Modeling Personality Traits by Predicting Questionnaire Responses as an Alternative Approach to Filipino Automatic Personality Recognition

Alessandra Pauleen I. Gomez, Ibrahim D. Kahil, Shaun Vincent N. Ong, Edward P. Tighe

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Modeling Personality Traits by Predicting Questionnaire Responses as an Alternative Approach to Filipino Automatic Personality Recognition

Alessandra Pauleen I. Gomez, Ibrahim D. Kahil,
Shaun Vincent N. Ong, Edward P. Tighe
Department of Software Technology and Center for Language Technologies
De La Salle University, Manila, Philippines
{alessandra_gomez,ibrahim_kahil,shaun_ong,edward.tighe}@dlsu.edu.ph

Abstract

Emerging research in Filipino Automatic Personality Recognition (APR) often utilizes social media data for its widespread availability and natural expression. However, current approaches focusing on direct personality trait modeling often yield subpar results, prompting exploration of alternative methods. Thus, we explored an APR framework where individual personality questionnaire item responses are predicted and then aggregated to estimate trait scores. Using text data from 2,168 Filipino X (formerly Twitter) users, we trained models for each item in the Big Five Inventory (BFI) related to Extraversion and Conscientiousness. We also experimented with multiple configurations of logistic regression, SVM, and XGBoost models using TF-IDF and term occurrence values. Findings highlight the challenges in predicting trait scores for both Extraversion and Conscientiousness. While implementing a hierarchical classification scheme at the item level showed some improvement, especially for Conscientiousness, overall trait-level performance remains lacking. Overall, while the original pipeline as well as the integration of a hierarchical approach show potential, significant improvements are needed before this item-based framework can be effectively used for APR.

1 Introduction

The extent of a person's individuality and identity encompasses a great number of factors, from their daily experiences all the way to their hobbies, interests, and way of interacting with others. Such traits are often considered part of one's personality—defined by the American Psychological Association as a collection of "enduring characteristics and behavior that comprise a person's unique adjustment to life." Numerous scientific theories and approaches have been created in order to deepen the world's understanding of personality into how it is today. As part of its evolution, personality psychology has been integrated into computational science; through the use of machine learning and natural language processing (NLP), personality recognition was made possible by incorporating data or signals from human-machine interaction, including but not limited to social media and telecommunication (Mushtaq and Kumar, 2022).

Works on text-based APR have branched out to include attempts to derive personality from social media posts within a specific regional context. There are a lot of cultural linguistic nuances that can serve as integral personality indicators, yet models are not always able to extract information that properly encapsulates these intricacies brought about by multilingualism.

With this new aspect of APR, studies on personality recognition on Filipino user data have begun to take place. From attempts at extraction methods (Agno et al., 2019; Chua Chiaco et al., 2022) to modeling Filipino personality traits using supervised learning models (Tighe and Cheng, 2018), Filipino APR studies are slowly breaking ground with the goal of applying techniques that can capture the rich linguistic diversity of the nation. However, since this particular branch of study is relatively new, there have been unsuccessful ventures as well; at present, existing studies on the use of higher complexity models such as neural networks (Tighe et al., 2020) failed to yield good results, especially considering that this was attempted when Filipino user data was scarce.

Given the current state of Filipino APR, it begs the question of whether it is possible to utilize another approach at modeling personality traits instead of directly generating user personality profiles from social media data. One such alternative is a questionnaire-based approach, wherein models trained on social media data will then predict how the user might answer a question from a personality inventory. By combining APR with

a questionnaire-based framework, it may reveal a new angle of extracting, processing, and analyzing data that will be able to account for the cultural linguistic cues found in the Filipino language—and by extension, can also be applied in the context of general, non-regional APR research.

