TweetClassifier
NLP-Driven Demographic Profiling

San Francisco State University
Lu-Ting Lee
Rex Cheung
Leyla Ozsen

Table of Contents

1. Executive Summary

This research focuses on identifying reliable methods to gather demographic data from Twitter users. Despite Twitter's policy of not requiring users to disclose personal information, such demographic insights are crucial for businesses seeking targeted marketing strategies. To address this gap, our study systematically examines and documents potential methods for predicting private user demographics based on publicly available Twitter data. The project spans the full scope of a data science project, including data collection, feature engineering, statistical analysis, and machine learning modeling.

In collaboration with an industry partner, we refined our research direction to ensure practical business applicability. The project's final deliverables include a detailed description of our data collection and labeling processes, along with a predictive machine learning model that accurately identifies a Twitter user's gender with an 85% success rate. Furthermore, our concluding section offers actionable recommendations for enhancing age prediction techniques, refining our methodology, and suggests avenues for future scholarly inquiry.

2. Introduction

Online marketing heavily relies on social media platforms, such as Twitter, Youtube, Instagram. Precise targeting the audience group is the key step to successful online marketing campaigns and increased advertising return of investment. With the massive volume and variety, Twitter data supports multi-prediction analysis tasks such as sentiment analysis, emotion analysis, topical analysis, and demographic analysis (Kursuncu, et al., 2018). In these analyses, customer demographic analysis is the most important factor that defines the target audience. In general, user demographic data includes gender, age, location, income, ethnicity, or education level. This research focuses on gender and age prediction, and utilizes location information build-in Twitter metadata.

Our partner company - Intuition Intelligence, a B2B solution services company specialized in social media platforms, aims to provide data insights supporting companies run marketing strategies and advertising campaigns efficiently. Intuition Intelligence runs a software as a service (SaaS) product embedded with several machine learning features to provide high-level marketing reports such as product review sentiment analysis, influencer branding, and predicting trends for digital content marketing. Twitter user demographic data prediction is a new trail to the company. By cooperation, we could receive technical support and knowledge from the industry, and the research results could increase the company's development process as well.

Beginning the research, we refer to a similar project available on Kaggle.com (Eight, 2016). We retrieve a dataset and research the whole process of the machine learning modeling. By this early-stage study, we further understand Twitter's metadata and figure out two text features that are feasible for our project. This study also supports us to draw down the project scope and define the research direction.

After discussions with Intuition Intelligence, we narrow down our research scope to the U.S. Twitter users who are interested in the entertainment industry only. The reasons are as follows. First, targeting specific user groups allows us to start research and collect data. Second, the entertainment industry is the domain business of Intuition Intelligence, which means the company could provide better technical support and future applications.

This research delivers several achievements: 1) Data labeling rule: Labeling is an extremely important step for the machine learning approach. Poor labeling quality prevents machine learning models from picking up information in the data properly. We develop reliable methods to label public data from the Twitter application programming interface (API). 2) A baseline model for gender prediction: We provide details of machine learning processes, including pre-processing methods, information extraction methods, and machine learning modeling. In this model, we deploy a two-layer model structure that contains eight week learners and one classifier to output predictions. And employ cross-validation methods, test accuracy score, recall score, precision score, F1 score, and area under the curve (AUC) to verify and fine-tune the model.

3. Related Work

In this section, we briefly review related author profiling works on gender and age. And then review feature extraction and machine learning modeling methods.

3.1. Related works on author profiling

For gender prediction, one research supports the idea that utilizing writing style as statistical features could predict author gender in binary categories(Bamman, David, Eisenstein, Jacob, & Tyler, 2012). Although the outliner exists, the statistical significance reveals short text data from Twitter could be used for gender detection. Another research applies a stacking machine learning structure approach from PAN (Agrawal, Madhulika, & Gonçalves, Teresa, 2016), which is an organization that provides a series of scientific challenges and shares research results in digital form.

