Social Quotient Feedback for Employment Fitness Based on Personality Analysis using Social Network Data

Amit Karyekar, Rohit Joshi
University of California, Los Angeles
{amit.karyekar, rohitjoshi}@cs.ucla.edu

Abstract

Social Networks provide a platform where users can connect with other people to share ideas, pictures, posts, activities, events and interests. It is a place where users can present themselves to the world, revealing their personality and insights about their life. Businesses are now exploiting this information to bolster their business growth. One nascent area where the social network data is being utilized is in the field of recruitment. Companies are now reviewing candidate's social network information to better predict how good a fit he is for the job. In today's world, companies expect academic proficiency to be complemented by excellent social skill set. In this paper, we are interested in analyzing user's personality based on his social network data and providing a feedback about the areas where the user can improve, to better match the social skill set required for a job. Through this paper, we describe a method to accurately predict a user's personality based on his Twitter data. We also provide our experimental results where we evaluate the performance of our system against the results given by Personality Recognizer system developed at MIT Computer Science and Artificial Intelligence lab.

1. Introduction

There has been a tremendous amount of growth in the use of social networks. Taking the case of Twitter, the number of registered users on Twitter grew from 1,000 in 2006 to 500,250,000 in 2012[1]. Through these social networking sites, users reveal a lot of personal information and give a glimpse of their real life in the form of status updates, pictures, frequency of communication with friends and their interests.

Recruiters are now beginning to consider social networks as a medium to gauge the real personality of the potential candidates. Traditionally, recruiters have relied on self-evaluations, self-reports and online questionnaires to identify a user's personality. However research has

shown that these tests are unreliable and do not necessarily give an accurate estimate of a user's personality as; self-evaluations are vulnerable to bias, users may not be available for self-evaluation reports and users can lie in these test [2]. As a result recruiters are now moving towards social networks to evaluate users' social behavior. Evaluations drawn based on social network data are less vulnerable to bias and it is difficult for a user to fake his profile over a long period of time.

As per 'Social Recruiting Survey Results 2012' conducted by Jobvite [3], 92% of the employers use or plan to use social media for recruiting as opposed to 83% in 2010. 73% of the employers have successfully hired a candidate through social networks. As of 2012, 54% of the employers use Twitter data, 66% of the employers use Facebook data and 93% of the employers use LinkedIn data to recruit a potential candidate. About 47% of the employers review the social network data of the applicant after receiving his application.

Previous work on analyzing social network data to judge user's personality has shown that user profiles on Facebook [4] and Twitter [5], [6] are reflective of their actual personalities. These studies extracted features from Facebook and Twitter, and analyzed them against the Big Five Personality Inventory [7]. The research conducted by the previous studies found out factors which accurately estimate the five domains of personality, Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism.

Since, Twitter data can be readily accessed through Twitter API, we decided to utilize Twitter as a platform to study user personality.

We aim to combine the work done in [5] and [6] in the field of analyzing Twitter data to analyze user personality, and further extend it by creating a system which provides a social quotient feedback in the form of areas where the user can improve to align him as per the job requirements. Through our system, we aim to make the user aware of how much a fit he is for a particular job and highlight the areas where he lacks so that he improve his personality in the future.

2. Related work

Research related to use of social network information to evaluate a user on the scale of the Big Five Factor Model has been predominantly done by [4], [5] and [6]. In [4], the authors had made use of Facebook profile of a user to analyze user personality. Since our work is related to Twitter, we would like to mention the work done in [5] and [6] since they had utilized Twitter as the social media.

In [5], the authors had used tweets of the users to analyze their personality. They ran the tweets through two tools namely, Linguistic Inquiry and Word Count (LIWC) [8] and MRC Psycholinguistic Database. LIWC produces statistics on 81 different features of text in five categories. These include Standard Counts (word count, words longer than six letters, number of prepositions, etc.), Psychological Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (such as occupation, financial issues, health), and Other dimensions (counts of various types of punctuation, swear words). MRC Psycholinguistic Database includes a list of 150,000 words with linguistic and psycholinguistic features of each word. They also used Twitter features like number of followers, number of following, density of social network, number of @ mentions, number of replies, number of hash-tags, number of links and words per tweets. In addition, they had performed a word by word sentiment analysis of each user's tweets. The prediction algorithm used in [5] had taken as input the LIWC features, MRC features, Twitter use features, structural features and sentiment analysis features. Their prediction model consisted of two machine learning algorithms namely, ZeroR and Gaussian process. Their algorithm was able to predict scores of each personality trait to within 11% to 18% of their actual scores.

