Multi-lingual Author Profiling on SMS Messages using Machine Learning Approach with Statistical Feature Selection

D. Thenmozhi, A. Kalaivani, and Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai
{theni_d,kalaivania,aravindanc}@ssn.edu.in

Abstract. Authorship profiling is the process of identifying authors' demographic traits by analyzing their written text. It has several applications in areas such as security and forensic analysis. People use social media as a major platform to share their thoughts and ideas. However, they use more than one language for their writings. It is a challenging task to perform author profiling from multilingual short text. In this paper, we present our methodology for a task to identify gender and age of the author from their SMS messages using a machine learning approach. We have used a statistical feature selection methodology to select the features that are significantly contributing for the gender and age classifications. We have performed paired t-test to show that the improvement in performance using feature selection is statistically significant. We have evaluated our methodology using the data set given by MAPonSMS@FIRE2018[1] shared task, and have obtained 85% and 63% accuracy for gender and age classifications respectively.

Keywords: Author profiling · Machine learning · Feature selection · Text mining

1 Introduction

Authorship profiling is the process of identifying the author's demographic features such as gender, age, occupation, native language by analyzing author's text. It has several applications like security, forensic analysis, marketing and identification of fake profiles on social media. Currently, people use several platforms such as Blogs, Tweets, YouTube, Facebook, and SMS on mobiles to share their thoughts, comments, and ideas. These platforms allow multiple languages that facilitate the user to write in their native languages. Several research work have been reported for author profiling from mono-lingual text [3, 13]. A very few research work [6, 5] have been carried out for author profiling from multi-lingual text. Several approaches [13, 15, 14] have been reported on author profiling tasks. In this work, we present a language agnostic approach without any language specific processing, along with a statistical feature selection technique, namely chi-square feature selection, for author profiling task. This approach may be very useful to perform author profiling in any language. The shared task MAPonSMS@FIRE2018 focuses on multi-lingual author profiling from SMS messages. This task aims to identify the gender and age of the author based on their writings using

[1] https://lahore.comsats.edu.pk/cs/MAPonSMS/index.html

two languages namely English and Urdu. MAPonSMS@FIRE2018 is a shared Task on Multi-lingual Author Profiling on SMS (MAPonSMS) collocated with FIRE-2018 (Forum for Information Retrieval Evaluation, 2018).

2 Related Work

Author profiling tasks are being organized by PAN every year since 2013. PAN 2013 [13] to PAN 2016 [15] focused on identification of age and gender in multiple languages. In this line, PAN 2017 [14] introduced gender and language variety identification tasks. Features such as bag of words [17, 1], character n-gram [2, 12], word n-gram [4, 10], and word embeddings [8, 11] have been used for author profiling tasks. Support Vector Machine (SVM) is the popular classifier that has been used [17, 10, 8] for author profiling tasks. Deep learning methods have also been employed [8, 11, 16] to identify gender and language variations of the author. These approaches utilize different feature extraction techniques with a classifier for author profiling tasks. We present a language agnostic approach without any language specific or linguistic related processing for author profiling task just by incorporating chi-square feature selection to extract the useful features. We have used all the terms as features and employed statistical feature selection that selects the most significant features for author profiling task.

3 Proposed Methodology

We have used a supervised approach with statistical feature selection for the author profiling task. The steps used in our approach are given below.

- Preprocess the data
- Extract bag of words (BOW) features [14, 1, 17] from training data
- Apply statistical feature selection namely χ^2 for both gender and age classification
- Build a model using a classifier from the selected features of training data
- Predict class label for the test instances as “male” or “female” as gender and “15–19”, “19–24” and “25–xx” as age group using the model

The steps are explained below in detail.

