“Fifty Shades of Yellow”
Predict Ethnic Origins of Asian Americans Using Surname

Firman M Firmansyah
Jing Ji

Stony Brook University

1. Project Overview

Many studies in the United States, including U.S. Census, categorize people having ancestor originally from Asian countries into one big group named Asians and Pacific Islanders (API). Not only it is less appropriate since some ethnicities are too distinct to be subsumed into one group, but also undermines opportunities to study and benefit from diverse cultures. Thereby, we are going to develop machine learnings that can be used to classify Asian Americans based on their ethnic origins using surnames. It should be noted that even though some researchers have attempted to predict ethnicity based on names [1, 2, 3], few of them have focused on Asian.

Furthermore, we give an example regarding conditions under which our work may be of help for researchers who study Asian racial identity on social media. In this respect, we argue that Twitter users who write Asia, Asian American, or Asian in their bios perceive Asian as their racial identity. However, it is not always the case because non-Asian may also do the same thing. For example, since they love Asian cuisine, then they write Asian in their bio. Both case can bee seen in the following pictures.

2. Team Contributions

We divided the work as follows. Ji was responsible for data collection and construction, as well as comparing classification models with Scikit Learn. Firmansyah was responsible for Twitter data and took charge of training model by TensorFlow and making predictions on display name extracted from Twitter dataset. We both then did the visualizations. For visualizations, we used MatPlotLib with the help of Pandas.

3. Tools and Techniques

We used various tools in the projects. For extracting data from Twitter dataset, we used SQLite with SQL queries. To retrieve current display name of Twitter users, we used R and 'r-tweet' package. To build machine learning models, we used Scikit Learn and TensorFlow on Jupyter notebook. For visualizations, we used MatPlotLib with the help of Pandas.

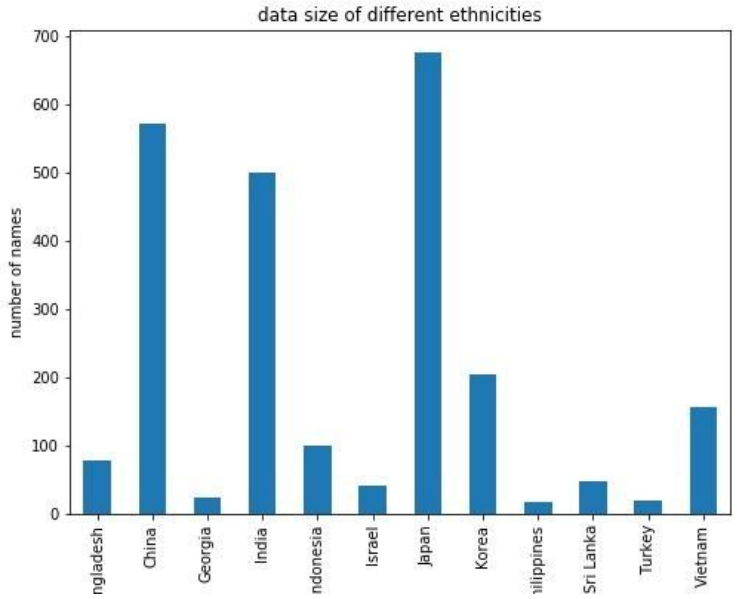
4. Process and Procedures

4.1 Source of Data

The whole dataset originates from Wikipedia, GitHub repository and other online resources. Data from Github projects include a csv file of more than 31 thousands of Indian names which are categorized into subareas in India and also a file of East Asian names, including Chinese, Japanese, Vietnamese and Korean. Besides, we also obtained data from Wikipedia list of most common Asian surnames manually.

4.2 Data Construction

Since the accuracy prediction is dependent on the data structure, it is necessary to make the data balanced and consistent. The final version of the dataset consist of two columns, surnames and ethnicity. As we focus on the basic ethnicity instead of subgroups, labels were transformed to the country name. In order to make the data comparatively balanced, we removed duplicates in India data, sorted the data by frequency, and extracted the most frequent 500 Indian names. Given too few names from Armenia and Azerbaijan (less than 15), they are excluded to from the dataset. The distribution of dataset is shown in the following chart.



