Mining HEXACO personality traits from Enterprise Social Media

Priyanka Sinha
Tata Consultancy Services Limited
Indian Institute of Technology Kharagpur
priyanka27.s@tcs.com

Lipika Dey
Tata Consultancy Services Limited
lipika.dey@tcs.com

Pabitra Mitra
Indian Institute of Technology Kharagpur
pabitra@gmail.com

Anupam Basu
Indian Institute of Technology Kharagpur
anupambas@gmail.com

Abstract

In this paper we introduce a novel computational technique of extraction of personality traits (HEXACO) of employees from Enterprise Social Media posts. We deal with challenges such as not being able to use existing survey instruments for scoring and not being able to directly use existing psychological studies on written text due to lack of overlapping words between the existing dictionary and words used in Enterprise Social Media. Using our approach we are able to infer personality traits (HEXACO) from posts and find better coverage and usage of the extended dictionary.

1 Introduction

It is well known that modern organizations rely heavily on unstructured information to capture expertise and knowledge that otherwise exist in the minds of its employees. Understanding the behavior and personality of the employees help in group formation and understanding group dynamics which could help predict project success. Among the many ways in which modern organizational psychology (Ashton et al., 2004) describes human personality, some important attributes that generally emerge can be summarized as follows:

Agreeableness being helpful, cooperative and sympathetic towards others

Conscientiousness being disciplined, organized and achievement-oriented

Extraversion having a higher degree of sociability, assertiveness and talkativeness

Emotionality the degree of emotional stability, impulse control and anxiety

Openness to Experience having a strong intellectual curiosity and a preference for novelty and variety

Honesty-Humility being a good person who is ethical and altruistic

These are collectively known as personality traits in the HEXACO (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, Openness) personality trait model as described in (Ashton et al., 2004). Intensity and polarity of each trait varies from person to person thereby capturing a person's personality. These traits are measured by trained psychologists using self rating or by being rated by the psychologist. These rating scales such as the HEXACO-PI-R as described in (Lee and Ashton, 2004) contain questions about the person that help in judging their traits. (Ashton et al., 2004) also identifies sets of personality describing words with loading factors that are related to each trait which forms a dictionary of such words.

Written text is a medium of communication within a group, when members communicate through emails and/or social media. Emails originate from individuals and are targeted towards a specified set of people. In social media, there are usually no targeted groups. Rather communication is meant for as many people to see, read and react. While emails are used for confidential information exchange within an enterprise, enterprise social networks are targeted towards rapid disbursement of information across large communities. They also encourage sharing of knowledge and information, flatten hierarchy, and enable speedy resolution through crowd-sourcing.

These text sources are observed to contain very few of the existing personality describing words. In our corpus of an Enterprise Social Media dataset, 0.22% percent of total word usage as well as 152 words out of total 185,251 distinct words

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 140–147,
Lisboa, Portugal, 17 September, 2015.

contain personality describing words from the set described in (Ashton et al., 2004). Our dataset has a total of 14,849 distinct users of which only 1,939 users use atleast one of these words atleast once. Whether they are at all used in the context of describing someone's personality or behavior is not studied.

These are a very low number and they do not capture all the implicit expressions in the text describing someone's personality or behavior. We could however infer the presence of personality describing words and other personality expressions from such formal and semi-formal text. As summarized in (Goldbeck et al., 2011a), personality traits are useful in predicting performance and success in enterprise context. Hence, the motivation to explore other techniques to infer personality and behavior expressions about each individual as well as group(s) from enterprise text sources.

2 Literature Survey

There are two different challenges in trying to assess HEXACO traits from enterprise social media as follows:

1. Psychologists have studied the problem of identifying personality traits from language usage. They have used various methods amongst which rating scales, both self reported and administered by trained psychologists are established techniques. The Big Five Factors, HEXACO and other such models of personality traits have been related to language usage by psychologists (Ashton et al., 2004; Tausczik and Pennebaker, 2010). Pennbaker has conducted very many studies relating how people of different demographics in different situations use language and how it relates to human behavioral traits. In particular there are a set of features which are identified as relevant to human behavior. Linking of words to personality traits/behavioral traits has been done by different groups of psychologists. A challenge here is that there are different lists used by different groups.

