Mood State and Behavior Predictions in Social Media through Unstructured Data Analysis

Abstract

For mood State and Behavior Predictions in Social Media through Unstructured Data Analysis, a new model, Behavior Dirichlet Probability Model (BDPM), which can capture the Behavior and Mood of user on Social Media is proposed using Dirichlet distribution:

$$P(Z_{d.n} = k) \alpha \left(n_{m,(.)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(.),v}^{k,-(m,n)} + \beta_v}{\sum_{r=1}^{V} n_{(.),r}^{k,-(m,n)} + \beta_r}$$

$Z_{d.n} = k$ is the topic which follows Multinomial distribution as described above. $n_{(.),v}^{k,-(m,n)}$ is the number of terms in the given document given that *(m,n)* dimension is removed from study. β_v is the *v*th component of β vector. βr is the *r*th component of β vector.

Context: *There is a colossal amount of data being generated regularly in the form of text from various channels by individuals in the form of posts, tweets, status, comments, blogs, reviews etc. Most of it belongs to some conversation where real-world individuals discuss, analyze, comment, exchange information. Deriving personality traits from textual data can be useful in observing the underlying attributes of the author's personality which might explain a lot about their behavior, traits etc. These insights of the individual can be utilized to obtain a clear picture of their personality and accordingly a variety of services, utilities would follow automatically.*

Keywords: Personality Trait or Behavior Predictions, Mood State Prediction,

1. Introduction

User engagement on social media websites has grown dramatically over the last decade. Activities of the user on social media websites provide valuable information of individual's interests, opinions, behavior and experiences, thus providing insights into his/her personality.

Personality is one of the most complicated attributes of a human being. It also describes how unique an individual is. Personality is one of the fundamental aspects, by which we can understand an individual's behavior. Behavior is a manifestation and amalgamation of the different underlying personality attributes of an individual. Our objective primarily is to access and analyze textual data to identify personality types of their respective authors.

Personality is formally described in terms of the Big Five personality traits from the Five Factor Model (1).

- **Extraversion (EXT) –** outgoing, talkative, energetic versus reserved, solitary
- **Neuroticism (NEU) –** sensitive, nervous versus secure, confident
- **Agreeableness (AGR) –** friendly, compassionate versus challenging, detached
- **Conscientiousness (CON) –** efficient, organized versus sloppy, careless
- **Openness (OPN) –** inventive, curious versus dogmatic, cautious

1.1 Utilities and Applications

In recent years, there has been a rapid growth in interest around personality prediction, particularly from social

media networks. Challenges such as multiple social networks pertaining to a single person or use of other regional languages apart from English are being faced by researchers.

There are numerous fields where one can leverage the findings from personality recognition, such as recommender systems, fraud detection, plagiarism detection, sentiment/mood state/opinion analysis etc.

Research and findings of this discipline are beneficial for many online activities such customer feedback, customer ratings and reviews, products/service recommendations etc. Another important potential of personality prediction lies in the recruitment domain of companies where personality mappings of potential future employees can be obtained well ahead from their job applications.

Personality detection models could also be helpful in fields like e-learning, information and collaborative filtering by a user interface that learns and changed itself accordingly based on the personality of the user.

1.2 Utilities and Applications

Written text by an author provides information of the various attributes that contribute to his/her personality. Our objective is to extract the attributes of an individual's personality from written text obtained from his/her social media accounts.

A brief overview of our approach is as follows:

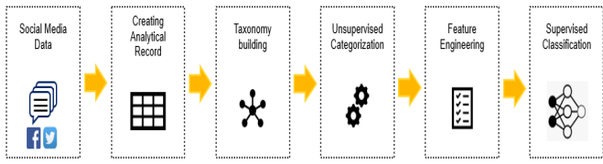


Fig 1.1: Approach Overview flow of control

2. Detailed Approach

The functional components of the approach are as follows:

- **Creation of Analytical Record –** Preparing and rolling up the data to make it usable for modeling
- **Taxonomy Creation –** To categorize the statuses based on frequently occurring words
- **Unsupervised Categorization –** To label each status in dataset to appropriate personality traits
- **Feature Engineering –** To create useful linguistically valuable KPIs which would be fed into classification model
- **Supervised Classification –** To train the model to predict suitable personality traits

2.1 Creation of analytical record

First, we roll up the social media posts data at an author and post level. For this exercise, we are using the myPersonality(2) dataset. The original publicly available database by myPersonality which contained Facebook statuses of ~250 users with number of statuses ranging from 1 to ~ 200, is no more available online anymore. However, a sample of ~ 10k statuses is still available.

To conduct this research, we have used the sample dataset consisting of the ~10k records.

