Gender and Age Prediction Multilingual Author Profiles Based on Comments

Ali Nemati[1]

University of Washington Tacoma, USA
[1]1900 Commerce St, Tacoma, WA 98402, United States of America
anemati@uw.udu

Abstract. Recently, several approaching been presented to detect automatically users' age and gender classification from multiple languages based on documents, text, and comments on the web or social media update status. The purpose of this task is determining and detecting information such as age, and gender from multilingual (Roman, Urdu and English) author profiles based on texts or documents. By using four machine learning techniques, my system derives an ensemble model for age and gender categories. The ensemble model is composed of a multinomial Naive Bayes classifier, a Gradient Boosting Classifier, a Logistic Regression CV and a Multi-Layer Perceptron classifier. The system can categorize and diagnose text source automatically with a sensitivity and specificity of age and gender with unknown testing data. The accuracy result is 83 percent for gender category, 60 percent for age, and accuracy 49 percent is for joint age and gender category.

Keywords: Age and gender prediction, Multilingual author Profile, Social media, Text analysis, Data mining

1 Introduction

Authors profile task helps to reach age and gender classification by the feature extractions from texts, documents and comments on the web or social media update status [1]. Recently, many researchers have investigated multilingual author comments to detect as much as possible and important information such as gender and age from an author. For example, business companies are gathering customers' age and gender in order to give better services in the future [2].

Furthermore, identifying gender and age about customers style, according to their comments on social media, helps them to recognize who their customers are. Therefore, they make decisions to improve their services in the future [3]. In case of developing and evaluating automatic author profiling system, the training dataset combines 350 separate text files. The training dataset contains documents that have accumulated over social media such as Facebook, Tweeter, other social media websites and authors' comments are based on multilingual languages such as Urdu, English, and Roman.

The dataset has collected with smartphones that are written by QWERTY keyboards and is available publically on the web address “Fire’18 MAPonSMS” [4]. A true CSV file has released with 350 records including age and gender that corresponds with each text files. An ensemble model [5][6] which is a combination of four classifiers is used in this study. The first classifier is called Logistic Regression CV Classifier. The second one is called Naïve Base Classifier. Multi-layer Perceptron Classifier (MLP) is another classifier and the last classifier is a Gradient Boosting Classifier. The goal of this task is to implement a system to recognize users’ information on social media. This system is trained according to authors’ Short Message Send (SMS) or documents. The result for accuracy metric based on unknown testing data reveals that for gender class 83 present, for age class 60 percent and joint age and gender class 49 percent accuracy is obtained. The findings of our dataset are:

1. Even though the dataset is very small, a better efficiency than the baseline result is achieved.
2. The results of the model improved when the ensemble model was used because having the specific model for analysis and process text data has not better performance or does not achieve the higher accuracy.

The python application is downloadable in https://goo.gl/D37Qii .

2 Related Works

As already mentioned, Author profile identification be used in serval areas such as psychology and natural language processing. In more recent studies, the interest in data mining has grown, and several papers have explored the developing age and gender prediction collected information over social media [7-10]. Gender identification was done by Burger and Henderson in 2006[11]. Another Author profile research was proposed by Pastor López-Monroy and his colleagues to detect a new document representation gender and age over social media in 2015. Furthermore, Monroy and el., were presented a new paper representation for author profiling detection in 2013 [12].

Marquardt and el., has published a paper about the predictive age and gender identification according to Social Media at University of Washington [13]. Similar work has been done on predicted task such as gender, and age from smaller dataset consists of social media comments on Twitter [14].

Compared to the dataset that is proposed in this paper, they have used the differing dataset. And all the prior works done in age and gender prediction have targeted the task of using ensemble model to obtain the higher achievement.

3 Dataset Description and Preprocessing

The dataset has gender and age class and consists of a binary classification with male and female. In addition, age class is a multiple classification that is based age group on such as 15-19, 20-24, 25-xx. This ensemble model is chosen to reduce the time during learning the system and to obtain a highly accurate result or at least close to the real outcome as much as possible.