2. In recent times, phenomenal rise in social media content has given birth to the sub-area of text mining where researchers analyze language usage to infer behavioral traits from social media content. Inferences are usually validated by self appraisal or voluntary revelation of identity or psychologists identify. Since language usage is substantially different in social media and the erstwhile controlled psychoanalytic methods used by psychologists, there has been efforts to generate mappings between social media text and personality traits.

Existing literature in each of the these above areas are reviewed in detail below.

2.1 Review of related work in text analysis for psychoanalysis

These have been used as features in most of the recent work in identifying personality traits from social media text. Most of these works have been validated by trained psychologists. There is not much work that has focussed on text which is from business enterprises where language used is more formal than on websites like Twitter and Facebook.

We discuss below some of the related literature with respect to the challenges mentioned in Section 2.

(Ashton et al., 2004) reports in tabular form a list of adjectives that relate to the each of the HEXACO personality traits. This paper explores the HEXACO personality trait model. It also explores Religiosity as an extra seventh factor and accepts that there may be more factors than six. It notes that the 1982 Goldberg 1,710 adjective set is nearly the entire population of English personality descriptive adjectives. We use the result of this study which results in a reduced set of the 1,710 personality descriptive adjectives in English with loading factor for each of the six factors of the personality trait model. The reduction and identification of the word set seem like an important work for psychologists as it would enable them to work with fewer words which may mean faster and concise analysis. Use of computational power relaxes this restriction. Now even with a much larger dictionary it would be possible to scalably analyze people's personalities using computational models of analysis.

(Tausczik and Pennebaker, 2010) and (Chung and Pennebaker, 2007) describe the LIWC software, its usage and relevance to psychological processes. It summarizes how different parts of speech used by people tell us about them and their behavior. For example, it has been studied that lots

of use of first person personal pronouns is an indicator of depression. Content words indicate where the person is focussing such as people thinking about death, sex, money, or friends will refer to them in writing or conversation. People experiencing physical or emotional pain use first person personal pronouns to draw attention to themselves. Greater use of first person personal pronouns correlates with higher rank and higher status. Higher ranked individuals ask fewer questions. First person singular pronoun usage can be used to predict lower status. Greater use of first person plural pronouns show group cohesion. Word count can be a proxy for amount of communication and more communication may promote better group performance. Analysis of tense of verbs indicate temporal focus of attention. "We" signals a sense of group identity. Sometimes "We" also refers to others. When lying, people tend to use more words, more negative words, more motion words, less first person singular pronouns. The use of "You" is important in predicting lower quality relationships.

2.2 Review of related work on text mining of social media content for behavior analysis

(Goldbeck et al., 2011b) gave questionnaires to twitter users to fill out. They used structural properties such as number of followers, number of following, density of social network, number of mentions, hashtags, replies, links. For linguistic features they used LIWC, MRC Psycholinguistic Database and sentiment analysis. Using Weka, regression analysis was done for each feature for personality prediction within 11-18 percent of their actual value. They did not make use of a psychological validation of their results.

(Yarkoni, 2010) reports correlations between LIWC categories and Big Five personality traits. It also reports correlations with lower order facets. 694 participants collected using email or word of mouth were given 100-question and 315-question questionnaires for Big Five, NEO-FFI, NEO-PI-R. Their dataset consists of participants blogs from Google blogger service which may contain more informal text and not enterprise social media. For language usage study, top 5000 unstemmed words (where each blog had more than 50,000 words) in the corpus were ranked with respect to their frequency. These words were correlated with each of the Big Five and other lower order facets. For example, Neuroticism correlated positively with words expressing negative emotion such as awful, lazy, depressing, terrible and stressful; while Extraversion correlated positively with words reflecting social settings or experiences such as bar, restaurant, drinking, dancing, crowd and sang; additionally Openness showed strong positive correlations with words associated with intellectual or cultural experience such as poet, culture, narrative, art, universe and literature. Therefore, we are motivated to explore language use, LIWC to study personality traits.

