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. There is a colossal amount of data being generated regularly on social media 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. Using Dirichlet probability distribution, the aim is to estimate the probability of each personality trait (or mood state) for an author and then model the latent features in the text which are not captured by the BDPM. As a result, the study can be helpful in prediction of mood state/personality trait as well as capturing the significance of the latent features apart from the ones present in the taxonomies, which will help in making an improved mood state or personality prediction.

Keywords: Personality Trait; Behavior Predictions; Mood State Prediction; Dirichlet Distribution; Linguistic Features

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 asses 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.

- **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

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 or mood state from written text obtained from his/her social media accounts using the Dirichlet probability distribution and then identifying latent features from text that are not captured by the taxonomies. Once the latent features are obtained, those can be incorporated with the taxonomy features to get an improved approach for prediction of mood state and personalities.

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

Source: wekipedia

$Z_{d.n} = k$ is the topic which follows Multinomial distribution.

m represents number of documents

n represents total number of words in all documents

β is a real number vector of dimension v

v is number of words in vocabulary

d is the document under study

k is number of topics

α is a positive real vector of dimension k

$n_{(.),v}^{k,-(m,n)}$ is the number of terms in the given document, given that (m,n) dimension is removed from study.

β_v is the vth component of β vector.

β_r is the rth component of β vector.

Once the text data has been tagged with a personality trait or mood state basis the Dirichlet approach, the latent features are engineered based on the Linguistic Inquiry and Word Count (LIWC) utility (Pennebaker et al., 2001) and MRC Features (Coltheart, 1981). These features are then tested for significance in its ability to explain the variability within the text in terms of mood state or personality trait. Lastly, by incorporating the significant features with the existing taxonomies will result in the final prediction.

A brief overview of our approach is as follows:

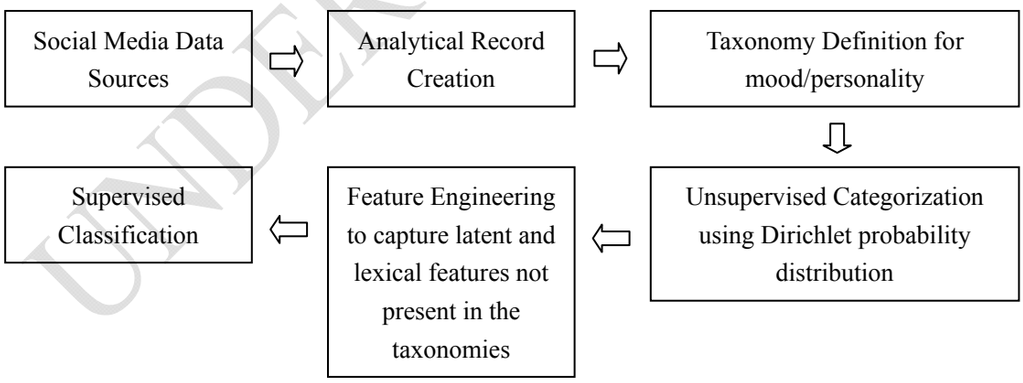

Fig 1.1: Approach Overview flow of control

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.

2. Materials and Methods

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 or mood states using Dirichlet probability distribution
- **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 to create an improved prediction model

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' (source: github.com) dataset. The original publicly available database by 'myPersonality', which contained Facebook statuses of ~250 users with number of posts ranging from 1 to ~200 per user. A sample of ~10,000 records are considered for this exercise.