Figure 1 exposes the distribution of age and gender in the true CSV file that releases into the training data. It illustrates that 40 percent of records are females and 60 present of gender are males. 50.28 percent of people aged between 20-24. and, 30.85 percent of them aged between 15-19. The rest of the recodes are 25 years old or above.

The baseline result for this dataset is 60 percent of gender (Male) and 51 percent of them aged between 20-24.

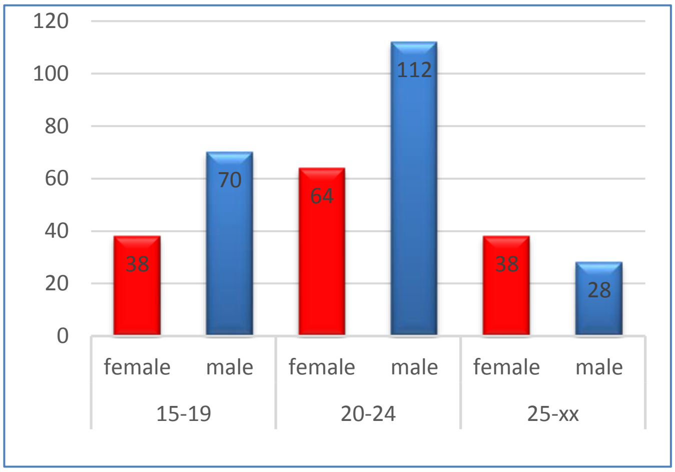

Fig. 1. Age and Gender on the true file

4 Methodology

This system can detect age and gender class based on author profiles. For text processing, Scikit-learn package is applied. This package is ubiquitous in order to use in machine learning with free libraries. The ensemble model is a supervised learning model by using Scikit-learn package.

First of all, the application receives the training file which has two columns, age and gender. The third column called that called transcripts_test is created to accumulate all comments of authors that have correlated with each person. Next, the dataset splits into 280 training instances (80 %) and 70 test instances(20%). The dataset has shuffled for reducing variance and avoiding overfit.

Afterward, for converting transcripts_test column to a vector of integer counts, the system requires to insert transcripts_test into CountVectorizer with all parameters as the default [7]. Finally, the application recalls the ensemble model, that is discussed in the paper, and fit and predict the results with five-fold cross-validation. There are multiple models to apply for this task, but these models are being well-designed for text classification, binary and Boolean features. The application runs on Google Colaboratory (Colab) that has free CPU cloud services by using TensorFlow. The Google Colab consists of 33 GB hard disk and 13 GB RAM and 2-core Xeon 2.2GHz. for running and testing the system Python 3.x (3.6) is used.

The system works with the ensemble model that combines four classifiers with high accuracy. For that reason, all models in one particular model are joined and to get votes for all models. As a result, the ensemble model achieves the results with high accuracy or can be close to real precision by applying 5 fold cross-validation. The classifiers in the ensemble model are listed such as the Logistic Regression CV Classifier, the Naïve Base Classifier, the Multi-layer Perceptron Classifier(MLP) and the Gradient Boosting Classifier. In the below section, these four machine learning classifiers are described:

4.1 Logistic Regression CV Classifier

The system has used the Logistic Regression CV classifier as one of the model with python 3.6. The Logistic regression model is a machine learning method for the analysis of high dimensional information and text dataset.

Similarly, it uses the logistic sigmoid function to achieve the result of text sources, and different parameters are experimented. Eventually, the factor that is the solver is modified and default solver ‘lbfgs’ to ‘linear’ are altered because it is the appropriate solver for the small dataset. Other parameters are regarded as defaults. The result with the five fold cross-validation for gender class is 84.27 present and for age class is 64.32 present.

4.2 Naïve Bayes Classifier

Naïve Bayes classifier is an excellent machine learning technique for text categorization. This model is very fast and sophisticated method in real-world events such as spam filtering, document categorization and text classification in our task. Naive Bayes classifier has three models such as multinomial, Gaussian and Bernoulli.

The Multinomial Naive Bayes (Multinomial NB) classifier is chosen to be able to extract features e.g., word counts for our task [8]. This specific model requires having integer counts for a numerical statistic. In order of having integer counts, it requires calling term frequency-inverse document frequency (tf-idf).