(Schwartz et al., 2013; Kern et al., 2014; Park et al., 2014) work with the myPersonality dataset which consists of about 19 million Facebook status updates from about 136,000 participants. Their motivation for studying social media as against a psychology lab is that social media language is written in natural social settings, and captures communication among friends and acquaintances. They take two approaches to study language usage in reference to personality traits. One experiment is closed vocabulary study where in for each category for each participant the ratio of sum of frequency of words used by participant in manually created category of language and sum of frequency of words used by participants is noted. Least squares regression is used to link word categories with author attributes, fitting a linear function between explanatory variables (LIWC categories) and dependent variables (such as a trait of personality, e.g. Extraversion). This approach is in some ways similar to earlier approaches. The new approach they take is the open vocabulary approach, where they extract words, phrases(1 to 3 n-grams) and topics (using LDA) via tokenization. The phrases with high pointwise mutual information are retained. Correlation analysis using least squares regression is carried out. They then find categories extending the LIWC category list corresponding to Big Five traits. They also do a predictive evaluation using SVM and ridge regression to predict personality traits using closed/open vocabulary approach. They identify words related to Big Five which are not present in LIWC and any previous analysis. Based on this study, they devise a prediction algorithm to identify personality traits. They do not report whether the myPersonality dataset suffers the challenges of a non-overlapping dictionary with LIWC or personality describing words.

(Banerjee, 2002) describes the lesk similarity algorithm that the software tool (Pedersen et al., 2008) implementation being used as a similarity algorithm is based on. The lesk algorithm uses the information contained in a dictionary to perform word sense disambiguation. Here the dictionary is WordNet. The intuition is that words co occuring in a sentence are being used to refer to the same topic, and topically related senses of words are defined in the dictionary using the same words. It suffers from the fact that lexicographers try to create concise definitions with as few words as possible so even related words may not have common words in their definitions. Using the WordNet relations this is addressed. Every synset in Wordnet has a gloss which is a definition explaining the meaning of the concept of the synset. It also has example sentences. Semantic relationships define a relationship between two synsets. Thus, the glosses of various synset relationships between the word being disambiguated are used as dictionary definitions to the original lesk algorithm. The similarity score between two words is a sum of overlap between the various glosses in Wordnet for each of the two words. The gloss in Wordnet is an approximation of the dictionary definition of the word. Examples of different kinds of glosses used would be example-gloss, gloss-gloss, hypo-gloss.

3 Methodology

Initially we have obtained data from our internal enterprise social network where approximately 300,000 people interact on various topics ranging from technical to work life. This contains different types of posts such as microblogs, blogs, questions, wikis and challenges over a period of 2 years. The other category of content include comments, answers and responses to challenges. Conventional statistical analysis was performed on the data and the following are observed.

One of the ways we identify personality traits is to use a similarity algorithm such as lesk (in Section 2.1) to include adjectives from the dataset that are similar to the adjectives in the HEXACO set for each of the personality traits. In order to increase our yield of personality descriptive words, we include other personality descriptive words similar to the HEXACO set before expanding our set with words similar to those in the dataset. There are 25,553 unique adjectives in the dataset,

Communities	#users	#blog- posts in dataset	#blog- posts of top 50 users	#uBlogs	#uBlogs of top 50 users	#comments per (blog, uBlog)
Technical 1	9,887	9,069	5,301	8,172	5,070	(7, 4)
Technical Sub 1	954	470	322	451	290	(5, 2)
Technical Sub 2	65	43	43	31	31	(3, <1)
Non Technical 1	9,088	2,167	1,053	4,060	2,372	(14, 6)
Non Technical 2	2,442	682	358	350	250	(19, 4)

