Modeling Personality Traits of Filipino Twitter Users

Edward P. Tighe and Charibeth K. Cheng
Software Technology Department
De La Salle University, Manila, Philippines
edward.tighe@dlsu.edu.ph
chari.cheng@delasalle.ph

Abstract

Recent studies in the field of text-based personality recognition experiment with different languages, feature extraction techniques, and machine learning algorithms to create better and more accurate models; however, little focus is placed on exploring the language use of a group of individuals defined by nationality. Individuals of the same nationality share certain practices and communicate certain ideas that can become embedded into their natural language. Many nationals are also not limited to speaking just one language, such as how Filipinos speak Filipino and English, the two national languages of the Philippines. The addition of several regional/indigenous languages, along with the commonness of code-switching, allow for a Filipino to have a rich vocabulary. This presents an opportunity to create a text-based personality model based on how Filipinos speak, regardless of the language they use. To do so, data was collected from 250 Filipino Twitter users. Different combinations of data processing techniques were experimented upon to create personality models for each of the Big Five. The results for both regression and classification show that Conscientiousness is consistently the easiest trait to model, followed by Extraversion. Classification models for Agreeableness and Neuroticism had subpar performances, but performed better than those of Openness. An analysis on personality trait score representation showed that classifying extreme outliers generally produce better results for all traits except for Neuroticism and Openness.

1 Introduction

Personality traits aim to describe the uniqueness of an individual in terms of their interactions within themselves, with other people, and in certain environments (Friedman and Schustack, 2014; Larsen and Buss, 2008). The most common representation or model of personality traits used today is the Five Factor Model (FFM; Norman, 1963; Goldberg, 1981; McCrae and Costa Jr). The FFM, sometimes referred to as the Big Five, measures an individual's personality on five dimensions or traits, namely *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. It is important to note that traits vary in terms of degrees. In other words, one might be considered an extravert; however, someone could be more extraverted.

The Big Five is typically assessed by administering questionnaires such as the Big Five Inventory (BFI; John et al., 1991); however, an alternative method to assessing an individual's Big Five is through analysis of one's writing style. The way a person writes is reliably stable over a period of time (Pennebaker and King, 1999; Mehl and Pennebaker, 2003) which is similar to the stability of one's Big Five (Cobb-Clark and Schurer, 2012). Multiple studies have also shown how certain writing styles correlate to certain degrees of personality from analysis of student essays and journal abstracts (Pennebaker and King, 1999) to emails (Gill and Oberlander, 2002) to web blogs (Gill et al., 2009; Li and Chignell, 2010) to posts from social network sites (Qiu et al., 2012; Schwartz et al., 2013; Marshall et al., 2015). It is through this link between personality and writing that the field of text-based personality recognition emerged.

Although the field has taken great strides in determining state-of-the-art techniques in data processing, feature extraction, and machine learning, little focus is given to exploring language use of a group of individuals, such as those defined by nationality, in modeling personality traits. Individuals of the same nationality share practices and are exposed to certain situations that can lead to the development of certain psychological tendencies (Markus and Kitayama, 1998). Con-

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 112–122
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

versations and discussions expose individual differences and these differences eventually become embedded into natural language (Goldberg, 1981). However, many nationals are not limited to speaking just one language, such as how Filipinos speak Filipino and English, the two national languages of the Philippines. The addition of a number of regional/indigenous languages, along with the commonness of code-switching, allow for a Filipino to have a rich and diverse vocabulary. This rich vocabulary presents an opportunity to create a text-based personality model based on how Filipinos speak, regardless of the language they use. In order to do so, a web application was constructed to collect personal and Twitter data in which there were 250 Filipino participants. Raw personality scores were then experimented upon in order to determine the representation (continuous or discretized) that would best capture information. Tweets were then processed using simple language-independent natural language processing techniques. Finally, personality was modeled using both regression and classification techniques. The contributions of this paper are as follows:

- A corpus was created consisting of $610,448$ tweets from 250 Filipino participants. Each participant's personality traits were also assessed using the Big Five Inventory. Although a relatively small dataset, it serves as a source of information in which further experimentation can be performed.
- In both regression and classification, Conscientiousness is consistently the easiest personality trait to model, followed by Extraversion. Classification models for Agreeableness and Neuroticism produced subpar performances and did not fare well in regression. Lastly, models for Openness generally struggled in performance.
- In experimenting with personality score representations, results show that Neuroticism and Openness did not benefit from modeling extreme outliers ($\pm 1SD$ from the mean). Both traits were better modeled with a relaxed cut off at $\pm 0.5SD$, implying that useful information was lost when removing participants between $\pm(0.5SD - 1SD)$. As for the remaining three traits, performance was best when dealing with extreme outliers, as originally expected.