For the age prediction, researchers develop a data collection and machine learning pipeline strategy to conduct age group prediction with DBpedia resources (Alan, Smith, & Manas, Gaur, 2018). The researchers utilize pre-identified celebrities' data including date of birth from DBpedia, and then utilize Twitter API to extract screen names of followers in the celebrities' network. Finally, match the follower list to DBpedia and take 50 features including age for machine learning training. They achieved 84% accuracy by Support Vector Regression with Radial Basis Function (RBF) Kernel.

Another age prediction approach (Morgan-Lopez, Kim, Chew, & Ruddle, n.d.) provides detail about data collection. The researchers apply Twitter API for the ‘Happy Birthday’ keyword searching and collecting data in two years' duration. Twitter API is a reliable and stable method for data extraction, but the keyword search method is not an effective method for our limited time. Eventually, we use the filtering method to obtain the user age information.

3.2. Related works on ensemble modeling

Stacking is a machine learning technique to improve overall performance (Wolpert, 1992). The main idea is to use several different classifiers as weak learners to produce multiple outputs. Since each classifier has different characteristics, each classifier extracts different information from the data during training. Statistically speaking, combining the predictions from several weak learners is possible to distill more information in the data, and then improve the overall performance. Another experiment proves that the Stacking model performed well in the English Twitter data and exceeds the other common machine learning structures. (Mihael, Nikic´, Martin, Gluhak, & Filip, Dzˇidic, 2017).

4. Methodologies

4.1. Assumption

The research is based on the data we extracted from Twitter. The sample size is relatively small compared to the whole Twitter activated account thus our data potentially have a bias to the population data. To start our experiments and eliminate factors we can not control, we set up two assumptions for this research.

Assumption 1: We assume Twitter users' writing style is different by gender or age groups. We decide to utilize text data as our features and build machine learning models to predict the gender and age group with such features. If the writing style is the same in all user groups, we are unable to predict user groups by machine learning methods.

Assumption 2: All Twitter users have an equal willingness to disclose demographic data in their Twitter account profile. We develop a labeling workflow to distill the gender and age group information from Twitter data, however, different user groups may have different wellness. In general, we could better define the patterns and bring consistency prediction results with the larger dataset. In this early stage exploration, we apply a sample dataset for analysis and baseline model building. This assumption allows us to treat all user categories with the same weights and utilize our sample dataset to represent the whole Twitter user population.

4.2. Data collection and labeling

4.2.1.Data collection

Although other relevant studies have provided labeling methods, we are not convinced to apply third-party databases or unverified estimation methods. Moreover, the trends and writing language on social media change rapidly over time. Compared to the research or dataset achieved a few years ago, we tend to use real-time Twitter data for research.

We utilize Tweepy library ("Streaming With Tweepy") with Python script for data streaming. Tweepy is an open-source library that integrates multiple functions for Twitter API connecting. We build a streamer with Tweepy and set filters by our scope to collect real-time tweets (Appendix 1). The original returns from Twitter are encoded by JSON with over 100 attributes. We select 12 attributes as candidate features and transform the original data to comma-separated values (CSV) file. The data extraction starts from 2020 March to May, a total amount of about 400k observations from individual U.S. domestic users.

4.2.2.Labeling

The machine learning approach requires labeled data for both training and validating processes. High quality labeled data is crucial for prediction modeling. Based on previous observation, we utilize user description from metadata and user profile picture to define user gender and age groups. For example, in the user description section, some users introduce themselves with specific pronouns or key phrases such as ‘she/her, dad/father, or 23 y.o. lady...etc ’. Based on these findings, we further conclude labeling rules as an instruction for manually labeling.

To reduce the workload of manual labeling, we label gender and age groups at the same time. Moreover, several filters apply to reduce the data before manual processing (Appendix 2). The labeling rules based on user descriptions. When the user description cannot confirm the gender, check the user profile picture instead. If neither of the two above steps works, we leave the data as unrecognizable (Fig. 1.). As a result, we labeled 2,029 observations for gender, and 1,628 observations for age groups (Table 1., Table 2.).

