Collection Methods and Data Characteristics of the PagkataoKo Dataset

Edward Tighe, Luigi Acorda, Alexander Agno II, Jesah Gano,
Timothy Go, Gabriel Santiago, Claude Sedillo
De La Salle University, Manila, Philippines
{edward.tighe, luigi_acorda, alexander_agno, jesah_gano,
timothy_go, gabriel_santiago, claude_sedillo}@dlsu.edu.ph

Abstract

We present the PagkataoKo Dataset, a new dataset for Filipino Automatic Personality Recognition (APR) containing demographic, personality trait, and social media data sourced from 3,128 Filipino Instagram and/or Twitter users. As APR is focused on processing an individual's observable actions, we improve upon the previous Filipino APR dataset by collecting multimodal data, as well as similar data expressed in different environments. In our paper, we describe our collection methodology and detail the general characteristics of the dataset. We also report the language characteristics of the posts and highlight the presence of multiple languages (e.g. English, Tagalog) and code-switching – two aspects that make Filipino APR difficult. Lastly, we discuss how our dataset provides future work with multiple options to explore in order to navigate around the complexities found in Filipinos' language usage.

1 Introduction

Automatic Personality Recognition (APR) is a computing task that focuses on personality traits (i.e. individual differences) and their externalizations (Vinciarelli and Mohammadi, 2014). The task is rooted in the idea that traits influence a person's interactions in different environments (Larsen and Buss, 2008); hence, a person's observable actions contain traces of their traits, such as how it is believed that important personality characteristics and individual differences are encoded into one's language (i.e. the lexical hypothesis) (Goldberg, 1981; Tausczik and Pennebaker, 2010). As a computing task, APR, therefore, involves the collection and processing of these observable actions and analysis (e.g. descriptive, causal, predictive) of any traces of personality left behind.

APR has received much attention over the past two decades – leading to the exploration of many different types of mediums (e.g. text, image, audio) in which personality may have been expressed upon. Early work in APR mostly focused on studying language usage – with data coming from conversation recordings (Mehl et al., 2006; Mairesse et al., 2007), emails (Gill and Oberlander, 2002), and essays (Pennebaker and King, 1999; Argamon et al., 2005; Mairesse et al., 2007). These early studies did not produce high-performing predictive models nor did they have a high volume of data to validate results with but were at least able to show that indicators of personality can be found in one's language. APR studies then branched out and explored other sources of observable actions with a vast number of studies gravitating towards social media platforms, such as (but not limited to) Facebook (Golbeck et al., 2011b; Gosling et al., 2011; Wald et al., 2012; Markovikj et al., 2013; Schwartz et al., 2013; Kosinski et al., 2014; Park et al., 2015; Segalin et al., 2017), Twitter (Golbeck et al., 2011a; Quercia et al., 2011; Rangel Pardo et al., 2015; Liu et al., 2016; Skowron et al., 2016; Ong et al., 2017; Samani et al., 2018), Instagram (Ferwerda et al., 2015; Skowron et al., 2016; Lay and Ferwerda, 2018), Sina Weibo (Gao et al., 2013; Guntuku et al., 2015), Flickr (Samani et al., 2018), and general blogs (Nowson and Oberlander, 2006; Nowson and Oberlander, 2007; Gill et al., 2009; Yarkoni, 2010). Social media platforms have since become a perfect source of data for APR given the many different ways a person might interact within the online environment.

While more recent APR studies have gravitated towards exploring neural network based methods (Mehta et al., 2019), an area of opportunity within APR that lacks attention is in studying social media data from Filipinos. Kemp (2021) noted that individuals from the Philippines spent the most time on social media – clocking in a little over four hours a day on social media versus the global average of roughly 2.5 hours. This high usage is an indicator that there is a high volume of observable actions that can be collected from online Filipino users. However, despite this upside, social media data from Filipinos can generally be considered hard to deal with when coming from a natural language processing perspective. Filipinos are known to speak multiple languages (e.g. English, Filipino, Cebuano, and a number of other Philippine languages) and code-switch between these languages (Caparas and Gustilo, 2017; Abastillas, 2018; Tighe and Cheng, 2018). A corpus containing these language characteristics – coupled with informal language usage usually found in social media and the low number of language resources available for Philippine language processing – presents quite a challenge when looking to extract useful linguistic information related to personality.