The tf-idf determines how many important words can be in the dataset. Multinomial NB with alpha equal 0.13 is used and the rest of the parameters are as defaults. The result achieves 86.08 percent accuracy for gender and 65.01 percent accuracy for age.

4.3 Gradient Boosting Classifier

To receive a high accuracy for text source, the Gradient Boosting Classifier has used and the following parameters are modified. At first, to explain this model, the learning rate default which is 0.1 is shifted to 0.2. Then, the max_depth default parameter is tuned from 3 to number integer 2 to achieve a better performance. At last on, random_State is modified to false because the dataset already has shuffled. Other factors are not change. The results have shown 84.29 for gender and 64.89 for age prediction.

4.4 Multi-Layer Perceptron Network Classifier (MLP Classifier):

Multi-Layer Perceptron Network Classifier (MLP Classifier) has derived from feed-forward artificial neural network. It uses a backpropagation method for training. The accuracy metric 86.76 percent for gender category and 66.54 percent for gender category has been obtained. Achieving this accuracy requires to modify the following parameters. Parameters are altered to deliver high performance and hidden_layer_size have been changed to 21. 21 hidden layers have been applied to avoid overfitting the model. The shuffle factor is false and random_state is zero (0) because the dataset already has shuffled.

For the training dataset, the parameter maximum numbers of iterations (max_itrr) tune to 1500. The max_iter default is 200 iteration. Tolerance for the optimization modifies to 0.012 with default 1e-4 (0.0001) and the rest features are as defaults.

5 Result and Analysis

In table 1, the five fold cross-validation displays for each classifier as above mentioned. The training data to 80 presents for training and 20 percent for testing are devied. To achieve higher accuracy or close to the real result, the system has calculated five fold cross-validation of the dataset. At that point, it computes the mean of that the cross-validation. Accuracy metric can be generalized in this text data. The ensemble model is voting for all the above classifiers. The ensemble voting has two types of voting, hard and soft. The default voting has applied which is hard voting. The system predicts the label gender or gender as a result label. This result label has the most frequency label from all four classifiers. In figure 2, the ensemble model is proposed.

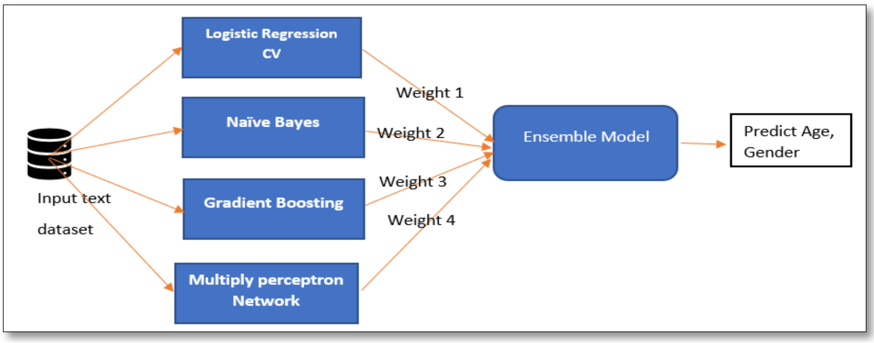

Fig. 2. Ensemble model

Table 1 declares the mean of five-fold cross-validation for 350 records of training and testing data. The best accuracy result is 88.54 percent for gender and 67.18 percent for gender by using MLP-Classifier.

The worst accuracy result displays for gender and age categories by having Logestic Regression CV and Gradient Boosting classifier. The system has 59 percent for gender and 84 percent for gender in the worst occurrence.

Moreover, the dataset with 150 hidden text files on MAPonSMS website does not introduce gender and gender. The system needs to apply the ensemble model to predict gender and gender by feeding hidden dataset. Results confirm as shown below that gender category predicts 83 percent and gender category predicts 60 percent. The joined age and gender predicts 49 percent by using the ensemble model.

