PACIS 2017 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 7-19-2017

Age and Gender Profiling of Social Media Accounts

Jan Kristoffer Y. Cheng
De La Salle University, jan_kristoffer_cheng@dlsu.edu.ph

Avril Ranezca L. Fernandez
De La Salle University, avril_fernandez@dlsu.edu.ph

Rissa Grace Marie M. Quindoza
De La Salle University, rissa_quindoza@dlsu.edu.ph

Shayane E. Tan
De La Salle University, shayane_tan@dlsu.edu.ph

Charibeth Cheng
De La Salle University, chari.cheng@delasalle.ph

Recommended Citation

Cheng, Jan Kristoffer Y.; Fernandez, Avril Ranezca L.; Quindoza, Rissa Grace Marie M.; Tan, Shayane E.; and Cheng, Charibeth, "Age and Gender Profiling of Social Media Accounts" (2017). *PACIS 2017 Proceedings*. 93.
http://aisel.aisnet.org/pacis2017/93

Age and Gender Profiling of Social Media Accounts

Research-in-Progress

Jan Kristoffer Y. Cheng
De La Salle University
Manila, Philippines
jan_kristoffer_cheng@dlsu.edu.ph

Avril Ranezca L. Fernandez
De La Salle University
Manila, Philippines
avril_fernandez@dlsu.edu.ph

Rissa Grace Marie M. Quindoza
De La Salle University
Manila, Philippines
rissa_quindoza@dlsu.edu.ph

Shayane E. Tan
De La Salle University
Manila, Philippines
shayane_tan@dlsu.edu.ph

Charibeth K. Cheng
De La Salle University
Manila, Philippines
chari.cheng@delasalle.ph

Abstract

The growth of social networking platforms such as Facebook and Twitter has bridged communication channels between people to share their thoughts and sentiments. However, along with the rapid growth and rise of the Internet, the idea of anonymity has also been introduced wherein user identities are easily falsified and hidden. Hence, presenting difficulty for businesses to give accurate advertisements to specific account demographics. As such, this study aims to identify gender and age group of Filipino social media accounts through analyzing post contents. Several features will be considered and various techniques will be adopted to process posts written in English, Filipino, and Taglish (Tagalog interspersed with English). The study will implement these techniques and record their compatibility and performance in a Filipino setting. A computational model capable of gender and age classification will be built as the final product.

Keywords: Profiling, Natural Language Processing, Machine Learning, Information Extraction, Language Resources

Introduction

Businesses invest heavily on social media, with 92% of marketers claiming its importance to their business (Stelzner, M. A., 2015), and social media advertising spending expected to hit $35 billion this 2017 (AMMEX iSupport, 2016). Businesses are concerned with the demographics of their target markets because it helps them understand their ideal customer and formulate marketing strategies (The Upfront Analytics Team, 2015). Online advertising also utilizes demographics in order to group audiences according to shared traits and focus an advertising campaign on these traits. However, businesses and advertisers who utilize the social media may find difficulty in accurately targeting account demographics due to the abundance of anonymous accounts that lack demographic information.

Gender and age are two of the basic information that compose the identity of a person, and are important for defining demographics. However, both can be easily hidden in social media. Exploring these information can be used in business intelligence where customer demographics are used for market studies in targeted advertising and product development.

Based on a 2016 report regarding digital statistics in the Philippines, Filipinos spend the most time on social media, with almost 4 hours spent daily (AMMEX iSupport, 2016). As one of the fastest growing countries in social media use and one of the top users of social media platforms (Chaffey, 2016), the demand to reduce anonymity in the country is high.

Language plays a vital part in profiling social media posts. Prior research have explored profiling social media posts written in English (Murkherjee and Liu, 2010; Cheng et.al, 2011; Newman et.al, 2008) and Spanish (Rangel and Rosso, 2013; Marquadt et.al, 2014). In the Philippines, the language in social media includes English, Filipino, Taglish, and the other Philippine languages and dialects. Taglish is the code-switching between the two most common language in the country, which are English and Tagalog. Posts written in Taglish normally follow the sentence structure of either English or Tagalog, then intersperses the sentence with words from the other language.

