਲ" etc.
17. Religion word count (relig): count of religion words e.g. "ਰੱਬ", "ਰੱਬਾ", "ੴ", "ਗਨੇਸ਼", "ਲੱਛਮੀ", "ਹਨੂਮਾਨ", "ਸ਼ਨੀ"etc.
18. Assent word count (assent): count of assent words e.g. "ਸਹਮਿਤ", "ਮਰਜ਼ੀ", "ਅਨੁਮਤੀ", "ਮਨਜੂਰੀ" etc.

6. Experimentation

6.1 Data Collection

Users post from publically available facebook accounts has been collected using facebook API. Posts of total 50 persons have been collected. 15 persons are female and rest are male. 25 Posts are collected for every person covering the time span ranging from last 3 months to 5 years.For collecting the statistical distribution of the features in the social media text, a separate corpus of posts in Gurumukhi script has been collected from various publically accessible facebook accounts. This corpus contains 25000 posts of varying number of lines and includes 3 lakh words. Mean of each feature has been extracted from this corpus. The average distribution of each feature has been shown in table 1.

Table 1 Mean of features

Feature	Mean	Feature	Mean
ap	2.0	wc	13.0
cm	0.0	we	0.0
em	0.0	yu	0.0
im	0.0	mf	1.0
np	0.0	WPS	12.0
ne	0.0	qntfc	0.0
nb	0.0	ordnalc	0.0
pe	0.0	fmc	0.0
pp	2.0	frc	0.0

pr	1.0	posfeel	0.0
qm	0.0	tentat	0.0
sl	1.0	certain	0.0
sr	0.0	money	0.0
sw	0.0	relig	0.0
tt	1.0	assent	0.0

6.2 Resources

For the calculation of above discussed features, some resources are needed. This includes list of pronouns, ordinal numbers, tentative words, religious words etc. As no such lists are available for Punjabi language so hand crafted lists has been used. Following is the statistics of these lists:

Table 2 Statistics of various resources used

List	Number of entries	List	Number of entries
Pronouns	110	Quantifier words	25
First person singular pronouns	15	Ordinal words	105
First person plural pronoun	6	Family words	60
Second person singular pronoun	8	Friend words	35
Negative particles	10	Tentative words	20
Prepositions	65	Certainty words	15
self-reference	15	Money words	40
Swear words	40	Religious words	80
Negative emoticons	128	Assent words	15
Positive emoticons	635		

Besides, we have also used Punjabi sentiment lexicon which has been developed by department of computer science, Punjabi University Patiala. This lexicon has more than 60,000 words along with their sentiment score.

6.3 Process

All the posts of a single person are passed to the system as input. All the features are extracted from the posts. The score of trait is increased or decreased depending upon whether it correlates positively or negatively with the feature. The correlation table provided by [12] has been adopted for this task (See table 3). The score changes only if the count of the feature is more than the average occurrence of that feature in representative corpus as shown in table 2. The score remain unchanged if feature occurrence is less than the mean value of that feature in representative corpus. After processing all posts, the final numerical values are converted to the nominal values ("y","n" and "o") by simply replacing positive value with "y", negative value with "n" and zero with "o". Finally these nominal values are converted into descriptive strings about the personality of a person in question. Table 4 provides the details of descriptions.

Table 3 Correlation Table between Linguistic cues and Big 5 Personality Trait as provided by[12]