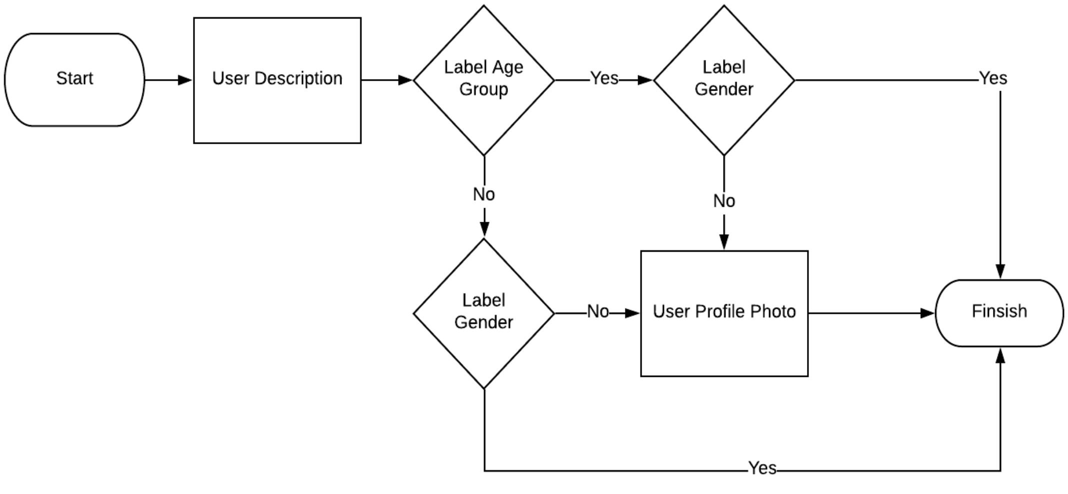


Fig. 1. Manually labeling workflow

	Gender	Age Groups (years)
Categories	Male Female	Under 18 18 - 24 25 - 34 35 - 49 Above 50
Metadata Reference	User description User profile photo	User description

Table 1. Labeling categories and reference

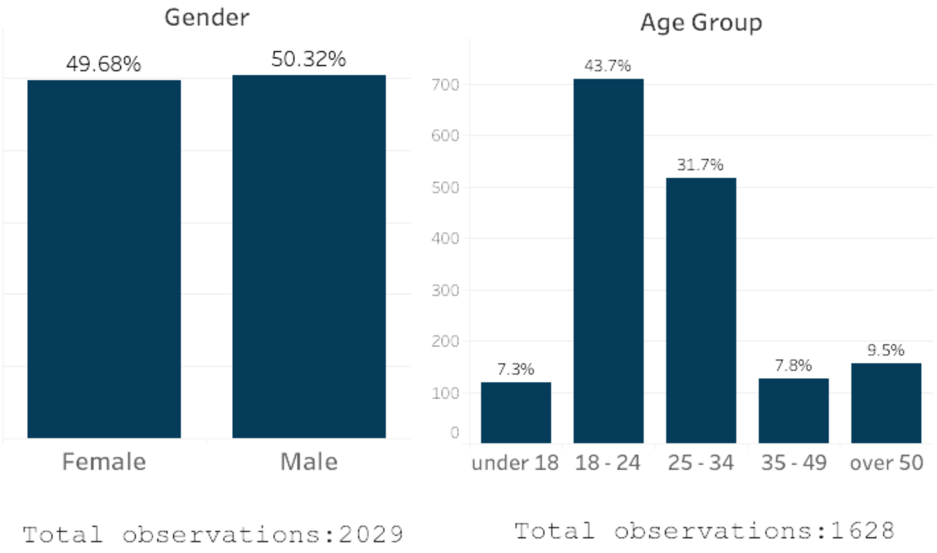


Table 2. Summary of the datasets used in this study.

4.3. Machine learning approach

The modeling section includes several subsections to approach prediction modeling. Since the age prediction part is still ongoing, the following paragraphs demonstrate with gender prediction only (Fig. 2.).

