Predicting Personality Traits using Multimodal Information

Firoj Alam
Department of Information Engineering and
Computer Science, University of Trento, Italy
alam@disi.unitn.it

Giuseppe Riccardi
Department of Information Engineering and
Computer Science, University of Trento, Italy
riccardi@disi.unitn.it

ABSTRACT

Measuring personality traits has a long story in psychology where analysis has been done by asking sets of questions. These question sets (inventories) have been designed by investigating lexical terms that we use in our daily communications or by analyzing biological phenomena. Whether consciously or unconsciously we express our thoughts and behaviors when communicating with others, either verbally, non-verbally or using visual expressions. Recently, research in behavioral signal processing has focused on automatically measuring personality traits using different behavioral cues that appear in our daily communication. In this study, we present an approach to automatically recognize personality traits using a video-blog (vlog) corpus, consisting of transcription and extracted audio-visual features. We analyzed linguistic, psycholinguistic and emotional features in addition to the audio-visual features provided with the dataset. We also studied whether we can better predict a trait by identifying other traits. Using our best models we obtained very promising results compared to the official baseline.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; I.5 [**Information Systems**]: Pattern Recognition

Keywords

Multimodal Personality Recognition, Behavioral Signal Processing.

1. INTRODUCTION

Researchers in psychology have been trying to understand human personality in the last century and it has become one of the sub-fields of psychology. Later, the study of personality became one of the central concerns in different fields such as business, social science, health-care and education [11]. Since the late 19th century, psychologists have been trying to define theories, rating scales and questionnaires by analyzing lexical terms or biological phenomena [11, 9]. In personality psychology, personality trait is defined as the coherent pattern of affect, behavior, cognition and desire over time and space, which are used to characterize a unique individual.

The advancement of personality traits theories and simplified inventories opened the window for its automatic processing. Hence, in the last few years, automatic personality traits recognition has become one of the mainstream topics in the field of speech and natural language processing to ease the process of interaction between human and virtual agents. This is because it adds value in different areas, such as virtual assistants, healthcare such as mood detection, detection of personality disorder, recommender systems such as customer profiling.

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. A typical approach is to use linguistic or acoustic features, or a combination of both. Linguistic features include lexical features using the bag-of-word approach and in some cases using Parts-Of-Speech (POS) or psycholinguistic features [12], whereas acoustic features include statistical functionals applied to low-level descriptors [1]. In most cases, the goal is to find the most relevant features, learning algorithms [1] or the correlation between the lexical features and traits [12].

Automatic processing of personality traits from different modalities is a challenging problem and there are many open research issues to solve, such as the types of features, long or short term history of a user, small datasets with imbalanced class labels, combination methods for multimodal information. In this study, we investigate the usefulness of different feature sets using a Youtube dataset released in the Workshop on Computational Personality Recognition (Shared Task) 2014 (WCPR14). The main contributions of this study are the following:

- *Studying audio-visual, lexical, POS, psycholinguistic and emotional features and their combinations*
- *Using predicted traits as features*

We used predicted traits as features to predict a trait in a cascaded classification system in order to show that traits can be used as predictors in automatic classification task.

Recent work relevant to personality traits theories and its automatic processing are described in Section 2. A concise description of the corpus that was used in this study is given in Section 3. A complete pipeline of the experimental method is given in Section 4. Details of the classification re-

WCPR'14, November 7, 2014, Orlando, FL, USA.
Copyright 2014 ACM 978-1-4503-3129-6/14/11 ...$15.00.
http://dx.doi.org/10.1145/2659522.2659531.

sults are given in Section 5. Finally, conclusions and future study are provided in Section 6.

2. LITERATURE REVIEW

In personality psychology, researchers have been interested in understanding how individuals differ. They have been trying to discover how to measure and map personality traits in accordance with theories. Among the several theories of personality traits "BIG-5", the five factor model (FFM) is the most widely used representation for automatic analysis. The "Big-5" factors of personality are five broad dimensions of personality that are used to describe unique individuals [11]. The "Big-5" dimensions are as follows:
O (Openness): Artistic, curious, imaginative.
C (Conscientiousness): Efficient, organized, responsible.
E (Extraversion): Energetic, active, assertive.
A (Agreeableness): Compassionate, cooperative, friendly.
N (Neuroticism): Anxious, tense, self-pitying. The opposite direction is referred to as Emotional Stability

There are different approaches to assessing personality traits and the most widely used approach is the use of questionnaires. The size of the questionnaires is also varies such as revised the 240-item NEO Personality Inventory (NEO-PI-R), and the Ten-Item Personality Inventory-(TIPI) [11]. For automatic recognition, the reference annotation is usually generated using such questionnaires. Several rating instruments for measuring each of these traits include self-report and observer-report. Self-report is used to rate oneself and observer-report is used to rate others [11]. The annotation of the corpus that we used in this study is based on the observer report [4].