Features	Extrovert	Stability	Agreeableness	conscientiousness	Openness
ap	-0.08	-0.04	-0.01	-0.04	-10
cm	-0.02	0.01	-0.02	-0.01	0.1
em	0	-0.05	0.06	0	-0.03
im	0.05	-0.15	0.05	0.04	-0.14
np	-0.08	0.12	0.11	-0.07	0.01
ne	-0.03	-0.18	-0.11	-0.11	0.04
nb	-0.03	0.05	-0.03	-0.02	-0.06
pe	0.07	0.07	0.05	0.02	0.02
pp	0	0.06	0.04	0.08	-0.04
pr	0.07	0.12	0.04	0.02	-0.06
qm	-0.06	-0.05	-0.04	-0.06	0.08
sl	-0.06	0.06	-0.05	0.02	0.1
sr	0.07	-0.14	-0.06	-0.04	-0.14
sw	-0.01	0	-0.14	-0.11	0.08
tt	-0.05	0.1	-0.04	-0.05	0.09
wc	-0.01	0.02	0.02	-0.02	0.06
we	0.06	0.07	0.04	0.01	0.04
yu	-0.01	0.03	-0.06	-0.04	0.11
mf	0.05	-0.06	0.03	0.06	-0.07
WPS	-0.01	0.02	0.02	-0.02	0.06
qntfc	-0.03	0.05	-0.03	-0.02	-0.06
ordnalc	-0.03	0.05	-0.03	-0.02	-0.06
fmc	0.05	-0.05	0.09	0.04	-0.07

frc	0.06	-0.04	0.02	0.01	-0.12
posfeel	0.07	-0.01	0.03	-0.02	0.08
tentat	-0.06	-0.01	-0.03	-0.06	0.05
certain	0.05	-0.01	0.03	0.04	0.04
money	-0.02	0.24	-0.13	-0.24	0.01
relig	0	0.03	0	-0.06	0.07
assent	0.01	0.02	0	-0.04	0.04