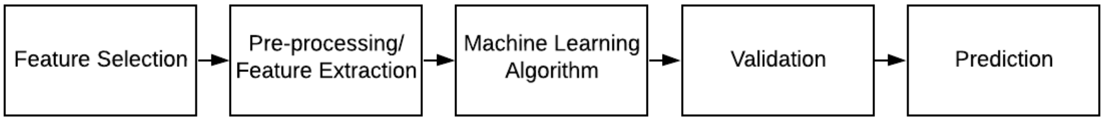

Feature Selection
Pre-processing/ Feature Extraction
Machine Learning Algorithm
Validation
Prediction

Fig. 2. Manually labeling workflow

4.3.1.Feature Selection

The research aims to find the gender and age groups with Twitter metadata, thus we select tweet and user description two text features by unique user files. Moreover, we aggregate two additional features ‘sidebar color’ and ‘emoji usage’ as candidate features for future research.

4.3.2.Pre-processing & Feature extraction

The original Twitter data contains several noises and is incompatible with machine learning models, therefore, pre-processing and feature extraction are required for data cleaning and transforming.

For pre-processing, we eliminate URLs, hashtags, emojis, and non-alphabet characters, and then we transform all plural nouns to singular nouns. Attempting with multiple combinations, we do not clean up stop-words, which is a common pre-processing method in natural language process cases for excluding high frequent pronouns and articles.

For feature extraction, we select Bidirectional Encoder Representations from Transformers (BERT) to convert text features to numerical matrices. BERT is a pre-trained language representation model structured by neural network techniques (Jacob, Devlin, & Ming-Wei, Chang,& Kenton, Lee,& Kristina Toutanova, 2019). BERT has been widely used in various natural language processing projects since it was introduced. In this research, we employed DistillBERT library for text transforming. DistilBERT is a compact version of the original BERT (Victor, Lysandre, Julien, Wolf, & Thomas, 2020). With specific distilling structure, DistilBERT performs 60% faster and retains 97% language transforming capabilities. Each sentence is converted into a 768-dimensional vector. To be more specific, each user's tweet and user description will be converted into two 768-dimensional vectors respectively.

4.3.3.Modeling

We define our supervised machine learning models with a stacking structure. In this two-layer structure, the output of the first layer will be input to the second layer, and then the second layer will generate the final prediction. The first layer has eight classifiers as weak learners, generating the probability of target categories. In the second layer, only one classifier is used to output the final prediction results.

Two reasons are forming this structure. 1) Improve accuracy: Compared with no stacking structure control group, the stacking structure improves the accuracy by 5.9%. 2) Dimension reduction: If we want to apply more features for modeling, the high-dimensional text vector will affect the strength of the extra features. Thus we transform the high-dimensional vector into a one-dimensional probability output at the first layer. Future new features can be imported here and then input into the second layer for prediction.

Regarding the data flow, the input data is split into training and validation sets from the beginning. And then we fit the training set to the first layer classifiers and collect prediction output for the second layer training. The dimension of first layer output depends on how many weak learner numbers are applied, for example, we apply eight weak learners in this case, thus one observation will expand to eight attributes. The validation set goes through the first layer for generating prediction only, that said the dataset never leaks for weak learners.

In the second layer, we apply the first layer output as a training set and apply validation set as test set for second layer classifier (Fig. 3.).