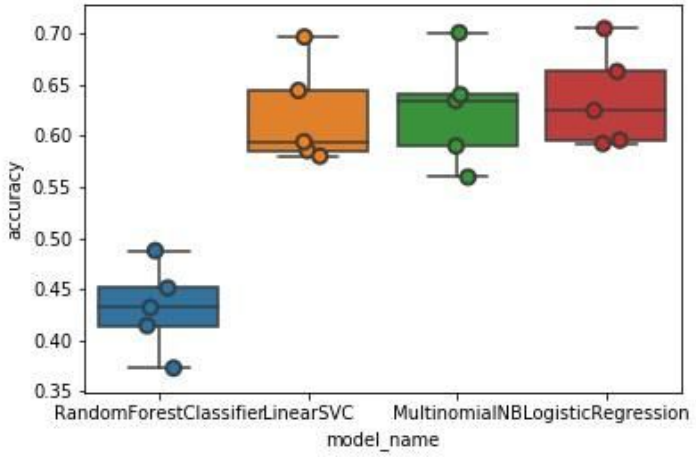
4.3 Machine Learning Models Development

4.3.1. Scikit Learn

Classifiers do not really work on strings like names. Therefore, we first converted names to n-gram counts by CountVectorizer Class. Since the position of n-grams (at the beginning or end) really matters, we add the edge maker at both size of a name string. For example, “bien” becomes “$bien$”. The range of n-grams was set to be 1 through 4.

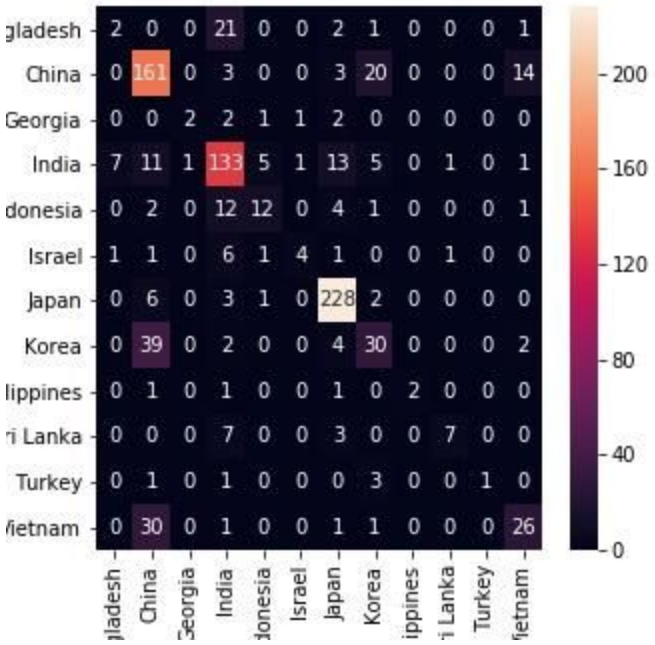
Since it is classification task, we benchmarked the following four models, Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Machine and Random Forest. Performance of four models were tested by cross validation score. As is shown in the chart below, Logistic Regression turned out to show to best performance so we are going to continue training with this model.

Model_name	Accuracy
LinearSVC	0.619895
LogisticRegression	0.635866
MultinomialNB	0.624740
RandomForestClassifier	0.431632



Considering the final dataset made up of only 2440 entries, bootstrap method was applied to enlarge the training and testing data. As there are different number of names in different ethnicity, we made a subset for each ethnicity and did bootstrap for every subset, generating 2440 training data and 865 testing data. The confusion matrix is shown below. The overall precision is 69%. In terms of each group, prediction is relatively better for larger groups such as China, India and Japan. But for minor name groups like Korea, the precision is less than 50%. It is possible that the result was limited by the size of dataset and the lack of frequency of tokens. Some Korean names are similar to Chinese names because of romanization, hence a lot of Korean names misclassified as Chinese names.

	precision	recall	f1-score	support
Bangladesh	0.00	0.00	0.00	27
China	0.67	0.87	0.75	201
Georgia	0.50	0.12	0.20	8
India	0.60	0.85	0.71	178
Indonesia	0.60	0.09	0.16	32
Israel	0.50	0.07	0.12	15
Japan	0.83	0.97	0.89	240
Korea	0.61	0.29	0.39	77
Philippines	1.00	0.20	0.33	5
Sri Lanka	0.86	0.35	0.50	17
Turkey	1.00	0.17	0.29	6
Vietnam	0.89	0.27	0.42	59
avg / total	0.69	0.70	0.65	865