Table 4 Description of Personality Traits

Extrovert	"ਤੁਸੀਂ ਆਪਣੇ ਆਲੇ ਦੁਆਲੇ ਲੋਕਾਂ ਨੂੰ ਦੇਖ ਕੇ ਜੋਸ਼ ਮਹਿਸੂਸ ਕਰਦੇ ਹੋ। ਤੁਹਾਨੂੰ ਪਾਰਟੀਆਂ ਵਿਚ ਜਾਣਾ, ਜਲਸਿਆਂ ਇਕੱਠਾਂ ਮੇਲਿਆਂ (ਧਾਰਮਿਕ, ਰਾਜਨੀਤਿਕ, ਸਮਾਜਿਕ, ਜਾਂ ਵਪਾਰਿਕ) ਵਿੱਚ ਸ਼ਾਮਿਲ ਹੋਣਾ ਪਸੰਦ ਹੈ। ਤੁਸੀਂ ਗਰੁੱਪ ਵਿੱਚ ਕੰਮ ਕਰ ਕੇ ਖੁਸ਼ ਹੁੰਦੇ ਹੋ ਪਰ ਇੱਕਲੇ ਹੋਣ ਤੇ ਬੋਰੀਅਤ ਮਹਿਸੂਸ ਕਰਦੇ ਹੋ।"
Introvert	"ਤੁਸੀਂ ਆਪਣੇ ਆਪ ਵਿੱਚ ਮਸਤ ਰਹਿੰਦੇ ਹੋ ਅਤੇ ਆਪਣੇ ਆਲੇ ਦੁਆਲੇ ਲੋਕਾਂ ਦੇ ਇਕੱਠ ਨੂੰ ਪਸੰਦ ਨਹੀਂ ਕਰਦੇ। ਤੁਹਾਨੂੰ ਅਜਿਹੇ ਕਾਰਜਾਂ ਵਿੱਚ ਦਿਲਚਸਪੀ ਹੈ ਜੋ ਇੱਕਲੇ ਹੁੰਦੇ ਹੋਣ ਜਿਵੇਂ ਕਿ ਪੜ੍ਹਨਾ, ਲਿਖਣਾ, ਕੰਪਿਊਟਰ ਤੇ ਕੰਮ ਕਰਨਾ, ਗਾਣੇ ਕੰਪੋਜ਼ ਕਰਨੇ ਵਗੈਰਾ। ਤੁਸੀਂ ਗਰੁੱਪ ਵਿੱਚ ਕੰਮ ਕਰ ਕੇ ਖੁਸ਼ ਨਹੀਂ ਹੁੰਦੇ ਹੋ। ਪਰ ਇੱਕਲੇ ਕੰਮ ਕਰ ਕੇ ਖੁਸ਼ ਹੁੰਦੇ ਹੋ। ਇੱਕ ਸਮੇਂ ਤੇ ਇੱਕ ਕੰਮ ਨੂੰ ਹੀ ਹੱਥ ਪਾਉਂਦੇ ਹੋ। ਬੋਲਣ ਤੋਂ ਪਹਿਲਾਂ ਸੋਚਣ ਵਿੱਚ ਵਿਸ਼ਵਾਸ ਰੱਖਦੇ ਹੋ।"
Emotionally stable	"ਸ਼ਾਂਤ ਸੁਭਾਅ ਅਤੇ ਤਣਾਅ ਮੁਕਤ ਹੋ।"
Emotionally Not Stable	"ਚਿੰਤਾ, ਟੈਂਸ਼ਨ ਲੈਣ ਵਾਲੇ, ਤਣਾਅ ਵਿੱਚ ਰਹਿਣ ਵਾਲੇ, ਦੁਜਿਆਂ ਤੋਂ ਜਲਣ ਵਾਲੇ, ਈਰਖਾ ਕਰਨ ਵਾਲੇ, ਬਹੁਤ ਛੇਤੀ ਚਿੜਚਿੜੇ ਅਤੇ ਗੁੱਸੇ ਵਿੱਚ ਆ ਜਾਂਦੇ ਹੋ।"
Agreeable	"ਦੋਸਤਾਨਾ ਜਾਂ ਮਿਲਾਪੜੇ ਸੁਭਾਅ ਦੇ ਮਾਲਕ ਹੋ। ਮਿਲਵਰਤਨ ਨਾਲ ਕੰਮ ਕਰਨ ਵਾਲੇ, ਦੁਜਿਆਂ ਲਈ ਆਪਣੇ ਕੰਮ ਪਿੱਛੇ ਛੱਡਣ ਵਾਲੇ ਹੋ। ਤੁਸੀਂ ਦੁਜਿਆਂ ਦਾ ਲਿਹਾਜ਼ ਕਰਦੇ ਹੋ ਅਤੇ ਉਦਾਰਵਾਦੀ ਹੋ।"
Non Agreeable	"ਤੁਸੀਂ ਕੋਰੇ ਸੁਭਾਅ ਦੇ ਹੋ। ਆਪਣੇ ਕੰਮਾਂ ਨੂੰ ਪਹਿਲ ਦੇਣ ਵਾਲੇ, ਦੁਜਿਆਂ ਦਾ ਲਿਹਾਜ਼ ਨਾ ਕਰਨ ਵਾਲੇ ਹੋ। ਤੁਸੀਂ ਮਿਲਵਰਤਨ ਨਾਲੋਂ ਅਲੱਗ ਹੋ ਕੇ ਕੰਮ ਕਰਨ ਵਿੱਚ ਭਲਾਈ ਸਮਝਦੇ ਹੋ। ਦੁਜਿਆਂ ਦੀ ਮਦਦ ਘੱਟ ਹੀ ਕਰਦੇ ਹੋ।"
Conscientiousness	"ਤੁਸੀਂ ਭਰੋਸੇਯੋਗ ਅਤੇ ਜ਼ਿੰਮੇਵਾਰ ਹੋ। ਹਰ ਕੰਮ ਪਲੈਨਿੰਗ ਨਾਲ ਕਰਨ ਵਾਲੇ, ਹਰ ਕੰਮ ਵਿੱਚ ਚੰਗੀ ਕਾਰਗੁਜ਼ਾਰੀ ਦਿਖਾਉਣ ਵਾਲੇ ਹੋ। ਦਿਆਨਤਦਾਰੀ, ਸਵੈ-ਅਨੁਸਾਸਨ ਅਤੇ ਸਾਵਧਾਨੀ ਤੁਹਾਡੀਆਂ ਖਾਸੀਅਤਾਂ ਹਨ। "
Non Conscientiousness	"ਮਸਤ, ਬੇਇਖਤਿਆਰ ਜਾਂ ਬਿਨਾਂ ਪਲੈਨਿੰਗ ਦੇ ਵਿੱਚ ਕੰਮ ਕਰਨ ਵਾਲੇ ਹੋ। ਲੋਕ ਤੁਹਾਨੂੰ ਘੱਟ ਭਰੋਸੇਯੋਗ ਸਮਝਦੇ ਹਨ।"
Openness	"ਤੁਸੀਂ ਰਚਨਾਤਮਕ ਬਿਰਤੀ ਦੇ ਮਾਲਕ ਹੋ। ਤੁਹਾਨੂੰ ਨਵੇਂ ਵਿਚਾਰ ਕੁੱਝ ਨਵਾਂ ਸਿੱਖਣ ਦੀ ਚੇਸ਼ਟਾ ਰਹਿੰਦੀ ਹੈ। ਕੁਝ ਨਵਾਂ ਕਰਨਾ ਲੋਚਦੇ ਹੋ। ਤੁਸੀਂ ਵਿਹਾਰਕ (practical) ਨਾਲੋਂ ਜ਼ਿਆਦਾ ਕਲਪਨਾਸ਼ੀਲ ਹੋ।",
Non Openness	"ਆਪਣੇ ਆਪ ਵਿੱਚ ਰਹਿਣ ਵਾਲੇ, ਹਰ ਗੱਲ ਵਿੱਚ ਵਿਸ਼ਲੇਸ਼ਣ ਕਰਨ ਵਾਲੇ ਅਤੇ ਤਬਦੀਲੀਆਂ ਤੋਂ ਬਚਣ ਵਾਲੇ ਹੋ। ਕੁਝ ਨਵਾਂ ਕਰਨ ਤੋਂ ਕਤਰਾਉਂਦੇ ਹੋ। ਤੁਸੀਂ ਵਿਹਾਰਕ(practical) ਸੁਭਾਅ ਦੇ ਹੋ। "