2 Related Literature

The early studies of the field mostly experimented with different feature extraction techniques on the Pennebaker and King (1999) Essay Dataset and utilized various Support Vector Machines for classification. Argamon et al. (2005) focused on determining high and low (top and bottom $\frac{1}{3}$) scoring individuals on the Extraversion and Neuroticism dimensions. Features were extracted based on a list of function words, along with other features based on Systemic Function Grammar. Their work showed that simple linguistic features contained information in determining personality traits – a task that requires "focused questions" such as those found in personality questionnaires. Soon after, multiple studies (Mairesse et al., 2007; Poria et al., 2013; Mohammad and Kiritchenko, 2013) utilized different linguistic resources in extracting information, including the Linguistic Inquire and Word Count (LIWC), MRC Psycholingusitic Database, NCR Emotion and Hashtag Lexicon, and SenticNet. Mairesse et al. (2007) conducted the first extensive study covering all five traits and treated personality recognition not just as a classification problem, but also as a regression and ranking problem as well. Their feature set is often referred to as the Mairesse baseline and consists of LIWC and MRC features. In another work, affect-related words were found to aid model performance when paired with LIWC and MRC (Mohammad and Kiritchenko, 2013). The method leading to the best improvement was where sentic computing was utilized in order to extract common sense knowledge with affective and sentiment information (Poria et al., 2013). Across the previously mentioned studies, Openness was found to be the easiest trait to model, while Agreeableness was the hardest to model.

As for studies that collected data from online sources, there was particular attention given to blogging sites. Blogs were an interesting source of data because of their personal nature. Oberlander and Nowson (2006) sourced their data from bloggers whom they administered a 41-item personality test. Classification was performed for all of the Big Five except for Openness due to non-normal distribution of personality scores. Once again, participants were grouped according to their scores based on varying levels of standard deviation (greater than 1SD, 0.5SD, and the mean). N-gram occurrence was utilized for extracting information and various feature selection techniques were employed. Nowson and Oberlander (2007) mirrored the previous study's methodology, but experimented with both the previous dataset and a new dataset. However, Iacobelli et al. (2011) produced the most notable results using the new dataset of the previous study. Although they tested with LIWC features, they found that using boolean scoring (present or not present) performed much better. Despite utilizing a coarse questionnaire, they managed to produced the best performing models with Openness being the easiest to model and Neuroticism being the most difficult.

Other early studies that sourced online data targeted social networking sites such as Twitter and Facebook in order to dealing with enormous amounts of data. Two studies (Golbeck et al., 2011a,b) were very similar as they used LIWC to process text from Twitter and Facebook, respectively. Their main difference was the use of site-specific information, such as internal Facebook stats or Twitter usage. The later study also utilized MRC as an additional means to extract information. But most noteworthy of all was of Schwartz et al. (2013) in which the biggest study on personality modeling was conducted with a total 75,000 Facebook volunteers. They highlighted the use of Differential Language Analysis as a means to generate open topics in comparison to the closed topics – categories generated by LIWC.

More recent developments involve the shift to analyzing non-English text. This could be seen in the PAN2015 (Rangel et al., 2015), where English, Spanish, Italian, and Dutch Tweets were made available to multiple research teams. One of the top performing submissions González-Gallardo et al. (2015) extracted n-grams of characters and utilized FreeLing, a language processing tool. FreeLing had resources for each of the languages in the dataset except for Dutch, so the English module was utilized despite possibly creating more errors. In Alvarez-Carmona et al. (2015), regarded as the top performing submission, focus was given to extracting discriminative and descriptive features. This was done by applying Second Order Attributes and Latent Semantic Analysis on a Term Frequency Inverse Document Frequency matrix. Outside of the PAN2015, Peng et al. (2015) focused on predicting Extraversion by segmenting Chinese characters from Chinese Facebook users. As Chinese characters are harder to delimit than other languages, they utilized Jieba, a Chinese character tokenizer. Lastly, Xue et al. (2017) focused on the use of Label Distribution Learning as an alternative to common machine learning algorithms while processing Chinese text. They extract information from posts from Sina Weibo users with TextMind, a Chinese language psychological analysis system similar to LIWC.