Currently, only the dataset of Tighe and Cheng (2018) is suitable for Filipino APR. Tighe and Cheng (2018) produced a dataset containing text data from $610,448$ tweets of 250 Filipino Twitter users and were able to show that there were indeed some traces of Conscientiousness and Extraversion from term frequency inverse document frequency (TFIDF) values. However, a follow-up study by Tighe et al. (2020) showed that tuned multilayer perceptron (MLP) models trained on word embedding data (pre-trained and trained-over) did not learn at all and performed poorly when compared against MLPs using TFIDF. One reason for the poor performance can be attributed to the limited size of the data given their train-test split led to an even smaller amount of data for learning. It should be taken into consideration that the methods of Tighe et al. (2020) performed a limited analysis of the usage of word embeddings – implying that more detailed studies need to be crafted to gauge the usefulness of embedding-based approaches on Filipino text data. Nevertheless, the dataset's small size poses a limitation when applying certain computing methods (e.g. neural network approaches). In addition to the small data size, the dataset only contains text data from one platform. Related literature has shown that image data can also contain personality traces and aid in modeling personality (Liu et al., 2016; Segalin et al., 2017; Lay and Ferwerda, 2018). Also, a fusion of data – whether from different types of data (e.g. image + text + account) and/or different sources of data (e.g. Twitter + Instagram) – has produced better personality models against models using a single modality or sources (Skowron et al., 2016; Samani et al., 2018). The potential benefits of exploring different data modalities and sources is an aspect that the current Filipino APR dataset cannot provide to any future studies.

To address the need for a larger and more flexible data resource for Filipino APR, we created the PagkataoKo Dataset. The dataset contains personality, demographic, text, image, and account data from $3,128$ Filipino Instagram and/or Twitter users. Participants were administered the Big Five Inventory (BFI-44) to assess their trait scores and were given the choice to share access to one or both of their social media accounts' data. The novelty of our dataset lies in that the data is multimodal – capturing more observable actions than the previous dataset – and is sourced from two different platforms – capturing actions expressed in two different environments. In our paper, we discuss the methodology for collecting the data and detail general data characteristics, as well as temporal and language characteristics of the collected posts. We also discuss design considerations based on the characteristics of the data.

2 Collection Methodology

We extend the methodology of Tighe and Cheng (2018) by also collecting image and account-related data, aside from text data. We also gave participants an option to share access to multiple social media accounts, instead of just one. We selected Twitter and Instagram as the sources of observable actions because of how each platform encourages different behavior – with Twitter being micro-blogging oriented and Instagram being media-sharing oriented.

2.1 Personality Trait Representation

To assess trait scores, we used the Big Five Inventory (BFI-44), a 44-item self-reported question-

naire that measures the five dimensions of the Big Five (John et al., 1991; John et al., 2008). These five dimensions – sometimes collectively referred to as OCEAN – include Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each dimension or broad trait is associated to about 8 to 10 Likert scale questions ranging from 1 (strongly disagree) to 5 (strongly agree). Computation of a trait's score involves reversing specific item scores and then computing the average score of a trait's associated items – resulting in a value that ranges from 1.0 to 5.0. We selected the BFI for our study mainly because numerous other APR studies utilized a questionnaire that measured the Big Five. By standardizing the representation of personality to the Big Five, any results from our data would be easier to compare with other studies.

2.2 Collection Tool

We created a web application that facilitated the collection of demographic data, social media data, and personality trait scores of our participants. The application first presented individuals with the initial collection formalities (e.g. consent form, directions). After agreeing, participants would be asked to grant our application permission to read their social media data from either or both Twitter and/or Instagram. The application would then collect social media data by interfacing with each platform's respective APIs. Participants were then presented a demographic questionnaire followed by a personality test. The demographic questions were presented to acquire the information we need to describe and eventually filter the participants. Only individuals 18 years old or above were permitted to proceed with the collection. As for the personality test, we implemented an online version of the BFI-44. The application collected answers to each of the 44 items, as well as the computed Big Five scores. Participants were then shown their personality scores compared against the trait scores presented in Tighe and Cheng (2018).

For Twitter, our application would utilize the Twitter API v1 to collect an individual's most recent $3,200$ tweets – a limitation of the API. The application then discarded all retweets as we were only concerned with tweets that were written by the user; however, quoted tweets were retained since these were written by the user. Aside from tweet-level data, the application also collected links to each participant's profile picture, as well as other account-related data (e.g. # of followers / following). As for Instagram, our application would utilize the Instagram Legacy API to collect as many of a user's posts as the API would return. For each post, our application collected both the caption and a link to the post's image. In the case that a post had multiple photos, only the main photo was collected. As for posts that contained videos, our application discarded them as video data was not initially factored into the research design. In addition to the post data, our application collected a link to the profile picture and other account-related data to their Instagram account.