7. Results

We have executed the personality prediction system for 50 users having 25 posts each. For the evaluation of results of algorithm used in our system, we did the manual assessment. The predicted personality has been shown to the subject themselves and subjects were asked to evaluate themselves. A five point scale has been provided to them for rating the prediction ranging from Strongly Disagree to Strongly Agree. The results are shown in table 5.

Table 5 Rating results of predictions.

Agreeability	Percentage of people
Strongly Agree	6%
Agree	44%
Neutral	28%
Disagree	22%
Strongly Disagree	0%

The most common personality type among those who were tested is "*NYNNY*" i.e. "Introvert, Emotionally Stable, Friendly, Careless and Insightful". Following is the list of common personality types and their percentage. [4] reported similar table stating the most frequent personality type in the Italian subset of Friend-Feed is represented by the model of an extravert, insecure, agreeable, organized and unimaginative person.

8. Discussion and Conclusion

As shown in table 5, 6% people agree with the prediction of our system and 44% peopleagree that system is able to approximate three or four of their personality traits. 28% people has the view that few points are predicted correctly whereas 22% believe that prediction is wrong. As personality is a complex combination of different traits, we believe that given the resources, overall performance of the system is quite good. Various factors may affect the performance of system. Foremost among them is the correlation table that we have employed. The table has been developed using essays which has sufficient amount of clean text. A similar correlation table need to be developed using data from social media which is limited in quantity and also noisy in nature. Due to noisy social media text, identification of feature also get affected which in turn affects the prediction capability. For example, some feature may get missed due to spelling variation or code mixed text. As in every region of India, minimum two languages are common-one regional language and other National Language. Like in Punjab, Punjabi is the mother tongue and Hindi is National language. Besides, people also use English language. This diversity also reflects in social media text. Posts are generally in code mixed form. So the resources which are purely in Punjabi, may not be helpful in identifying the features from code mixed text. Thus multilingual resources need to be developed. Besides, there are number of other features for which we have no correlation table. It is worth checking how other social media parameters like number of user's followers, number of likes or dislikes received by user on his/her various posts, number of friends and time of posting etc correlates with the big 5 personality traits. Due to lack of these correlations, we are unable to check these factors this time. One of

