Comparative Study of Data Sources, Features, and Approaches for Automatic Personality Classification from Text

Jayshri Patil
Assistant Professor
SRIMCA, Uka Tarsadia University

Jikitsha Sheth, PhD
Assistant Professor
SRIMCA, Uka Tarsadia University

ABSTRACT
Personality is a concern with individual differences in characteristic patterns of thinking, feeling, and behavior. Computational recognition of user personality is likely to be useful in many computational applications and technologies such as career counseling, relationship, and health counseling, human resource management, forensics, and mental health diagnosis. It involves understanding, prediction, and analysis of human behavior. The different methods have been proposed to automatically infer the user's personality from their user generated content. The paper discusses state-of-the-art personality recognition on various data sources, features, and their impact on different application areas.

Keywords
Data Sources, Automatic Personality Classification

1. INTRODUCTION
A personality is a combination of all the attributes-behavioral, temperamental, emotional, and mental that forms an individual's distinctive character. Human personality has a great impact on their life, which influences their life choices, well-being, and many other factors [3]. The personality is typically described in Big Five Personality Traits [14]:

- **Extraversion-** Extroversion measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. People scoring high on Extroversion are more likely to be friendly and socially active.
- **Neuroticism**- Neuroticism measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression. Neurotic peoples are more likely to experience stress and nervousness.
- **Agreeableness**- Agreeableness relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. People scoring high on Agreeableness trust others and adapt to their needs.
- **Conscientiousness**- Conscientiousness measures preference for an organized approach to life in contrast to a spontaneous one. People scoring high on Conscientiousness are tends to be well organized, consistent, and reliable.
- **Openness** – Openness is related to imagination, creativity, curiosity, tolerance, political liberalism, and appreciation for culture. People scoring high on Openness appreciate new and uncommon ideas, and have a good sense of aesthetics.

This Five Factor model of personality is most useful for describing personality and for assessing and describing personality disorders. Personality influences numerous facets of tasks related to individual behavior. For example personality traits influence leadership capability. In forensic, it is useful in analyzing conversations of suspected terrorists. Personality recognition would help psychologists to understand human personality and its impact on human behavior for identifying personality disorder and depression. It is important to automatically recognize user personality from spoken words and written text as it expresses huge information about the speaker or author.

In the field of automatic personality recognition, various datasets are used as resources for personality identification and analysis. The following section describes various data sources.

2. DATA SOURCES
The datasets for recognition of human personality are either collected from social media platforms or even individuals are asked to write text, which is further collected and treated as a data source.

2.1 Written Text and Conversation
The spoken words or written text convey immense information about the speaker or author. The first dataset of this data source is Stream-of-consciousness Essays is a large dataset written by psychology students who were said to write whatever comes into their mind thoughts, feeling, and sensations for 20 minutes. It contains 2,479 essays with 1.9 million words. This data was collected and analyzed by the authors in [9]. Texts have been written by students who took the Big5 test. This dataset has been used by research scholars in their research work [2],[3]. Another source of data collected by the authors in [13] consists of 96 participant's conversation extracts recorded using an Electronic Activated Recorder (EAR). It contains 97,468 words and 15,269 utterances.

2.2 Social Media
Social Media is a place where users share their views, information, and ideas, and they do many activities like posting, status updating, and commenting. User-generated content on social media provides an excellent opportunity to recognize user personality. Many researchers have been taken affords for utilizing data collected from Facebook and Twitter to infer personality from it.

A. Facebook
Many approaches have been proposed to automatically infer the user's personality from the content of Facebook. The MyPersonality corpus was collected from Facebook. It is released by organizers of the "Workshop on Computational

Personality Recognition(Shared Task) [13] and it has been collected by David Stillwell and Michal Kosinski. It contains a Facebook status message, author information (network size, betweenness, nbetweenness, density, brokerage, nbrokerage, and transitivity), gold standard labels(classes and scores) obtained using self-assessment questionnaire. The classes have been obtained from scores with a median split. This has been collected from 250 users and the number of statuses per user ranges from 1 to 223. This corpus has been utilizing by the various researcher in their work [4], [5], [6].