The general objective of this study is to investigate the effectiveness of a questionnaire item-based prediction approach to automatic personality recognition on social media text data. The specific objectives of the study are defined below:

1. To define a list of qualification criteria for deriving a subset of the *PagkataoKo* dataset;
2. To extract text-based information from users' social media posts;
3. To build and train prediction models for each personality questionnaire item using the generated user embeddings;
4. To evaluate and analyze the performance of the item-based prediction models at an individual item level and an overall trait score level; and
5. To compare the item-based prediction approach to automatic personality recognition against baseline prediction models

The results of this study represent the output of a different approach to APR, specifically predicting users' Likert scale-type answers to the BFI questionnaire instead of predicting their personality trait scores directly. Due to the uniqueness of the approach, it offers the viability of utilizing the approach to conduct APR and introduces the idea of predicting questionnaire items for other models as well.

2 Methodology

This section provides a step-by-step breakdown of the individual processes undertaken to achieve the objectives of this study. As seen in Figure 1 that shows the overall research pipeline, using the original *PagkataoKo* dataset, a smaller subset of data was derived by filtering based on a set of defined qualification criteria. Then, preprocessing and feature extraction were done on the data of each user from their X (formerly Twitter) posts. After, feature reduction was performed to further trim down the number of features. Machine learning models were then built for each questionnaire item under the Extraversion and Conscientiousness traits, which were trained and tested. The mentioned traits were chosen among the Big Five in accordance with Tighe and Cheng's (2018) findings about the two being the easiest to model.

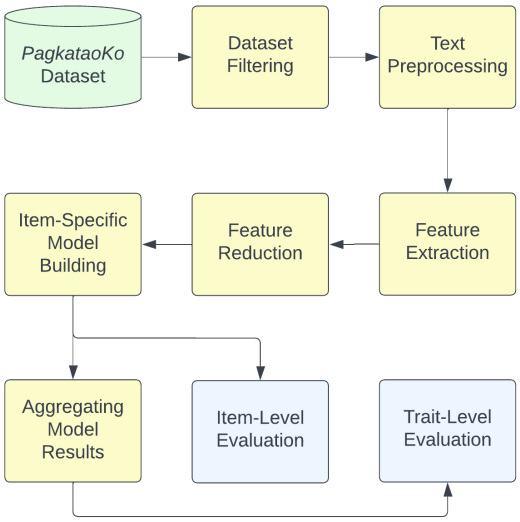

Figure 1: Diagram of the Overall Research Pipeline Following Our Proposed Item-Based Approach

The resulting predictions for each questionnaire item were then aggregated to estimate the Extraversion and Conscientiousness trait scores of each user. Evaluation of the machine learning models were conducted for each individual item, along with a separate trait-level evaluation to assess the performance of the overall approach of utilizing questionnaire item predictions for estimating personality trait scores.

2.1 Data Source

The dataset used in the study is the *PagkataoKo* dataset curated by Tighe et al. (2022). Collected starting the first week of June 2019 up until the second week of February 2020, the study was able to gather a total of 3,128 records and contains information about Filipino X (formerly Twitter) and/or Instagram users such as demographic data, account metadata, post data, and personality data.

The primary information utilized from the dataset includes the X (formerly Twitter) post data such as the actual post text and the data containing BFI responses and overall score per dimension which are needed for ground truth comparisons and evaluation.

To align with the scope of the study, the data was

filtered according to set qualification criteria. First, the users must be of Philippine legal age; that is, they must be at least 18 years old. Second, as the study is focused on text-based data, the users must have X (formerly Twitter) with at least 50 posted tweets.

A simple demographic and summary statistic analysis was conducted on the original curated dataset as well as the filtered qualifying dataset. These statistics are reported on Table 1..

Demographics	Universal Set	Twitter Subset	Qualified Subset
Count	3,128	2,283	2,168
Age			
Mean	21.2	21.0	21.0
SD	3.9	3.9	3.6
Age Range			
18-20	53.9%	55.9%	56.0%
21-23	29.3%	29.0%	29.2%
24-26	9.3%	8.5%	8.5%
$\geq$ 27	7.5%	6.6%	6.3%
Sex			
Male	21.0%	22.0%	21.5%
Female	76.1%	75.0%	75.5%
Intersex	0.5%	0.6%	0.6%
Declined[1]	2.4%	2.5%	2.4%
Nationality			
Filipino	99.2%	99.1%	99.2%
Mixed[2]	0.8%	0.9%	0.8%

[1] Declined to disclose their sex
[2] Filipinos with one or more other nationalities