In [6], the authors aimed at identifying a set of measures besides using public data from Facebook profiles and text from Twitter streams. They identified a number of features based on the number of interactions a user has with his friends along dimensions like reciprocity and priority. By using Forward subset based regression, the authors identified a set of features which are most predictive of each personality trait. As per their research, extraversion is best captured by the text length of tweets, the mean number of propagations and amount of attention given by their friends. Agreeableness can be best understood by the number of hash-tags used in tweets, balance of response time, mean response time of user's friends and amount of attention received by friends of user. Studying the timing between tweets, the divergence in propagation and mean response time among friends gives us an estimate of Openness. Conscientiousness can be measured by mean degree of similarity among friends of users, mean time between tweets among the friends, divergence in the number of messages sent by the user to all his friends and divergence of response time of user to his friends' messages. The features highlighting Neuroticism are deviation in the length of tweets of the user, number of days the user has been on Twitter, attention received from friends, amount of favoritism to friends, deviation in the response times and fraction of re-tweets of the user. To predict a user's personality, the authors had made use of ZeroR and Gaussian process regression models. Their results showed similar performance to the results obtained by analyzing users' tweets. However they mentioned that their performance can be improved for personality traits like Neuroticism and Extraversion.

3. Challenges

The first challenge in our pursuit of analyzing user's Twitter data is the privacy constraints enforced by Twitter due to which data of users is not publicly available. To overcome this privacy constraint, we created an application under Twitter to utilize the features provided by Twitter Rest API 1.1. For collecting data, we started with the user under whose account we had created the Twitter application. From this start node, we first crawled all the friends of the user. We also went into two depths further from the friends of the authenticating user.

Second challenge in our data collection phase was the limit of 180 requests per time window (15mins) issued by Twitter Rest API 1.1 to fetch the timeline of the user and the limit of 15 requests per time window to fetch the friends list of the user . This implies that, for a user A with 200 friends on an average and 100 friends of friends on an average would require approximately 24 hours to gather the timeline data of user network. To elaborate, the time to get friend list of friends of user would require approximately 3 hours (Number of friends/ Number of requests per hour = 200/60). And the time required to get timeline containing 200 tweets of these friends of friends would require approximately 24 hours (Number of friends of friends / Number of requests per hour = 200*100/ 720). Considering the fact that above time requirement is associated with single user at 200 tweets per user in the network, we can imagine the time requirement associated with all the users at 2000 tweets per user in network. (This meant that we could not access all user data for calculating the features at runtime due to the factors mentioned above. Hence to overcome this challenge, we downloaded all the data we would require for our analysis beforehand in flat files using the method as described in section 5.)

The third challenge was to compute the values for all the parameters mentioned in [6]. The authors in [6] had made use of following features to compute the user score for the five behavioral traits:

- Extraversion:
    - Length of tweets
    - Mean number of propagations
    - Standard deviation of attention received from friends
- Agreeableness:
    - Standard deviation of attention received for the friends of user
    - Mean number of hash-tags used in tweets
    - Balance of response time
    - Mean response time among friends of user
- Openness:
    - Standard deviation of time among tweets

    - KL-Divergence of mean propagation reciprocation
    - Mean response time among friends of user
- Conscientiousness:
    - Mean degree of similarity among friends of user
    - Mean time between tweets of friends of user
    - KL-Divergence of mean number of messages exchanged
    - KL-Divergence of balance of response time
- Neuroticism:
    - Standard deviation of length of tweets
    - Number of days the user has been on Twitter
    - Fraction of tweets that are re-tweets among friends of user
    - Standard deviation of mean number of propagations among friends of user
    - Standard deviation of mean response time of the user
    - Attention received from friends
    - Number of messages favorited by friends of user

Among all the features mentioned above, we could compute the values for the following features:

- Extraversion:
    - Length of tweets
    - Mean number of propagations
    - Standard deviation of attention received from friends
- Agreeableness:
    - Standard deviation of attention received for the friends of user
    - Mean number of hash-tags used in tweets
- Openness:
    - Standard deviation of time among tweets
    - KL-Divergence of mean propagation reciprocation
- Conscientiousness:
    - Mean time between tweets of friends of user
- Neuroticism:
    - Standard deviation of length of tweets
    - Fraction of tweets that are re-tweets among friends of user
    - Standard deviation of mean number of propagations among friends of user
    - Attention received from friends

Feature Dependency

Feature	Dependent Parameter having restricted access in Twitter	Reason for leaving out the feature
Balance of response time, Mean response time among friends of user, Mean response time among friends of user, Mean degree of similarity among friends of user, KL-Divergence of mean number of messages exchanged, KL-Divergence of balance of response time, Standard deviation of mean response time of the user	Response time of user which can be accesses using Directed Messages a user sends and receives	Twitter Rest API 1.1 gives access only to directed messages of the user who has authenticated the application. Hence we are unable to access directed messages of friends, and friends of friends of target user
Number of days the user has been on Twitter	No parameter of Twitter Rest API 1.1 gives direct access	As no parameter in Twitter Rest API gives directed access, we had to leave out the feature
Number of messages favorited by friends of user	Directed Messages a user sends and receives	Twitter Rest API 1.1 gives access only to directed messages of the user who has authenticated the application. Hence we are unable to access directed messages of friends, and friends of friends of target user

Hence we could compute the values of 12 features out of 21 features mentioned in [6].

4. Analysis of Approach

In this section, we give a brief description of the methods we have used in our approach. In section 4.1, we describe the 'Big Five Personality Inventory' model and give a brief idea about what characteristics each of the five personality trait displays. In section 4.2, we describe how LIWC and MRC can be used to extract features from tweets. Section 4.3 describes the features extracted from the social interaction of users in Twitter. Section 4.4 deals with the WEKA

models which we have used for training and testing. In section 4.5, we describe how we went about doing the training of our system using WEKA.

4.1. Big Five Personality Inventory

The "Big Five" model of personality analysis has been widely researched and used models of personality structure. As per [9], the Big Five personality traits have been widely used due to their consistency in interviews and self-descriptions when observed.

The model characterizes personality using five traits namely [10]:

- Extraversion: Outgoing, amicable, assertive, friendly and energetic, tend to draw inspiration from social situations
- Agreeableness: Cooperative, helpful, nurturing, compassionate
- Openness to experience: Appreciation for art, emotion, adventure, unusual arts, curiosity, variety of experience
- Conscientiousness: A tendency to be self-disciplined, responsible, persevering, reliable, high-achievers
- Neuroticism: Anxious, insecure, sensitive, moody, tense and easily tipped into experiencing negative emotions

4.2. Textual Analysis of Tweets

As per the method used in [5], we used two tools to analyze the textual features in tweets.

- **Linguistic Inquiry and Word Count (LIWC)**
  LIWC produces statistics on 81 different features of text in five categories namely [5],
  - **Standard counts:** word count, words longer than six letters, number of prepositions
  - **Psychological process:** emotional, cognitive, sensory, social processes
  - **Relativity:** words about time, the past, the future
  - **Personal Concerns:** occupation, financial, issues, health
  - **Other dimensions:** counts of various types of punctuations, swear words

- **MRC Psycholinguistic Database**
  MRC Psycholinguistic Database is a machine usable dictionary containing 150837 words with up to 26 linguistic and psycholinguistic attributes for each [12]. Some of these features are: Kucera-Francis written frequency, number of categories, and number of samples; Brown verbal frequency; Familiarity rating; Meaningfulness via Colorado norms and via Paivio Norms; Concreteness; age of acquisition; Thorndike-Lorge written frequency; and the number of letters, phonemes, and syllables.

We calculated the above features for each user under consideration for our study.