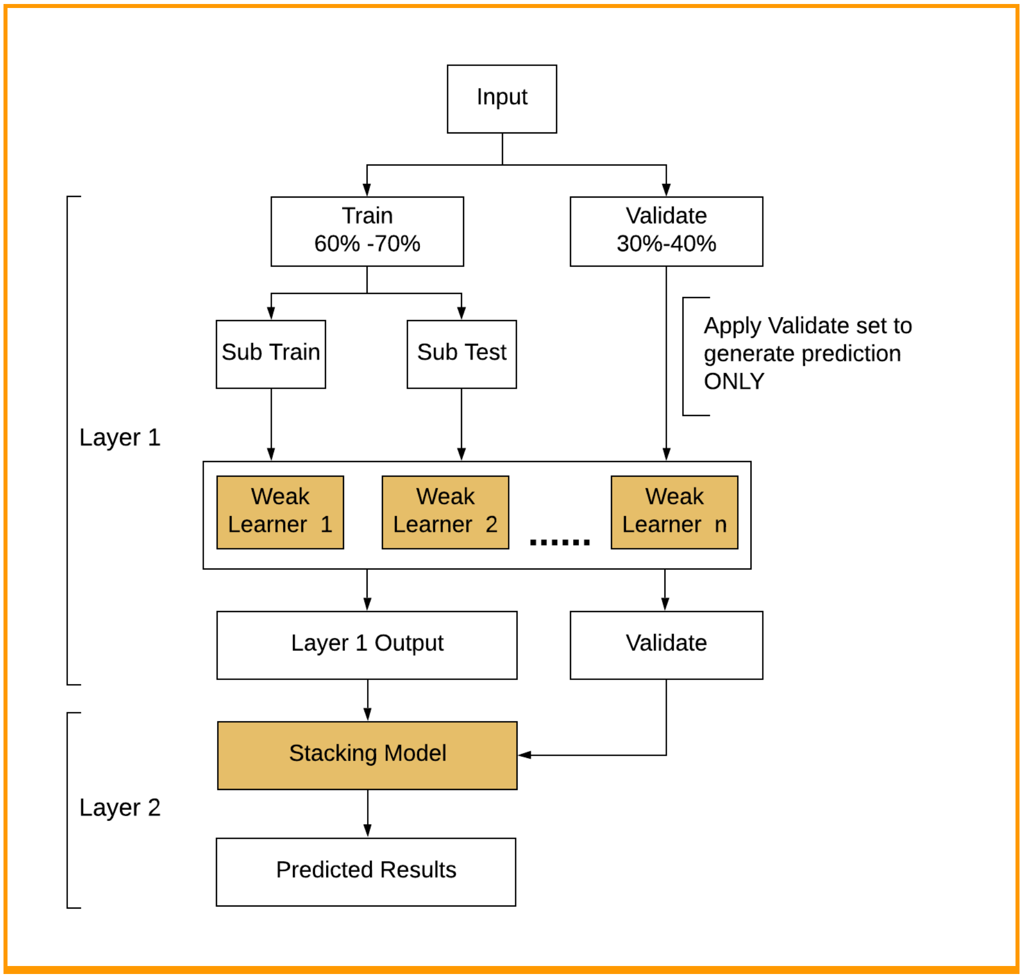


Fig. 3. Stacking structure

5. Analysis Results:

The prediction output is evaluated with 5 folds cross validation and GridSearch to approach the global optimization. We define RandomForest Classifier as the stacking model in layer 2 based on the highest test accuracy 84% and AUC 91% score (Table 3, Table 4.).

Category	Layer 1 Classifiers	Layer 2 Classifier
Gender	LogisticRegression	RandomForestClassifier
	SGDClassifier	
	RandomForestClassifier	
	AdaBoostClassifier	
	SVC - RBF kernel	
	LinearSVC	
	GaussianProcessClassifier	
	MLPClassifier	