Much work has been done on automatic recognition of personality traits in different domains and by using different modalities such as text [5], speech [2] and facial expressions [4]. The domains include personality of the blogger [8], dialogue system, social media, behavioral analytics, and marketing (see [2], [5] and the references therein).

For automatic processing, researchers use acoustic, lexical and audio-visual features and have very recently started to use emotional categories [15] and traits [7] as features. Personality plays a role in emotion, and this has been discussed in several literatures in psychology [9]. For the automatic prediction of personality, Mohammad et al. [15] studied emotional features for personality traits prediction and showed that fine-grained emotions are more relevant predictors. Later, Farnadi et al. [6] found a correlation between emotion and personality traits using Facebook status updates and showed that users' posts of *openness* traits convey emotions more frequently than other traits. Motivated by the results of these studies we used emotional features as predictors.

3. TASK DESCRIPTION AND DATASET

In WCPR14, systems are required to recognize "Big-5" personality traits from Youtube [4] and/or Mobile [17] datasets. The shared task consists of two tracks: 1) close task with two competitions - participants are allowed to use multimodal information using one of the datasets and transcriptions from the Youtube dataset and 2) open task - participants can use any external resources. Tasks also include solving both classification and regression problems. Our contributions are comprised of both tracks, however, we focused on only solving the classification problem using the Youtube dataset. The corpus consists of vlogs collected from Youtube, where a single person talks by looking at the camera with their face and shoulders showing and the vloggers talk about a product or an event. Annotation of the vloggers' personality traits has been obtained using Amazon Mechanical Turk . For the shared task, the dataset has been released in the form of extracted audio-visual features, along with automatic transcription. It contains 348 training, 56 test instances, consisting of 404 vlogs in total, where 194 (48%) are male and 210 (52%) are female vloggers.

4. EXPERIMENTAL DESIGN

For the study, we experimented with audio-visual features that had been released with the dataset and also extracted lexical, POS, psycholinguistic and emotional features from the transcription. In the following sub-sections, we describe the details of each feature set, feature selection and classification method.

4.1 Features

Audio-visual features (AV): Different groups of audio-visual features are acoustic, visual and multimodal [4]. The acoustic features include *speech activity* - speaking time, average length of the speaking segments and number of speaking turns, and *prosodic cues* - voice rate, number of autocorrelation peaks, spectral entropy, energy, Δ-energy and different variation of pitch. The visual features include *looking activity and pose* - looking time, average length of the looking segments, number of looking turns, proximity to the camera and vertical framing and *visual activity* - statistical descriptors of the body activity. The multimodal features are the combination of speaking and looking ratio.

Lexical features (Lex): From the transcription we extracted lexical features (tokens) and then transformed them into a bag-of-words, vector space model. This is a numeric representation of text that has been introduced in text categorization [10] and is widely used in behavioral signal processing [1]. We computed frequencies and then transformed them into logarithmic term frequency (tf) multiplied with inverse document frequency (idf). To use the contextual benefit of n-grams, we extracted token trigram features, which eventually results in a very large dictionary, however, we reduced them by selecting the top 10K frequent features and filtering out lower frequent features.

POS features (POS): To extract POS features we used Stanford POS Tagger [19] and used similar approach of lexical features for the transformation and reduction of the POS feature set.

Psycholinguistic features (LIWC): Pennebaker et al. designed psycholinguistic word categories using most frequent words and developed the Linguistic Inquiry Word Count (LIWC) [13]. It has been used to study gender, age, personality and health in order to understand the correlation between these attributes and word uses. The word categories include family, cognitive mechanism, affect, occupation, body, article, and function words. We extract 81 features using LIWC and also include gender information with this feature set, is available with the dataset.