3.1 Feature Extraction and Feature Selection

The given data is with UTF encoding. We have not used any language specific (or linguistic related) processing to extract the features. Thus, we did not preprocess the text with stop word removal and stemming techniques. The given text consists of CR-LF and space as line delimiter and word delimiter respectively. These delimiters are used to tokenize the text. Our approach is completely language agnostic which consider the text as sequence of Unicode bytes, so that the approach may be applied for any language. We have removed only five punctuations namely ’.’, ’,’, ’?’, ’;’, and ’:’. Such punctuations do not contribute to authorship analysis. Thus, we simply took bag of words to consider all the words in the text. However, the number of features that are

extracted may be huge. This may be reduced by applying feature selection techniques. We have observed from our previous research [20, 21] that chi-square feature selection significantly improves the performance of text analysis tasks. Hence, in the current task also, we have employed χ^2 feature selection technique to extract the useful features that are contributing towards the gender and age classifications.

Features Selection for Gender Classification: Gender Classification has two categories namely "male" and "female". Thus, we have constructed a 2×2 contingency table (Table 1) or CHI table [9, 7, 20, 21] for every feature f_x. This table contains the observed frequency (O) of f_x for every category G and $\neg G$.

Table 1. Feature-Category CHI Table for Gender Classification

	G	$\neg G$
f_x	$O(f_x, G)$	$O(f_x, \neg G)$
$\neg f_x$	$O(\neg f_x, G)$	$O(\neg f_x, \neg G)$

The expected frequencies (E) for every feature f_x can be calculated from the observed frequencies (O) using Equation 1 [9].

$$E(x,\ y) = \frac{\Sigma_{a\in\{f_x,\neg f_x\}}O(a,y)\Sigma_{b\in\{G,\neg G\}}O(b,y)}{n} \tag{1}$$

where n is the total number of instances, x represents whether the feature f_x is present or not, y represents whether the instance belongs to G or not.

The expected frequencies namely $E(f_x, G)$, $E(f_x, \neg G)$, $E(\neg f_x, G)$ and $E(\neg f_x, \neg G)$ are calculated using Equation 1 for gender classification. Then the χ^2 statistical value for each feature f_x is calculated using Equation 2 [9].

$$\chi^2_{stat} f_x = \Sigma_{x\in\{f_x,\neg f_x\}}\Sigma_{y\in\{G,\neg G\}}\frac{(O(x,y) - E(x,y))^2}{E(x,y)} \tag{2}$$

Since, gender classification is a two class problem (n = 2), χ^2 critical value with degree of freedom 1 (n-1) is 2.706 (from chi-square table) and that was chosen as a threshold for gender classification. The set of features whose χ^2_{stat} value is greater than $\chi^2_{crit(\alpha=0.01,df=1)}$: 2.706 are considered to be significant features for gender classification and those selected features are used for building a model using a classifier.

Features Selection for Age Classification: Age Classification has three categories namely "15–19", "19–24" and "25–xx". We form a 2×3 CHI table (Table 2) similar to Gender classification for every feature f_x.

The expected frequencies (E) namely $E(f_x, A_1)$, $E(f_x, A_2)$, $E(f_x, A_3)$, $E(\neg f_x, A_1)$, $E(\neg f_x, A_2)$ and $E(\neg f_x, A_3)$ can be calculated using Equations 3 [21].

$$E(x,\ y) = \frac{\Sigma_{a\in\{f_x,\neg f_x\}}O(a,y)\Sigma_{b\in\{A_1,A_2,A_3\}}O(b,y)}{n} \tag{3}$$

Table 2. Feature-Category CHI Table for Age Classification

	A_1	A_2	A_3
f_x	$O(f_x, A_1)$	$O(f_x, A_2)$	$O(f_x, A_3)$
$\neg f_x$	$O(\neg f_x, A_1)$	$O(\neg f_x, A_2)$	$O(f_x, A_3)$

Then the χ^2 statistical value for each feature f_x is calculated using the Equation 4 [21].

$$\chi^2_{stat} f_x = \Sigma_{x \in \{f_x, \neg f_x\}} \Sigma_{y \in \{A_1, A_2, A_3\}} \frac{(O(x,y) - E(x,y))^2}{E(x,y)} \tag{4}$$

Age classification is a three class problem (n=3). χ^2 critical value with degree of freedom 2 (n-1) is 4.605 (from chi-square table) and that was chosen as a threshold for age classification. The set of features whose χ^2_{stat} value is greater than $\chi^2_{crit(\alpha=0.01, df=2)} : 4.605$ are considered to be significant features for age classification and those features are used for building a model using a classifier.