This paper focuses on profiling the age and gender of Filipino social media accounts. The research not only has significance in business but also provides insights on profiling using multilingual text, specifically in English, Filipino, and Taglish. The accounts to be profiled will be from the top active social networks used in the country, including Facebook with 26% and Twitter with 13% (AMMEX iSupport, 2016).

Review of Related Literature

There have been numerous studies regarding age and gender classification, and the performance of their solutions differed primarily based on the selected feature sets and learning technique. Most of the feature sets previously used include function words that are both utilized by Cheng et.al (2011) and Newman et.al (2008), f-measures for formality measurement and POS sequence patterns that are both introduced by Mukherjee and Liu (2010), stylistic features that are used by Rangel and Rosso (2013), and other features such as POS n-grams, gender preferential features, word-based features, and structural features. Aside from only using tweets in classifying gender, various combinations of gender-specific keywords, which are essentially free-text fields (i.e. full name, username, description, tweets) have also been used for gender classification (Burger et.al, 2012).

The performance of certain feature sets was found to differ between age and gender classification. Rangel and Rosso (2013) observed that stylistic features performed better for age identification than gender identification while Newman et.al (2008) have discovered that word count approach does not work well with age classification as it did with gender classification.

There are also researches that explore different model structures for simultaneously profiling both age and gender. Rao et.al (2010) explored the performance of a stack model using the predictions from both sociolinguistic and n-gram features along with their prediction weights and discovered that while the sociolinguistic model performed better than the n-gram model, the stacked model performed the best at an accuracy of 72.33%.

Various statistical classifiers were utilized by previous studies. Cheng et.al (2011), Mukherjee and Liu (2010) and Rao et.al (2013) explored the use of Support Vector Machines (SVM) as one of their learning algorithm aside from Naive Bayes and Adaboost decision tree algorithm where the results showed that SVM performed better in the gender classification process with at least 85% accuracy among other learning algorithms. Meanwhile, Halteren et.al (2014) used Support Vector Regression on character and token n-grams which achieved an accuracy of 95.5% with 5-fold cross validation.

In terms of age identification, Nguyen et.al (2011) noted that there are different ways to classify age: age groups, life stages, and exact age. Their results show that using life stages or exact age were more meaningful than categorizing by age groups. Additionally, Rao et.al (2010) used a binary approach in identifying age by dividing them into users below 30 and users above 30. By using a stacked model with SVM, they reached an accuracy of 74.11%.

Methodology

Figure 1 shows the general overview of the system architecture which is divided into Training and Runtime processes. The training process will build the computational model while the runtime process will present the capability of the final model in identifying the age and gender of accounts.