Currently, trends in the field of text-based personality recognition revolve around the use of Deep Learning, as the learning algorithm, and word embedding, as the way to represent text. Studies typically do not vary from using the two techniques, but distinguish themselves through their data source, such as how Yu and Markov (2017) experiments using a small subset of Facebook status posts. Another study (Majumder et al., 2017) considered adding the Mairesse baseline to their feature set in the analysis of the Essay Dataset. Tandera et al. (2017) used two Facebook datasets, one from MyPersonality and the other manually collected. Aside from word embedding, they included features from LIWC and SPLICE, another linguistic analysis tool. Lastly, Arnoux et al. (2017), although utilizing Gausian Process regression instead of Deep Leaning, still made use of word embedding. Their results showed that it was possible to reduce a dataset significantly while still achieving comparable model performances.

3 Methodology

This research collected data and approached modeling of personality traits through different combinations of data pre-processing, feature extraction, feature reduction, and machine learning techniques. Figure 1 shows an overview of the methodology.

3.1 Data Collection

A web application was developed to interface with Twitter and administer both a personal information sheet and a personality test. The information sheet asked for information such as sex, age, and nationality, while the personality test was the Big Five Inventory (BFI; John et al., 1991, 2008), a 44-item self-report questionnaire that measures the Big Five on a 5-point scale.

Recruitment of participants was mainly performed through postings on Facebook and Twitter. Friends and colleagues were targeted first which then later expanded to their social networks by word-of-mouth. However, a majority of the re-

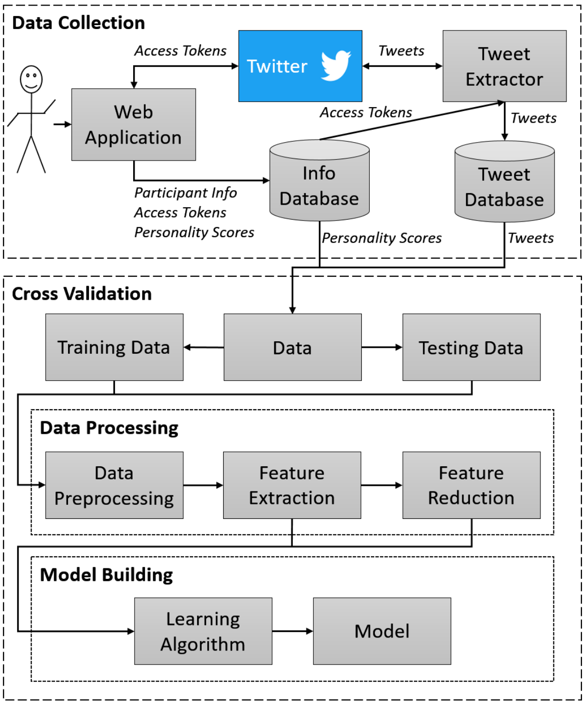

Figure 1: The methodology of this research

cruitment was focused on large Facebook groups into order to reach individuals outside of the the researcher's social network. Twitter Ads was also utilized to increase the reach of the web app, but it only resulted in a hand full of participants. Participants received no incentives for taking part in the data collection except for seeing the output of the personality test.

After recruitment, it was important to filter the participants based on their personal and Twitter account information as anyone could access the web application. Individuals were removed if they were non-Filipino or had less than 100 tweets. The filters in place ensured that the participants were at least Filipino, whether pure or mixed, and had a suitable amount of text data to process.

Each participant's Twitter account was then crawled using a Python script which retrieved up to 3,200[1] of their most recent tweets. If participants had less than 3,200 tweets, then as many tweets as possible were retrieved. Any retweets found were removed as they were not written directly by the participant. An exception was made for quoted tweets because a portion of the tweet is written by the participant. Lastly, participants whose tweet count fell below 100 because of the removal of retweets were removed.

After all of the filtering, a total of 250 individuals qualified as participants for this research. Table 1 shows the demographics of the participants and Table 2 shows the statistical characteristics of the participants' personality trait scores.

This research managed to collect $712,762$ tweets, but after retweets were removed, the total tweet count stood at $610,448$ with an average of $2,441.79$ tweets (SD=723.8) per participant. The participant with the lowest tweet count had 107 and the highest had $3,196$.

Table 1: Participant demographics.

Total Participant Count	250
Age	
Mean	22.34
Standard deviation	3.57
Min	19
Max	51
Sex	
Male	79
Female	169
Intersex	1
Decline to disclose	1
Nationality	
Filipino	234
Mixed-Filipino[1]	16

[1] Mixed-Filipinos are those who declared themselves Filipino and one or more nationalities

Table 2: Statistical characteristics of participants' personality trait scores.