The data is rolled at an author and post level and looks somewhat like the following:

#AUTHID	STATUS
00419a4c96b32cd63b2c7196da761274	back in cali!!!
02c37028a782cfda660c7243e45244bb	Supervisor: *PROPNAME* (second preference) Research Area: Re
02c37028a782cfda660c7243e45244bb	Tentative Examination Schedule, Semester 1, 2009//10. .
02c37028a782cfda660c7243e45244bb	abcdefghijklmnopqrstuvwxyz qwertyuiopasdfghjklzxcvbnm mnb
02c37028a782cfda660c7243e45244bb	Pa><dol x 2
02c37028a782cfda660c7243e45244bb	cartography+Select. Topics in the Geography of China (????F?a
02c37028a782cfda660c7243e45244bb	DEA DEA DEA :D
03133a828cd0cf52e3752813ce5d818f	Did Cindy 30 times in 20 minutes and GI jane in 12:11. Save the se
03133a828cd0cf52e3752813ce5d818f	Did *PROPNAME* in 16:37, made money in Vegas, and just had hi
03133a828cd0cf52e3752813ce5d818f	Did *PROPNAME* in 16:37, made money in Vegas, and just had hi
03133a828cd0cf52e3752813ce5d818f	Feels aweful right now. Why do you get sick like this on weekenc
03133a828cd0cf52e3752813ce5d818f	Did *PROPNAME* in 16:37, made money in Vegas, and just had hi

03e6c4eca4269c183fa0e1780f73faba	"Those who criticize our generation forget who raised it."
03e6c4eca4269c183fa0e1780f73faba	"In awe I watched the waxing moon ride across the zenith of the
03e6c4eca4269c183fa0e1780f73faba	"I refuse to answer that question on the grounds that I don't kno
03e6c4eca4269c183fa0e1780f73faba	"Parents spend the first part of our lives teaching us to walk and
03e6c4eca4269c183fa0e1780f73faba	"Ah, *PROPNAME*. The eyes are open, the mouth moves, but *P
03e6c4eca4269c183fa0e1780f73faba	"Death and famine stalk the land like two great stalking things."
03e6c4eca4269c183fa0e1780f73faba	"Arguments are to be avoided; they are always vulgar and often
03e6c4eca4269c183fa0e1780f73faba	podkanva vseki, koito ima da i prashta snimki ot razlichni meropr
06b055f8e2bca96496514891057913c3	is enjoying the cricket...comfy boxers and rainy weather...is it too
06b055f8e2bca96496514891057913c3	can't believe that her son walked up to the Christmas tree stating
06b055f8e2bca96496514891057913c3	got home from shopping all evening (*PROPNAME* and kids in to
06b055f8e2bca96496514891057913c3	- so, this morning *PROPNAME* gets up to play and he goes over
06b055f8e2bca96496514891057913c3	needs beer...actually, any kind of alcohol will do...

Fig 2.1: Sample Analytical Record

Here one author can have multiple entries in the status column. However, for simple and easy computation, we concatenate the multiple posts of an author to one single post and treat it as an essay. At the same time, we also keep track of the frequency of posts made by the author. Thus, now our rolled up analytical record would contain one author in every row with all posts or statuses concatenated and the frequency or count of statuses made.

2.2 Taxonomy Creation

To easily categorize the statuses of the authors into personality traits, we have used lists or collections of frequent and commonly used words corresponding to each personality trait. These words were compiled from the *World Well-Being Project data for 'The Language of Personality'.(3)*

OPENNESS	CONSCIEN	EXTRAVEF
cant wait	bored	gym
had a good	meh	out with
cant believe	ftw	great wee
listening to	dead	a great nig
society	emo	had an am
awesome	kill	love u
humanity	omfg	love my li
the sky	wtf	ready to
dream	dammit	excited fc
the universe	i hate	a great
.	.	.
.	.	.

Fig 2.2: Sample words for each Personality Trait

2.3 Unsupervised Categorization

Once we have the set of words for taxonomy corresponding to each of the personality traits, we then create variables to count the number of words corresponding to each personality the author uses. To normalize, we divide the frequency or word counts by the number of statuses the respective authors had posted.

avg_EXT_freq	avg_NEU_freq	avg_AGR_freq	avg_CON_freq	avg_OPN_freq
5.6	1.3	4.8	0.6	1.3
0.9	2.4	1.6	5.5	2.7
4.5	3.3	1.8	1.7	3.4
3.4	1.4	2.9	2.7	2.7
3.3	3.1	1.9	5.3	3.9
5.1	4.1	3.2	2.1	1.9

Fig 2.3: Sample normalized frequencies for words

The above table means – the first author used on an average 5.6 EXTRAVERSION words per status, 1.3 NEUROTICISM words per status, 4.8 AGREEABLE words per status and so on. From these normalized frequencies, we would be tagging each author to respective personality traits based on the frequency scores.

2.4 Feature Engineering

The features used in the analysis are inspired by prior psychological studies about correlations personality traits and linguistic factors.

We started off with extracting frequency counts of word categories from the Linguistic Inquiry and Word Count (LIWC) utility (Pennebaker et al., 2001) (4). Creating these features helps in capturing both syntactic (e.g., ratio of individual parts of speech) and semantic information (e.g., positive sentiment/mood state words). Some of

these features are illustrated below. Pennebaker and King (1999) (5) had earlier found significant correlations between these features and each of the Big Five personality traits.

