

Application of Natural Language Processing on Character Computing : A Short Review

Koushik Deb (Assistant professor)
Computer Science and Engineering,
IEM, Salt Lake, India
koushikdeb2009@gmail.com

Sayan Banerjee
Computer Science and Engineering,
IEM, Salt Lake, India
sayanbanerjee5263@gmail.com

Rohan Sharma,
Computer Science and Engineering,
IEM, Salt Lake, India
980rohan2702@gmail.com

Sana Bano Khan,
Computer Science and Engineering,
IEM, Salt Lake, India
sanakhanbano@gmail.com

Sk Saqlain Ali,
Computer Science and Engineering,
IEM, Salt Lake, India
saqlainrocks7@gmail.com

Abstract—Character Computing consists of not only personality trait recognition, but also correlation among these traits. Tons of research has been conducted in this area. Various factors like demographics, sentiment, gender, LIWC, and others have been taken into account in order to understand human personality. In this paper, we have concentrated on the factors that could be obtained from available data using Natural Language Processing. It has been observed that the most successful personality trait prediction models are highly dependent on NLP techniques. Researchers across the globe have used different kinds of machine learning and deep learning techniques to automate this process. Different combinations of factors lead the research in different directions. We have presented a comparative study among those experiments and tried to derive a direction for future development.

Keywords: *NLP · SVM · Deep Learning · Neural Networks · character computing · personality trait recognition · psychoanalysis*

I. INTRODUCTION

Personality traits play a significant role in a person's psychoanalysis. It also defines how a person behaves, like their response, reaction and way of addressing an issue. This analysis gives us the knowledge to predict an individual's preferred reaction to the circumstances and scenarios. Usage of such predictions have been successfully exploited in predicting students' performance, the recommendation for buyers, patient diagnostics, etc. Previously this analysis used to be done by survey with multiple choice question answers. Over the last decades, there have been several efforts to automate this process. Not only automate but to derive psychological patterns without asking any direct questions. Thriving data in social media adds fuel to it. Different blogs and transcripts of vlogs [2] have been used with permission from subjects. Studies to date used a different technique on several types of features.

It has been noticed that most of the features have been derived by Natural Language Processing. Different types of machine learning processes like SVM, Naive-Bayesian Classifier, Logistic Regression, RNN have been used to classify an individual with their rating in Big Five personality trait spectrum [11][8]. These are Extraversion(EXT), Emotional stability (EMS), Agreeableness (AGR), Conscientiousness (CON), Openness to experience (OPN) [3]. Most of the experiments used supervised machine learning techniques to predict a person's behavior which assigns a high or low score in each of these traits.

II. DIFFERENT PREDICTION METHODS

A. Conditional Probability on Personality Trait

Motivation-

In several studies, it has been shown that personality traits are correlated. To exploit these characteristics many scientists addressed this issue with a structured classification. In this approach, dependencies among the output variable have been considered in the model. As an example, we can say a decrease in emotional stability is strongly correlated with a decrease in agreeableness[15].

Existing Solutions and Discussion -

When scientists start with machine learning, the first challenge comes with feature selection. Experiments on feature engineering show a variety of results on a different dimension. Word-based n-gram feature is very commonly used in character computing. Jon Oberlander and Scott Nowson used the WMatrix tool (Rayson, 2003) for CLAWS tagging of a proper noun and collapsed them into a single marker[1]. This gives a specific view of bi and tri-gram word features. They put different constraints to derive 5 levels of features and applied Naive-Bayes Classifier. In some cases, they reached 100% accuracy[15]. Francisco Iacobelli and Aron Culotta used NB in a different way. They used Conditional Random Fields (CRFs) (Lafferty, McCallum, and Pereira 2001). For a specific sequence of words and sequence of labels, CRF will denote the probability of the defined sequence of output labels, given a sequence of words. Conditional probability can be shown as follows

$$P(x|y) = \frac{1}{Z(x)} \prod_k^Y f_k(x, y, \lambda)$$

Where $f_k: X \times Y \rightarrow R^+$, is factor with λ as parameter. $Z(x)$ is a normalized constant.

They have used it in pairwise label feature factor and achieved the highest accuracy in Agreeableness 54.83%. With simple Naive Bayes Classification, they achieved 55.11% in Conscientiousness and 68.81 in Openness[16]. Corpora produced by Pennebaker and King (1999) and myPersonality corpus had been used in these experiments. It can be observed that one corpus has been taken from before the 20th century and another is facebook status. The effort is to find the linguistic style which is dominant all throughout.