Links to images were collected instead of downloading the image itself to lessen the strain on the web application especially as Instagram allowed for the collection of all a user's posts. Our initial plan was to wait until the collection was over before downloading all images; however, this was a costly decision as we did not realize that the URLs from both platforms changed over time. The changing of an image's URL could be due to a user uploading a new profile picture after their participation or the platform periodically refreshing links. Because of this, we lost access to image data from around 350 Instagram and 804 Twitter users. Once we discovered this issue, we opted to continue retrieving image links but we would download images at the end of each day until the end of the collection.

2.3 Time Frame and Sampling Methods

Our collection started on the 1st week of Jun. 2019 and ended on the 2nd week of Feb. 2020. We utilized a mixed sampling approach centered on volunteer sampling. We approached individuals with idea that participating and disseminating word of the research would generally be beneficial to the research project. While not always highlighted in the invites/advertisements, we would compute and present the results of the personality test to each participant who finished the entire collection – a factor that proved useful in encouraging individuals to share the collection with their own networks as we did not offer incentives.

From the start of the collection until the 1st week of Oct. 2019, we performed convenience sampling by posting information about the recruitment within our immediate networks. Postings

were made on different online platforms and by reaching out to individuals in person. We also achieved a minor snowball effect as our networks helped propagate the recruitment to their own respective networks. These initial efforts resulted in the collection of data from 362 participants. After which, we started online advertisement campaigns to reach out to Filipinos outside of our immediate network. We create ad campaigns that targeted individuals from the Philippines and had these ads run on Facebook, Instagram, and Twitter. We ran ad campaigns intermittently across the 2nd to 4th week of October 2019. This resulted in an additional $1,393$ individuals. The initial success of recruiting participants through ads led us to run additional ads throughout the 1st week of Dec. 2019 and the 2nd week of Feb. 2020. By the end of our collection, we were able to recruit and collect data from $3,186$ participants.

2.4 Participant Filtering

Participants were included in the final dataset if they signified that their nationality was Filipino. We also included individuals who were mixed Filipinos or individuals who stated they were Filipino and had one or more additional nationalities. After filtering, we were left with a total of $3,128$ individuals. We discarded data from those who did not qualify based on this filter.

2.5 Ethical Clearance

Our methods were reviewed and given clearance by the Research Ethics Office of De La Salle University, Philippines. Individuals voluntarily gave electronic consent to participate in our study and their social media data was collected in accordance with the developer policies of Twitter and Instagram.

3 The PagkataoKo Dataset

The PagkataoKo[1] Dataset is composed of demographics, personality trait scores, user-generated data (e.g. post, tweet, profile pictures), and other account-related metadata from $3,128$ Filipino Instagram and/or Twitter users.

3.1 Subgroups of Participants

We organize participants into four subsets based on the social media account(s) they provided as follows:

- **I** – All participants with Instagram accounts,
- **T** – All participants with Twitter accounts,
- **I**$\cup$**T** – All participants (i.e. the union between **I** and **T** or the universal set of participants), and
- **I**$\cap$**T** – All participants with both Instagram and Twitter accounts (i.e. the intersection between **I** and **T**).

3.2 Participant Demographics

We report the participant demographics across all four subsets in Table 1. We note that among all participants, 17.1% gave access to both their Twitter and Instagram accounts – leaving a majority (82.9%) of the participants unique to one of the two social media platforms. In terms of age, 80.6%-84.9% of the participants across all subsets were between 18-23. The Twitter subset has a slightly younger age distribution in comparison to the Instagram subset. As for sex, 75.0%-78.0% of the participants across all subsets are female. While most of the statistics on sex are relatively stable across subsets, we note a slightly higher percentage of females on Instagram and that there were fewer people who decline to disclose their sex if they granted access to both of their social media accounts. Lastly, only 0.8%-1.3% of participants across all subsets declared their nationality as Filipino and one or more nationalities.

3.3 Personality Trait Score

We report descriptive statistics of the personality trait scores of all participants in Table 2 and visualize the score distributions in Figure 1. All trait score distributions are unimodal and approximately symmetric with skewness values > -0.40 and < 0.04. We also report the Cronbach's alpha values for each trait in order to how consistent our participants were answering the Big Five Inventory (i.e. internal consistency). The alpha values indicate good reliability for Extraversion and Neuroticism, acceptable reliability for Conscientiousness and Agreeableness, and questionable reliability for Openness. As for correlation coefficients, most values showed negligible correlation. When coefficients weren't negligible, values showed low correlations, such as with Agreeableness and Conscientiousness, Neuroticism and Conscientiousness, Neuroticism and Extraversion, and Neuroticism and Agreeableness.

[1] *Pagkatao* is Filipino for *personality*, while *ko* refers to one's self or the word *my*. Hence, *PagkataoKo* is a play on the popular dataset, *myPersonality*.