Emotional features (Emo): We considered emotional categories and sentiment predictions as emotional features extracted from different resources. These resources include NRC lexicon [15], WordNet-Affect[18], SentiWordNet [3] and

Stanford-sentiment tool [16]. To extract information for emotional categories, we used NRC lexicon and WordNet-Affect where list of words are annotated with emotional categories. We calculated the frequency of an emotional category by matching the words belonging to this category with the words in the instance of the transcription. The NRC categories include *anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise* and *trust* whereas the WordNet-Affect categories include *anger, disgust, fear, joy, sadness* and *surprise.* There are overlaps between categories of these two lexicons. However, we have not combined them as the designing processes of these two lexicons are different. We computed sentiment scores using the SentiWordNet, which computes scores based on the positive and negative sentiment scores defined in the lexicon and sentiment decision using the Stanford-sentiment tool. Apart from that, we also use two additional *neutral* categories. One *neutral* category is composed of the list of words from NRC that do not belong to any of the NRC emotional categories and the other *neutral* category includes the words of an instance that do not belong to any emotional category. Therefore, we have 20 features - 10 NRC, 6 WordNet-Affect, 1 SentiWordNet, 1 Stanford-sentiment and 2 neutral.

Traits as features (Traits): To design a model for a trait we used other four traits' labels as features and to obtain the traits labels for the test set we designed a two-level cascaded classification system. In the cascaded system, first level model is selected from the models we generated using different feature sets and by using that we generated the features (traits labels) for the test set. Then, we designed the second level model.

Feature selection: We extracted high-dimensional features for lexical and POS sets, which is one of the reasons of overfitting. Therefore, to avoid high variance and overfitting and to improve the performance, we performed feature selection using Relief (see [1] and the reference therein) algorithm with 10-fold cross-validation on the training set, following the same approach used in [1]. Before the feature selection, feature values were discretized into 10 equal frequency bins.

4.2 Classification and Evaluation

We generated our classification models using Sequential Minimal Optimization (SMO) for Support Vector Machine (SVM) [14] for each feature set as described above. SMO is a variant of SVM, which solves the Quadratic Optimization (QP) problems analytically and avoids time consuming numerical QP optimizations. We used different kernels for different feature sets, such as linear kernel for lexical (Lex) and POS features and polynomial kernel for audio-visual (AV), psycholinguistic (LIWC), emotional (Emo) and traits (Traits) features. Linear kernel was chosen in order to alleviate the problem of higher dimensions for lexical and POS feature sets. Sometimes, however, it also gives optimal results for small set of features. We have tuned the parameters to obtain a better performance on each feature set using 10-folds cross-validation on the training set. Feature selection has been applied for lexical and POS feature set as mentioned earlier (see Section 4.1). The performance of each classification model has been measured in terms of average precision, recall and F1, which are the evaluation metrics specified for the shared task. However, for the reasons of brevity, we only present F1 scores.

For the combination of different models of the feature sets we used decision fusion and combined the decisions from the models of five feature sets. As a combiner we applied majority voting. We first designed a model by combining the decisions from the models generated using five feature sets, named it as **Maj-5** model - majority voting of the five models of five feature sets. After that, we designed another model discarding the model of emotional features from the combination, named this model as **Maj-4** - majority voting of the four best models out of the models of five feature sets.

To understand the usefulness of the traits as features, we designed a two-level cascaded classification system. In the cascaded system, we generated the traits labels for the test set using the best combined model (**Maj-5**) as the average performance of this model is best among the models. We designed the second level model using the predicted traits as features (see Section 4.1) and used SMO with its default parameters, named this model as **Maj-5-Traits**. To obtain the baseline of this feature set, we trained models using the traits labels of the training set and then evaluated them using the traits labels on the test set as shown the results in Table 2, named it as **Ref**.

5. RESULTS AND DISCUSSION

Here, we present the performance of the classification models designed using different feature sets, decision combination, traits features and their best F1 on each trait in Table 1, in addition to the official baseline. In the close shared task, using audio-visual features, we obtained an average of F1: 1.6% better than using lexical features and an average of F1: 21% better than official baseline. We obtained comparative results among the AV, Lex, POS and LIWC feature sets. The emotional feature set (Emo) does not perform well individually and we will investigate that in future by examining the representation of these features in the vector form, either as frequency or relative frequency or any other transformation.

The decision combination provides better results compared to the results of any single feature set. We obtained an average of F1: 65.6% using the model Maj-5 and F1: 64.8% using the model Maj-4, which implies that emotional features also contribute to improve the performance in combination.

The performance of traits features is lower compared to the Maj-5 model, however, we obtain better results on *extraversion* category.

In Table 2 we show the performance of the traits feature set using the reference labels and Maj-5-Traits. The results of Maj-5-Traits model are better in *agreeableness* category compared to the model using reference labels. We will investigate the traits features further on different datasets to understand their significance.