4.3. Measures of Social Interaction in Twitter

To consider the amount of social interaction a user has on twitter, we included the following features for studying each of five traits under the 'Big Five Factor' model [6]:

- Extraversion:
  - **Len:** Mean Text Length
  - **Prop#:** Mean number of propagations A -> B -> X
  - **STD-Attn:** Standard deviation of attention received from friends
- Agreeableness:
  - **FF-STD-Attn:** Standard deviation of attention received for the friends of user
  - **Hash:** Mean number of hash-tags used in tweets
- Openness:
  - **STD-Time:** Standard deviation of time among tweets
  - **KL-PropH:** KL-Divergence of mean propagation reciprocation
- Conscientiousness:
  - **FF-Time:** Mean time between tweets of friends of user
- Neuroticism:
  - **STD-Len:** Standard deviation of length of tweets
  - **FF-Rtw:** Fraction of tweets that are re-tweets among friends of user
  - **FF-STD-Prop#:** Standard deviation of mean number of propagations among friends of user
  - **Attn:** Attention received from friends

We computed the above features for all the users who were considered for the study.

Refer Appendix A for detailed information about the computation of these parameters.

4.4. WEKA models

On obtaining the values of above parameters, we made use of WEKA models to train and test the system. We made use of the following models:

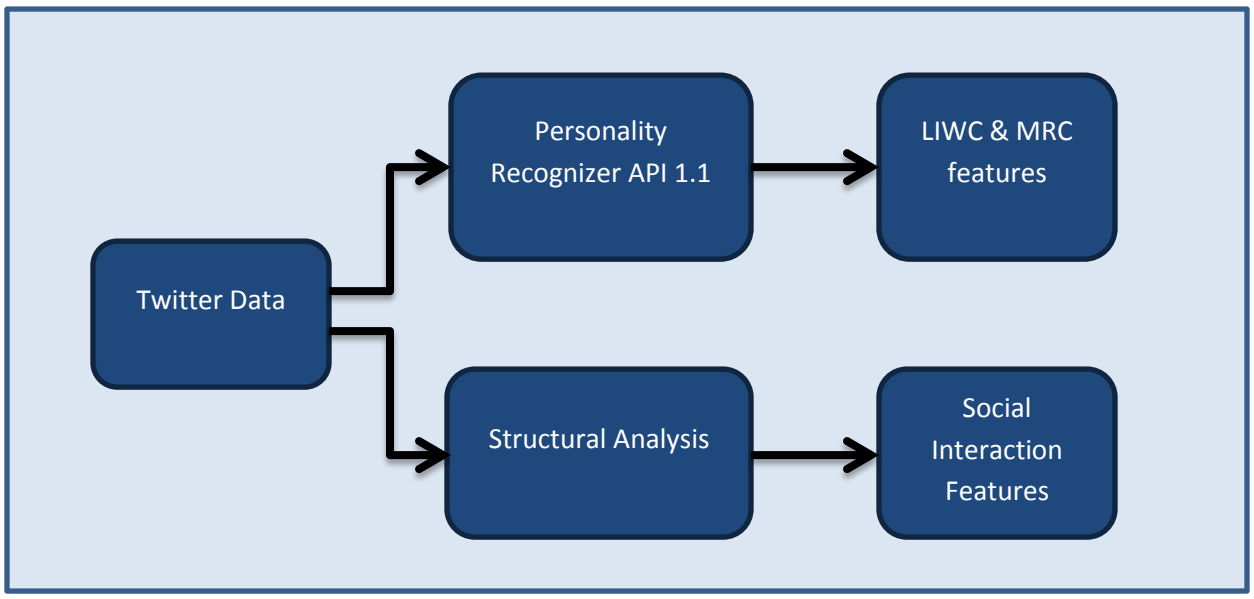
- ZeroR

We have trained our system using each of above mentioned models and have calculated the accuracy for each.