Table 1: Demographic statistics across the universal set of all participants (U), the subset of participants with Twitter accounts (T), and the subset of participants with Twitter accounts that satisfied the qualification criteria (QT)

2.2 Text Preprocessing

Preprocessing was first performed on the text corpus. The study mainly utilized tokenization and N-Grams. For tokenization, Marges's (2019) Pinoy TweetTokenizer will be used, which is a modified TweetTokenizer for the Filipino language. The tokenizer features are as follows:

1. Replacing usernames with a placeholder (i.e. USERNAME);
2. Hashtag tokenization;
3. Limiting repeating syllables;
4. Emoticon tokenization;
5. Replacing URLs with a placeholder (i.e. URL); and
6. Lowercasing

For N-Grams, the study utilized NLTK's *nltk.lm* package to extract *n*-grams of different lengths needed (Bird et al., 2009). It should be noted that only unigram and bigram features were tested.

2.3 Formulating User Documents

Concurrently, while performing text preprocessing, user documents were constructed wherein all tweets of a user were combined into one document for analysis. To do this, the study utilized the technique of concatenation of strings in each tweet of a particular user which then forms the user document. To implement this, tokenization was first performed on the text at the tweet level, followed by applying *n*-grams to the tokens of each tweet, outputting a group of tokens per tweet. From there, we concatenate the arrays of tokens together, formulating a user document for a particular user where these tokens are treated as terms.

2.4 Feature Extraction

Feature extraction was performed on the preprocessed text data to extract the necessary information from the text. The study utilized TF-IDF and Term Occurrence as the extraction methods. Due to the *PagkataoKo* dataset containing multiple languages (i.e., English and Filipino), both TF-IDF and Term Occurrence are among the more viable methods as these can handle multilingual text and terms. There are two parameters in the tfidfVectorizer that were included as experiment parameters, which are *min_df* and *max_df*. Both *min_df* and *max_df* are document frequency filters that remove features depending on the percentage of documents they are found in.

2.5 Feature Reduction

In order to retain only the most relevant features as input for model building, feature reduction techniques were employed on the training set. Note that this was also treated as an experiment parameter, testing between the use of the chi-square test and principal component analysis (PCA). Using the chi-square (X^2) test, we only retained the features that fall within the top 20% of results and these features were selected for training the machine learning models.

2.6 Model Building

The study made use of the following supervised machine learning models that focused on solving

a classification problem centered around the prediction of BFI item responses based on their social media data:

- Logistic Regression
- Support Vector Machine with a Non-Linear Kernel
- XGBoost

These three models were chosen because in the context of the study, they may perform best given the amount of data available.

It is worth noting that since the study focuses on predicting responses to BFI questions, individual models were created for each of the 17 BFI items under either Extraversion or Conscientiousness. In addition to the approach of directly classifying the specific Likert scale-type responses for each item, the study also experiments with a two-phase, hierarchical classification scheme. This alternative method involves training initial models that broadly classify users' responses into one of three categories: (a) 1-2, (b) 3, or (c) 4-5. Then, for the second phase, a set of binary models is trained for each item to further distinguish users' responses within each category, thus obtaining the specific item responses.

2.7 Aggregating Item-Level Model Results

Once the individual item-level models were used to predict the responses of a given user, these results were then be aggregated to estimate their raw personality trait scores. This may be accomplished by following the pseudocode depicted in Algorithm 1, which is patterned after the actual scoring metric of the BFI. The algorithm shows how to calculate each trait score by obtaining the average of the predicted responses for all question items that fall under a particular personality trait. In doing so, it should also be kept in mind that questions tagged as reversed should have their responses converted accordingly.

Algorithm 1 Aggregating Item-Level Model Results

Input: Predicted item responses for a given user
Output: List of estimated personality trait scores

```
initialize empty trait score list
for each personality trait do
    sum = 0
    for each question item under current trait do
        if question item is reversed then
            sum += REVERSE(predicted response)
        else
            sum += predicted response
        end if
    end for
    trait score = sum / number of questions under current trait
    append current trait score to trait score list
end for
return trait score list
```

3 Experiment Setup and Evaluation

3.1 Experiment Setup

This study experimented with multiple combinations of feature extraction, feature reduction, and machine-learning techniques to identify the configurations that yield the most optimal results.