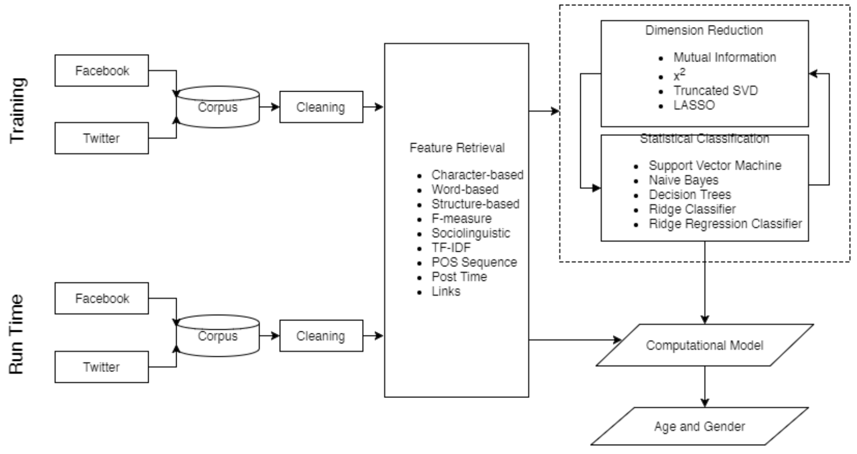

Figure 1. General Overview of the System's Training and Running Processes

Data Gathering

The data would largely composed of posts and tweets from Facebook and Twitter users. Twitter API and the myPersonality Project (Stillwell & Kosinki, 2016) corpus will be used for data collection. Posts will be anonymized by replacing the usernames with a generic word (e.g. USERNAME) to protect the privacy of the accounts.

Feature Retrieval

The system will obtain the different features using various tools specific for each feature. Features to be used are divided into 9 sets, namely character-based, word-based, structure-based, function words, sociolinguistic, Term Frequency-Inverse Document Frequency (TF-IDF) of Text, POS sequence pattern, post time, and links.

Character-based Features

Character-based features include the total number of lowercase and uppercase letters, numbers, spaces, special characters (e.g. %,$,#,&), repeated alphabetical characters and punctuation marks. Previous research have provided evidence that suggests the correlation between the character's frequency and an individual's age or gender (Burger and Henderson, 2006; Rao et al., 2010).

Word-based Features

Word-based features include frequency of Linguistic Inquiry and Word Count (LIWC) features, word-count features, and several statistical measures such as Yule's K measure. Frequencies such as the ratio of the number of unique words and the total number of words, the length of the words, and the number of words with repeated letters etc., are good indicators of age (Rangel and Rosso, 2013).

Structure-based Features

Cheng et al. (2011) used structural features to identify user gender. These features include the total number of lines, sentences, and paragraphs, average number of sentences per paragraph, number of sentences beginning with uppercase, etc.

Sociolinguistic Features

Nguyen et al. (2011) and Rao et al. (2010) used sociolinguistic features that reflect the culture and environment that a user is in when they post. These features include function words, disfluent and agreeing words, contextual features and emoticons or emojis.

Term Frequency-Inverse Document Frequency (TF-IDF) of Text

The TF-IDF of words in posts and tweets provide indications of a person's age and gender. Nguyen et. al (2011) determined that younger people used more informal words like lol, hmm, like and kinda, and Corney et al. (2002) also suggests that females make more use of emotionally intensive adjectives and adverbs such as terribly, so, and awfully. Mukherjee and Liu (2010) also identified that word formality affects gender identification.

Part-of-Speech Sequence

Mukherjee and Liu (2010) introduced POS sequences that are retrieved from the POS sequence-pattern mining algorithm that captures all part of speech sequences that satisfy certain thresholds. Since the system will involve multilingual analysis, different taggers for Filipino and English will be used. For consistency in analysis, Filipino POS tags will be mapped to its equivalent English POS tags if applicable; otherwise, the tag will be retained and used.

To identify which tool to utilize, the text will undergo language detection. English tags will be used for English text and mapped Filipino-to-English tags will be used for Filipino text. If Taglish or unknown, the post will be separated into sentences and each sentence will undergo language detection for appropriate tagging.

Post Time

Burger and Henderson (2006) determined that younger people post more often between 9PM and midnight, while those aged 24 and above post more often during the afternoon.

Links

Marquadt et al. (2014) and Nguyen et al. (2013) studied the relevance of the number of links that appear in their datasets. Websites also have categories and keywords tags that summarize their contents, and these may show biases for certain gender or age group to some contents and topics.

Dimensionality Reduction

Due to the high number of features to be considered, it is only necessary to determine feature subsets that would have better performance in classification tasks. Tighe et al. (2016) demonstrated that removing non-relevant and redundant features allow the classifier to focus only on relevant features, resulting in an increase in the classifiers' performance. This is achieved by identifying the most important features and reducing the number of features being considered. Four methods of dimension reduction will be explored and utilized: Information Gain, χ^2 statistics, Truncated Singular Value Decomposition, and Least Absolute Shrinkage and Selection Operator (LASSO).

