Exploring Personality Prediction from Text on Social Media: A Literature Review

Veronica Ong, Anneke D. S. Rahmanto, Williem and Derwin Suhartono

Abstract—Personality assessment can provide insight on what a certain individual is like, which can be used to evaluate the individual on different aspects. Traditional personality assessments are done by having individuals participate in personality tests. There are several weaknesses to this approach, namely that it is time consuming, and test participants could have made up their answers. A new approach of personality prediction is explored by merely evaluating the contents of a user’s social media account. This paper provides an overview on the development of personality prediction from text on social media, the common issues faced in performing said task, and further improvements that can be applied in the future.

Index Terms—linguistic analysis, natural language processing, personality prediction, social media

I. INTRODUCTION

According to the Merriam-Webster dictionary [1], social media is defined as forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos). Social media is an inevitable part of the Internet, as statistics [2] show that people spend 1 in every 4 minutes of their Internet usage on social media. An observation [3] regarding Facebook usage reported that users log in to their Facebook accounts from 2 to 5 times a day, with an average of 5 to 15 minutes per session.

The different kinds of social media on the Internet accommodate each user's needs. A framework was proposed [4] which divides the functions of social media into seven blocks. One of these blocks is the identity block which represents how users consciously and unconsciously reveal their identities on social media. Through social media, a user is able to display different sorts of information about themselves. They may consciously fill out information about themselves, or unconsciously show their own identities through their behavior on social media.

II. PERSONALITY PREDICTION FROM TEXT ON SOCIAL MEDIA

Personality prediction is a task where information about an individual’s personality trait is identified, given a set of data. There have been several approaches on automated personality prediction based on different kinds of dataset, such as essays, social media posts, videos, and social media behavior. This paper will only focus on studies of personality prediction from text based on social media posts.

There are several tools and corpora which are widely used in personality prediction studies, including the ones mentioned in this research. The first tool is called LIWC (Linguistic Inquiry Word Count), which is a text analysis tool used to evaluate psychological properties from language. The LIWC tool supports several languages such as Arabic, Chinese, Dutch, English, German, Italian, Korean, Norwegian, Portuguese, and Spanish. Approaches using the LIWC tool are often referred to as closed-vocabulary approaches or category-based analysis. Some other tools used in the closed-vocabulary approach are MRC, NRC, SentiStrength and SPLICE. The second, while not exactly a tool, which is commonly used in the personality prediction task is the MyPersonality corpus. It contains records about psychometric scores and social media posts from Facebook users. There are also other related corpora available for other social medias, such as...
Another research was conducted by using the Dark Triad personality model with 2,927 Twitter users [11]. 337 features were selected for the personality prediction task, consisting of Twitter statistic data and frequency of pre-defined words for each individual. The prediction task was then run with 4 algorithms from WEKA: Support Vector Machine, Random Forest, J48, and Naïve Bayes.

Personality prediction has also been conducted in a semi-supervised way, with Brazilian TV shows as an additional label [12]. In their study, they used a list of meta-attribute features, which was then run using a Naïve Bayes classifier with a supervised and semi-supervised learning approach. Results showed that the semi-supervised learning outperforms supervised learning, with 0.8415 as their highest accuracy.

C. Blogger Dataset

Further research on personality prediction via social media was attempted on Blogger [13]. This study also utilizes texts from bloggers, but further improves Yarkoni’s study by doing a personality prediction task based on the selected features. In addition, he also did a comparison of performances achieved between different approaches of the personality model. The different approaches used are: (a) different n-grams (n=1 or n=2), (b) utilization of stop words (using stop words or omitting stop words), and (c) term weighting (Boolean weighting or TF-IDF weighting). These approaches were also compared to the performance when using the LIWC tool. These features were then classified in Weka using Support Vector Machine (SVM). The experiment’s results showed that the best accuracy, with a value of 84.36%, was achieved by using bigrams (n=2), utilizing stop words and implementing Boolean weighting. This proves that the open vocabulary approach (by extracting n-gram tokens) can be used to predict personality, since it outperforms the closed vocabulary approach (using LIWC). However, they also mention that the classification may have overfitted, due to the few amount of bigrams in each personality trait.

D. Facebook Dataset

Personality prediction was also attempted on Facebook. One of these studies utilized a Facebook dataset named MyPersonality corpus [14]. This study attempted to perform the personality prediction task using an open-vocabulary approach. Significant features were found between n-grams (n=1 to 3), extracted topics with Latent Dirichlet Allocation and personalities. The models created with these features outperformed the model created based on LIWC.