Table 1: Posting statistics of dataset

Proposed Method	#Words in dictionary	#Words in vocabulary	% unique in dataset	% usage in dataset	% coverage of users
HEXACO	245	152	0.08	0.22	13
HEXACO Extension	2,108	1,999	1.07	3.95	50.18
LIWC	4,487	3,993	2.16	43.77	90.51

Table 2: Coverage statistics of proposed methods

Type	#Unique words	#Total words
Adjectives	25,553	590,910
Nouns	136,592	2,397,410
Pronouns	73	536,813
Verbs	23,033	1,059,940
Total	185,251	4,585,073

Table 3: Part of speech statistics of dataset. Words with different capitalization and spelling are treated as unique.

Trait	Overall % usage	Overall Intensity in dataset	Usage % in Posts	Intensity in Posts	Usage % in Comments	Intensity in Comments
Honesty	0.9	-1424	0.58	-870	0.33	-554
Emotionality	0.66	553	0.45	457	0.21	96
Extraversion	0.96	438	0.58	245	0.39	193
Agreeableness	1.15	-3486	0.75	-2253	0.40	-1233
Conscientiousness	0.75	1539	0.48	958	0.26	581
Openness	0.97	1536	0.67	1071	0.30	464

Table 4: In a work life related community with 7745 people posting 37263 items

Correlation Between	Honesty	Emotionality	Extraversion	Agreeableness	Conscientiousness	Openness
Normalized usage score in posts versus score in feedback	0.0882	0.1770	**0.3362**	**0.3287**	**0.3471**	**0.4005**
Normalized usage score in posts versus positive emotions in posts	-0.0914	0.2040	0.0873	0.0419	**0.3476**	-0.0937
Normalized usage score in posts versus negative emotions in posts	-0.1316	0.1963	-0.0520	-0.0810	**0.2556**	0.0064
Positive emotions in posts versus positive emotions in feedback	**0.3171**	**0.2700**	**0.3844**	**0.4334**	**0.3368**	**0.4689**
Negative emotions in posts versus negative emotions in feedback	**0.3147**	**0.5070**	0.1544	**0.4677**	**0.3845**	**0.3438**

Table 5: Correlations between authored posts and feedback comments received by top users in work life related community of extended HEXACO scores and LIWC emotion categories

	Received Comment Score	+ve Emotion in Posts	-ve Emotion in Posts	+ve Emotion in Received Comments	-ve Emotion in Received Comments
Agreeableness Post Score	0.23	-0.29	-0.58	-0.30	-0.30
Openness Post Score	0.16	0.23	0.54	0.22	0.30
Extraversion Post Score	0.18	0.36	0.43	0.19	0.14
Honesty Post Score	-0.07	-0.35	-0.73	-0.19	-0.48
Conscientiousness Post Score	0.13	0.41	0.21	0.05	0.33
Emotionality Post Score	0.16	0.23	0.54	0.22	0.30

Table 6: Correlations between LIWC processes of emotion in received comments and posted posts in work like related community

which account for 13.79% of the vocabulary. We create a similarity score matrix between the seed set and adjectives in the dataset. In the lesk algorithm using Wordnet, given a set of strings from the gloss' of each word, in order to calculate the overlap score we need the longest common substrings or phrases between them. For each such overlapping substring, the individual score is number of words in the substring squared multiplied by the number of times this substring repeats in the definitions. This score is then weighted with the weight of the type of gloss entry. For example, undemanding is a personality describing adjective of the trait agreeableness and lenient is an adjective in the dataset that has similarity with it and is part of the extended HEXACO set. The words undemanding and lenient have glosses "posing no difficulty requiring little effort" and "demanding little effort not burdensome". The overlapping substring here is "little effort" so the overlap score between these two strings is 2*2*1 = 4. Sum over all the glosses results in a score of 94 for undemanding as an adjective in sense 1. For easy comparison amongst various pairs of words, we normalize the scores by dividing the similarity score of a pair of words with the highest score between the different senses of the pair of words. We threshold the minimum similarity we consider to include the word as similar.