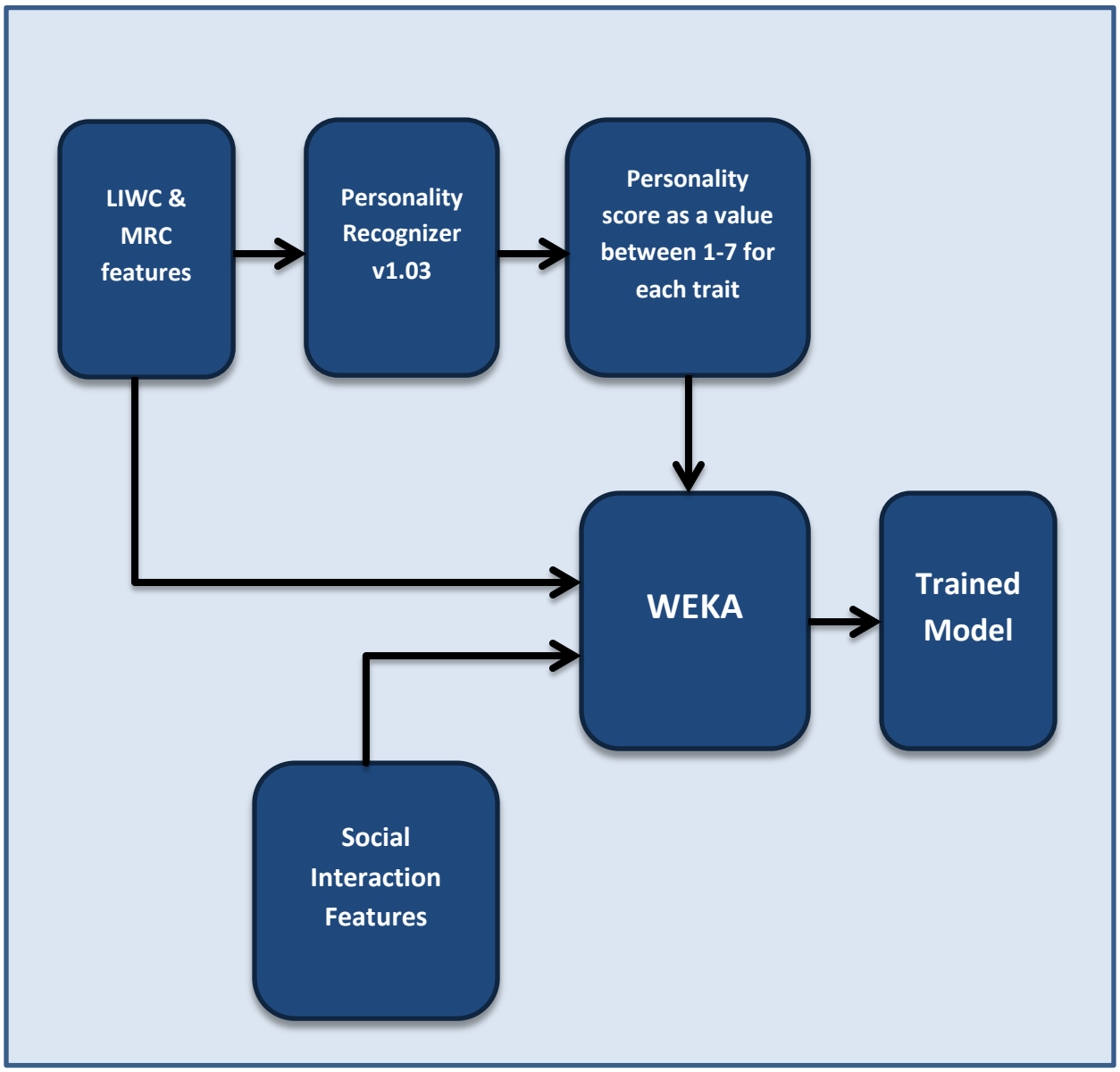
4.5. Training the system

For training our system, we have used a dataset of 12 users. For these 12 users, we computed their scores for the LIWC features, MRC features and features based on their social interaction on Twitter. We have used WEKA models as described in section 4.4 for training. For training purposes, it was imperative to label the features with corresponding personality scores so that the system gets trained to analyze the personality of a new user. We used the Personality Recognizer v1.03 API to compute the personality scores for our training set. Personality Recognizer v1.03 is based on the models analyzed in [13]. The API is available at the MIT Computer Science and Artificial Intelligence laboratory. Personality Recognizer outputs a score between the ranges of 1-7 for each of the personality trait for each user.

Following is a diagrammatic representation of our feature extraction system:



Following is a diagrammatic representation of our training model using 12 users:



5. Data Collection

For data collection, we have used the Twitter Rest API 1.1. We had registered an application a under Twitter account. So the registered Twitter account served as the starting node for our data collection process. From the starting node, we crawled 3 levels deep into the friendship network.