Statistical Classification

After obtaining the important features, statistical classifiers will be used to classify gender and age. Four statistical classifiers will be considered: SVM (Cheng et al., 2011), Naive Bayes (Mukherjee & Liu, 2010), Decision Trees (Cheng et al., 2011), and Ridge Classifier, a classifier based on Ridge Regression (Schwartz et al., 2013).

Iterative Computational Model Building and Evaluation

Multiple models will be made to test different combinations of different dimension reduction algorithms and statistical classifiers. In addition, because the model has two outputs, age groups and gender, multiple model structures will be tested, namely Age to Gender Stacked Model Structure, Gender to Age Stacked Model Structure, Parallel Model Structure, and Combined Model Structure. Given that there are three different dimension reduction algorithms, three model structures, six feature combinations, and 2 outputs, there will be a total of 396 models to be generated.

Each model will be evaluated on accuracy, precision, recall, and f-measure. The model that receives the best results on these evaluation metrics will be the final computational model with the best combination of features, dimension reduction techniques, statistical classifiers, and model structure.

Results and Discussion

For preliminary results, Decision Trees, SVM, Naive Bayes, and Ridge Classifier algorithms were initially used together with Information Gain, x^2, and SVD as dimensionality reduction techniques. Fifty Twitter accounts amounting to 50,254 posts were used for Decision trees, Naive Bayes, and Ridge Classifier, while 10 Twitter accounts amounting to 8935 posts were used for SVM. For the features, post time, POS sequence patterns, term frequency-inverse document frequency (TF-IDF) of text, and select LIWC features were used. In order to evaluate the results, 10-fold cross validation was utilized and the average accuracies were calculated.

A summary of preliminary results can be found in Tables 1, 2, 3 & 4. Results from the computational model using Naive Bayes with SVD are all labeled N/A because SVD does not work with Naive Bayes due to negative values.

The accuracies of the models for age were immensely higher compared to the models for gender. This can be attributed to the data used to build the computational model for age mainly composed of users with the same age ranges (i.e. 18-25). Comparing the statistical classifiers, Decision Trees performed better than Naive Bayes and Ridge Classifier. The model structures do not affect the accuracy of the results since they are relatively similar.

Table 1. Results using decision trees

	Information Gain		χ2		SVD	
Model Structure	Gender	Age	Gender	Age	Gender	Age
Parallel	78.33%	92.00%	78.33%	92.00%	78.33%	92.00%
Stacked 1	78.33%	-	80.00%	-	78.33%	-
Stacked 2	-	92.00%	-	92.00%	-	92.00%
Combined	78.33%	92.00%	78.33%	92.00%	80.00%	92.00%

Table 1. Results from computation model using decision trees

Table 2. Results using Naïve Bayes

	Information Gain		χ2		SVD	
Model Structure	Gender	Age	Gender	Age	Gender	Age
Parallel	42.83%	68.50%	33.33%	70.50%	N/A	N/A
Stacked 1	40.83%	-	68.49%	-	N/A	N/A
Stacked 2	-	68.50%	-	68.50%	N/A	N/A
Combined	42.83%	68.50%	45.33%	70.50%	N/A	N/A

Table 2. Results from computation model using Naive Bayes

Table 3. Results using Ridge Classifier

	Information Gain		χ2		SVD	
Model Structure	Gender	Age	Gender	Age	Gender	Age
Parallel	78.33%	92.00%	33.33%	78.33%	78.23%	92.00%
Stacked 1	78.33%	-	68.49%	-	78.33%	-
Stacked 2	-	92.00%	-	78.33%	-	92.00%
Combined	78.33%	92.00%	45.33%	78.33%	78.33%	92.00%

Table 3. Results from computation model using Ridge Classifier

Table 4. Results using SVM

	Information Gain		χ2		SVD	
Model Structure	Gender	Age	Gender	Age	Gender	Age
Parallel	30.00%	80.00%	30.00%	80.00%	30.00%	80.00%
Stacked 1	40.00%	-	40.00%	-	40.00%	-