Personality	**Mean**	**SD**	**Min**	**Max**
Openness	3.45	0.44	2.00	4.50
Conscientiousness	3.08	0.62	1.44	4.67
Extraversion	3.13	0.80	1.25	5.00
Agreeableness	3.59	0.67	1.56	5.00
Neuroticism	3.39	0.75	1.25	4.88

As this research focused on how Filipinos tweeted regardless of language, tweets in all languages were retained. 58.14% of the total tweets were labeled as English, while 31.89% were labeled as Tagalog[2]. The remaining tweets were either labeled as undefined (5.09%; unable to determine the language) or other languages (4.89%).

[1] The most recent 3,200 tweets is a limitation of Twitter's API; More information can be found in https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

[2] Tagalog is a Philippine language that served as the basis for Filipino, the national language

Top among the other labels included Indonesian (1.22%) and Spanish (0.07%) – two languages that share words commonly used in Filipino. Language labels were taken from the metadata of a tweet. Table 3 shows a breakdown of the languages present in the corpus.

Table 3: The breakdown of languages present in the corpus as well as their usage per participant.

Lang	Count		Mean	SD
Eng	354889	58.14%	1419.56	585.61
Tag	194644	31.89%	778.58	516.81
Und	31062	5.09%	124.25	78.82
Oth	29853	4.89%	119.41	75.70

[Abbr] Eng - English, Tag - Tagalog, Und - Undefined, Oth - Others

3.2 Data Pre-processing

Data pre-processing is performed in order to prepare raw text and personality trait scores for classification. This research defines the Term-Document Matrix as the following:

1. Term (t): an n-gram of tokens extracted from a single tweet of a participant
2. Document (d): all terms derived from all tweets of a participant
3. Collection (C): a set of documents of all participants

3.2.1 Tokenizing

This research utilizes Tweetokenize (Suttles, 2013), a regular expression based tokenizer for Twitter, to parse each character in a tweet to properly identify words/terms and social media entities (usernames, hashtags, or emojis). The default settings were kept when processing the tweets and are as follows:

1. Uppercase letters were converted to lowercase; but tokens, where all letters are capitalized, are not converted to lowercase,
2. Repeating letters are limited to 3 (e.g. *hmmmmm* and *hmmmm* are both reduced to *hmmm*),
3. Identified usernames and urls were replaced with *USERNAME*,
4. Identified urls were replaced with *URL*,
5. Identified hashtags are not replaced with a token, and
6. Stop words are not removed.

3.2.2 N-Grams

An n-gram is a sequence of n tokens. This research experimented with only 1-grams. N-grams were extracted through the use of Natural Language Toolkit (NTLK; Bird et al., 2009).

3.2.3 Document Frequency Filtering

Document frequency filtering is applied to remove terms that are either too common or too unique. The document frequency of a term t in a collection C is defined as

$$DF(t, C) = \frac{N_{t,C}}{N_C}, \tag{1}$$

where $N_{t,C}$ is the number of documents in C wherein t occurs at least once and N_C is the total number of documents found in C. Different combinations of minimum and maximum thresholds were experimented upon, but this research limits the combinations to:

1. min: 1%, max: 99%, and
2. min: 10%, max: 70%.

3.2.4 Personality Trait Score Representation

Personality trait scores are continuous values and instantly fit as input for regression models; however, these scores must be discretized in order to perform classification. This research modifies Oberlander and Nowson (2006)'s idea of partitioning the participants based on their personality scores' mean (μ) and standard deviation (SD). Therefore, five different methods are experimented upon and are defined given a personality trait score s as

1. Continuous - refers to the natural form of personality trait scores and will be the sole trait score representation for regression
2. LAH - Stands for *Low Average High*; Groups all participants into low, average, and high; Participants nearest to a boundary between two partition have similar scores; Defined as:

$$\text{LAH}(s) = \begin{cases} high, & \text{if } s > \mu + \frac{SD}{2}; \\ low, & \text{if } s < \mu - \frac{SD}{2}; \\ average, & \text{otherwise.} \end{cases} \tag{2}$$

3. LH - Stands for *Low High*; Groups all participants into low and high, but participants nearest to the boundary still have similar scores; Defined as:

$$\text{LH}(s) = \begin{cases} high, & \text{if } s > \mu; \\ low, & \text{if } s < \mu. \end{cases} \tag{3}$$

4. LHNA - Stands for *Low High, No Average*; Creates distinction between high and low scorers by removing all average; Results in the removal of ∼38.2% of the participants; Defined as:

$$\text{LHNA}(s) = \begin{cases} high, & \text{if } s > \mu + \frac{SD}{2}; \\ low, & \text{if } s < \mu - \frac{SD}{2}; \\ \text{omit}, & \text{otherwise.} \end{cases} \quad (4)$$