Future Work -

Researchers need to establish parallel-ground truth values for personality traits on collected data. There can be a gap between perception and true correlation among personality traits and text. We need more clear understanding among the projected and perceived character traits of humans. Feature analysis and feature selection will play a great role in accuracy. This is an open area to improve computing. A small change in any feature will affect the result in a big way.

B. Mapping Psycho-Lingual Features into Hyper-Plane

Motivation -

Automatic recognition of personality traits from speech, visual expressions or textual content is aimed at building classifiers generated using a supervised machine learning approach that learns the patterns from data. Currently, this task is most effectively handled using Supervised Machine Learning Approaches. SVM is used as a classifier in relation extraction. It simply classifies the features to different relation classes [16]. SVM is used in many state of the art classification tasks and on a very high dimensional feature space. It is a great model for avoiding overfitting. Prepare Your Paper Before Styling

Existing Solutions and Discussion -

The goal of personality trait recognition is assigning a personality profile to the individual. This personality profile is usually conceptualized as consisting of five different traits (Big Five) that are present to a certain degree in an individual. Researchers have shown how

content-independent features for gender prediction can also achieve good results in personality recognition. B. Verhoeven et al. classified the entire features in five different groups: Character-based, Word-based, Sentence based, Dictionary-based and Syntactic [3]. By using ten-fold cross-validation B. Verhoeven et al. have shown that we can beat both baselines for EXT, NEU, CON and we can beat weighted random baselines for OPN [3] [6]. While AGR has a more skewed distribution and appears harder to learn. It reached an accuracy of 54%. Riccardi et al. presented on automatic personality recognition from a video-blog corpus [2] by studying different types of features sets such as audio-Visual Features, Lexical Features, POS Features, Psycho-linguistic Features (LIWC) and Emotional Features (Emo) [5]. This set of scientists generated a classification model using Sequential Minimal Optimization (SMO) for SVM. In this case Linear Kernel was chosen. The performance of the classification model was designed using different feature sets, decision combination, trait features, and their best F1 score was 67.3%. They have also used predicted traits as features and designed a cascaded classification system. It obtained a promising result compared to the official baseline. It has an average success rate of 62.6%. T. Smedt et al have reported on a proof of concept using ensemble learning as a way to address the data acquisition problem which is a bottleneck in the development of robust personality recognition system. They have used Ensemble methods (Seni and Elder 2010) which is an approach in Machine Learning that has been successful in improving the accuracy of systems by combining the predictions of different component classifiers[3]. These

scientists showed that an ensemble method based on meta-learning allows the productive use of out-of-genre data in a more effective way than simpler approaches. They have reached an accuracy level of 74.2%.

Future Work-

In future research, it needs to be investigated further in a complete cross-validation set up and using error analysis to provide deeper insights into the condition under which this approach is useful. Performance of a few models using a certain feature set is very low compared to the other feature sets, however, it helps in combination. We plan to experiment with the traits and those features with other datasets in the future.

C. Approaches to Put Text and Behavior on a Line : Regression Technique

Motivation -

Logistic regression is one of the most important analytic tools in the social and natural sciences. In natural language processing, logistic regression is the baseline supervised machine learning algorithm for classification and it also has a very close relationship with neural networks. A neural network can be viewed as a series of logistic regression classifiers stacked on top of each other. Logistic regression can be used to classify an observation into one of two classes (like positive sentiment and negative sentiment), or into one of many classes.

Existing Solutions and Discussion-

Logistic Regression treats the task as binary classification problems (one for each personality trait) and involves experimentation with different combinations of the features. Scientists have been in this

research to detect significant features which is an essential part of a personality recognition system. Chandrima Sarkar et al. performed an in-depth analysis for multi-modal personality recognition tasks using YouTube personality dataset by the help of Logistic Regression[7]. They presented a study on diverse features categories such as Audio-visual cues, Text features, Word statistics features, Sentiment features, Demographic features. This set of researchers used a separate training set for building a classification model with 86% of the data and the rest as test instances. Chandrima Sarkar et al. found that non-verbal features such as Audio-visual and Sentiment features help in the identification of personality types of Extraversion, Conscientiousness, Emotional Stability. Text features are more significant in predicting Openness to experience personality type. Using the Logistic Regression model it reached different levels of accuracy for different personality types, such as for EXT (77%), AGR(67.9%), CON(64.3%), EMS(59.5%), OPN(64.8%).