A total of 17 item-level models were created for each configuration or combination of techniques as described above to correspond to each of the items in the Big Five Inventory that correspond to either Extraversion or Conscientiousness.

Furthermore, it should also be noted that a train-validation-test split was applied on the dataset, with a split ratio of 70%, 15%, and 15%, respectively. This was implemented by utilizing scikit-learn's *train_test_split* function to ensure objective and black-boxed splitting.

3.2 Item-Level Evaluation

This phase of the experiments centers on building models for the 8 items under Extraversion and the 9 items under Conscientiousness.

Experiment parameters came in the form of multiple combinations of feature extraction and reduction techniques as well as machine learning algorithms and configurations, all utilized to derive the best performing model for each item. Taking into account all of the experiment parameters except for the two-phase hierarchical classification scheme, there are a total of 96 configurations generated for each item (*2 feature extraction methods* × *2 feature reduction methods* × *3 machine learning algorithms* × *2 min_df values* × *4 max_df values*). Additionally, the set of 96 experiment configurations is conducted using the two-phase hierarchical classification approach, resulting in a final total of 192 models per questionnaire item (*96 models us-*

ing direct approach + 96 models using two-phase hierarchical classification approach).

Following model training and hyperparameter tuning, the primary metric that was used to determine the best model configuration for each item was the validation F1 score, as this takes into consideration the class imbalance present in the source dataset's distribution of item responses. In the case of the models created following the two-phase hierarchical classification approach, the validation F1 score of the initial broad classification models is the metric used as the basis for determining the best configurations. These best models then make the final predictions of the test users' answers, which are then compared to their ground-truth responses for each item.

Baseline models were implemented using majority class classifiers to serve as benchmarks for comparing the proposed best item models. These classifiers were trained using the responses for each item, identifying the majority class as a constant predictor.

3.3 Trait-Level Evaluation

This second phase of the experiment focused on acquiring the predicted item responses for each trait from the best item models in the previous phase and computing for the users' trait-level scores using the designated formula of the BFI.

Once the personality trait results were aggregated for each user in the test set and compared against their ground-truth trait scores, evaluation was performed with the use of root mean squared error (RMSE) and R^2 score.

Similar to the previous phase, baseline models were employed to have a further comparison and performance evaluation of the proposed approach. These baselines included a mean regressor, a simple linear regression model, and a multi-layer perceptron (MLP) regressor.

The mean regressor was trained using the raw personality trait scores from the dataset, with the average score for each trait serving as a constant predictor. Meanwhile, the pipeline for both the mean regressor and the MLP regressor follows a process similar to the proposed approach up until the feature reduction stage. However, instead of proceeding to item-specific model-building and aggregation, the pipeline for these baseline models directly transitions to trait-specific model building and trait-level evaluation. This divergence stems from their trait-based approach of training directly on the raw personality trait scores of each user, rather than on the individual item responses as in the proposed approach.

4 Results

4.1 Evaluation of Initial Proposed Approach

4.1.1 Item-Level Evaluation Results

Out of all the item-level models constructed and tested during experimentation, only the configurations that achieved the best validation results for each individual questionnaire item are reported.

Table 2 and Table 3 provide overviews of the best-performing models for each Extraversion item and each Conscientiousness item, respectively. The results of these item models are also juxtaposed with the results of baseline majority class classifiers, as illustrated in Figure 2 and Figure 3.

Across all of the Extraversion and Conscientiousness item models, there appears to be a fair amount of variance in the optimal configurations identified for almost all of the parameters included in the experiment. The one exception, it seems, is the feature type for the Extraversion item models, as most seem to favor the use of Term Occurrence, possibly due to its potential to aid in model generalization.

As seen in Table 2, the overall test F1 scores of the best item models for Extraversion fall between 0.3000 to 0.5000, with Item 31R achieving the highest test F1 score at 0.4334. Conversely, the weakest performing model belongs to Item 36, which has a test F1 score of approximately 0.3196. A comparison of these F1 scores with those obtained on the train-validation set suggests a possibility that the models overfitted on the training data.