Table3 Classifiers for stacking model

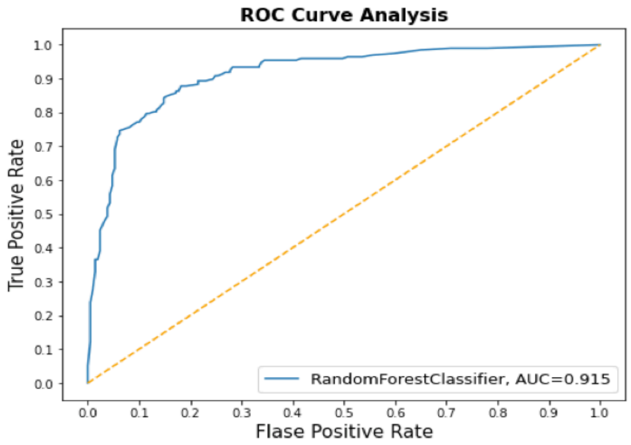
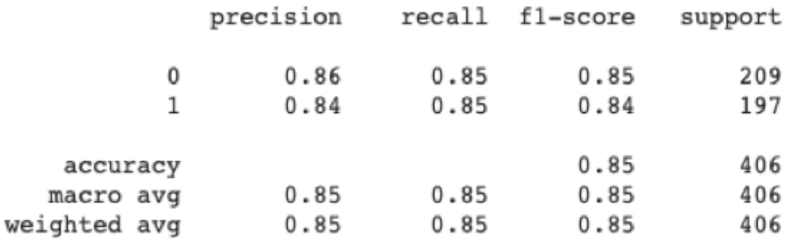


Table4 Classifiers for stacking model

Then we employ an evaluation matrix to further examine the prediction performance. According to the output below, we receive balanced precision, recall, adn F1 score in both categories, which means the model has sufficient ability (>0.5) to predict gender without over weighted on specific categories (Table 5.).



```
              precision    recall  f1-score   support

           0       0.86      0.85      0.85       209
           1       0.84      0.85      0.84       197

    accuracy                           0.85       406
   macro avg       0.85      0.85      0.85       406
weighted avg       0.85      0.85      0.85       406
```

Table 5 Evaluation matrix

6. Conclusion & Future Work:

In this article, we build a machine learning model to predict Twitter user gender data through public Twitter metadata. The results include data collection and labeling methods, models and analysis results for future research. Each part is equally important and directly contributes to product development. We obtained 84% accuracy in predictive modeling, which indicates that there is a significant relationship between the user's writing style and gender.

For the age prediction, we can repeat this experiment with the same preprocessing method and model structure. The only difference is that the age group is a multi-category classification task. We can utilize the current model with an outer loop to approach the OnevsAll method for obtaining predictions without rebuilding the model from scratch.

In terms of data collection, we can utilize other models such as face recognition to accelerate the manually labeling process. The model could exclude non-human profile photos such as animated characters or pets in the first place. Of course, the unsupervised face recognition model has the potential to label data automatically as well.

However, it is still quite difficult in age labeling. In related research, a team uses social media network connection to cluster the age by social group. This seems to be a research direction, but we could not verify the labeling quality for the case. More efficient age labeling methods require further research.

In terms of feature selection, we extracted emoji usage statistics that may further improve gender prediction performance (Appendix 3). According to our dataset, women have a significantly higher frequency of emoji usage than men, and 30% of Twitter text data include at least one emoji usage. Besides emoji, we also found eight attributes related to sidebar, background, and other user interface color code in the metadata. Based on the research time limit, we do not apply further statistics to find out the significance. Finally, the hashtags and URLs that we removed during preprocessing may also potentially form new features.

In business applications, our predictive model has been able to identify the gender of specific user groups. And we have filtered out the user's tweet location during the data collection process. Combined with age prediction, we will be able to provide comprehensive market reports. Through this information, we can not only better estimate the target audience for marketing purposes, but also provide more valuable research in combination with other public information. For example, we can use these demographic data to match US personal income information to provide advanced user insights.

7. Reference

1. Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2018). Predictive Analysis on Twitter: Techniques and Applications. *Lecture Notes in Social Networks Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, 67–104. doi: 10.1007/978-3-319-94105-9_4

2. Eight, F. (2016, November 21). Twitter User Gender Classification. Retrieved from https://www.kaggle.com/crowdflower/twitter-user-gender-classification

3. Bamman, D., Eisenstein, J., & Schnoebelen, T. (2012). Gender in twitter: Styles, stances. *andsocial networks. Technical Report 1210.4567, arXiv*.

4. Morgan-Lopez, A. A., Kim, A. E., Chew, R. F., & Ruddle, P. (n.d.). Predicting age groups of Twitter users based on language and metadata features. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183537

5. Agrawal, M., & Gonçalves, T. (2016). Age and Gender Identification using Stacking for Classification* Notebook for PAN at CLEF 2016.