3.2 Model Building and Prediction

The models for gender and age classifications are built from training data using Multi Layer Perceptron (MLP) and Multinomial Naive Bayes (MNB) classifiers by considering the selected feature set. The class label either "male" or "female" is now predicted for the test data instances by using the gender model. Similarly, the age model is used to predict the class label "15–19", "19–24" or "25–xx" for the test data instances.

4 Implementation and Results

We have implemented our methodology in Python for the Shared Tasks on Multilingual Author Profiling on SMS (MAPonSMS). The data set used to evaluate the task consists of 350 author profiles as training data and 150 profiles as test data. The ground truth for the training instances has been provided along with the data. The bag of word features are extracted from the training instances by tokenizing the text based on line and word delimiters after removing the punctuation marks, namely '.', ',', '?', ';', and ':', which do not contribute to authorship analysis. We have obtained a total of 23956 features from training data. We have implemented our χ^2 feature selection algorithms to extract the significant features for both gender and age classifications. We have obtained 1343 and 1091 features for gender and age classifications respectively. We have used Scikit–learn machine learning library to vectorize the training instances for the selected features and to implement the classifiers for the classification tasks. CountVectorizer of sklearn with selected features as vocabulary is used for vectorization. We have employed several classifiers namely, Multinomial Naive Bayes, Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), Extra Trees (ET), Ada Boost (AB), Gradient Boosting (GB), Support Vector Machines (SVM), Stochastic Gradient Descent (SGD) and Multi Layer Perceptron, and measured 10-fold cross validation to

select the best classifier for gender and age predictions. Table 3 shows the cross validation output of various classifiers without feature selection. This table also shows the cross validation output of various classifiers using χ^2 feature selection.

Table 3. 10-fold cross validation accuracies.

Classifier	Cross Validation Accuracy (%)			
	Gender Classification		Age Classification	
	Without Feature Selection	With χ^2 Feature Selection	Without Feature Selection	With χ^2 Feature Selection
MNB	86.00	**91.14**	61.60	**69.18**
GNB	73.14	74.29	59.11	41.17
DT	74.00	73.43	48.92	53.39
RF	72.86	79.43	54.63	59.73
ET	76.57	81.14	54.87	56.98
AB	83.43	84.57	56.87	52.84
GB	82.29	84.29	64.35	64.71
SVM	60.00	70.57	50.30	52.56
SGD	82.86	88.86	61.49	64.41
MLP	84.29	**92.57**	62.63	**69.52**

It is evident from the results that all the classifiers except DT perform better using selected features for gender classification. Also, all the classifiers except two namely GNB and AB perform better using selected features for age classification. To show that the feature selection significantly improved the performance for author profiling, we have performed statistical test namely paired t-test between the approach without feature selection and the approach using feature selection. We have obtained p value as 0.0015 for gender classification which is lesser than 0.05 and this shows that the improvement is statistically significant. For age classification, we have obtained p value as 0.55 which is not lesser than 0.05 and it shows that the improvement is not statistically significant. However, we have chosen MLP as the best classifier, because it improved the accuracy by more than 6.89% for feature selection approach. More hidden layers may be added in future to incorporate deep learning approach with enhanced data set. To show that the variations between folds are not significant, we have performed one-way Anova test (statistical test) for 10 random runs on training data. Table 4 shows the 10-fold cross validation accuracies we have obtained using MLP and selected features for 10 random runs.

We have obtained a p-value of 0.991 for one-way Anova test for gender classification based on cross validation of 10 random runs. This is greater than 0.05 which shows that the variations on different folds are not statistically significant. Similarly, we have performed one-way Anova test for age classification based on cross validation of 10 random runs. Table 5 shows the 10-fold cross validation accuracies we have obtained using MLP with selected features for age classification.

For age classification, we have obtained a p-value of 0.999 for one-way Anova test based on cross validation of 10 random runs. This is greater than 0.05 which shows that the variations on different folds are not statistically significant in age classification.