B. Twitter

The user generated content on Twitter also provides an important source of information for inferring the user's personality. One of the Twitter datasets is collected through myPersonality project, only a few hundred users among thousands of participants of this project posted links to their Twitter accounts, which forms the content of this dataset [14]. This dataset has been utilized by the researcher for the task of automatically personality recognition, as well as for user behavior analysis [19]. In [19] authors have found that both popular users and influentials are Extroverts and emotionally stable and also found that popular users are 'imaginative' means high in Openness, while influentials tend to be 'organized' means high in conscientiousness. On the other hand authors in [14] collected the Twitter dataset which contains 102 Twitter user and gold standard personality type labels in range of [-0.5, 0.5]. An Author in [15] has been collected a dataset from Twitter. They created a Twitter application with 45- question Big Five Personality Inventory. The dataset contains the latest 2,000 tweets. Authors have also collected a set of statistics of user account and their tweets. It includes the number of followers, number of following, the density of the social network, Number of "@mentions", number of replies, number of hashtags, number of links, and word per tweet. The work in [15] considers a connection between personality and actual social network for this author has considered two structural features number of friends and network density. This work has significant inference on the marketing and interface design area.

C. FriendFeed

The FriendFeed social media dataset was sampled by Celli et al. [9]. It has been collected from the FriendFeed application, where recent posts are available. This work aims to analyze social interaction that takes place in a social network site. This dataset was used in [10] for personality recognition from a social network site. The author has a sampled dataset of 748 Italian FriendFeed users contains 1065 posts.

2.3 Other Resources

Data sources discussed so far are used for psycho-linguistic features, lexical level analysis, emotion words, and lexical clues to recognize personality. In paper [17] authors have been employing common sense knowledge with sentiment polarity scores and affective labels using resources SenticNet, ConceptNet, and EmoSenticNet. The SenticNet resource is useful for opinion mining and sentiment analysis. It is a collection of commonly used 'polarity concepts' with strong positive and negative polarity[20]. In SenticNet each concept is associated with one value float [-1, 1] represent their polarity. It includes more than 5700 polarity concepts and it is freely available. The ConceptNet[21] is a semantic network represent information from the Open Mind corpus. It contains nodes as concepts and labeled edge are commonsense assertions that interconnect concepts. The EmoSenticNet [22]comprises about 5,700 common-sense knowledge concepts, including Wordnet Affect list concepts, along with their affective labels in the set {anger, joy, sadness, surprise, fear}. The authors in [16] have been combining common-sense knowledge-based features with psycho-linguistic features and frequency-based features for their studies.

The data sources discussed so far in this section utilized by the researchers and extracted linguistic and emotional features from these data sources [2], [3], [4], [5], [6], [9], [10], [16]. Psychological studies [9] shown that there exist links between linguistic features and users' personality traits. This finding is demonstrated by the correlations between features and personality traits. The following section describes features that are utilized by researchers in their work.

3. FEATURES

The features adopted by many researchers in their experiments are motivated by prior findings related to the correlation between measurable linguistic factors and personality traits [2].

3.1 LIWC Features

The linguistic features extracted from text using the LIWC text analysis program [23] that counts words in psychologically meaningful categories. It comprises two main components- the processing component and the dictionary. The dictionary is a collection of words that define a specific category [Positive emotion, Social process, Anger words, sadness]. The processing component goes through each word by word in the text and then each word in the text compared with the dictionary. If a word in the dictionary having three categories then all three categories will be incremented. After processing all words in the text LIWC calculate the percentage of each word category. The output of the LIWC program is a list of all categories and the rates that each category used in the text. The LIWC features have been used by several researchers in their work [2] [3] [4].

3.2 MRC Features

The MRC Psycholinguistic database [17], contains psychological and distributional information about words. It includes 150,837 entries with information about 26 properties, such as number of phonemes, number of letters, frequency of use, and familiarity [14]. This database consists of three files, DIST file (a dictionary of information about syntactic, semantic, orthographic, and phonological properties of a large set of words), S-R file (word association responses to a large set of stimulus words), and the R-S file (a large set of response word). The MRC features have been utilized by the authors [2] [14].

3.3 SPLICE(Structured Programming for Linguistic Cue Extraction) Features

The SPLICE is used to extract linguistic features, including cues that relate to the positive or negative self-evaluation of the speaker [14]. It includes various features categories like Quantity, Part of Speech, Immediacy, Pronouns, Positive Self Evaluation, Negative Self Evaluation, Influence, Deference, Complexity, Tense, Senticwordnet, etc. The SPLICE features have been used in several studies in this field [4] [14] [18].

3.4 SNA Features

The Social Network Analysis features are provided by the myPersonality dataset which gives a detailed information of the user's friendship network [4]. It contains social network information such as Networksize, Betweenness, NBetweenness, Density, Brokerage, NBrokerage, and Transitivity. These features have been utilized in several

studies for personality detection from social media content [4], [6].