Following is the summary of our data collection process at different levels into the friendship network:

- **Level 1:** At level 1, we collected data from 26 friends of the starting user.
  Data consisted of:
  - Up to 2000 tweets of each user
  - List of all friends of these users

- **Level 2:** At level 2, we crawled all the friends of above 26 users.
  - For these friends, we collected the following data:
    - Up to 200 tweets of each user
  - For a random sample of 50 friends (from the friend list of 26 users), we collected following data:
    - List of all the friends. So in the end, we collected a list of friends for 26 * 50 = 1300 users at level 2.

- **Level 3:** At level 3,
  - For a random sample of 5 friends (selected from the random sample of 50 friends selected from the friend list of 50 users), we collected the following data:
    - Up to 200 tweets of each user

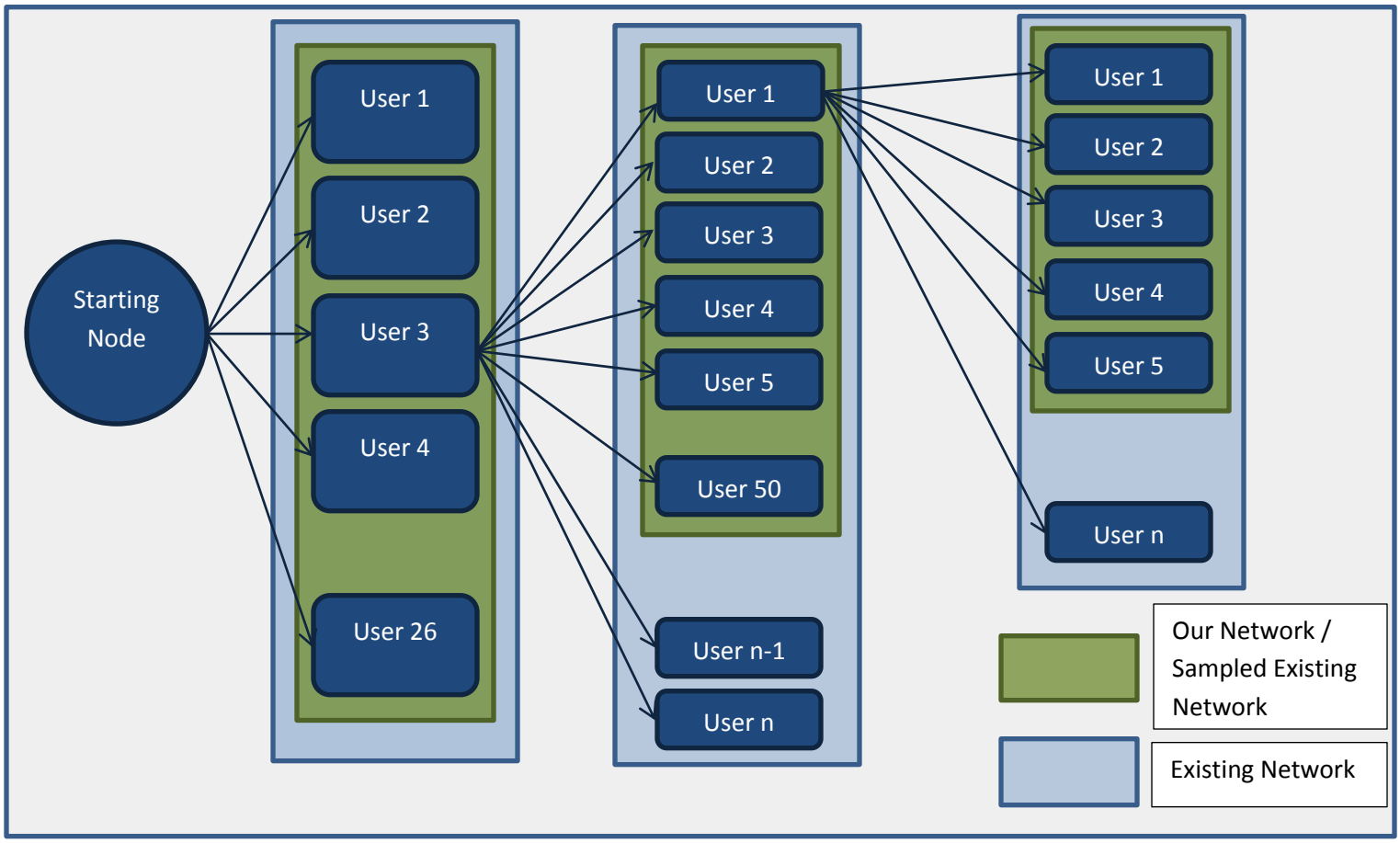
Use of Sampling Theorem

While downloading tweets using Twitter Rest API 1.1, there is a rate limitation which allows a maximum of 200 tweets to be downloaded per request. Also there is a limit of 180 requests per 15 minute window for downloading tweets.