Item-Level Results for Extraversion							
Item	**Min_df**	**Max_df**	**Feature Reduction**	**Algorithm**	**Feature**	**Train-Val F1**	**Test F1**
Item 1	0.1	0.9	PCA	LR	TO	1.0000	0.3450
Item 6R	0.05	0.7	CHI	XGB	TF-IDF	1.0000	0.3740
Item 11	0.05	0.9	CHI	LR	TO	1.0000	0.3311
Item 16	0.1	0.7	CHI	LR	TF-IDF	1.0000	0.3586
Item 21R	0.05	0.6	PCA	LR	TO	1.0000	0.3386
Item 26	0.1	0.6	CHI	XGB	TO	1.0000	0.3785
Item 31R	0.05	0.8	CHI	SVM	TO	0.9875	0.4334
Item 36	0.1	0.9	PCA	SVM	TO	0.9962	0.3196

Table 2: The performance and configurations of the best performing classification models per Extraversion item. Models were selected based on validation F1 score.

Compared to the results produced by the Extraversion item models, the range of values for the test F1 scores of the best performing Conscientiousness item models is generally broader, both on the lower and higher ends of the scale. Table 3 reveals

that the best performing item model for Conscientiousness produced a test F1 score of 0.5416, while the worst performing model had a test F1 score of 0.2426.

Item-Level Results for Conscientiousness							
Item	**Min_df**	**Max_df**	**Feature Reduction**	**Algorithm**	**Feature**	**Train-Val F1**	**Test F1**
Item 3	0.05	0.9	CHI	XGB	TO	0.7207	0.4574
Item 8R	0.05	0.9	CHI	XGB	TO	0.9902	0.5416
Item 13	0.1	0.6	CHI	XGB	TF-IDF	0.2761	0.2426
Item 18R	0.1	0.6	PCA	SVM	TO	0.8959	0.2534
Item 23R	0.1	0.6	PCA	LR	TO	1.0000	0.4373
Item 28	0.05	0.7	PCA	LR	TF-IDF	0.9680	0.4152
Item 33	0.1	0.7	CHI	LR	TF-IDF	1.0000	0.3534
Item 38	0.05	0.6	PCA	LR	TF-IDF	1.0000	0.2750
Item 43R	0.1	0.9	PCA	XGB	TF-IDF	1.0000	0.3921

Table 3: The performance and configurations of the best performing classification models per Conscientiousness item. Models were selected based on validation F1 score.

As evidenced by the side-by-side comparisons of the test F1 scores for the item models of both traits against the baseline majority classifiers in Figure 2 and Figure 3, it becomes apparent that all of the proposed item models consistently underperform. This disparity in classification performance may potentially be caused in part by the disproportionate number of samples for the majority class label of each questionnaire item. The degree to which this class imbalance exists can be seen from how most of the majority class classifiers exhibited test F1 scores above 0.5.

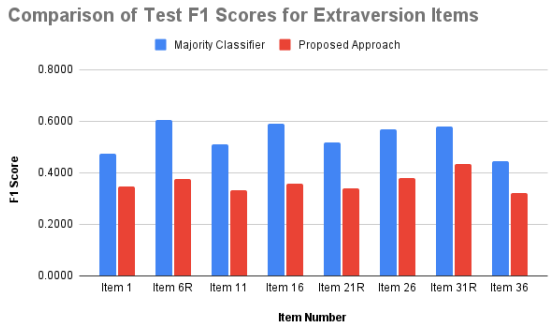

Figure 2: A comparison of test F1 scores between baseline majority class classifiers and the best item models for Extraversion

4.1.2 Trait-Level Evaluation Results

Table 4 and Table 5 present the trait-level results comparing the aggregated predictions against the ground-truth personality trait scores for Extraversion and Conscientiousness, respectively. The results of the proposed approach are also compared to that of 3 different baselines, particularly, a mean

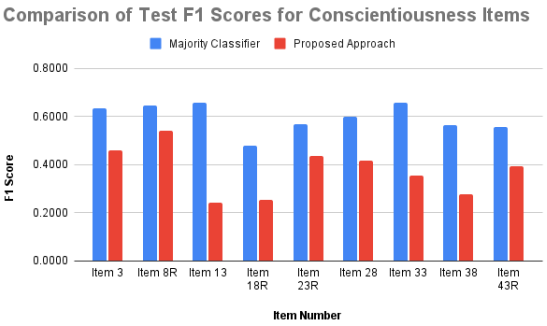

Figure 3: A comparison of test F1 scores between baseline majority class classifiers and the best item models for Conscientiousness

regressor, a linear regression model, and a multi-layer perceptron regressor.