We also added additional features from the MRC Psycholinguistic database (Coltheart, 1981) (6), which contains statistics for over 150,000 words, such as estimates of the age, acquisition, familiarity and frequency of use.

Generally, introverts would take longer to reflect on the words they say, Heylighen and Dewaele (2002) (7) suggest that an introvert's vocabulary is more precise, implying a lower frequency of use. The MRC set was also earlier used by Gill and Oberlander (2002) (8), who demonstrated that extraversion is negatively correlated with concreteness. Concreteness also indicates neuroticism, as well as the use of more frequent words (Gill & Oberlander, 2003) (9).

2.4.1 LIWC Features (Pennebaker et al., 2001):

- **Standard counts:**
  - Word count, words per sentence, words captured, words with more than 6 letters, type/token ratio, assents, negations, prepositions, articles, numbers
  - Pronouns (Pronoun): 1st person singular, 1st person plural, total 1st person, total 2nd person, total 3rd person (Other)
- **Psychological processes:**
  - Affective or emotional processes: positive emotions, positive feelings, optimism and energy, negative emotions, anxiety or fear, anger, sadness
  - Cognitive Processes: causation, insight, discrepancy, inhibition, tentative, certainty
  - Sensory and perceptual processes: seeing, hearing, feeling
  - Social processes: communication, references to people, friends, family, humans
- **Relativity:**
  - Time, past tense verb, present tense verb, future tense verb
  - Space: up, down, inclusive, exclusive
  - Motion
- **Personal concerns:**
  - Occupation: school, work and job, achievement
  - Leisure activity: home, sports, television and movies, music
  - Money and financial issues
  - Metaphysical issues: religion, death, physical states and functions, body states - symptoms, sexuality, eating, drinking, sleeping, Grooming
- **Other dimensions:**
  - Punctuation: period, comma, colon, semi-colon, question, exclamation, dash, quote, apostrophe, parenthesis, other - Swear words, non-fluencies, fillers

2.4.2 MRC Features (Coltheart, 1981):

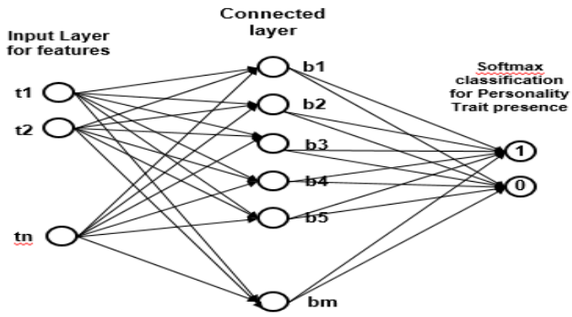
Number of letters, phonemes, syllables, Kucera-Francis written frequency, Kucera-Francis categories count, Kucera-Francis sample count, Thorndike-Lorge written frequency, Brown verbal frequency, familiarity rating, concreteness rating, imageability rating, meaningfulness Colorado Norms, meaningfulness Paivio Norms, age of acquisition

2.4.3 Utterance Type Features:

Ratio of commands, prompts, back-channels, questions, assertions.

2.5 Supervised Classification

For binary classification, we use a two-layer perceptron consisting of a full connected layer of size 20 and the final Softmax layer of size two, representing the 1-0 or yes-no classes.



Connected layer - We multiply the document-feature d $\in$ R_m by the weight matrix W*(connected layer)* $\in R^{n\times m}$ and add a bias B*(connected layer)* $\in$ R^n to obtain the vector d*(connected layer)* $\in$ R^n. We introduce nonlinearity with Sigmoid activation, where $\sigma(x) = 1/(1 + \exp(-x))$.

Softmax output - We use the Softmax function to determine the probability of the document belonging to the classes 1 or 0.

For this, we build a vector (x_1, x_0) = d*(connected layer)* W*(softmax)* + B*(softmax),*where W*(softmax)* $\in$ $R^{m\times 2}$ and the bias B*(softmax)* $\in$ R^2, and we calculate the class probabilities as

$$P(i \mid features) = \exp(x_i) / \exp(x_1) + \exp(x_0) \text{ for } i \in \{1,0\}.$$

3. Conclusion

Social network usage has gone up exponentially and there is huge amount of data available for analysis and insights generation today. Deriving personality traits from text written by an author holds immense potential in a variety of domains ranging from product recommendation to recruitments to virtual dating services. Our approach is mostly dependent on linguistic features of the text. Further refinement can be done by layering in social network metrics such as friends' networks, influence scores etc. which should help further improve the predictions and accuracy as well as provide more valuable insights about the same.

References

1. J. Digman, "Personality Structure: Emergence of the Five-Factor Model," Ann. Rev. Psychology
2. https://sites.google.com/michalkosinski.com/mypersonality

   https://github.com/dbrehmer/Knowself
3. http://www.wwbp.org/
4. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum, Mahwah, NJ.
5. Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296–1312.
6. Coltheart, M. (1981). The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A, 497–505.
7. Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: an empirical measure. Context in Context, Special issue of Foundations of Science, 7 (3), 293–340.
8. 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).
9. Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Processes, 42, 239–270.https://doi.org/10.1 080/1462220041 0001676305