Another open vocabulary approach on The Five Factor Model personality prediction using the MyPersonality dataset was conducted using Support Vector Regression and Latent Dirichlet Allocation models [15]. Their results show that the LDA models, sLDA (supervised Latent Dirchlet Allocation) and PT-LDA (Probabilistic Topic model-Latent Dirichlet Allocation) outperforms the Support Vector Regression model (topics and N-grams). Furthermore, they also proved that PT-
LDA is more robust and improves computational efficiency up to 64%.

Meanwhile, another attempt at closed-vocabulary approach was done by using the LIWC tool [16]. This study used 81 LIWC features, 7 social network features, 6 time-related features, and 6 other features that can be extracted from the posts’ content. The learning algorithms tested on this study are Support Vector Machine, K-Nearest Neighbor, and Naive Bayes. The highest precision achieved by combining all these features using K-Nearest Neighbor is 0.54, while highest precision was obtained by merely using social network features, reaching a precision of 0.71.

E. YouTube Dataset

This task was also implemented on YouTube personality datasets [17]. The dataset consisted of audio-video features, speech transcripts, gender, and personality impression scores from a total of 404 YouTube vloggers. 7 types of features were used in this study: gender, 25 audio-video features, 81 LIWC features, 10 NRC features, 14 MRC features, 3 SentiStrength scores, and 74 SPLICE features. Different kinds of multivariate regression algorithms with mostly decision trees as the base learner were applied to compare their performances. The system achieved the lowest root-mean square error (RMSE) in predicting the Conscientiousness trait, with a value of 0.64, using Multi-Target Stacking Corrected and Multi-objective random forest multivariate regression algorithms.

F. Non-English Dataset

The previously mentioned researches are conducted with social media datasets in English. Recent studies have attempted to predict personality based on non-English datasets. One of these studies was conducted with a dataset which consists of English, Spanish, Dutch, and Italian tweets [18]. Tokenized terms are matched with the an enhanced LIWC tool. Afterwards, the personality prediction task is executed with a multivariate regression technique called Ensemble of Regressor Chains Corrected (ERCC). They were able to achieve best results in predicting the Openness trait for English and Spanish languages, with an MAE value of 0.0811.

Another attempt was done for Twitter texts in Indonesian language. An attempt in personality prediction on Twitter text using the Indonesian language was done by translating the MyPersonality dataset contents into said language [19]. Similar to other open-vocabulary approaches, they extracted the top 750 words that frequently appear in the dataset, and compared the Naive Bayes, K-Nearest Neighbors, and Support Vector Machine algorithms for the classification process. While they obtained 72.29% as their highest accuracy, they noted that a native Indonesian language dataset may be more reliable for classification.

A Chinese dataset consisting of 222 Taiwan Facebook users was also utilized for the personality prediction task, by using an open-vocabulary approach [20]. They conducted the prediction task with different methods to compare their performances: (a) weighting scheme (term frequency (TF) or term frequency-inverse document frequency (TF-IDF)), (b) tokenization tool (Jieba segmentation tool or scikit-learn tokenizer), and (c) feature selection algorithm (chi-squared test or recursive feature elimination). A 73.5% value was achieved as their highest accuracy by using the Jieba Chinese text segmentation as their tokenizer, using the TF weighting scheme, chi-squared test as the feature selection algorithm, and SVM as the machine learning algorithm. One thing worth noting from this study is that the Jieba tokenizer improves precision up to 60% in the case of utilizing a Chinese dataset.

An experiment using a different Chinese dataset from Sina Weibo was also used to predict personality [21]. They used 2 tools to execute the task: LIWC2007 Chinese Simplified Dictionary and IKanalyzer, a Chinese word segmentation tool. There are a total of 4 user behavior features, 3 interaction behavior features, and 71 text features which were used in this system. The Logistic Regression and Naive Bayes were chosen as the machine learning algorithms for this study, with Logistic Regression yielding the highest precision in identifying the Agreeableness trait, with a value of 75.2%.

G. Cross-media Dataset

There have also been attempts to predict personality by utilizing datasets from more than one social media. One of those attempts was done by combining features from Twitter and Instagram [22]. They extracted linguistic and Twitter statistics data from Twitter, while image and linguistic features were extracted from Instagram. These features were matched with different combinations, but the best result was achieved by using both linguistic and Twitter statistics data from Twitter, and image and linguistic features from Instagram, yielding an average root-mean square error of 0.66 for all personality traits.

Another cross-media personality prediction attempt was conducted by utilizing datasets from Facebook, Twitter, and YouTube [23]. They extracted LIWC, NRC, MRC, SentiStrength, SPLICE, demographics, and audio-video features for the YouTube dataset. The same features, except NRC and audio-video features, were applied to both the Facebook and Twitter dataset. The learning algorithm applied to the classification task were univariate and multivariate techniques with a decision tree or SVM algorithm as the base learners, although results showed that there wasn’t significant difference between univariate and multivariate techniques. They achieved the best result with techniques that applied the decision tree algorithms. Even though they managed to perform cross-media personality prediction, they noted that expanding their model with training samples from different sources didn’t improve the learning performance.