4. RELATED WORK

This section presents a comprehensive review of automatic personality recognition from the text. The general framework for personality recognition includes the collection of datasets labeled with personality scores gathered through questionnaires, linguistic features selection, construction, and evaluation of the recognition algorithm. In [3] the experiment performed over an Essay dataset to extract user personality from it. The author used three Convolutional Filters to extract unigram, bigram, and trigram features from each sentence. This experiment performed using a Convolutional Neural Network, author trained five different networks for the five personality traits. The authors have used a two-layer perceptron comprising of a fully connected layer of size 200 and a final softmax layer of size two, denote yes and no classes. The accuracy ranged between 50% and 62% depending on the filter, personality trait, and classification. The best performance was achieved for Openness using Multiple Layer Perceptron (MLP).

The data source essay used [3] also utilize in [17]. In the paper, authors have been employing common sense knowledge with sentiment polarity scores and affective labels using resources SenticNet, ConceptNet, and EmoSenticNet. For personality recognition authors have combined common sense knowledge-based features with psycho-linguistic features and frequency-based features (LIWC, MRC) and then the features were used in supervised classifiers. In this work, five Sequential Minimal Optimization (SMO) classifiers have been designed for five personality traits. In this experiment, authors have shown that the use of common sense knowledge with affective and sentiment information enhances the accuracy of the existing work which uses only psycho-linguistic features and frequency-based analysis at the lexical level. In this work performance evaluation is done by 10 fold cross-validation. This experiment showed that the Openness trait is easiest to identify as its F-score is 0.662 and the Agreeableness trait is most difficult to identify among all traits as its F- score id 0.615. Authors have reported that the new approach proposed in this work performs much better than previously reported state-of-art methods on the same dataset.

In recent times the use of social networking has increased extremely. It has become a popular application for information sharing and social interaction. It is a place where users represent their information, ideas, career interests, views, etc, and therefore it is an excellent source for the research on personality computing [4], [5], [6], [7].

In [4], the experiment aimed at predicting personality from Facebook user statuses. The author used two datasets in this experiment myPersonality and manually collected dataset. The task performed with traditional machine learning algorithm Naïve Bayes, SVM, Logistic Regression, Gradient Boosting, Linear Discriminant Analysis(LDA) and Deep Learning architecture Multi-Layer Perceptron(MLP), Long Short Term Memory(LSTM), Gated Recurrent Unit(GRU) and 1-Dimensional Convolutional Neural Network(CNN 1D). In this experiment, several features were used such as LIWC, SPLICE, and SNA. For traditional machine learning, they used a closed vocabulary approach (Predefined features) such as 85 features from LIWC, 74 features of SPLICE, and SNA features. And for deep learning implementation, they used linguistic features of open vocabulary approach(not predefine feature) such as word embedding using Glove. Deep learning architecture MLP has the highest average accuracy in myPersonality and LSTM+CNN 1D architecture has the highest average accuracy in the manually collected dataset. Deep learning architecture gave a better result.

The experiments performed in [5] aimed at automatic recognition of Big-5 personality traits on Social networks using the user's status text. The experiment was performed on myPersonality corpus. This corpus was collected from Facebook. The authors have used bag of words approach for the feature extraction with unigram as features. They utilized various classification methods such as Sequential Minimal Optimization for Support Vector Machine(SMO), Bayesian Logistic Regression(BLR), and Multinominal Naïve Bayes (MNB) sparse modeling. The result shows that the MNB sparse generative model performs better than discriminative models SMO and BLR.

The approach proposed in [15] analyzes the user's Twitter profile. The authors were performed experiments over the 2000 latest Tweets of 279 users collected from the Twitter application. The features included not only the LIWC and MRC categories but also measurements of Twitter such as Number of followers and following, Density of Social Network, Number of "@mentions", Number of replies, Number of “hashtags”, Number of links, and Words per Tweet. Regression experiments were performed on data to access user Big Five personality. In this study two regression algorithms have used Gaussian Process and ZeroR both had similar performance over the personality features. The authors shown result analysis that Twitter data yielded similar results for Openness and agreeableness but less impressive results in other traits.

5. COMPARATIVE STUDY OF AUTOMATIC PERSONALITY CLASSIFICATION

Language psychology indicates that the choice of words reflects not only the meaning of words but also driven emotions, relational attitudes, power status, and personality traits [1]. Thus, due to incorporation of sociolinguistic in techniques for automatic personality classification task, it is possible that researchers can infer personality traits from the written text. The studies discussed so far in this paper adopted lexical approaches in their personality classification task. The results and details of the task performed in these studies are summarized in Table-1. It reports from left to right, dataset, the number of the subject involved in the task, features, type of task, and performance over different traits. The performance for the classification tasks is presented in terms of Accuracy, F-Measure, and Mean Absolute Error.