To conduct this research, we have used the sample dataset consisting of the ~10,000 records, the screenshot below shows the first 25 records which are the different status updates of the same author:

#AUTHID	STATUS
b7b7764cfa1c523e4e93ab2a79a946c4	likes the sound of thunder.
b7b7764cfa1c523e4e93ab2a79a946c4	is so sleepy it's not even funny that's she can't get to sleep.
b7b7764cfa1c523e4e93ab2a79a946c4	likes how the day sounds in this new song.
b7b7764cfa1c523e4e93ab2a79a946c4	is home. <3
b7b7764cfa1c523e4e93ab2a79a946c4	www.thejokerblogs.com
b7b7764cfa1c523e4e93ab2a79a946c4	saw a nun zombie, and liked it. Also, *PROPNAME* + Tentacle!Man + Psychic Powers = GREAT Party.
b7b7764cfa1c523e4e93ab2a79a946c4	is in Kentucky. 421 miles into her 1100 mile journey home.
b7b7764cfa1c523e4e93ab2a79a946c4	is celebrating her new haircut by listening to swinger music and generally looking like a doofus.
b7b7764cfa1c523e4e93ab2a79a946c4	has a crush on the Green Lantern.
b7b7764cfa1c523e4e93ab2a79a946c4	has magic on the brain.
b7b7764cfa1c523e4e93ab2a79a946c4	saw Transformers, Up, and Year One this week. Good movie overload. :D
b7b7764cfa1c523e4e93ab2a79a946c4	Who wants to meet up on schedule pick-up day at Oviedo?
b7b7764cfa1c523e4e93ab2a79a946c4	desires the thrill of inspiration. Also, money.

b7b7764cfa1c523e4e93ab2a79a946c4	is going to bed at 9:30! Yeah!
b7b7764cfa1c523e4e93ab2a79a946c4	is reading, admiring her permit, and occasionally glancing at her ner McDonald's uniform.
b7b7764cfa1c523e4e93ab2a79a946c4	thinks intangibility should be an option in reality settings.
b7b7764cfa1c523e4e93ab2a79a946c4	is tired. *PROPNAME*, let me go to sleep pl0x.
b7b7764cfa1c523e4e93ab2a79a946c4	is discovering the many flavors of insomnia.
b7b7764cfa1c523e4e93ab2a79a946c4	is watching cousin play computer game on televison box thing. Also, sleepy.
b7b7764cfa1c523e4e93ab2a79a946c4	Why is it I'm only getting the urge to draw when I have stuff to do for school? D;
b7b7764cfa1c523e4e93ab2a79a946c4	Who'da thought a single text message could be enough to change my mind?
b7b7764cfa1c523e4e93ab2a79a946c4	wishes to develop a super power that prevents her from needing to sleep.
b7b7764cfa1c523e4e93ab2a79a946c4	TELL ME WHAT TO DRAW, PLOX.
b7b7764cfa1c523e4e93ab2a79a946c4	found a bunny, bunny died, buried bunny, now is drawing.

Figure 2.0: Sample dataset

As each author has multiple statuses ranging from 1 to 200, thus all records for all the authors are rolled up to have a single record per author. Below figure shows the analytical record for 3 authors. Author #1 had 5 status updates, and all are concatenated to make a single record. The same has been performed for all the remaining authors.

#AUTHID	ALL STATUSES CONCATENATED
06b055f8e2bca96496514891057913c3	needs beer...actually, any kind of alcohol will do... is sick of being sick! is enjoying the cricket but has no fingernails left. would like the rain back! had a great day yesterday thanks to aunties babysitting :0)
0724fe854bd455061ba84efecdeff469	Getting ready for the fun! Escapada espiritual al corazón del oráculo... Riviera Maya! Meus somnierum! I want to reconcile the violence in your heart! I want to recognise your beauty is not just a mask! I want to exorcise the demons from your past! I want to satisfy the undisclosed desires in your heart! (Gracias a Sofi por la canción tan genial!) I just... Don't know what to think anymore...
0737e4e4980f56c9fb1cb5743001c917	might have to trade in bastille day celebrations for a nice quiet night in my apartment...if anyone wants to hang tonight let me know. would rather have too many dreams and reach half of them then not enough dreams to reach for. wonders if it's a better idea to stay home and not celebrate Hemingway's birthday with free absinthe drinks since I think I'm getting sick :(has the Camino De Santiago on my mind again...the effects of receiving emails from my Spaniard friends. is a sucker for academics

Fig 2.1: Sample Analytical Record

In the original dataset, 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 a collection of text. At the same time, we also keep track of the frequency of posts made by the author. Thus, now the rolled up analytical record contains 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, lists or collections of frequent and commonly used words corresponding to each personality trait has been used. These words were compiled from the *World Well-Being Project data for 'The Language of Personality'* (source: http://www.wwbp.org/data.html). For each personality trait, there are ~ 200 words that has been considered in the taxonomy.