Mark T.Tomlinson et al. presented a method for detecting an individual level of conscientiousness based on an analysis of the content of their Facebook status update [9]. Primarily they were interested in two different thematic roles, the agent role, and the patient role. For each individual, they calculated an average depth and average objectivity for predicates that had a First-person agent, predicates which had a Non-first person agent, predicates which had the First-person patient and finally predicates with Non-first person patients. The model achieved a correlation of $r = 0.27$ on a subset of Facebook data published for myPersonality workshop with an accuracy of 58.13% for detecting if an individual is above or below the median. It reached 68.03% for those outside of one standard deviation. Utilizing a similar analysis with a 10 fold cross validation approach it showed an average raw correlation of 0.26 to the held out test.

Future Work-

We encountered that most individuals have a rating around the mean with relatively few outliers. This creates a problem with regards to framing a binary classification task, trying to detect if an individual is above or below the mean is always going to be a very difficult problem because most people are very close to the mean. As future work, we plan to develop a dedicated feature selection strategy for a more improved personality recognition system.

D. Classification Using Neural Network

Motivation –

Deep Learning is an umbrella where most of the NLP problems find their best solution. There are several aspects of these technologies that let scientists address their issues in a sophisticated way. Word Embedding is a salient feature. It preserves the closeness of meaning by its vector representation. It reduces the gigantic dataset to a dictionary of unique words. Word2vec, GloVe are commonly used, word embedding models. In some cases, morphological information in words need to be preserved. In today's world social media plays a significant role as data source of NLP study. People post lots of expressions or we can say micro text which can not be treated as a

defined word but they have their meaning. Also, it can deal with unknown words and keep account of them. Also features like n-gram can be used to extract higher level features. It helps in problems like sentiment analysis, detecting emotional expression. These all contribute to personality trait recognition.

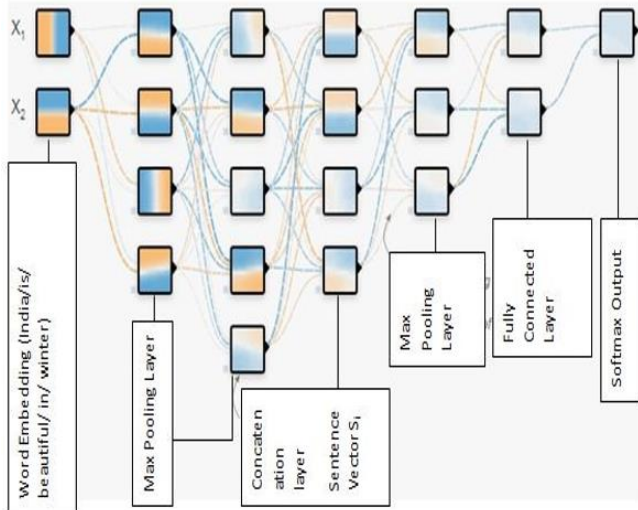
Existing Solutions and Discussion-

Scientists have a tendency to imitate the human brain to solve the most complicated problems. Deep learning comes into play whenever it is a neural network.

T. Tandra et al. used various techniques like MLP (Multi-Layer Perceptron), LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit), and CNN 1D (1-Dimensional Convolutional Neural Network) [10]. These experiments have been conducted on myPersonality project sample data. It has been found that deep learning shows the highest accuracy over other machine learning methods but there is no single architecture found that can give good results for all five traits of personality. Using the combination of MLP, LSTM+CNN 1D architecture produced comparatively best result among all[12]. It reached an accuracy of 93.33%. They used multiple features like EmoLex, Mairesse. NRC Emotion Lexicon is a corpus to extract emotional words. 14,182 words are present in this lexicon. It will generate a vector of 10 dimension. Each dimension represents an attribute like anger,disgust,anticipation,fear etc. Francois Mairesse et. al. studied personality in various levels[11]. They proposed features from text to address those levels like Conversational Behaviour,Topic Selection,Style,Speech etc.Features like Content and Syntax, Utterance Type,Prosody are used to measure different traits.[12].

N. Majumder et al. conducted a very intelligent pre-processing technique to boot the training. They pruned those sentences which are not carrying any personality clue. These sentences were treated as noise. They used word2vec embedding to train network. In figure 1 , the process has been narrated in a simple manner. They have shown different success rates for different traits based on features selection and model used. CNN with Mairesse feature gave the highest accuracy rate of 62.68%in detecting openness of subject[12]. F. Liu et al. focused on the character level expression in text. As they have used twitter data, there are lots of character-level expressions and emoticons. They used a character-level bi-directional RNN(Char-Bi-RNN) . Next word level Bi RNN will use the constructed word vector. The subsequent layer will take sentences as input and produce a scalar as output. This is the measure of personality trait. The model was named after different layers as Character to Word to Sentence for Personality Traits (C2W2S4PT). As it is growing through character, the model is language independent. Gated Recurrent Unit has been used as a unit in Bi-RNN. They executed their experiment on Italian, Spanish and English. RMSE results for English on predicting Emotional Stability was highest-17.1 percent[13].