Table 4. 10-fold cross validation accuracies using MLP and feature selection for gender classification.

Fold	Cross Validation Accuracy (%)									
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	97.14	94.29	97.14	94.29	97.14	94.29	97.14	97.14	97.14	94.29
2	94.29	97.14	94.29	97.14	94.29	97.14	94.29	94.29	97.14	94.29
3	85.71	82.86	85.71	85.71	80.00	88.57	82.86	82.86	85.71	88.57
4	91.43	91.43	94.29	91.43	91.43	91.43	91.43	94.29	94.29	91.43
5	97.14	94.29	97.14	94.29	94.29	97.14	97.14	97.14	97.14	94.29
6	97.14	91.43	91.43	91.43	94.29	94.29	94.29	94.29	91.43	91.43
7	94.29	94.29	100.00	97.14	94.29	97.14	97.14	97.14	97.14	97.14
8	97.14	94.29	97.14	97.14	97.14	94.29	97.14	97.14	97.14	91.43
9	91.43	94.29	91.43	94.29	91.43	94.29	94.29	91.43	91.43	91.43
10	88.57	88.57	88.57	88.57	88.57	88.57	88.57	88.57	88.57	88.57

Table 5. 10-fold cross validation accuracies using MLP with feature selection for age classification.

Fold	Cross Validation Accuracy (%)									
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	63.89	61.11	63.89	58.33	63.89	61.11	61.11	63.89	58.33	61.11
2	61.11	61.11	61.11	58.33	66.67	61.11	61.11	61.11	63.89	61.11
3	75.00	66.67	63.89	63.89	63.89	66.67	69.44	69.44	69.44	72.22
4	77.78	75.00	80.56	75.00	80.56	75.00	77.78	77.78	75.00	77.78
5	63.89	69.44	69.44	63.89	72.22	66.67	75.00	75.00	69.44	72.22
6	75.00	69.44	69.44	69.44	66.67	69.44	72.22	69.44	66.67	75.00
7	79.41	85.29	82.35	82.35	82.35	82.35	82.35	82.35	85.29	85.29
8	64.71	64.71	64.71	64.71	67.65	64.71	64.71	67.65	64.71	67.65
9	60.61	60.61	63.64	66.67	63.64	66.67	60.61	63.64	63.64	57.58
10	72.73	72.73	69.70	75.76	75.76	72.73	75.76	72.73	78.79	66.67

Hence, we have chosen MLP to build models for gender and age classifications. These models are used to predict gender and age for the test instances.

We have submitted two runs for the shared task based on the top two classifiers namely MLP and MNB. For the first run, we have used MLP classifier to build models for gender and age classifications and predicted the class labels for the 150 test instances. MNB classifier was used for the second run. The code to reproduce the results is available in Github link[2].

Table 6 shows the result we have obtained for these two runs. Our machine learning approach with statistical feature selection ranked second among the approaches presented by several teams.

Table 6. Results comparison.

Team	MAPonSMS Performance (Accuracy)		
	Gender	Age	Joint
Sharmila Devi et al.	0.87	0.65	0.57
Our Approach	0.85	0.63	0.52
Ali Nemati	0.83	0.60	0.49
Deepanshu Gaur	0.75	0.64	0.47
Kosmajac and Keselj	0.74	0.59	0.43
Oscar Garibo	0.77	0.57	0.43
Imran and Iqbal	0.73	0.53	0.38
Safdar et al.	0.69	0.53	0.35
Baseline	0.60	0.51	0.32
Sittar and Ameer	0.55	0.37	0.23

Since the ground truth for test data is not provided by the organizers, we are unable to measure other metrics namely precision, recall, F-measure and rejection and perform detailed analysis including statistical tests such as paired t-test and McNemar test on the test data results.