For the Extraversion trait, Table 4 shows that the proposed approach produced the best results, with the lowest test RMSE of approximately 0.6714, and the highest R^2 score of around 0.1240. However, when taking these values on their own, the R^2 value can be considered relatively low. This may suggest that the variance in the Extraversion trait scores is still not explained very well by the predictor using the given features.

Trait-Level Results for Extraversion				
Model	**Train-Val RMSE**	**Train-Val R^2**	**Test RMSE**	**Test R^2**
Mean Regressor	0.7499	0.0000	0.7175	-0.0003
Linear Regression	0.2650	0.8751	0.6747	0.1154
MLP Regressor	0.7500	-0.0004	0.7174	0.0000
Proposed Approach	0.0382	0.9974	0.6714	0.1240

Table 4: The trait-level results for Extraversion using the proposed approach as well as baseline models

Compared to Extraversion, the results produced by all of the models for the Conscientiousness trait are considerably worse. The proposed approach performs the worst with a test RMSE of 0.6760 and a test R^2 value of -0.2273, while the linear regression model performs the best with a test RMSE of 0.6010 and a test R^2 value of 0.0298. These results show that the initial item-based approach for Conscientiousness leaves much to be improved, as direct trait modeling still works better in predicting overall trait scores.

Interestingly, despite generally having better test RMSE scores, the Conscientiousness models appear to have poorer test R^2 scores across the board, which may suggest that with the given feature set, Conscientiousness trait scores are more challenging to predict compared to Extraversion.

Trait-Level Results for Conscientiousness				
Model	**Train-Val RMSE**	**Train-Val R^2**	**Test RMSE**	**Test R^2**
Mean Regressor	0.6108	0.0000	0.6105	-0.0010
Linear Regression	0.2499	0.8326	0.6010	0.0298
MLP Regressor	0.6144	-0.0120	0.6162	-0.0199
Proposed Approach	0.2033	0.8892	0.6760	-0.2273

Table 5: The trait-level results for Conscientiousness using the proposed approach as well as baseline models

4.2 Evaluation of Proposed Approach with Hierarchical Classification

Another experiment was done with the proposed approach, particularly the integration of a hierarchical classification scheme. As mentioned previously, hierarchical classification attempts to classify the data into broader classes (e.g. Class *1-2*, Class *4-5*) on the first classification layer, then classifies the data in a more specific class (e.g. Class *1*, Class *2*) on the second layer. This experiment was done to attempt to classify data points better by grouping classes that were closer to each other first and then differentiating them later on.

Extraversion							
Train-Val RMSE		0.2097		**Test RMSE**		0.7126	
Train-Val R^2		0.9218		**Test R^2**		0.0131	
Item	**Val F1 (Broad)**	**Val F1 (Specific)**	**Val F1 (Binary 1)**	**Val F1 (Binary 2)**	**Val F1 (Binary 3)**	**Train-Val F1**	**Test F1**
Item 1	0.5685	0.3502	0.6520	1.0000	0.5399	0.9519	0.3892
Item 6R	0.5359	0.3990	0.7825	1.0000	0.6313	0.9822	0.3138
Item 11	0.5220	0.3431	0.6040	1.0000	0.5613	1.0000	0.3905
Item 16	0.5560	0.3350	0.7307	1.0000	0.5815	0.7085	0.3205
Item 21R	0.5567	0.3643	0.7508	1.0000	0.5445	0.7209	0.2999
Item 26	0.4956	0.3913	0.6427	1.0000	0.7402	1.0000	0.3230
Item 31R	0.6579	0.4650	0.6269	1.0000	0.5986	0.9412	0.4284
Item 36	0.5317	0.3236	0.5018	1.0000	0.5692	0.6096	0.2848