Demographics	I∪T	I	T	I∩T
Count	3,128	1,380	2,283	535
Age				
Mean	21.2	21.4	21.0	21.2
SD	3.9	3.7	3.9	3.5
Age range				
18-20	53.9%	49.3%	55.9%	50.1%
21-23	29.3%	31.3%	29.0%	33.3%
24-26	9.3%	10.7%	8.5%	9.5%
≥27	7.5%	8.8%	6.6%	7.1%
Sex				
Male	21.0%	20.0%	22.0%	22.6%
Female	76.1%	78.0%	75.0%	76.3%
Intersex	0.5%	0.3%	0.6%	0.4%
Declined[1]	2.4%	1.7%	2.5%	0.8%
Nationality				
Filipino	99.2%	99.1%	99.1%	98.7%
Mixed[2]	0.8%	0.9%	0.9%	1.3%

[1] Those who declined to disclose their biological sex.
[2] Those who were Filipinos and had one or more other nationalities.

Table 1: The demographic statistics across all four subsets of participants: the universal set of all participants (**I**∪**T**), the set of participants with Instagram accounts (**I**), the set of participants with Twitter accounts (**T**), and the set of participants with both Instagram and Twitter accounts (**I**∩**T**).

3.4 General Data Characteristics

We summarize general statistics of user-generated data (e.g. posts, profile pictures) and account-related metadata (e.g. number of followers / following) across each of subgroup of participants in Table 3.

For the Instagram subset, the distribution of total collected posts per user is positively skewed with 72% of the subset having a total post count less than the mean (i.e. < 149.55) and 89% of the subset having fewer than one standard deviation plus the mean (i.e. < 377.03). There are 71 Instagram users with 0 posts and 75 Instagram users with total posts more than two standard deviations plus the mean (i.e. > 597.55). As for the posts themselves, only 83% of all collected posts have an image and 91% of the posts have a caption. Ideally, each post should have an associated image as one cannot post on Instagram without an image; however, we incurred a 17% loss in collectable image data due to the image link download issue discussed in Section 2.2. This issue also explains the missing 350 profile pictures. As for the missing 9% of posts without captions, we note that Instagram treats captions as an optional field when posting; hence, these posts really did not contain any text data. As for account-related data, we were only able to collect the total account recorded posts and the user's following count. Instagram's Legacy API was in the process of depreciating at the time of collection and did not allow for other metrics to be collected.

As for the Twitter subset, the distribution of total collected tweets per user is bimodal with roughly 49% of the subset falling between the 2200-3100 tweet count range and roughly 22% of the subset falling between the 0-600 range. There are 28 users with 0 tweets and 32 users with 3100-3200 tweets. Unlike posts on Instagram, all tweets contain text data. However, similar to the case with the Instagram subset, 804 users are missing a profile picture due to the image link download issue. As for account-related data, we were able to collect and report the total account recorded posts, following count, followers count, and favorite count.

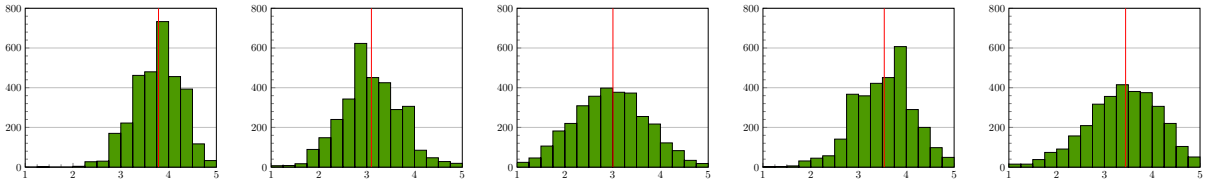
Of the total $3,128$ participants, only 17% of the participants granted access to both their Instagram and Twitter accounts. This subset retained 39% and 25% of the Instagram and Twitter posts, respectively. Despite the significant reduction in size, the subset's statistics are comparable to each platform's own subsets when factoring in that there are fewer outliers.

3.5 Temporal Characteristics of Post Data

We report the distribution of collected posts/tweets created per year for both social media platforms in Figure 2. As we collected as many posts/tweets as allowed, the Twitter data contains tweets made within an eleven-year period (2009 to 2020), while the Instagram data contains posts made within a ten-year period (2010 to 2020). The distribution of Instagram posts is approximately symmetric and peaks in 2016 ($n = 37,717$), while for Twitter, the distribution of tweets is left-skewed and peaks in 2019 ($n = 1,560,201$). Both distributions show a sharp drop off in 2020 due to the collection ending in Feb 2020.