4.3.2. TensorFlow

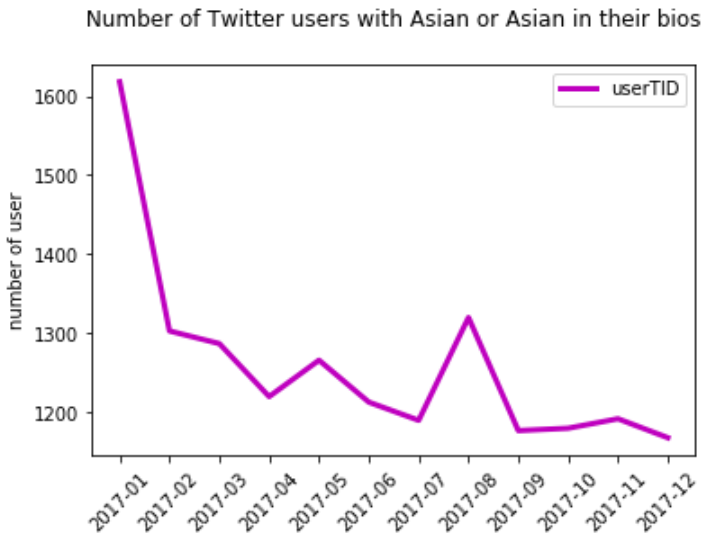
As mentioned earlier, we also used TensorFlow to build a machine learning model. Different from previous steps, we first made a dictionary to count frequency of occurrence of given surname in given ethnicity. Then we created conditional probability based on this data and converted to a matrix. Following this, we divided the dataset to training and test data and implement deep neural networks technique. The result was better than previous models as can be seen in this following table:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	22
1	0.90	0.93	0.92	170
2	1.00	1.00	1.00	9
3	0.99	1.00	1.00	166
4	1.00	1.00	1.00	31
5	1.00	0.92	0.96	13
6	0.99	0.98	0.99	193
7	0.82	0.98	0.89	59
8	1.00	1.00	1.00	9
9	1.00	1.00	1.00	8
10	1.00	1.00	1.00	8
11	1.00	0.68	0.81	44
avg / total	0.96	0.96	0.96	732

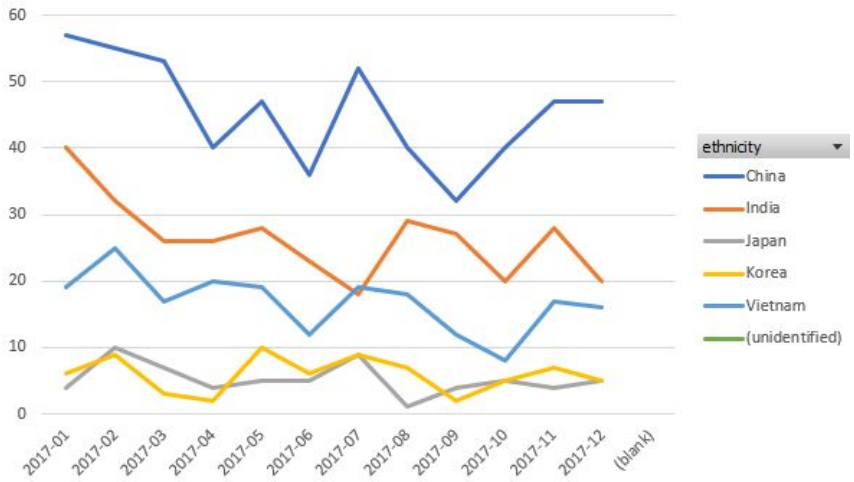
However, since the frequency for each surname is unbalanced and our data is very few to be used to predict bigger dataset, we should mention that this model is subject to bias.

5. Results

Before we developed our models using machine learnings on Python, the dataset from Twitter that contain Asian racial identity word on bios show a declining trend as can be seen in this following graph.



After we developed the machine learning models, we could look the dataset from Twitter differently. However, we faced a challenge in implementing the models. Not all display names from Twitter have the same format as the database and they needed further process. Therefore, we tried another way and did an inner join of the Twitter dataset and the surname database using R. In this case, we predicted the ethnic origins based on surname similarity. As can be seen in the following graph, the trends were relatively stable through the year and across ethnic origins. Regarding the size, the majority of Twitter users who perceive Asian as their identity are Chinese while the opposite are Japanese.



6. Conclusion, Limitation, and Future Work

The machine learning models, even though they can be used to predict given dataset and some of them show a promising result, need to be elaborated further. Our databases are very limited and they do not cover all ethnic origins. Also, it is important to validate the database with various methods such as survey. As for the study case, we learned that dataset from Twitter seemed contain many noises if we only rely on one or two signals. We also learned that working on text data from social media may need longer process that we previously thought. We suggest that for next project, it will be useful to include other signals such as geolocation, profile pictures, and followers to build the machine learning models.

7. References

[1] Ambekar, A., Ward, C., Mohammed, J., Male, S., & Skiena, S. (2009, June). Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (pp. 49-58). ACM.

[2] Imai, K., & Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, *24*(2), 263-272.

[3] Sood, G., & Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109*.