Table 6: Extraversion Results with Hierarchical Classification

Conscientiousness							
Train-Val RMSE		0.2015		**Test RMSE**		0.6270	
Train-Val R^2		0.8911		**Test R^2**		-0.0560	
Item	**Val F1 (Broad)**	**Val F1 (Specific)**	**Val F1 (Binary 1)**	**Val F1 (Binary 2)**	**Val F1 (Binary 3)**	**Train-Val F1**	**Test F1**
Item 3	0.6373	0.6281	0.8617	1.0000	0.5824	0.8263	0.5742
Item 8R	0.6366	0.5419	0.5513	1.0000	0.6123	0.8297	0.5078
Item 13	0.7167	0.4909	0.8526	1.0000	0.5480	1.0000	0.4366
Item 18R	0.5135	0.4036	0.4775	1.0000	0.5090	0.9859	0.3380
Item 23R	0.7327	0.4451	0.7957	1.0000	0.5514	0.9712	0.4388
Item 28	0.6344	0.5052	1.0000	1.0000	0.5099	0.8611	0.4555
Item 33	0.5780	0.4435	0.9033	1.0000	0.6314	0.9925	0.3608
Item 38	0.5016	0.4434	0.7528	1.0000	0.6323	0.6317	0.3406
Item 43R	0.6583	0.4156	0.7148	1.0000	0.5480	0.7604	0.5399

Table 7: Conscientiousness Results with Hierarchical Classification

Tables 6 and 7 show the results of the item models with hierarchical classification, along with the validation F1 scores for each layer for both *broad* and binary classification.

The *broad* F1 scores represent classification accuracy in the first layer of classes, specifically in Classes *1-2*, *3*, and *4-5*, respectively. These aforementioned scores for both traits show generally higher values, meaning that on the *broad* level of classification, the models are able to classify more accurately compared to previous scores.

The validation F1 scores labeled *specific*, on the other hand, are not as high as the *broad* F1 scores. The *specific* F1 scores pertains to the accuracy of classifying the data to the actual response prediction classes (i.e. Class *1, 2, 3, 4, 5*).

The validation F1 scores labeled *Binary* represent the accuracy of predicting the right binary class after the first classification layer has been done (i.e. Binary 1 - Class *1 and 2*, Binary 2 - Class *3*, Binary 3 - Class *4 and 5*). Although the F1 scores for each Binary are generally high, this only deals with classifying the data into one or two classes.

Trait-Level Results for Extraversion		
Version	**Test RMSE**	**Test R^2**
Original	**0.6714**	**0.1240**
Hierarchical Classification	0.7126	0.0131

Table 8: Extraversion Trait-Level Results for Original and Hierarchical Experiments

Trait-Level Results for Conscientiousness		
Version	**Test RMSE**	**Test R^2**
Original	0.6760	-0.2273
Hierarchical Classification	**0.6270**	**-0.0560**

Table 9: Conscientiousness Trait-Level Results for Original and Hierarchical Experiments

Overall, observing the results found in Table 7, the validation scores look somewhat promising, with predictions that look more accurate after passing through two layers as opposed to the original proposed approach for Conscientiousness. It can be observed that the approach with hierarchical classification is a potentially viable method in classifying as it produced more accurate results at the item-level. This difference in metric scores may likely be attributed to the step-by-step process of classifying the data, where data is classified in a broader threshold of similar classes and then further differentiated on the second level. By breaking the modeling process into two phases, this approach

better accounted for the inherent ordinality of the data and showed that the models still had potential for distinguishing between high and low responses, which was particularly beneficial for the Conscientiousness trait. However, despite an improved item-level performance, the trait-level results still much to be desired. That said, it is still a step in the right direction to be able to classify the item-level data more accurately at least at the *broad* level.