6. Smith, A., & Gaur, M. (2018). What's my age?: Predicting Twitter User's Age using Influential Friend Network and DBpedia. *arXiv preprint arXiv:1804.03362*.

7. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. doi: 10.1016/s0893-6080(05)80023-1

8. Nikic, M., Gluhak, M., & Dzidic, F. An Ensemble-Based Approach to Author Profiling in English Tweets. *Text Analysis and Retrieval 2017 Course Project Reports*, *18*(24), 51.

9. Streaming With Tweepy¶. (n.d.). Retrieved from http://docs.tweepy.org/en/latest/streaming_how_to.html

10. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from https://arxiv.org/abs/1810.04805

11. Victor, Lysandre, Julien, Wolf, & Thomas. (2020, March 1). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Retrieved from https://arxiv.org/abs/1910.01108

12. sklearn.model_selection.GridSearchCV¶. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

13. Li, L., Sun, M., & Liu, Z. (2014, August). Discriminating gender on Chinese microblog: a study of online behaviour, writing style and preferred vocabulary. In *2014 10th International Conference on Natural Computation (ICNC)* (pp. 812-817). IEEE.

Appendix 1

Keywords applied for data collection

[‘movie', 'new movie', 'netflix', 'hulu', 'youtube', 'hbo', 'tiger king', 'onward', 'the plaftform', 'trolls world tour', 'unorthodox','gentlemen', 'the loop', 'extraction', 'once upon a time', 'hollywood', 'parasite', 'mana lorian', '1917', 'invisible man', 'birds of prey','harley quinn', 'mulan', 'sonic', 'artemis fowl', 'kings man', 'Disney+','Netflix', 'Warner', 'HBO', 'Amazon Prime Video', 'Hulu', 'HBO', 'YouTube','Google Play','Vudu', 'movie']

['NowWatching', 'Oscars', 'AppleTV', 'TV', 'films', 'Netflix','Disney Plus',
'AMC', 'HBO', 'Amazon Prime', 'Hulu', 'Adult Swim', 'YouTube',
'GoldenGlobe','GoldenePalme', 'festivaldeCannes', 'Hollywood',
'rotten tomatoes', 'IMDb', 'SONY crackle', 'Yahoo! Movies',
'movie time', 'movie to watch', 'movies out now', 'movie theater', 'movies',
'movie enthusiast', 'movie nerd', 'movies 2020', 'movies 2019',
'Marvel', 'Showtime', 'DC Comics',
'ZoeysPlaylist', 'feelgoodnetflix', 'Schitt', 'BetterCallSaul',
'Blow the Man Down', 'SEX EDUCATION', 'OZARK','Farmageddon',
'UncorkedNetflix', 'TigerKingNetflix', 'TigerKing', 'Westworld','awkwafina',
'ActionMovie', 'Drama','Comedy', 'horror', 'sci-fi', 'trailer']

['meangirls', 'Taylor Swift', 'Kim Kardashian', 'SharronTownsend', 'ImRhondaPaul',
'Rhonda Paul','Haley Cureton', 'haley.cure', 'nicole.ob', 'Nicole O’Brien',
‘Harry Jowsey', 'HarryJowsey', 'Chloe Veitch', 'chloeveitchofficial',
'Matthew Smith', 'matthewstephensmith', 'Kelechi Dyke','kelechidyke',
'David Birtwistle', 'DBirtwistlePT', 'Francesca Farago', ' Naomi Scott',
'Margot Robbie', 'Sophie Turner', 'Kaitlyn Dever', 'Emilia Clarke', 'Cate Blanchett',
'Amy Adams', 'Marion Cotillard', 'Anne Hathaway', 'Kate Winslet',
'Scarlett Johansson','Halle Berry', 'Jennifer Lawrence', 'Nicole Kidman',
'Tom Hanks', 'Samuel L. Jackson', 'Christian Bale', 'Hugh Jackman',
'Matthew McConaughey', 'Anthony Hopkins', 'Robert Downey','Benedict Cumberbatch', 'Christopher Walken']

Appendix 2

Filters

There are 9 filters that present with regular expression.
To accelerate the labeling process, we apply different filter combinations for each labeling batch.