Traits	Mean	SD	Alpha	Pearson Correlation Coefficient				
				O	C	E	A	N
O	3.7893	0.4837	0.6780	1.0000				
C	3.0984	0.6130	0.7844	0.1700	1.0000			
E	3.0066	0.7550	0.8249	0.1677	0.1379	1.0000		
A	3.5383	0.6075	0.7269	0.1309	0.2916	0.1907	1.0000	
N	3.4427	0.7462	0.8102	(0.1126)	(0.3695)	(0.2327)	(0.3016)	1.0000

Table 2: The mean, standard deviation, Cronbach's alpha, and Pearson correlation coefficients of each personality trait – **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism – with respect to all $3,128$ participants.



(a) Openness (b) Conscientiousness (c) Extraversion (d) Agreeableness (e) Neuroticism

Figure 1: Personality trait score distribution of all $3,128$ participants for each of the Big Five. The x-axis measures the raw trait scores, while the y-axis measures the number of individuals per bin. The red line represents the mean.

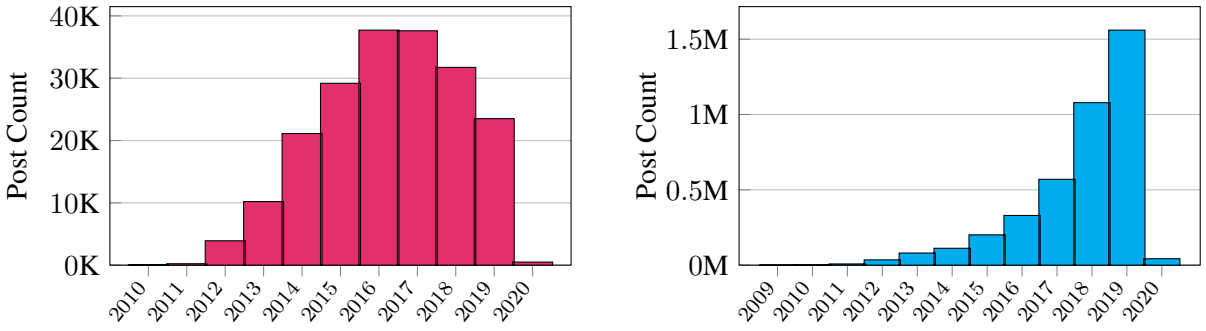


Figure 2: Histograms of dataset's posts made per year from Instagram (left) and Twitter (right)

3.6 Language Characteristics of Post Data

To highlight the multilingual aspect of our dataset, we observe language information at two levels: the entire post (*document-level*) and the individual words found in each post (*word-level*). In its raw form, the dataset only contains document-level language information for the Twitter data as this information was provided when collecting tweets via Twitter's API. To observe language information across the entire dataset, we process the text data to extract language information. We limit our observation to the following eight Philippine languages: *Bikol*, *Chavacano*, *Cebuano*, *English*, *Iloko*, *Kapampangan*, *Tagalog*, and *Waray*. While there are more Philippine languages than listed, these eight languages have some support from available language resources.

Data Pre-processing. For each individual post, we remove tokens with questionable or no language information, such as emojis, hashtags, usernames, URLs, and punctuation. We also lowercase all characters to reduce typical noise found in social media data. After the initial cleaning, we discarded all documents that were empty strings or contained only white space characters. This left us with $168,723$ posts from Instagram and $3,979,010$ tweets from Twitter. We would like to note that pre-processing was done solely to extract

Platform	Data	Statistics	**I** $n = 1,380$	**T** $n = 2,283$	**I∩T** $n = 535$
Instagram	Collected Posts	Total	195,757	-	76,697
		Average	141.85	-	143.36
		SD	224.00	-	232.65
		Min / Max	0 / 1,902	-	0 / 1,902
		# w/ Caption	178,650	-	72,266
		# w/ Image	162,500	-	60,070
	Acct. Recorded Posts	Average	146.60	-	144.59
		SD	267.99	-	232.99
		Min / Max	0 / 5,680	-	0 / 1,902
	Following Count	Average	470.39	-	449.47
		SD	442.85	-	373.69
		Min / Max	0 / 6,338	-	0 / 4,427
	Profile Pictures	Total	1,030	-	385
Twitter	Collected Tweets	Total	-	4,018,628	1,033,089
		Average	-	1,760.24	1,931.01
		SD	-	1,016.71	987.62
		Min / Max	-	0 / 3,185	0 / 3,185
	Acct. Recorded Tweets	Average	-	8,003.57	9558.45
		SD	-	12,260.68	13,338.02
		Min / Max	-	0 / 162,738	0 / 103,381
	Following Count	Average	-	289.48	321.69
		SD	-	335.83	318.62
		Min / Max	-	0 / 7,079	0 / 3,501
	Followers Count	Average	-	333.93	320.58
		SD	-	1,023.46	393.51
		Min / Max	-	0 / 29,328	0 / 3,433
	Favorites Count	Average	-	8,517.47	8,886.03
		SD	-	12,252.05	10,734.24
		Min / Max	-	0 / 193,119	0 / 91,012
	Profile Pictures	Total	-	1,479	333

Table 3: Data characteristics of user-generated and account-related data across three participant subsets: all participants with Instagram accounts (**I**), all participants with Twitter accounts (**T**), and all participants with both Instagram and Twitter accounts (**I∩T**). The size of each subset (n) is also indicated.

language characteristics and that the raw data still contains this information.