5. LHNASD - Stands for *Low High, No Average, whole Standard Deviation*; Creates the most distinction between high and low scorers by increasing the threshold to $\pm 1SD$; Results in the removal of ∼68.2% of the participants; Defined as:

$$\text{LHNASD}(s) = \begin{cases} high, & \text{if } s > \mu + SD; \\ low, & \text{if } s < \mu - SD; \\ \text{omit}, & \text{otherwise.} \end{cases} \quad (5)$$

A visualization of the different representations can be seen in Figure 2

3.3 Feature Extraction

In order to extract information from raw text, two feature extraction techniques are used in this research: Term Frequency Inverse Document Frequency (TFIDF) and Term Occurrence (TO). Language independent approaches are preferred due to the presence of English and Filipino, among other langauges.

3.3.1 TFIDF

Term Frequency Inverse Document Frequency (TFIDF) captures the frequency of use of a term in a given document, while factoring the importance of the term in relation to the overall collection of documents. TFIDF was computed for each term in each document to construct a TFIDF word-matrix. All values were then normalized. The features in TFIDF dataset consists of the terms that appear throughout the entire collection of Twitter users.

TFIDF is computed by multiplying the Term Frequency (TF) with the Inverse Document Frequency (IDF). Given a term t of a document d of a collection C, TFIDF is defined as:

$$TFIDF(t,d,C) = \frac{N_{t,d}}{N_d} \cdot \frac{N_C}{N_{t,C}}, \quad (6)$$

where $N_{t,d}$ is the number of t in d, N_d is the total number of terms in d, N_c is the total number of documents in C, and $N_{t,C}$ is the number of documents in C wherein t occurs at least once.

3.3.2 Term Occurrence

Term occurrence (TO) is a binary representation of whether a particular term was used or not – occurred or not occurred. The TO of a term t given a document d can be defined as:

$$TO(t,d) = \begin{cases} 1, & \text{if } N_{t,d} > 0; \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where the output is 1 if where $N_{t,d}$, the number of t in d, is greater than 0, and 0 if otherwise.

3.4 Feature Reduction

Even with the utilization of document frequency filtering, there would still be a good number of features that could contain both relevant and irrelevant information. Feature reduction would reduce a dataset, while retaining the most relevant features. Therefore, reduction is applied on the training set and would consist of the top 20% of the results of univariate linear regression test for regression and chi-square (χ^2) for classification. Experiments were performed with and without feature reduction in order to properly observe the effects.

3.5 Machine Learning Algorithms

Multiple learning algorithms were experimented upon, but this research highlights the following algorithms:

1. Linear Regression (LIN),
2. Ridge Regression (RID),
3. Support Vector Machines (linear SVM), and
4. Logistic Regression (LOG).

The algorithms were highlighted because they performed better than other the algorithms during the experiments of this research. Those that produced subpar models were not reported. The algorithms were implemented using Scikit-Learn (Pedregosa et al., 2011), a general purpose machine learning Python library. All settings were kept to Scikit-Learn's default settings.

3.6 Model Evaluation

Data was split into training (60%) and testing (40%) sets in order to have enough data for learning, while having enough data remaining for testing. As the sample count for the classes was not balanced, 10-fold stratified cross validation was performed to ensure that each class was well represented in each fold. For classification models, both F_1 score and kappa statistic are observed in

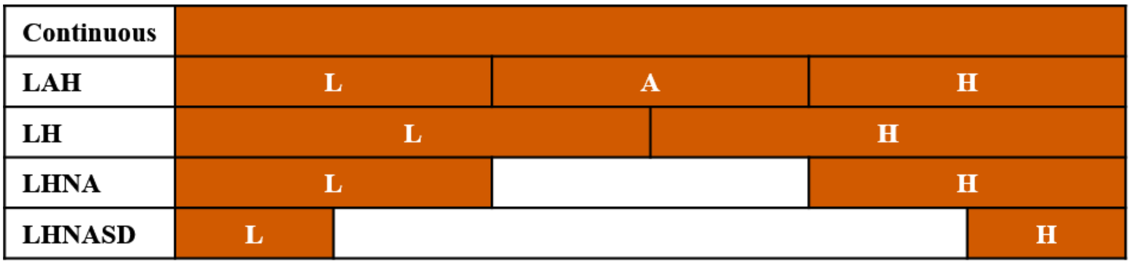

Figure 2: The different ways personality trait scores are represented in this research. Boxes filled with color represent partitions of participants.

evaluating a model's performance. For regression models, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 are observed.