Table 1: Automatic Personality Classification from textual data

Ref.	Dataset	Samples	Features	Approach	Extraversion	Agreeableness	Conscientiousness	Openness	Neuroticism	Result Analysis
[3]	Essay	2,467 written essay	Mairesse, N-gram	Classification-Convolutional Neural Network	58.09 ACC	56.71 ACC	57.30 ACC	62.68 ACC	59.38 ACC	The Mairesse features improved the result as compared to N-gram features. The best performance was achieved for Openness using Multiple Layer Perceptron (MLP).
[17]	Essay	2400 written essay	LIWC, MRC, common sense knowledge features(SenticNet, ConceptNet, EmoSenticNet)	Classification-Five Sequential Minimal Optimization (SMO)	0.634 F-score	0.615 F-score	0.633 F-score	0.661 F-score	0.637 F-score	Inclusion of Common sense knowledge features improved the performance of classifier. The classification task with LIWC, MRC and common sense knowledge features gives better result for Openness Personality trait as compare to other trait.
	myPersonality	myPersonality-250 users 10,000 statuses	LIWC, SPLICE, SNA	Classification-Machine Learning Algorithms	68.80 ACC	63.20 ACC	59.20 ACC	70.40 ACC	60.80 ACC	The authors have achieved highest accuracy by using SVM, Logistic Regression Algorithm and LDA Algorithm for Openness Trait. Openness has highest accuracy in myPersonality dataset.

[4]		myPersonality-250 users 10,000 statuses	Word Embedding	Classification-Deep Learning Algorithms	78.95 ACC	67.39 ACC	62.00 ACC	79.31 ACC	79.49 ACC	The authors have achieved highest accuracy by using MLP architecture and Word Embedding method.
	Manually collected dataset	150 Facebook users	LIWC, SPLICE, SNA	Classification-Machine Learning Algorithms	79.33 ACC	60.67 ACC	67.33 ACC	67.33 ACC	70.00 ACC	Authors utilized manually collected dataset and achieved highest accuracy by using LDA algorithm and SVM algorithm. The highest average accuracy of classification task achieved for Extraversion trait.
		150 Facebook users	Word Embedding	Classification-Deep Learning Algorithms	93.33 ACC	70.37 ACC	68.00 ACC	76.19 ACC	80.00 ACC	The highest accuracy obtained by using MLP and LSTM+CNN 1D architecture Extraversion trait has highest accuracy in manually collected dataset.
[5]	myPersonality	250 users	Unigram	Classification-Machine Learning Algorithms	58.57 ACC	59.16 ACC	59.4 ACC	69.48 ACC	63.00 ACC	The classification task performed better by using MNB sparse generative model, discriminative models SMO and BLR. Classifier obtained highest accuracy for Openness trait.

[15]	Twitter dataset	2000 Tweets	LIWC, MRC, Profile information	Regression Algorithms	0.16 Mean Absolute Error	0.13 Mean Absolute Error	0.14 Mean Absolute Error	0.12 Mean Absolute Error	0.18 Mean Absolute Error	Authors have utilized two regression algorithms: Gaussian Process and ZeroR. These two algorithms performed similar over the personality features. Openness is the easiest to compute and neuroticism was the most difficult trait to compute.

The Table-1 shows that the Openness personality trait gives a better result. Also it reflects that the result in [4] obtained by manually collected dataset gives the better result as compared to the MyPersonality dataset. In [3] authors have used Deep Learning algorithms, Mairesse and N-gram feature, adding Mairesse feature has been proved beneficial in experiments. Due to insufficient training data, CNN alone without the document level features underperformed the Mairesse baseline. In [5], experiment semantic features have not been utilized, including these features may provide more information to recognize personality traits. The study in [16], incorporated common sense knowledge with psycholinguistic features, which led to an effective result.

6. DISCUSSION AND CONCLUSION

The paper represents the study of data sources which was either collected from social media or individuals were asked to write a text. We also showed a study of features and approaches utilized by the researchers in their personality recognition task. Furthermore, the paper represents a comparative analysis of personality recognition tasks. Incorporating personality recognition models in another task viz., detection of deception, mood, point of view, opinion mining, and dominance in the meeting may improve accuracy. Existing studies also reported that the computational recognition of a user's personality could be useful in many applications such as identification of personality disorder and depression, identification of leaders, tutoring system, predicting job satisfaction; references for different interfaces, and it also improves the performance of recommender system. Also personality recognition task can assist the suicide prevention system.