Figure 1 : CNN Based Deep Learning Model



Future Work-

Though different methods in deep learning gained success to some extent, it still raises lots of concern. The relation between language and personality trait is much more complicated than we can perceive. Most of the experiments prune stop words like prepositions that play a great role in detecting human character. When scientists are dealing with social media data, lots of text is coming from regional language, colloquial terms, and slangs. These are generally treated as unknown words but these carry huge significance. If we follow the result, all experiments produce different success rates for different datasets and different personality traits. This also raises the question : how dependable really are these methods?

Approach	EST	STA	AG R	CO N	OPN
C2W2S4PT 5-fold cross-validation	0.13 1	0.17 1	0.14 0	0.12 4	0.10 9
MLP(with Doc Vector CNN + Mairesse)	55.5 4	58.4 2	55.4 0	56.3 0	62.6 8
sMLP/MP(with Doc Vector CNN + Mairesse)	58.0 9	57.3 3	56.7 1	56.7 1	61.1 3

Table 1. Comparison Among Different Classifier

III. Remaining Challenges and Opportunities

Although notable progress has been made in applying NLP to personality recognition from social media, barriers still exist with respect to developing practical solutions and detection methods that are both generalized and sufficiently accurate. Below we discuss a few such barriers and potential directions to overcome them. Human behavior and their reactions are at times situational. This necessitates to take into account a lot many other psychological factors before we can claim the prediction results to be accurate based on social network site behaviour. In this sense behavior-based approach is dependent on psychological and sociological researches as well. There are millions of fake profiles in Facebook which might impact the accuracy of the analysis. Another limitation is the algorithm used for getting prediction results. Although there is already a lot of research done on prediction algorithms from simpler regression algorithms to complex machine learning and artificial neural network; however right use of the algorithm is very important which might affect the result. User behavior in social networking sites might be very dynamic and for the research work where we are taking structural information like the number of likes and number of friends etc. in account, these factors might change quite heavily in a very short span of time thus making the results less reliable.

IV. Conclusions

We are now also aware of the possibilities to apply reinforcement learning, un- supervised methods, and deep generative models to complex NLP tasks such as visual QA and machine translation. Attention mechanisms and memory augmented networks are powerful at extending the capacity of neural-based NLP models. Combining all these powerful techniques provides a compelling method to understand the complexity of language. The next wave of language-based learning algorithms promises even greater abilities such as common sense and modeling complex human behaviors.

REFERENCES

- [1] Jon Oberlander, Scott Nowson : Classifying author personality from weblog text
- [2] Alastair J. Gill : Projection of Extraversion Across Different Social Contexts
- [3.] Giovanni Seni and John Elder: Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
- [4.] Ben Verhoeven: Evaluating Content-Independent Features for Personality Recognition
- [5.] Giuseppe Riccardi, Firoj Alam : Predicting Personality Traits using Multimodal In- formation
- [6.] Ben Verhoeven and Walter Daelemans: Ensemble Methods for Personality Recog- nition

[7.] Chandrima Sarkar, Sumit Bhatia: Feature Analysis for Computational Personality Recognition Using YouTube Personality Data set

[8.] Yoram Bachrach, Michal Kosinski: Personality and Patterns of Facebook Usage

[9.] Marc T. Tomlinson, David Hinote and David B. Bracewell: Predicting Conscientiousness through Semantic Analysis of Facebook Posts

[10.] Tommy Tandra, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetyo(2017)Personality Prediction System from Facebook Users Procedia Computer Science Volume 116, 2017, Pages 604-611

[11.] F. Mairesse, M. A. Walker, M. R. Mehl, R. K. Moore(2007)Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research 30 (2007) 457-500

[12.] Navonil Majumder ; Soujanya Poria ; Alexander Gelbukh ; Erik Cambria (2017)Deep Learning-Based Document Modeling for Personality Detection from Text. IEEE Intelligent Systems (Volume: 32 , Issue: 2 , Mar.-Apr. 2017)

[13] Fei Liu, Julien Perez, Scott Nowson (2017)A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.

[14] Mairesse et al. 2007; Luyckx and Daelemans 2008; Noecker, Ryan, and Juola 2013

[15] Francisco Iacobelli and Aron Culotta(2013) Too Neurotic, Not Too Friendly: Structured Personality Classification on : AACL Technical Report WS-13-01 Computational Personality Recognition (Shared Task) Textual Data