Document-level Language Information. We extract document-level language tags using two language identifiers: *Polyglot* (Al-Rfou et al., 2013) and *FastText* (Joulin et al., 2016b; Joulin et al., 2016a). FastText supports all languages within our scope but forces a language tag even when it's uncertain. On the other hand, Polyglot covers all languages except Bikol, Chavacano, and Iloko and includes an *Undefined* tag when it lacks in confidence. Using the identifiers' output, we label a document based on the language tag with the highest confidence. When a language tag is out-

side of our scope, we assign the *Others* tag. Specific to Twitter data, we utilize the language metadata tag returned by Twitter's API, referred to as *Twitter Tag*, as a third language tag. The Twitter Tag only covers English and Tagalog and includes an undefined tag. After language extraction, we measure agreement among the assessed document-level tags through a *Majority Vote* (i.e. agreement $> 50\%$). If there is agreement among the tags, we assign the language tag. Otherwise, we assign a *Conflict* tag.

We summarize the results of our document-level language extraction in Table 4. The results show that English and Tagalog are the top two languages found on both platforms. English has a significantly higher usage on Instagram with a majority vote at almost 75.6% compared against the 47.9% majority vote on Twitter, while Tagalog has a significantly higher usage on Twitter (majority vote at 32.7%) versus that on Instagram (majority vote at 4.3%). As for the other Philippine languages, we note they occur significantly less often with Cebuano and Waray coming in third and fourth most used on both platforms. While this might indeed be true for the dataset, we take into consideration that different language identifiers do not align with each other – causing the majority vote to be low or result in zero. We also note that the extracted tags did not reach an agreement for 19.4% of the Instagram data and 17.0% of the Twitter data. Documents that fall under this category typically contain textspeak or some form of code switching between English and Tagalog.

We note that the Polyglot and FastText numbers are relatively similar despite the differences in their respective outputs; however, one glaring issue we would like to highlight is how Twitter Tag vastly differs from the two language identifiers. Twitter Tag indicates that there are 9.9% more Tagalog tweets than English, while numbers from Polyglot and FastText indicate that there are around two times more English than Tagalog tweets. Unfortunately, there are no specifics on how Twitter's language identifier works, but we speculate that Twitter uses different preprocessing techniques from our methods or uses information only accessible to Twitter.

To gain a better understanding of the issue, we performed a brief inspection of the Twitter documents. When agreement was reached, we note that 75% of the English tweets and 53% of the Tagalog tweets had perfect agreement across the three tags. These documents have a dominant language with respect to both the grammar and vocabulary. When there is disagreement between Twitter Tag and the other language identifiers, we note that it is rare ($< 0.05\%$ of the total tweets) for Twitter Tag to output English when Polyglot and FastText output Tagalog. On the other hand, almost 9% of the total tweets are labeled Tagalog by Twitter Tag when the the other two language identifiers agree on English. We observe that it is generally harder to determine these documents' language due to multiple factors, such as a balanced mix of words from both languages, multiple sentences following different grammar structures, and noise usually found in social media text. Based on our manual observation, we gained greater confidence in Twitter Tag particularly when it comes to the Tagalog labels. We also view Polyglot and FastText as sufficient off-the-shelf language identifiers but that they have a tendency to favor the English label. Hence, we caution interpreting the numbers too strictly and advise to keep in mind that the language characteristics of the data can be quite complex. Additionally, while these issues were solely observed on the Twitter data, we assert that the same issues with Polyglot and FastText may apply to the Instagram data but to a lesser extent.

Word-level Language Information. To observe word-level language information, we extracted the tokens and word types from our text data. We then compared how many tokens and vocabulary were found in a Philippine language word reference or dictionary. To serve as our reference, we used the words found in FastText's pre-trained word vectors (Grave et al., 2018) as there are resources for the languages within our scope except for Chavacano. We note that while FastText is a convenient resource, the word vectors' vocabularies are not unique from each other as they were trained on data from Wikipedia and CommonCrawl, which most likely included words from other languages.