Below are the few sample words considered for each of the 5 personality traits in this study.

Trait	Taxonomy terms (sample)
OPN	sigh, apparently, zombie, into the, strange, epic, dreams, poetry, that's, i've been, death, they're, dream, music, soul, the universe, writing, i've, art, universe
CON	had a great, so excited, wonderful, excited, vacation, long day, workout, at work, ready to, back to work, relaxing, thankful, a great, ready, blessed, great day, to work, ready for
EXT	lovin, cant wait to, great night, goin, hit me up, night with, i love my, ya, dont, chillin, lil, gettin, love you, baby, cant, im, girls, cant wait, party
AGR	psalm, thanksgiving, in christ, thank you, excited for, great day, an amazing, had a great, the lord, an awesome, a wonderful, prayers, amazing, wonderful, a great, blessed, excited
NEU	blessed, the lord, praise, chillin, beautiful day, fam, blessings, soccer, beach, workout, basketball, lakers, smh, success

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.

AuthorNum *	Average EXT Count	Average NEU Count	Average AGR Count	Average CON Count	Average EXT Count
1	5.6	1.3	4.8	0.6	1.3
2	0.9	2.4	1.6	5.5	2.7
3	4.5	3.3	1.8	1.7	3.4
4	3.4	1.4	2.9	2.7	2.7
5	3.3	3.1	1.9	5.3	3.9

Fig 2.3: Sample normalized frequencies for each personality trait
(* - sensitive author information, hence author ids changed to numbers)

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. For each author, term-author

frequency matrix with words included in taxonomies is created. That matrix is considered as an input to the Dirichlet probability distribution for predicting the probabilities of five personality traits for each user.

Below is the distribution used:

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

Below are the results obtained using Dirichlet probability distrubition:

AuthorNum *	Probability EXT Count	Probability NEU Count	Probability AGR Count	Probability CON Count	Probability EXT Count
1	0.41	0.10	0.35	0.04	0.10
2	0.07	0.18	0.12	0.42	0.21
3	0.31	0.22	0.12	0.12	0.23
4	0.26	0.11	0.22	0.21	0.21
5	0.19	0.18	0.11	0.30	0.22

Fig 2.4: Sample probabilities basis Dirichlet distribution

Comparing the above figures 2.3 and 2.4, it can be observed that the authors who has used more EXT words in their text, has higher probability of EXT personality trait basis the Dirichlet distribution. Same holds valid for other traits as well.

2.4 Feature Engineering

The features used in the analysis are inspired by prior psychological studies about correlations between 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). 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) 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), 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) 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), 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).

Features tested in unsupervised model are as belows:

- **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

o **MRC Features (Coltheart, 1981):**

• *Dimensions:*

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

o **Utterance Type Features:**

• *Dimensions:*

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

2.5 Supervised Classification

Based on above feature vectors and personality traits predicted using Dirichlet probability distribution, a supervised classification model is trained with binary target variable which reflects the presence or absence of a particular personality trait.

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. The below setup is for EXT trait, but the same can be replicated for the remaining four traits in similar fashion.