After applying the above algorithm, the earlier list of 245 words was extended to include 2108 words out of which 1,999, i.e., 95% of the words now appeared in the social media content. It was found that 50% of the users have used one of these 1,999 words atleast once. In the next section we propose an algorithm for deriving personality traits of people from their written content based on the usage of this extended set.

We propose a computational means of assigning HEXACO personality trait scores to people based on their posts on enterprise social media. For each person in our dataset, we consider all the posts authored by the person. For each post, for words from the extended HEXACO set, we sum their contribution to the corresponding personality trait and normalize using total words used by the author. Contribution of a word already in the HEXACO set is the loading factor as given in (Ashton et al., 2004). Contribution of a word is the sum of the product of its similarity to a word in a trait and the loading factor of that word in the trait normalized by the total number of words in that trait it is similar to.

4 Observations

From the tables depicting the intensity of each trait in different communities, we can see that openness and agreeableness are well represented and their cummulative intensity in each community is high.

In taking a deeper look into the higher order elements in enterprise social media content we use LIWC2007 (Pennebaker et al., 2007a) on the dataset. 2.1% of our enteprise social media dataset vocabulary are indicative of LIWC processes that account for 43.7% of total enterprise social media content used by 90.51% of the users. This indicates the importance of LIWC processes that are indicative of behavioral traits.

LIWC usage is not directly linked to HEXACO properties, although as reviewed in section 2.1 there have been attempts at using LIWC processes as features that contribute to prediction of Big Five personality traits from web social media. Dataset variability makes it infeasible in many cases to do this mapping as datasets vary in the linguistic features that are indicative of behavior. It is particularly applicable in our case where there are restrained expressions unlike other social media.

We study a subset of users from two communities who have posted atleast a few blogs over the period of 2 years and also have received atleast a few comments so that we may be able to make meaningful observations. We see that openness correlates positively with positive emotion expressed in posts and honesty correlates strongly negatively with negative emotions expressed in posts. We do see that people scoring of their posts on each of the hexaco traits using our method receive near about the same score on the comments they receive indicating that they are possibly perceived as they appear in the posts. From Table 5 and 6, we observe that people's extended HEXACO trait scores on their posts using our methods are strongly correlated with those on the comments they receive, indicating that they are possibly perceived as they appear in their posts. It is slightly lower for Honesty and Emotionality traits but high for Extraversion, Agreeableness, Conscientiousness and Openness. It indicates that people who are more open, agreeable, extraverted, conscientious evoke similar traits from people responding to them in an organization. Another in-

Algorithm 1 Personality scoring algorithm

procedure LESK ADAPTATION
  for each trait of HEXACO **do**
    for each pair of trait adjective and dataset adjective **do**
      for each sense pair in Wordnet **do**
        for each pair of gloss **do**
          gloss_sim = count number of words in overlapping substring * weight of type of gloss
        end for
        total_gloss_sim = $\sum gloss\_sim$
      end for
      score = MAX(total_gloss_sim) over all sense pairs
    end for
    sim = score/(MAX(score) over all dataset adjectives)
    threshold sim by minimum similarity (usually greater than 0.9) and add to extended HEXACO trait
  end for
end procedure
procedure LOADING FACTOR
  for each dataset adjective in extended HEXACO set **do**
    for each trait **do**
      loading_factor = SUM(similarity with each trait adjective * loading factor of that trait adjective)/total number of trait adjectives
    end for
  end for
end procedure
procedure HEXACO SCORING
  for each employee **do**
    for each HEXACO trait **do**
      score = SUM(adjectives used from extended HEXACO set * loading factor of adjective)/number of words used by employee
    end for
  end for
end procedure

teresting observation is that there is a low correlation between openness scores of a person posting and the use of emotive words, which indicates that use of positive emotive words or negative emotive words is largely independent of how open and straightforward a person is and evokes that sentiment. We also see that use of a lot of emotion words positive or negative evokes the same kind of emotion in received comments as well.