5 Conclusion

Following initial item-level and trait-level evaluations of the approach, it was inferred that due to data imbalance, substantial results became hard to derive because models performed poorly in terms of item-level prediction, and were even outperformed by baseline classifiers and regression models. In hopes of addressing this issue, a hierarchical classification approach was integrated, which involved breaking down the modeling process into two phases. Implementing this method showed a somewhat distinct advantage, most notably for the Conscientiousness trait. However, while the hierarchical approach worked relatively better for Conscientiousness, the original pipeline still reigned for Extraversion. This difference in model inclination may be attributed to the difference in feature significance between the two traits.

It is also worth noting that when compared against baseline models, the original pipeline still performed best for Extraversion, whereas the baselines performed better for Conscientiousness even with the slight improvement provided by the hierarchical approach. This supports the deduction that Conscientiousness items responses may be harder to predict, particularly with the given data.

With these results, it is evident that this particular field of APR study, especially in a Filipino context, leaves much room for pondering and experimentation. Some models indeed showed promise, but even the so-called best performing models have very low test metric scores. The overall results of this study signify that more tuning for both data and models needs to be done for this item-based approach to manifest improvements and become a framework that can prove beneficial to APR.

6 Recommendation

Future works that will choose to build up on the results from this study are encouraged to focus more on the best performing approaches for each trait. They can delve into more experimentations that aim to determine how the data qualitatively correlates to model performance, and what can be changed during preprocessing, extraction, and reduction in order for models to learn better from them and attain the most optimal performance results. Another angle of interest is examining trait-level result correlations with feature tokens, as this may help in identifying trends or patterns in terms of how each trait's best performing approach assigns weights or significance to certain terms or phrases, especially considering the mix of English and Filipino linguistic nuances.

At a more general level, future studies may opt to focus on a wider scope. Recommendations include exploring multimodal approaches that make use of images alongside textual data, testing the item-based approach on a high-resource language like English to more accurately assess the impact of data quantity, and investigating methodologies on how to properly structure social media data.

Future works may also address the identified issues from the results of the study, mainly data imbalance leading to model overfitting, hyperparameter limitations, and data quality or weight assignments on features. This can be done by increasing hyperparameter search space and number of iterations for the models, as well as attempting to experiment only with the unigram data instead of including bigrams.

The potential of the hierarchical approach can also be expounded upon; with proper data balancing methods and the right set of configurations, this approach may prove to be integral and beneficial to the overall pipeline.

Other recommendations include exploring other feature extraction and reduction techniques, as well as utilizing the remaining three traits of the Big Five (Openness, Agreeableness, and Neuroticism) to determine if the proposed approach could work equally or better as compared to its Extraversion and Conscientiousness results. Future works are also recommended to test the proposed approach against diverse datasets and different social media platforms and contexts in order to have a better benchmark for performance and generalizability.

References

Alexander H. II Agno, Jesah R. Gano, and Claude Kristoffer Sedillo. 2019. Instagram vs Twitter: Analyzing the manifestation of personality through the writing style of Filipino SNS users. Bachelor's thesis, De La Salle University.

American Psychological Association. Personality.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python. O'Reilly Media Inc*.

Ronn Christian Chua Chiaco, Howard Montecillo, Ronell John Roxas, and Bryan Ethan Tio. 2022. Application of word embeddings on automatic personality recognition using Filipino Twitter data. Bachelor's thesis, De La Salle University.

Andrew Marges. 2019. pinoy_tweetokenize.

Sumiya Mushtaq and Neerendra Kumar. 2022. Text-based automatic personality recognition: Recent developments. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, pages 537–549. Springer.

Edward Tighe, Luigi Acorda, Alexander Ii Agno, Jesah Gano, Timothy Go, Gabriel Santiago, and Claude Sedillo. 2022. Collection methods and data characteristics of the PagkataoKo dataset. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 513–524, Manila, Philippines. Association for Computational Linguistics.

Edward Tighe, Oya Aran, and Charibeth Cheng. 2020. Exploring neural network approaches in automatic personality recognition of Filipino Twitter users. In *Proceedings of the 20th Philippine Computing Science Congress*, pages 137–145.

Edward Tighe and Charibeth Cheng. 2018. Modeling personality traits of Filipino Twitter users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 112–122, New Orleans, Louisiana, USA. Association for Computational Linguistics.