1. All description must start with 2-digit integer
   r'(^[0-9][0-9])'
2. All description must contain 2-digit integer
   r'([0-9][0-9])'

----

3. Following A), the rest of description must contain one or more keyword below:
   he|she|him|his|her|girl|gal|dude|woman|female|male
4. Following B), the rest of description must contain one or more keyword below:
   he|she|him|his|her|girl|gal|dude|woman|female|male

----

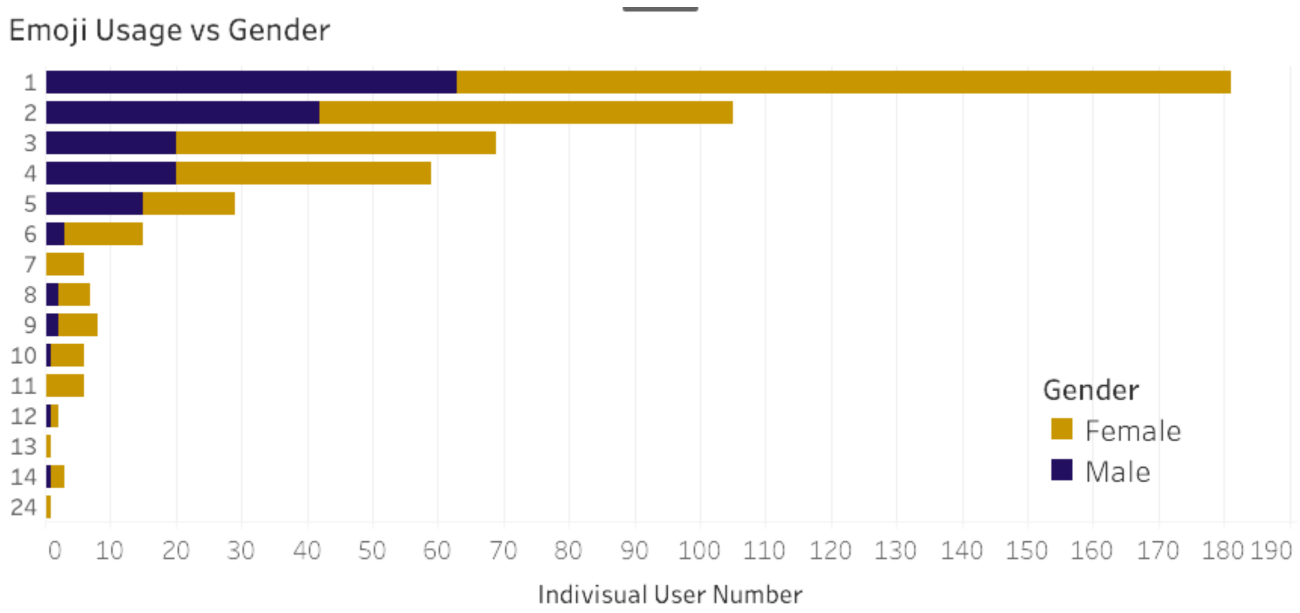
5. Following A), the rest of description must contain one or more keyword below:
   yr|years|year|old|born
6. Following B), the rest of description must contain one or more keyword below:
   yr|years|year|old|born
7. **Without** any filter, the description must follow keyword order below:
   ([1-9][0-9]) (year|yr|years|yo) (boy|girl|gal|wife|husband|man|woman)

----

8. Following A), the rest of description must contain one or more keyword below:
   husband|dad|mon|wife
9. Following B), the rest of description must contain one or more keyword below:
   husband|dad|mon|wife

Appendix 3

X: Emoji count from tweet and user description
Y: Sum of individual user


Emoji Usage vs Gender
Indivisual User Number
Gender
Female
Male