5 Conclusions

We have presented a methodology for multi lingual author profiling to identify the gender and age of the author based on his/her writings in SMS. We have proposed a machine learning approach which is language agnostic with statistical feature selection to identify the gender as either “male” or ”female” and to identify the age as “15–19”, “19–24” or “25–xx” without any language specific processing. In our method, we have selected features using χ^2 method from bag of word features and feature vectors are constructed from training data for the selected features. We have employed various classifiers to build the models for gender and age classification. We have performed

[2] https://github.com/ThenmozhiDurairaj/SSN_SMS

a statistical test namely paired t-test which shows that the approach using feature selection significantly improved the performance for gender classification. The best two models namely MLP and MNB were chosen based on the cross validation accuracy to build the models. We have performed one-way Anova test to show that the variation among fold results are not statistically significant. We have used the data set given by MAPonSMS@FIRE2018 shared task to evaluate our methodology. The performance has been measured using the metric accuracy. We have obtained the accuracy of 85% and 63% for gender and age predictions respectively. We have shown from our earlier research [18, 19] that the clause-based features (predicates) improve the performance in information extraction and classification tasks. The performance of author profiling may also be improved further if we incorporate the predicate information of the text as features.

References

1. Adame-Arcia, Y., Castro-Castro, D., Bueno, R.O., Muñoz, R.: Author profiling, instance-based similarity classification. In: CEUR Workshops Proceedings. vol. 1866 (2017)
2. Alrifai, K., Rebdawi, G., Ghneim, N.: Arabic tweeps gender and dialect prediction. Cappellato et al.[13] (2017)
3. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM **52**(2), 119–123 (2009)
4. Ciobanu, A.M., Zampieri, M., Malmasi, S., Dinu, L.P.: Including dialects and language varieties in author profiling. arXiv preprint arXiv:1707.00621 (2017)
5. Fatima, M., Anwar, S., Naveed, A., Arshad, W., Nawab, R.M.A., Iqbal, M., Masood, A.: Multilingual sms-based author profiling: Data and methods. Natural Language Engineering pp. 1–30 (2018)
6. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Information Processing & Management **53**(4), 886–904 (2017)
7. Janaki Meena, M., Chandran, K.: Naive bayes text classification with positive features selected by statistical method. In: In International Conference on Autonomic Computing and Communications, ICAC 2009. pp. 28–33. IEEE (2009)
8. Kodiyan, D., Hardegger, F., Neuhaus, S., Cieliebak, M.: Author profiling with bidirectional rnns using attention with grus: notebook for pan at clef 2017. In: CLEF 2017 Evaluation Labs and Workshop–Working Notes Papers, Dublin, Ireland, 11-14 September 2017. vol. 1866. RWTH Aachen (2017)
9. Li Yanjun, C.L., Chung, S.M.: Text clustering with feature selection by using statistical data. IEEE Transactions on Knowledge and Data Engineering **20**(5), 641–652 (2008)
10. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. Working Notes Papers of the CLEF (2017)
11. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author profiling with word+ character neural attention network. In: CLEF (Working Notes) (2017)
12. Ogaltsov, A., Romanov, A.: Language variety and gender classification for author profiling in pan 2017. Cappellato et al.[13] (2017)
13. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)

14. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF (2017)
15. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. pp. 750–784 (2016)
16. Sierra, S., Montes-y Gómez, M., Solorio, T., González, F.A.: Convolutional neural networks for author profiling in pan 2017. In: CLEF (Working Notes) (2017)
17. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D.: Gender and language variety identification with microtc. Cappellato et al.[13] (2017)
18. Thenmozhi, D., Aravindan, C.: An automatic and clause based approach to learn relations for ontologies. The Computer Journal **59**(6), 889–907 (2016)
19. Thenmozhi, D., Aravindan, C.: Paraphrase identification by using clause based similarity features and machine translation metrics. The Computer Journal **59**(9), 1289–1302 (2016)
20. Thenmozhi, D., Mirunalini, P., Aravindan, C.: Decision tree approach for consumer health information search. In: FIRE (Working Notes). pp. 221–225 (2016)
21. Thenmozhi, D., Mirunalini, P., Aravindan, C.: Feature engineering and characterization of classifiers for consumer health information search. In: Forum for Information Retrieval Evaluation. pp. 182–196. Springer (2016)