5 Conclusion and Future Work

Though the set has increased, however, these words still account for only 1.1% of the vocabulary contributing to 3.95% of total word usage. So it can be concluded that though both usage and coverage have gone up still there is a large volume of enterprise social content which remains untapped. Hence, we propose to look at higher order linguistic elements like phrases, interaction patterns and also LIWC processes, as detailed in (Pennebaker et al., 2007b), in text for better coverage.

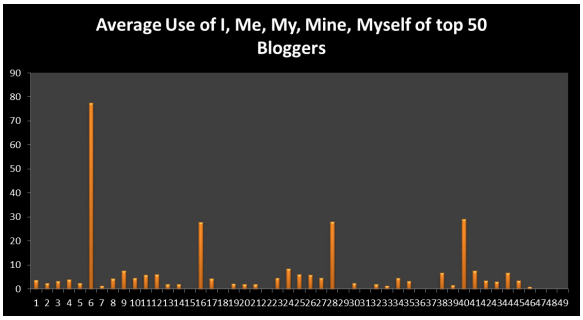

Figure 1: Average usage of first person personal pronouns

Figure 1 is a profile of the average usage of first person personal pronouns by top 50 bloggers. We see that 4 people score significantly higher than others and it is suspected (Tausczik and Pennebaker, 2010; Chung and Pennebaker, 2007) that they are neurotic and depressed. On reading their posts, we find that the highest scorer posts original depressing short stories which have a fan following that encourage the author through positive comments. Therefore, we see that just word usage without communication and other structural aspects do not capture the context in which the words have been used and hence may wrongly identify the author as depressed or neurotic.

As of now we do not have scoring annotations of HEXACO scores using employee completed

(Lee and Ashton, 2004) but we intend to gather text and annotations from employees using surveys to compare our results.

Acknowledgments

We would like to thank Tata Consultancy Services Limited for use of the enterprise social media dataset for research purposes.

References

Michael C. Ashton, Kibeom Lee, and Lewis R. Goldberg. 2004. A hierarchical analysis of 1,710 english personality-descriptive adjectives. *Journal of Personality and Social Psychology*, 87(5):707–721.

Satanjeev Banerjee. 2002. Adapting the lesk algorithm for word sense disambiguation to wordnet. Master's thesis, University of Minnesota.

Cindy K. Chung and James W. Pennebaker. 2007. The psychological function of function words. In *K. Fiedler (Ed.), Social communication: Frontiers of social psychology*, pages 343–359.

Jennifer Goldbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011a. Predicting personality from twitter. In *Proceedings of the 3rd IEEE International Conference on Social Computing*, pages 149–156, Boston, Massachusetts, USA.

Jennifer Goldbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011b. Predicting personality from twitter. In *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011*.

Margaret L. Kern, Johannes C. Eichstaedt, H. Andrew Schwartz, Lukasz Dziurzynski, Lyle H. Ungar, David J. Stillwell, Michal Kosinski, Stephanie M. Ramones, and Martin E. P. Seligman. 2014. The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169.

K. Lee and M.C. Ashton. 2004. Psychometric properties of the hexaco personality inventory. In *Multivariate Behavioral Research*, volume 39, pages 329–358.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2014. Automatic personality assessment through social media language. In *Journal of Personality and Social Psychology*.

Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. 2008. WordNet::SenseRelate::WordToSet. `http://www.d.umn.edu/~tpederse/senserelate.html`.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007a. LIWC2007 for Mac OSX. `http://liwc.net`.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007b. The development and psychometric properties of LIWC2007. `http://liwc.net`.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 09.

Yla R. Tausczik and 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.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. In *Journal of Research in Personality*, volume 44, pages 363–373.