We summarize the results of our word-level language information in Table 5. For the Twitter data, we note that the Tagalog word vectors provide the best coverage – providing vectors to 93.4% of our tokens, as well as 15.4% of the vocabulary. English comes in at a close second place covering 88.8% of the tokens and 15.1% of the

Language	Instagram ($n = 164,044$)			Twitter ($n = 3,870,153$)			
	PG	**FT**	**MV**	**PG**	**FT**	**TT**	**MV**
Bikol	-	0.00%	0.00%	-	0.01%	-	0.00%
Chavacano	-	0.01%	0.00%	-	0.01%	-	0.00%
Cebuano	0.48%	0.81%	0.12%	3.15%	2.99%	-	0.86%
English	82.56%	81.29%	75.58%	54.13%	54.15%	40.41%	47.94%
Iloko	-	0.04%	0.00%	-	0.22%	-	0.00%
Kapampangan	0.00%	0.00%	0.00%	0.00%	0.02%	-	0.00%
Tagalog	6.68%	5.88%	4.32%	26.89%	24.12%	50.26%	32.66%
Waray	0.40%	0.14%	0.01%	1.27%	0.77%	-	0.04%
Others	8.29%	11.82%	0.56%	13.96%	17.71%	6.26%	1.08%
Undefined	1.59%	-	0.00%	0.60%	-	3.06%	0.38%
Conflict	-	-	19.41%	-	-	-	17.01%

Table 4: The document-level language information of Instagram and Twitter documents. Language identifiers used were Ploylglot (**PG**) and FastText (**FT**). Twitter's language metadata tag, referred to as Twitter Tag (**TT**), was also reported. Tag agreement was measured using a Majority Vote (**MV**) approach. Blank marks ('-') indicate the language identifier does not support the category.

vocabulary. As for Instagram, we are less certain of the word vector that provides the best coverage as English word vectors cover the vocabulary the most at 48.0% (versus Tagalog's 46.2% coverage), while the Tagalog word vectors cover the tokens the most at 94.7% (versus English's 94.6% coverage). As for the other Philippine languages, we note that the Waray word vectors come in a definitive third place both in terms of tokens and vocabulary coverage across the two platforms. After Waray, the ranking starts to vary across platforms. However, one observation we would like to point out is that the Cebuano word vectors have a higher token coverage (72.1%) on Twitter in comparison to the numbers of Iloko (68.1%) and Kapampangan (71.3%) despite having the second lowest vocabulary coverage at 2.9%. We speculate that the word vectors of Iloko and Kapampangan may have a sizeable overlap with other languages and that the higher token coverage may be an indicator that the Cebuano word vectors are able to capture a number of Cebuano function words.

4 Discussion and Future Directions

The PagkataoKo Dataset is a novel dataset for Filipino Automatic Personality Recognition that contains demographics, personality trait scores, and social media data from $3,128$ Filipinos. The dataset is an improvement over the dataset of Tighe and Cheng (2018) having 12.5 times more participants and sourcing social media data from more than one platform. It has yet to be seen how much personality information is present in the dataset and how well personality models can compare against that of Tighe and Cheng (2018) and Tighe et al. (2020); however, solely based on the amount of data present, our current dataset provides a wider foundation to study how personality can manifest in the social media data of Filipinos – particularly as there are different forms of observable actions (e.g. text and image data from posts) and similar types of data expressed in different environments (e.g. language usage on Twitter and Instagram, profile picture usage on Twitter and Instagram).

While there is much potential in the dataset, there are a number of challenges that need to be carefully studied. First, descriptive statistics of the personality trait scores show that Openness has questionable reliability – raising the issue of whether or not the questionnaire is appropriately capturing the dimension. While collecting data using different personality instruments is indeed an option for future studies, we believe there may be additional insights that can be extracted by conducting APR by studying individual questionnaire item answers aside from the computed trait score. Second, the temporal characteristics of the posts show data spanning multiple years.

Platform	Token Count / Vocabulary Size		% Found in FastText Word Embeddings						
			BCL	**CEB**	**EN**	**ILO**	**PAM**	**TL**	**WAR**
Instagram	Tokens	1,786,309	74.8%	77.0%	94.6%	76.6%	80.5%	94.7%	82.0%
	Vocab	92,030	11.5%	11.9%	48.0%	13.3%	15.2%	46.2%	18.2%
Twitter	Tokens	33,451,199	68.1%	72.1%	88.8%	68.1%	71.3%	93.4%	75.0%
	Vocab	719,927	2.6%	2.9%	15.1%	3.2%	3.1%	15.4%	4.4%

Table 5: The percentage of tokens and vocabulary of each platform found in FastText's pre-trained word vectors across Bikol (**BCL**), Cebuano (**CEB**), English (**EN**), Iloko (**ILO**), Kapampangan (**PAM**), Tagalog (**TL**), and Waray (**WAR**). The total number of tokens and vocabulary size per platform are also indicated.