Due to this rate limitation, we downloaded 2000 tweets per user at level 1 as these were the users whose personality we wanted to study. Hence, there should be sufficient number of tweets to compute the features based on LIWC and MRC for these users. Also out of the 12 features based on the social interaction of the users on Twitter, 4 features required us to go into the level 2 and level 3. Hence we decided to use sampling theorem at level 2 and level 3. At

level 2 we sampled 50 friends from the friend list of each of the 26 users. At level 3, we selected 5 friends from each of the 50 users selected at level 2.

Hence we predicted the properties of friends of users and of friends of friends of user by studying the properties of the sample.



6. Experimental Analysis

The experimental analysis comprises of:

Analysis:

1. Using ZeroR:

Determining the accuracy of prediction of five factors trained using ZeroR Model in WEKA as illustrated in Section 4.4

Motivation: Determining the reliability of proposed model

Dataset: Dataset for training encompasses 12 users.

The accuracy of testing/prediction is computed as follows:

- For each factor F:
    - determine count = the number of users under testing having predicted value for factor equal to the value computed using Textual Analysis
- accuracy for factor F = count/ the total number of users under testing
- end

Range of Five Factors is classified as follows

- Low: 1 to 2.5
- Medium: 2.5 to 5.5
- High: 5.5 to 7

Following is the chart of accuracy tested for 4 users:

In the following table: A: actual value, P: predicted value

Users	Agreeableness		Openness		Neuroticism		Extraversion		Conscientiousness	
	A	P	A	P	A	P	A	P	A	P
User1	5	4	5	5	5	5	4	5	6	4
User 2	5	4	4	5	5	5	4	5	5	4
User 3	4	4	4	5	5	5	6	5	5	4
User 4	3	4	5	5	5	5	4	5	6	4

Legend: Conflict of range classification

2. Using Naïve Bayes:

Users	Agreeableness		Openness		Neuroticism		Extraversion		Conscientiousness	
	A	P	A	P	A	P	A	P	A	P
User1	5	3	5	5	5	5	4	5	6	5
User 2	5	3	4	5	5	5	4	5	5	5
User 3	4	3	4	5	5	5	6	5	5	5
User 4	3	3	5	5	5	5	4	5	6	5

Legend: Conflict of range classification

Future Analysis: Determining the weight of feature extraction using social interaction in predicting the five factors

Dataset for this analysis encompass 4 users. The accuracy of determining the weight of feature extraction is computed using

- For each user U
    - Determine list of friends= friends who have high degree of correlation i.e. similar personality computed using factors determined through feature extraction only
    - Determine overlap = total number of friends from the list of friends who have high degree of correlation i.e. similar personality computed using factors determined through textural analysis only
- end

7. Conclusion

We have predicted the division of five factors as low (value between 1 and 2.5), medium (value between 2.5 and 5.5) or high (value between 5.5 and 7) for the four users based on training set of twelve users with 85% accuracy in both the models, ZeroR and Naïve Bayesian. In future, the accuracy can be better computed using large dataset of 1000 users and at least 10 different machine learning models.

To conclude, we firmly believe that although the reduction to practice of the proposed system may be constrained by privacy issues or result into tailoring of profiles, a large percentage of users in the field of education will be benefited. This belief is due to the fact that current professional network like LinkedIn suggest the users what job domains are suitable for them, however they do not suggest the characteristics required for the users in order to be apt for job domain.

8. Future Scope

At present we have created a system which gives feedback to the user about his employment fitness for a particular job. For creating this system, we have combined the features used in [5], [6]. Our goal was to combine the features obtained by using textual analysis of tweets as mentioned in [5] with the features obtained by studying the social interaction of user on Twitter as mentioned in [6]. In the future, we would like to look for additional parameters which would help in increasing the accuracy of the overall system. Introducing sentiment

analysis would further help in better classification of words into different categories thereby helping to accurately compute the behavioral traits.