The task of cross-media personality prediction was further improved by applying Heterogeneity Entropy Neural Network (HENN) to extract features from Renren and Sina [24]. The HENN learning algorithm was used to overcome the semantic and heterogeneity gap caused by cross-media platforms. The
CCA-based and Corr-AE learning methods were also applied to compare their performances with the HENN method. Results showed that HENN successfully outperformed other learning methods with 0.0723 as their smallest MAE value in predicting the Openness trait.

A summary of the mentioned personality prediction attempts in this paper is provided in Table 1 and Table 2.

Table 1 consists of the range of dataset size used in various social media which are mentioned in this paper. The dataset size in Table 1 is based on the number of users contained in the dataset. [6] and [12] are excluded from this table as its dataset size are based on the number of tweets. Table 2 is a list of studies with information about which social media is personality prediction performed on, the authors of the study, the features used in the study, and the best results that are achieved by the study.

III. ISSUES IN PERSONALITY PREDICTION TASK

With recent developments, the task of personality prediction still proves to be difficult. This section covers the issues and problems that probably might be faced in said task. The issues are as follows.

The first issue which is commonly encountered is the difficulty in finding an annotated dataset. So far, there are 2 approaches to obtain personality scores. The first method is to have participants answer a questionnaire which reveals insight on their personality. The second method is to have volunteers rate the personality of a user. [15] and [23] argued that this approach is difficult, time-consuming and expensive. There is also a risk of obtaining biased samples, as mentioned in [14]. Furthermore, it is stated in [11] that people could easily manipulate their self-assessment questionnaire to produce different results.

The second issue is the difficulty in identifying significant features in certain languages, especially those which are not supported by the LIWC dictionary. Among the 17 studies mentioned in this study, 10 of them utilized LIWC. Figure 2 shows the comparison between the number of mentioned studies using the open-vocabulary approach, closed-vocabulary approach, and both approaches. The figure shows that more than half of the mentioned studies in this paper used LIWC in their personality prediction system. The frequent usage of LIWC indicates that it provides great insight on personality and language. Despite being informative, LIWC only supports few languages.

Another problem that might be encountered is the difficulty in identifying the needed preprocessing methods. The language of social media is very noisy because users are able to freely express themselves. The informal style of social media language can be in form of spelling errors, abbreviations, uncommon acronyms, or slang words [25]. In [12], it is mentioned that one of the social media services named Twitter, has a different difficulty level for automatic analysis compared to formal texts. [9] also argued about how misspellings and language features become a challenge.

IV. POSSIBLE DEVELOPMENTS FOR PERSONALITY PREDICTION TASK

This section provides an outline about further improvements that can be applied to the personality prediction task from text on social media.

The first improvement that can be made to the personality prediction task is developing methods of said task for non-English language. This is in accordance with the second issue mentioned in the previous section, where not all languages are supported by LIWC.

Secondly, improvements to the research can be done by exploring more methods to achieve higher accuracy than the current state-of-the-art research. The improvements may include more suitable machine learning algorithms, feature selection on more significant features on social media posts or methods to preprocess the dataset. The improvement on methods to preprocess the dataset is in accordance with the third issue mentioned in the previous section. As mentioned in [18], dealing with multilingual, noisy, short, and informal social media posts can result in a better personality prediction model.

Lastly, while the mostly used model for personality prediction is the Five Factor model or The Big Five, further developments may include taking the Five Factor Model 30 facets into consideration, or conducting personality prediction of other personality models, such as the Dark Triad personality model which was implemented in [11].
### TABLE II
SUMMARY OF PERSONALITY PREDICTION ATTEMPTS MENTIONED IN THIS STUDY