While personality traits are known to be relatively enduring across time (Larsen and Buss, 2008), traits aren't immune to change even through adulthood (Roberts and Mroczek, 2008). Additionally, Arnoux et al. (2017) was able to show that there is merit in exploring a shorter number of documents for personality prediction; however, their work mainly focused on English data. We speculate that more data might be needed to account for the noise brought about by code-switching found in our dataset. Hence, we encourage future work on APR to explore how the recency of one's posts might have an effect on APR prediction models. Third, the data characteristics show that there are a number of participants with either zero or a very low number of data points. Coupled with missing image data and, specific to Instagram, missing text data, future studies in Filipino APR would need to conduct experiments to find appropriate thresholds that determine when there's enough data to analyze one's personality and/or design a framework for APR that can handle missing data. Lastly, the language characteristics of the post data show that future studies working on the PagkataoKo dataset should primarily focus on extracting information from Tagalog and English text data because these languages were the most prevalent. While it would be of particular interest to study manifestations of personality in the other Philippine languages, the collection methods did not result in a sizeable amount of text data to study the other Philippine languages. Regardless, the nature of how Filipinos write on social media poses a serious challenge to text processing given multiple posts of a user could be in different languages or contain code-switching. We encourage future Filipino APR studies to focus on experimenting with methods to handle multilingual data – whether through combining language resources or by exploring language-independent approaches.

Acknowledgments

We would like to express our gratitude to the University Research Coordination Office of De La Salle University for supporting our research project (project number 52FU2TAY18-2TAY19). We would also like to thank the anonymous reviewers for spending their time to provide us with useful feedback.

References

Glenn Abastillas. 2018. You are what you tweet: A divergence in code-switching practices in cebuano and english speakers in philippines. In *Language and Literature in a Glocal World*, pages 77–97. Springer.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pages 1–16.

Pierre-Hadrien 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 International AAAI Conference on Web and Social Media*, volume 11.

Pilar Caparas and Leah Gustilo. 2017. Communicative aspects of multilingual code switching in computer-mediated communication. *Indonesian Journal of Applied Linguistics*, 7(2):349–359.

Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, pages 7–10.

Rui Gao, Bibo Hao, Shuotian Bai, Lin Li, Ang Li, and Tingshao Zhu. 2013. Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 355–358.

Alastair 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.

Alastair Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? personality, topic and motivation in blogs. In *Third International AAAI Conference on Weblogs and Social Media*.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011a. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 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.

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

Samuel Gosling, Adam Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Sharath Chandra Guntuku, Lin Qiu, Sujoy Roy, Weisi Lin, and Vinit Jakhetiya. 2015. Do others perceive you as you want them to? modeling personality based on selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 21–26.

Oliver John, Eileen Donahue, and Robert Kentle. 1991. The big five inventory–versions 4a and 54.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Simon Kemp. 2021. Digital 2021: Global overview report. *DataReportal*.

Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. 2014. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95(3):357–380.

Randy Larsen and David Buss. 2008. *Personality: Domains of Knowledge About Human Nature*. Boston: McGraw-Hill.

Alixe Lay and Bruce Ferwerda. 2018. Predicting users' personality based on their 'liked' images on instagram. In *The 23rd International on Intelligent User Interfaces, March 7-11, 2018*. CEUR-WS.

Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Tenth International AAAI Conference on Web and Social Media*.

François Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David Stillwell. 2013. Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

Scott Nowson and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167. Palo Alto, CA.

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

Veronica Ong, Anneke DS Rahmanto, Derwin Suhartono, Aryo E Nugroho, Esther W Andangsari, Muhamad N Suprayogi, et al. 2017. Personality prediction based on twitter information in bahasa indonesia. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 367–372. IEEE.

Gregory Park, H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934.

James Pennebaker and Laura King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 180–185. IEEE.

Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Brent Roberts and Daniel Mroczek. 2008. Personality trait change in adulthood. *Current Directions in Psychological Science*, 17(1):31–35.

Zahra Riahi Samani, Sharath Chandra Guntuku, Mohsen Ebrahimi Moghaddam, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2018. Cross-platform and cross-interaction study of user personality based on images on twitter and flickr. *PloS one*, 13(7).

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie 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).

Cristina Segalin, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri. 2017. What your facebook profile picture reveals about your personality. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 460–468.

Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing social media cues: Personality prediction from twitter and instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 107–108. International World Wide Web Conferences Steering Committee.

Yla Tausczik and James Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

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.

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.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Randall Wald, Taghi Khoshgoftaar, and Chris Sumner. 2012. Machine prediction of personality from facebook profiles. In *2012 IEEE 13th International Conference on Information Reuse & Integration*, pages 109–115. IEEE.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.