4 Results and Discussion

A total of 600 models were created based on the different combinations of pre-processing, feature extraction, feature reduction, and ML techniques. All combinations were experiment on and only the best models are reported. To determine the best models per trait, goodness of fit was prioritized over minimizing error; therefore, R^2 is the basis for regression models and kappa statistic is the basis for classification. Table 4 and Table 5 shows the best regression and classification models, respectively. Each of the best performing models is compared against a baseline model of the same configuration and can be seen in Figure 3 for regression and Figure 4 for classification. Additionally, the effects of discretizing trait scores in relation to the performance of personality models is analyzed. The best classification models per personality score representation are found in Table 6.

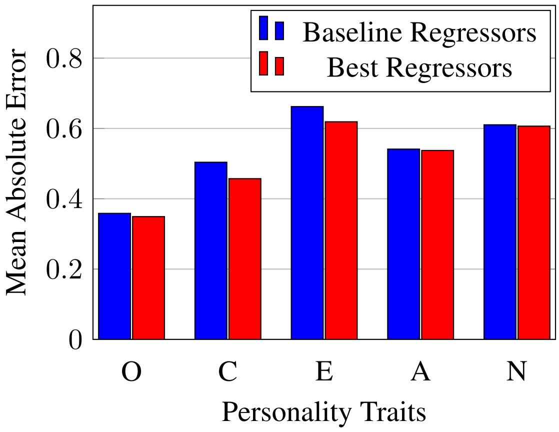

Figure 3: A comparison of the MAE between baseline mean regressors and the best regression models (as found in Table 4) per personality trait

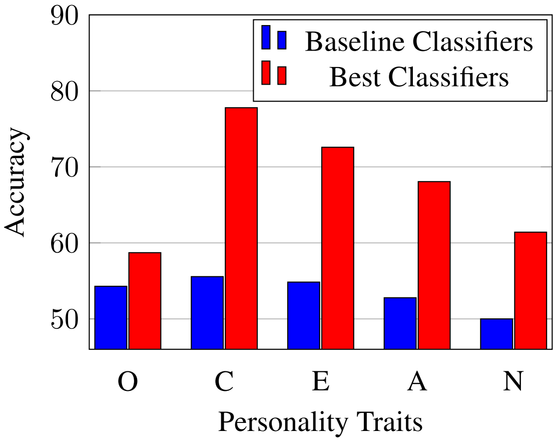

Figure 4: A comparison of the accuracies between baseline majority classifiers and the best classification models (as found in Table 5) per personality trait

General Findings. Out of all the Big Five, Conscientiousness is the easiest to model. Both in regression and classification, Conscientiousness had models with the best R^2 (0.1523) and kappa (0.5516), respectively. Extraversion came in second, again both in regression and classifications, with its R^2 of 0.1035 and a kappa of 0.4376. Results for both Conscientiousness and Extraversion indicate that simple TFIDF or TO features were able to extract useful information from a corpus of Filipino and English tweets. The remaining three traits performed poorly for regression, but Agreeableness and Neuroticism fared better in classification. The improvement in performance can mainly be attributed to excluding average scoring participants and looking for patterns in how the outliers generally tweet.

As for Openness, it can be considered the hardest trait to model, particularly because it performed worst in classification ($F_1 = 0.5669$ and $\kappa = 0.1438$). Models for openness are seen to utilize the softer document frequency filter (min=1%;max=99%) more often than in other traits. This indicates that strong patterns are not present and that in order to make appropriate predictions, most, if not all, information is needed.

Table 4: The performance and configuration of the best performing regression models per personality trait. Models were selected based on R^2.

Trait	Features	Doc Freq	Regressor	MSE	MAE	R^2
O	TO	1%-99%	LIN	0.1890	0.3493	0.0143
C	TFIDF	1%-99%	LIN	0.3174	0.4572	0.1523
E	TFIDF	10%-70%	LIN	0.5719	0.6190	0.1035
A	TFIDF	1%-99%	RID	0.4393	0.5374	-0.0088
N	TFIDF	1%-99%	LIN	0.5558	0.6066	-0.0031

Note: Although there were experiments with and without feature reduction, all the best performing models utilized all features; therefore, feature reduction was not included in the table.

Table 5: The performance and configuration of the best performing classification models per personality trait. Models were selected based on kappa statistic.