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Author</th>
<th>Features</th>
<th>Best Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Golbeck, Robles, Edmondson &amp; Turner</td>
<td>LIWC &amp; MRC text features, Twitter usage, structural, sentiment</td>
<td>MAE 0.1192333</td>
</tr>
<tr>
<td></td>
<td>Sumner, Byers, Boochever, &amp; Park</td>
<td>Twitter usage and pre-defined words frequency (LIWC)</td>
<td>Accuracy 0.919</td>
</tr>
<tr>
<td></td>
<td>Lima &amp; de Castro</td>
<td>Twitter usage meta-attributes, Brazilian TV shows</td>
<td>Accuracy 0.8415</td>
</tr>
<tr>
<td></td>
<td>Arroju, Hassan, &amp; Farnadi</td>
<td>LIWC (enhanced)</td>
<td>MAE 0.0811</td>
</tr>
<tr>
<td></td>
<td>Celli</td>
<td>Cross-linguistic features</td>
<td>Co-occurrence 0.6651</td>
</tr>
<tr>
<td>Blogger</td>
<td>Iacobelli, Gill, Nowson, &amp; Oberlander</td>
<td>n-grams</td>
<td>Accuracy 0.8436</td>
</tr>
<tr>
<td></td>
<td>Yarkoni</td>
<td>LIWC, n-grams</td>
<td>ρ (Spearman’s rank correlation coefficient) 0.32</td>
</tr>
<tr>
<td>Facebook</td>
<td>Schwartz et al.</td>
<td>n-grams, extracted topics</td>
<td>R (square root of coefficient determination) 0.42</td>
</tr>
<tr>
<td></td>
<td>Liu, Wang, &amp; Jiang</td>
<td>Latent topics from n-grams</td>
<td>RMSE 0.479*</td>
</tr>
<tr>
<td></td>
<td>Farzadi, Zoghbi, Moens, &amp; De Cock</td>
<td>LIWC, social network feature, time-related feature, content feature from posts</td>
<td>Precision 0.54**</td>
</tr>
<tr>
<td></td>
<td>Pratama &amp; Sarno</td>
<td>n-grams</td>
<td>Accuracy 0.7229</td>
</tr>
<tr>
<td></td>
<td>Peng, Liou, Chang, &amp; Lee</td>
<td>n-grams</td>
<td>Accuracy 0.735</td>
</tr>
<tr>
<td>YouTube</td>
<td>Farzadi et al.</td>
<td>Gender, audio-video features, LIWC, NRC, MRC, SentiStrength, SPLICE</td>
<td>RMSE 0.64</td>
</tr>
<tr>
<td>Sina Weibo</td>
<td>Wan, Zhang, Wu, &amp; An</td>
<td>LIWC, user behavior, interaction behavior</td>
<td>Precision 0.752</td>
</tr>
<tr>
<td>Instagram + Twitter</td>
<td>Skowron, Tkáčič, Ferwerda, &amp; Schedl</td>
<td>Linguistic (Twitter, Instagram)</td>
<td>RMSE 0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meta features</td>
<td>Number of followers and followings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Image (Instagram)</td>
<td>PAD, brightness, saturation, hue-related, content-based features</td>
</tr>
<tr>
<td>Sina Weibo + Renren</td>
<td>Xianyu, Xu, Wu, &amp; Cai</td>
<td>Bag-of-textual words</td>
<td>MAE 0.0723</td>
</tr>
<tr>
<td>Facebook + Twitter + YouTube</td>
<td>Farzadi et al.</td>
<td>Facebook</td>
<td>RMSE 0.115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Twitter</td>
<td>LIWC, MRC, SentiStrength, SPLICE, demographics, user behavior</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YouTube</td>
<td>LIWC, NRC, MRC, SentiStrength, SPLICE, audio-video features, gender</td>
</tr>
</tbody>
</table>

Summary of personality prediction attempts mentioned in this study containing the social media where personality prediction is conducted, authors of study, features used in study, and best result achieved in study.

*result from SLA, but isn’t as robust as PT-LDA.

**best result is 0.71, but doesn’t involve any linguistic features.
V. CONCLUSION

This paper provided an insight on existing attempts of the task of personality prediction from text on social media to-date, along with the various kinds of social medias which have been utilized for said task. Some of these methods use a closed-vocabulary approach with psycholinguistic tools such as LIWC, while other methods made use of an open-vocabulary approach by extracting n-grams and topics. While most personality prediction studies to-date require a dataset to perform supervised learning, it is costly to obtain a dataset labelled with personality traits of social media users. Recent studies have tried applying semi-supervised and unsupervised learning to tackle this problem. Further improvements to the existing state of personality prediction can be made by expanding the target language, applying more suitable algorithms or preprocessing methods to achieve higher accuracy, and implementing said task to other personality models.

REFERENCES


Veronica Ong is an undergraduate student from Bina Nusantara University, located in Jakarta, the capital city of Indonesia. She majors in computer science. She is currently studying the application of machine learning on personality prediction.

Annee D. S. Rahmanto is an undergraduate student majoring in computer science. She is taking her undergraduate study in Bina Nusantara University, Indonesia. Currently, she is working on applying machine learning in the field of personality prediction.

Williem is a computer science undergraduate student studying in Bina Nusantara University, Indonesia. He is participating in a study of applying machine learning to personality prediction.

Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He is currently taking his PhD in computer science, specifically in the natural language processing (NLP) field. Recently, he is continually doing research in the area of argumentation mining and personality recognition.