Personality recognition and psychology studies reported the correlation between personality traits and language cues. Some traits are more revealed through the language, like extraversion. Thus, some traits have formed more findings than others. Extraversion is highly correlated with spoken language. Extraverts talk more, louder and more repetitively, they have higher speech rates, shorter silence, fever pauses, and hesitations than introverts. So extraversion is the easiest traits to model from spoken language. Concerning to written language Openness is the easiest traits to recognize. The openness to people preferences use longer words and words expressing tentatively, as well as the avoidance of first-person singular pronouns and present tense forms. Neurotics person uses more first-person singular pronouns, more negative emotion words, and less positive emotion words. Existing studies reported that the observers don't use such clues correctly thus, observer reports of Neuroticism negatively correlate with self-reports. Agreeable people likely to be have more positive and fewer negative emotions. On the other hand, conscientious people avoid negations, negative emotion words, and words reflecting discrepancies.

7. REFERENCES

[1] Vinciarelli, and G. Mohammadi, "A survey of personality computing." IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 273-291, 2014.

[2] F. Mairesse et al., "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text,"J. Artificial Intelligence Research, vol.30, pp. 457–500, 2007.

[3] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning base document modeling for personality detection from text," IEEE Intelligent Systems, vol. 32, no. 2, pp. 7479, 2017.

[4] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso and Yen Lina Prasetio "Personality Prediction System from Facebook Users" 2nd International Conference on Computer Science and Computational Intelligence, Bali, Indonesia 2017, ICCSCI 2017, 13-14 October 2017.

[5] Firoj Alam, Evegeny A. Stepanov, Giuseppe Riccardi "Personality Traits Recognition on Social Network-Facebook" Computational Personality Recognition (Shared Task), 2013.

[6] Jianguo Yu, Konstantin Markov, "Deep Learning based Personality Recognition from Facebook Status Updates" IEEE 8th International Conference on Awareness Science and Technology (iCAST 2017)..

[7] S. Bai, T. Zhu, and L. Cheng, "Big-five personality prediction based on user behaviors at social network sites," Cornell University, Tech. Rep., 2012.

[8] Golbeck, J. and Robles, C., and Turner, K. "Predicting Personality with Social Media", In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, pp. 253–262. 2011.

[9] Pennebaker, J.W. and King L.A., "Linguistic style: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296-1312,1999.

[10] Celli, F. and Di Lascio and F.M.L. and Magnani, M. and Pacelli, and B., Rossi, L. Social Network Data and Practices: the case of Friendfeed. Advances in Social Computing, pp. 346–353. Series: Lecture Notes in Computer Science, Springer, Berlin. 2010.

[11] F.Celli, "Unsupervised personality recognition for social network sites," in Proceedings of the International Conference on Digital Society, 2012, pp. 59–62.

[12] Max Coltheart, "THE MRC PSYCHOLINGUISTIC DATABASE", Quarterly Journal of Experimental Psychology (1981) 33A, 497-505.

[13] Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology, 90, 862–877,2006.

[14] Golnoosh Farnadi, Geetha Sitaraman Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos Marie-Francine Moens Martine De Cock, "Computational personality recognition in social media" User Model User-Adap Inter DOI 10.1007/s11257-016-9171.

[15] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in Proceedings of IEEE International Conference on Social Computing, 2011, pp. 149–156.

[16] F. Celli, "Unsupervised personality recognition for social network sites," in Proceeedings of the International Conference on Digital Society, 2012, pp. 59–62.

[17] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, Newton Howard, "Common Sense Knowledge Based Personality Recognition from Text"

[18] E. Cambria, R. Speer, C. Havasi, and A. Hussain. SenticNet: A publicly available semantic resource for opinion mining. In: AAAI CSK, pp. 14-18, Arlington (2010).

[19] Cambria, E., Havasi, C. and Hussain, A. "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis." Proceedings of FLAIRS, Marco Island, pp. 202–207 (2012).

[20] Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: RANLP (2007).

[21] Poria, S.; Gelbukh, A.; Hussain, A.; Das, D.; Bandyopadhyay, S., "Enhanced SenticNet with Affective Labels for Concept-based Opinion Mining," Intelligent Systems, IEEE, vol. 28, no. 2, pp. 31–38, March-April 2013; doi: 10.1109/MIS.2013.4 (2013).

[22] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[23] Coltheart, M. (1981). The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A, 497–505.