Trait	Personality Rep	Features	Doc Freq	χ^2 Selection	Classifier	F_1	κ
O	LHNA	TO	1%-99%	top 20%	SVM	0.5669	0.1438
C	LHNASD	TFIDF	10%-70%	top 20%	LOG	0.7764	0.5516
E	LHNASD	TO	10%-70%	top 20%	LOG	0.7165	0.4376
A	LHNASD	TO	10%-70%	n/a	LOG	0.6767	0.3547
N	LHNA	TFIDF+TO	10%-70%	top 20%	LOG	0.6086	0.2281

This is also supported by the small differences in evaluation metrics found across the different personality trait score representations as seen in Table 6. In other words, retaining extreme outliers (LHNSSD) did not help in classification of Openness and actually performed slightly worse than having all participants presents across 3 trait groupings (LAH).

Configurations in Regression Models. The best regression models, as seen in Table 4, indicate that there are no relatively strong features in the prediction of an individual's trait score. Four traits utilized the softer document frequency filter (min=1%;max=99%) with Extraversion using the harsher one. In terms of features, TFIDF values are preferred over TO. And interestingly, none of the best models utilized feature reduction. However, despite the generally low performances, the findings show that simple TFIDF values contain some information about one's personality, at least for Conscientiousness and Extraversion. TFIDF values can be considered shallow information, so further investigation using more in-depth feature extraction techniques could yield better results.

Configurations in Classification Models. All of the best classification models, as seen in Table 5, utilized personality representations that removed average scoring users and focused on outliers - LHNASD and LHNA. As for features, TO was more useful than TFIDF as it was used in four out of the five traits; however, TFIDF was utilized by Conscientiousness, the best overall performing model. The features remaining after the harsher document frequency filter (min=10%;max=70%) proved to be more useful than the softer filter in all traits, except for Openness. This indicates that patterns indeed emerge when comparing individuals on the opposite ends of a personality dimension. Lastly, unlike in regression, feature selection was more useful than simply allowing the ML algorithms find patterns in the data.

Personality Trait Representation. As personality trait questionnaires typically output a numerical value, it is important to look at different ways to represent the scores – whether in continuous or discrete form. Continuous values provide the best coverage as they match the raw values output by questionnaires (e.g. 1.0 to 5.0 for the Big Five Inventory) and include all participants for testing and training purposes. Problems arise as features may not be highly correlated to the whole personality dimension or possible be correlated to a subset of individuals. On the other hand, discrete values allow for the grouping of individuals based on the mean and standard deviation of their scores. Grouping individuals makes classification possi-

Table 6: The F1 scores and kappa of the best performing classifiers per personality score representation.

Traits	LAH		LH		LHNA		LHNASD	
	F_1	κ	F_1	κ	F_1	κ	F_1	κ
O	0.4176	0.1233	0.5691	0.1388	0.5669	**0.1438**	0.5530	0.1222
C	0.4505	0.1693	0.6646	0.3295	0.7497	0.5010	0.7764	**0.5516**
E	0.4526	0.1680	0.6178	0.2359	0.6492	0.3033	0.7165	**0.4376**
A	0.3796	0.0743	0.5635	0.1298	0.5595	0.1329	0.6767	**0.3547**
N	0.3651	0.0475	0.5347	0.0711	0.6086	**0.2281**	0.5707	0.1469

ble, but problems can arise with individuals nearest to the boundary of a group as they would have similar scores to individuals in the groups next to them. A solution to this would be to create space in between classes; however, participants would have to be removed resulting in possible information loss. Because of the pros and cons of each method, analysis is performed on how personality scores affect personality modeling of Filipino Twitter users.

As seen in Table 6, LHNASD (*Low, High, No Average, whole Standard Deviation*) produced the best performing classifiers for three out of the five traits, namely Conscientiousness, Extraversion, and Agreeableness. This was expected because useful information was most likely found when comparing extreme high and low outliers, and not when including those who scored nearer to the mean. This is apparent by the gradual increase in evaluation metrics as the classes are reduced in size and the distances between outliers expands. However, it is important to note that Neuroticism and Openness fared best when utilizing the LHNA (*Low, High, No Average*) representation – the other representation that places space between outliers. LHNA has almost double the training data than LHNASD. Training instances of LHNA range from 88 to 103 across all traits, while LHNASD ranges from 46 to 53. This implies that there isn't strong discriminative information between extreme outliers and that the removal of participants also removed information useful for Neuroticism and Openness. Interestingly, models for Openness do not vary so much in terms of kappa statistic across all personality representation. The model for LAH (*Low, Average, High*), the hardest representation to predict because it has three class, has a kappa of (0.1233), while the model of LHNA has a kappa of 0.1438. In fact, LAH actually has better agreement than that of LHNASD (0.1222) indicating that the outliers of Openness are not easily distinguishable, at least with respect to the features extracted.