Table 1. 5-fold Cross Validation Accuracy of training and testing dataset

Classifier	5-fold CV Acc-Age	5-fold CV ACC-Gender
Logistic Regression CV:	59	84
Naïve Base:	62	86
Gradient Boosting:	59	84
Multi-layer Perceptron:	65	86
Ensemble:	**67.18**	**88.54**
Result of MAPonSMS:	**60**	**83**
Baseline:	**51**	**60**

In addition, table 2 indicates mean square error (MSE) in all four models. The smallest MSE offers be the best fitted for the data points. The best MSE is 0.15 for gender by using the Gradient Boosting classifier and the Logestic Regression CV classifier. Also, the best MSE is 0.49 for age by using the ensemble model. The worst MSE is 0.25 for gender by using the Naïve base classifier and 0.62 for age by using the Logestic Regression CV classifier.

Table 2. Mean Square Error (MSE)

Classifier	MSE - Age	MSE Gender
Logistic Regression CV:	0.62	0.15
Naïve Base:	0.54	0.25
Gradient Boosting	0.50	0.15
Multi-layer Perceptron:	0.50	0.17
Ensemble:	**0.49**	**0.17**

6 Conclusions and Future Work

In this task the model has described text features with predictive influence. It can be extended across online social media. The task aims to assist companies to have better services. Many classifiers have been tried to predict gender as the binary classification and age as the multi-class classification. Finally, this system applies the ensemble model and this machine learning technique is able to predict with 60 percent accuracy metric for age category and 83 percent accuracy metric for gender category on the hidden text files. The training results are shown in table 1 in details.

This task that displays the ensemble model leads the researcher to have better results. For the future work, the researcher need to work on text prediction to achieve high accuracy than to use the pre-trained machine learning models. In addition, the system can make possibly offer in real-time classification over smart-phones or websites by improving the ensemble model.

References

1. L´opez-Monroy, A. P, Montes-y-G´omez, et al. Discriminative subprofile-specific representations for author profiling in social media, Knowledge-Based Systems, Vol 89, 2005 , Pages 134-147, ISSN 0950-7051, doi:10.1016/j.knosys.2015.06.024 (2005).
2. Farnadi, G., Sitaraman, G., Sushmita, S. et al. User Model User-Adapt Inter 26:109. doi:10.1007/s11257-016-9171-0 (2016).
3. Marquardt, J, et al. Age and Gender Identification in Social Media. CEUR Workshop Proceedings, vol. 1180 , pp. 1129–1136, doi:10.1145/1871985.1871993 (2014).
4. FIRE'18 MAPonSMS, https://lahore.comsats.edu.pk/cs/MAPonSMS/index.html, Accessed 26 Aug. 2018.
5. Ensemble Model — scikit-learn, http://scikit-learn.org/stable/modules/ensemble.html, Accessed 26 Aug. 2018.
6. Tsoumakas, G., Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. In Machine Learning: ECML 2007, Springer, pp 406–417. doi:10.1007/978-3-540-74958-5_38 (2007)
7. D. Murray and K. Durrell. Inferring demographic attributes of anonymous internet users. In Web Usage Analysis and User Profiling, Springer, pp 7–20. (2000).
8. Mislove, Alan, et al. You Are Who You Know: Inferring User Profiles in Online Social Networks. Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp 251–260 , doi: 10.1145/1718487.1718519 (2010).
9. Rao, Delip, et al. Classifying Latent User Attributes in Twitter. Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents , pp 37–44 (2010).
10. Smith, James. Gender Prediction in Social Media. (2014).
11. Burger, John D. D, et al. Discriminating Gender on Twitter. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1301–1309 (2011).
12. L´opez-Monroy, A. P., Montes-y-G´omez, M., Escalante, H. J., Villase˜nor-Pineda, L., Villatoro-Tello, E. INAOE's participation at PAN'13: Author profiling task. (2013).
13. Marquardt, James F, et al. Age and Gender Identification in Social Media. CEUR Workshop Proceedings, vol. 1180, pp. 1129–1136 (2014).
14. B. Rao, D., et al. Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user generated contents, ACM, pp 37–44. (2010).