In the present scope of the system, we have fed the thresholds for each behavioral trait required for a particular job. We would like to create a system which would allow recruiters to feed in thresholds themselves.

Since, we are analyzing the personality of a user; the system can be broadly applicable to different fields like advertising on social media websites. The social media sites can publish advertisements on a user's homepage based on his nature. Also music stations can suggest songs to user based on his overall personality.

References

- **[1]** http://dstevenwhite.com/2013/02/09/social-media-growth-2006-to-2012/
- **[2]** Jennifer Dodorico McDonald. "Measuring Personality Constructs: The Advantages and Disadvantages of Self-Reports, Informant Reports and Behavioral Assessments". ENQUIRE, Volume 1, Issue 1, June 2008.
- **[3]** http://web.jobvite.com/rs/jobvite/images/Jobvite_2012_Social_Recruiting_Survey.pdf
- **[4]** Jennifer Golbeck, Cristina Robles, Karen Turner. "Predicting Personality with Social Media". PP 253-262, CHI EA '11.
- **[5]** Jennifer Golbeck, Cristina Robles, Michon Edmondson, Karen Turner. "Predicting Personality from Twitter". PP 149-156, IEEE Socialcom '11.
- **[6]** Sibel Adali, Jennifer Golbeck. "Predicting Personality with Social Behavior". PP 302-309, IEEE ASONAM '12.
- **[7]** M. Barrick and M. Mount. "The Big Five personality dimensions and job performance: A meta-analysis". Personnel psychology, 44(1):1-26, 1991.
- **[8]** J. PenneBaker, M.Francis, and R. Booth. Linguistic Inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 2001.
- **[9]** Schacter. D., Gilbert, D., and Wegner, D. (2011). "Psychology". Worth Publishers. Page 474-475
- **[10]** Atkinson, Rita, L.; Richard C. Atkinson, Edward E. Smith, Daryl J. Bem, & Susan Nolen-Hoeksema (2000). *Hilgard's Introduction to Psychology* (13 ed.). Orlando, Florida: Harcourt College Publishers. p. 437.
- **[11]** http://www.kovcomp.co.uk/wordstat/LIWC.html
- **[12]** http://www.psych.rl.ac.uk/

- **[13]** Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, Roger K. Moore, "Using linguistic Cues for Automatic Recognition of Personality in Conversation and Text", PP-457–500 Journal of Artificial Intelligence Research

Appendix A

Computation of features used to compute social interaction in Twitter:

Sr. No	Feature	Computation
1	Len	It is mean length of all the tweets of a user
2	Prop#	It is the mean number of propagations, i.e. A -> B -> X. **Explanation:** Propagation means out of all the tweets of user A, how many tweets are propagated by user B to user X
3	STD-Attn	It is the standard deviation of attention received from friends. **Explanation:** Suppose we want to calculate the attention received by user A from user B. Then attention is computed as the number of propagations of B that are from A.
4	FF-STD-Attn	It is Standard deviation of attention received for the friends of user. It is similar to feature 3 except that it is computed at the level of friends of friends of the given user, i.e. at level 3.
5	Hash	It is the mean number of hash tags used by the user in tweets.
6	STD-Time	It is the standard deviation in the mean time between tweets.
7	KL-PropH	It is computed as the Kullback-Leibler divergence in the balance of propagation reciprocation. **Explanation:** Suppose we want to compute the personality of user A and user B is his friend. Propagation reciprocation is computed as the measure of the tweets of user B that are propagated by user A.
8	FF-Time	It is the mean time between tweets calculated at the level of friends of friends of the given user, i.e. at level 3.
9	STD-Len	It is the standard deviation in the length of tweets of a given user.
10	FF-Rtw	Rtw is the fraction of tweets that are retweets. So FF-Rtw is this fraction computed at the level of friends of friends of a given user, i.e. at level 3.
11	FF-STD-Prop#	It is the standard deviation in the mean propagation of tweets computed at the level of friends of friends, i.e. at level 3.
12	Attn	It is the mean attention received by the user from his friends.

Formulae used:

- **Standard Deviation**

  Standard Deviation = $\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}$

- **Kullback-Leibler Divergence**

$$KL(X||U) = \sum_{i=1}^{k}\left(x_i * ln\left(\in + \frac{x_i}{1/k}\right)\right)$$