the drawbacks of current system worth mention here is that the system is dependent on number of posts and will not produce accurate results with few posts.

References

1. Allik, J., Realo, A., & McCrae, R.R.:Universality of the Five-Factor Model of Personality. In P.T. Costa & T. Widiger(Eds.) Personality Disorders and the Five Factor Model of Personality. Washington: American Psychological Association.61-74 (2013)

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

3. Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D.J.: Personality and Patterns of Facebook Usage. In Proceedings of Web Science 36–45, 2012.

4. Celli, F.: Unsupervised Personality Recognition for Social Network Sites. In Proceedings of Sixth International Conference on Digital Society, Valencia. 2012.

5. Alam F., Evgeny A. Stepanov, Riccardi G.: Personality Traits Recognition on Social Network Facebook, In Workshop of ICWSM-13 on Computational Personality Recognition (Shared Task),Cambridge, MA, USA pp: 6-9. 2013.

6. FarnadiG. , ZoghbiS., MoensM., De CockM.:Recognising Personality Traits using Facebook Status Updates, In Workshop of ICWSM-13 on Computational Personality Recognition (Shared Task), Cambridge, MA, USA. 2013

7. Golbeck, J., Robles, C., and Turner, K.: Predicting Personality with Social Media. In Proceeding of the extended abstracts of conference on Human factors in computing systems. 253–262.2011a.

8. Golbeck, J. R., Edmondson C. M. and Turner, K.: Predicting personality from twitter.In proceedings of IEEE 3rd International Conference on Social Computing, 149-156, 2011b.

9. Goldberg, L. R.: The development of markers for the big-five factor structure. *Psychological Assessment, Vol4*, 26-42, 1992.

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

11. Kralj N., SmailovićP., SlubanJ., &Mozetič I.: Sentiment of Emojis. *PLoS ONE*, *10*(12),(2015).doi:http://doi.org/10.1371/journal.pone.0144296

12. Mairesse F., Walker M. A., MehlM. R. and Moore R. K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. J. Artif. Intell. Res.(JAIR), vol. 30, pp. 457-500, 2007.

13. Matthews G., Deary I., and Whiteman, M: Personality traits.Cambridge University Press, Cambridge, UK, 2003

14. Mehl, M. R., Gosling, S. D., &Pennebaker, J. W.: Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology, 90, 862–877.2006.

15. Oberlander, J., & Gill, A. J.: Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Processes, 42, 239–270.2006.

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

17. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J.: The development and psychometric properties of LIWC2007 [LIWC manual]. Austin, TX: LIWC.net. 2007.

18. Pennebaker, J. W., James, Boyd R. L., Jordan K., and Blackburn K.:The development and psychometric properties of LIWC2015. UT Faculty/Researcher Works.Austin, TX: University of Texas at Austin.2015.

19. Quercia D., Kosinski M., Stillwell D., and Crowcroft J.: Our Twitter Profiles, Our Selves: Predicting Personality withTwitter. In Proceeding of International Conference on Social ComputingSocialCom. Boston, MA, USA180–185. 2011

20. PoriaS., Gelbukh A., Agarwal B., Cambria E., H.Newton: Common Sense Knowledge Based Personality Recognition from Text. In proceedings of 12th Mexican International Conference on Artificial Intelligence, Vol. 8266,pp 484-496, Springer 2013.