5 Conclusion and Recommendations

This research was able to collect text and personal data from 250 Filipino Twitter Users and use the way they tweet, regardless of language, to create personality trait models. In the process, different combinations of data processing and machine learning techniques were experimented upon to identify the best configurations and produce the best models. Findings show that Conscientiousness is an easy trait to model, directly followed by Extraversion. On the other hand, Openness is the hardest trait to model. Experiments in regression did not produce suitable models, but at least indicated that simple TFIDF values contain some information for Conscientiousness and Extraversion. Classification models had better results and generally benefited from modeling the outliers instead of classifying all of the participants. Lastly, Neuroticism and Openness also did not benefit from modeling of extreme outliers ($\pm 1SD$ from the mean) implying that outliers for the trait are not easily distinguishable.

As the participants were all Filipinos, further analysis of the content could provide insights into how personality traits manifest through the language use of Filipino Twitter users. The addition of more in-depth feature extraction techniques, such as topic modeling or the integration of multiple language-specific resources, might also help in improving the models' performances. Lastly, creating specific models of groups of individuals defined by demographics – such as by age, gender, or nationality – regardless of the number of languages used, proves to be a useful approach in personality modeling and can serve as a starting point for understanding their linguistic style.

References

Miguel A Alvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villasenor-Pineda, and Hugo Jair Escalante. 2015. Inaoes participation at pan15: Author profiling task. *Working Notes Papers of the CLEF*.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James Pennebaker. 2005. Lexical predictors of personality type.

Pierre Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Deborah A Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15.

Howard S. Friedman and Miriam W. Schustack. 2014. *Personality: Classic theories and modern research*. Pearson.

Alastair J Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? personality, topic and motivation in blogs. In *ICWSM*.

Alastair J Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011a. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011b. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262. ACM.

Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165.

Carlos E González-Gallardo, Azucena Montes, Gerardo Sierra, J Antonio Núnez-Juárez, Adolfo Jonathan Salinas-López, and Juan Ek. 2015. Tweets classification using corpus dependent tags, character and pos n-grams. In *CLEF (Working Notes)*.

Francisco Iacobelli, Alastair J Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Affective computing and intelligent interaction*, pages 568–577. Springer.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. The big five inventoryversions 4a and 54.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158.

Randy J. Larsen and David M. Buss. 2008. *Personality psychology: Domains of knowledge about human nature*. McGraw Hill Education.

Jamy Li and Mark Chignell. 2010. Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies*, 68(9):589–602.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Hazel Rose Markus and Shinobu Kitayama. 1998. The cultural psychology of personality. *Journal of cross-cultural psychology*, 29(1):63–87.

Tara C Marshall, Katharina Lefringhausen, and Nelli Ferenczi. 2015. The big five, self-esteem, and narcissism as predictors of the topics people write about in facebook status updates. *Personality and Individual Differences*, 85:35–40.

Robert R McCrae and Paul T Costa Jr. A five-factor theory of personality. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*. The Guilford Press, New York, NY.

Matthias R Mehl and James W Pennebaker. 2003. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857.

Saif M Mohammad and Svetlana Kiritchenko. 2013. Using nuances of emotion to identify personality. *Proceedings of ICWSM*.

Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574.

Scott Nowson and Jon Oberlander. 2007. Identifying more bloggers. *Proceedings of ICWSM*.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Kuei-Hsiang Peng, Li-Heng Liou, Cheng-Shang Chang, and Duan-Shin Lee. 2015. Predicting personality traits of chinese users based on facebook wall posts. In *Wireless and Optical Communication Conference (WOCC), 2015 24th*, pages 9–14. IEEE.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. 2013. Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence*, pages 484–496. Springer.

Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, page 2015. sn.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Jared Suttles. 2013. Tweetokenize. https://github.com/jaredks/tweetokenize.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. 2017. Personality prediction system from facebook users. *Procedia Computer Science*, 116:604–611.

Di Xue, Zheng Hong, Shize Guo, Liang Gao, Lifa Wu, Jinghua Zheng, and Nan Zhao. 2017. Personality recognition on social media with label distribution learning. *IEEE Access*, 5:13478–13488.

Jianguo Yu and Konstantin Markov. 2017. Deep learning based personality recognition from facebook status updates. In *Proceedings of 8th International Conference on Awareness Science and Technology (iCAST)*, pages 383–387. IEEE.