Stacked 2	-	70.00%	-	80.00%	-	80.00%
Combined	30.00%	80.00%	30.00%	80.00%	30.00%	80.00%

Table 4. Results from computation model using SVM

Conclusion

This study aims to develop a computational model that identifies the age range and gender of the users of social media accounts based on post contents. Datasets will be analyzed and different features will be used in the classification. There are many users who post their opinion regarding topics such as entertainment, politics, and daily life etc. Users post their opinion regarding a specific product or service that most businesses need for better consumer studies. For this study, English, Filipino, and Taglish will be used in processing and building the statistical model as these are the primary languages used in the Philippines. This research is relevant for businesses and advertisers for better consumer demographics.

References

AMMEX iSupport. 2016. "10 Eye-opening Facts about Social Media in PH", *AMMEX iSupport*, (available at http://isupportworldwide.com/blog/archive/socialmediaphilippines/; retrieved February 7, 2017).

Burger, J. and Henderson, J. 2006. "An exploration of observable features related to blogger age", in *2006 AAAI Spring Symposium*, Menlo Park, California: The AAAI Press, pp. 15-20.

Chaffey, D. 2016. "Global Social Media Statistics Summary 2016", *Smart Insights*, (available at http://www.smartinsights.com/social-media-marketing/social -media-strategy/new-global-social-media-research/; retrieved December 16, 2016).

Cheng, N., Chandramouli, R., and Subbalakshmi, K. 2011. "Author gender identification from text", *Digital Investigation* (8:1), pp. 78-88(doi: 10.1016/j.diin.2011.04.002).

Marquadt, J., Farnardi, G., Vasudevan, G., Moena, M., Davalos, S., Teredesai, A., and De Cock, M. 2014. "Age and Gender Identification in Social Media", in *CLEF 2014 Conference and Labs of the Evaluation Forum*, Springer International Publishing AG, pp. 1129-1136.

Mukherjee, A. and Liu, B. 2010. "Improving Gender Classification of Blog Authors", in *EMNLP 2010, Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 207-217.

Newman, M., Groom, C., Handelman, L., and Pennebaker, J. 2008. "Gender Differences in Language Use: An Analysis of 14,000 Text Samples", *Discourse Processes* (45:3), pp. 211-236(doi: 10.1080/01638530802073712).

Nguyen, D., Smith, N., and Rosé, C. 2011. "Author Age Prediction from Text Using Linear Regression", in *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 115-123.

Rangel, F. and Rosso, P. 2013. "Use of language and author profiling: Identification of gender and age", in 10th International workshop on natural language processing and cognitive sciences NLPCS 2013, Marseille, France, pp. 177–186 (available at http://www.kicorangel.com/ wp-content/uploads/2013/10/Use-of-Language-and-Author-Profiling-Identification-of-Gender-and-Age.pdf).

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. 2010. "Classifying latent user attributes in twitter", in 2nd international workshop on Search and mining user-generated contents, New York: ACM, pp. 37-44.

Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., and Ungar, L. 2013. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach", PLoS ONE (8:9), p. e73791(doi: 10.1371/journal.pone.0073791).

Stillwell, D., and Kosinski, M. "Start [Mypersonality Project]". *Mypersonality.org*. 2016. Web. 12 Apr. 2017.

Tighe, E., Ureta, J., Pollo, B., Cheng, C., and Bulos, R. 2016. "Personality Trait Classification of Essays with the Application of Feature Reduction", in International Joint Conference on Artificial Intelligence, New York: IJCAI Organization, pp. 22-28.

The Upfront Analytics Team,. 2017. "Understanding the Importance of Demographics in Marketing - Upfront Analytics", Upfront Analytics, (available at http://upfrontanalytics.com/understanding-the-importance-of-demographics-in-marketing).