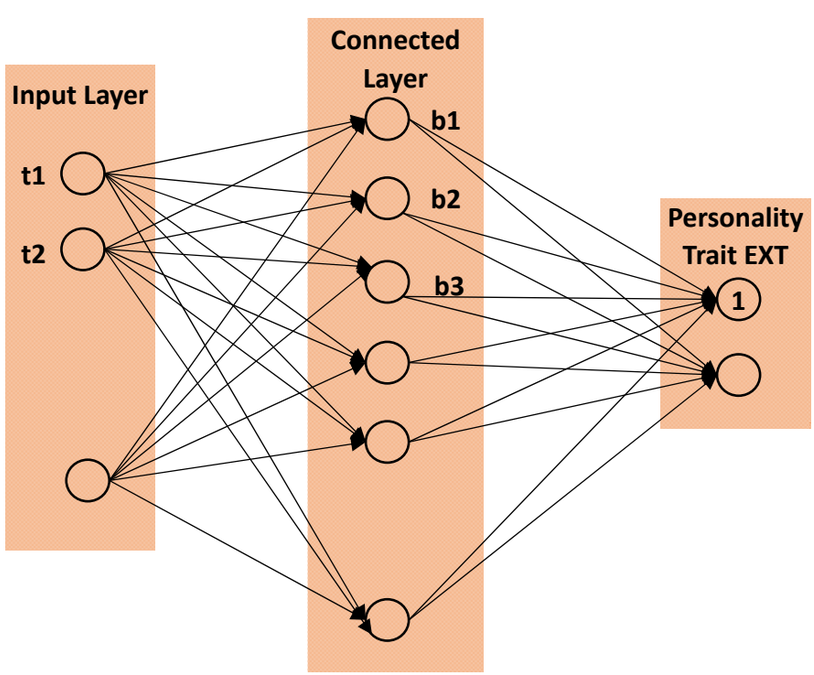

Figure 2.5: Fitted Neural Network model with 20 nodes in connected layer and ~280 features in input layer

b20

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 activation 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

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

The above implementation of the neural network was done using the open source tool R.

3. Results and Discussions

With the setup discussed in above section, the model was trained for EXT trait and the following results are observed. It can be clearly seen in the below figure that for EXT trait, there are latent features from the text which are not included in taxonomy but have significance in explaining the variability of EXT trait. Hence, if those latent features are incorporated in a classification model along with the defined taxonomies, it will result in predicting the traits better, even in situations where a large percentage of words in text are not present in the defined taxonomy.

Trait	Feature Type	Feature	Relative Importance
EXT	Taxonomy	lovin, cant wait to, great night, hit me up, i love my, ya, chillin, gettin, love you, baby, cant wait, party	63%
	Latent	Number of words, words per sentence, prepositions	6%
	Latent	Financial Topics	8%
	Latent	Count of communication threads involved in	11%
	Latent	Out-degree or number of friends in network	13%

Fig 2.6: Result showing variable importances for EXT

The features present in taxonomy contribute 63% to overall variability and remaining 37% are explained by the latent features from the text, which shows that the added features which are not present initially in the taxonomy

can explain a significant variability for the traits.

4. 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 depends on finding value from the taxonomic features of the text and then using latent features to explain an additive variability of personality traits. Further refinement can be done by layering in social network metrics such as friends' networks, influence scores etc. which should help further improve the prediction accuracy and provide more valuable insights about the same.

References

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3: 993–1022

Chang, M. W., Ratinov, L., Roth, D., and Srikumar, V. 2008. Importance of semantic representation: dataless classification. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. AAAI Press, Chicago, IL, 830–835.

Coltheart, M. 1981. The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A, 497–505.

Digman, J. 1990. Personality Structure: Emergence of the Five-Factor Model, Ann. Rev. Psychology, 41, 417-440

Dumais, S. T. 1994. Latent semantic indexing (lsi) and trec-2. In Proceedings of the Second Text REtrieval Conference (TREC-2). NIST, Gaitherburg, MD, 105–116. Egozi, O., Gabrilovich, E., and Markovitch, S. 2008. Concept-based feature generation and selection for information retrieval. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. AAAI Press, Chicago, IL, 1132–1137.

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.

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.

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).

Pennebaker, J. W., Francis, M. E., & Booth, R. J. 2001. Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum, Mahwah, NJ.

Pennebaker, J. W., & King, L. A. 1999. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296–1312.