Our observation is that performance of each trait varies for different feature sets, which implies that the same feature set or architecture might not work for all traits. We might have to use the model which performs best for a particular trait. The best models are marked in bold-form in Table 1 for the traits and the last row of Table 1 shows the best results where we obtained an average of F1: 67.3%.

Significance test: We conducted statistical significance test of our best models with the second best models using the binomial test. The test revealed that the results of the best models are statistically significant with p<0.05 for extraversion and with p <0.01 for other categories.

Table 1: Results on test set using different feature sets. Baseline: Official baseline, AV: Audio-Visual, Lex: Lexical, POS: Parts-Of-Speech, LIWC: psycholinguistic, Emo: Emotion, Maj-5: Majority voting of the five models, Maj-4: Majority voting of the four best models, Maj-5-Traits: Generated traits labels using Maj-5 model

Model	**O**	**C**	**E**	**A**	**N**	**Avg**
Baseline	40.4	42.9	41.1	33.3	37.1	39.0
AV	63.4	42.9	70.4	67.7	55.7	60.0
Lex	59.9	49.4	60.4	65.8	56.7	58.4
POS	57.3	54.3	57.8	69.6	**61.9**	60.2
LIWC	55.0	56.0	66.2	71.4	46.8	59.1
Emo	49.3	52.5	53.5	59.4	40.1	51.0
Maj-5	**65.0**	57.4	69.4	**76.7**	59.4	65.6
Maj-4	61.5	**61.9**	68.8	74.7	57.1	64.8
Maj-5-Traits	59.2	41.7	**71.0**	62.2	52.6	57.3
Best model	**65.0**	**61.9**	**71.0**	**76.7**	**61.9**	**67.3**

Table 2: Results on test set using traits as features. Ref: Reference labels of the test set. Maj-5-Traits: Generated traits labels using Maj-5 model

Model	**O**	**C**	**E**	**A**	**N**	**Avg**
Ref	77.1	41.7	77.1	58.6	58.6	62.6
Maj-5-Traits	59.2	41.7	71.0	**62.2**	52.6	57.3

6. CONCLUSIONS AND FUTURE STUDY

In this paper, we presented our contribution to the automatic recognition of personality traits from a video-blog corpus by studying different types of feature sets. The feature sets include audio-visual, lexical, POS, LIWC, emotional features and their combinations using majority voting. In addition, we also used predicted traits as features and designed a cascaded classification system. We obtained very promising results compared to the official baseline. Performance of the model using emotional feature set is very low compared to the other feature sets, however, it helps in combination. We plan to experiment with the traits and emotional features with other datasets in the future.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610916 ■ SENSEI - http://www.sensei-conversation.eu/.

8. REFERENCES

[1] F. Alam and G. Riccardi. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In *INTERSPEECH*, 2013.

[2] F. Alam and G. Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *ICASSP2014 - SLTC*, May 2014.

[3] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[4] J. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. on Multimedia*, 15(1):41–55, Jan 2013.

[5] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski. Workshop on computational personality recognition: Shared task. In *AAAI*, 2013.

[6] G. Farnadi, G. Sitaraman, M. Rohani, M. Kosinski, D. Stillwell, M.-F. Moens, S. Davalos, and M. De Cock. How are you doing? study of emotion expression from facebook status updates with users' age, gender, personality and time. In *In Proc. of the EMPIRE Workshop on UMAP*, 2014.

[7] F. Iacobelli and A. Culotta. Too neurotic, not too friendly: Structured personality classification on textual data. In *AAAI*, 2013.

[8] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *ACII*, pages 568–577. Springer, 2011.

[9] C. E. Izard, D. Z. Libero, P. Putnam, and O. M. Haynes. Stability of emotion experiences and their relations to traits of personality. *Journal of personality and social psychology*, 64(5):847, 1993.

[10] T. Joachims. *Text categorization with support vector machines: Learning with many relevant features.* Springer, 1998.

[11] O. P. John, R. W. Robins, and L. A. Pervin. *Handbook of personality: Theory and research.* G. Press, 2010.

[12] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500, 2007.

[13] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[14] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. In *Advances in Kernel Methods- Support Vector Learning*, 1998.

[15] M. Saif. and K. Svetlana. Using nuances of emotion to identify personality. In *ICWSM-WCPR*, 2013.

[16] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642. Citeseer, 2013.

[17] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland. Friends don't lie: inferring personality traits from social network structure. In *Ubicomp'12*, pages 321–330. ACM, 2012.

[18] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[19] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 173–180. ACL, 2003.