Multi-lingual Author Profiling Using Stylistic Features

Abdul Sittar*, Iqra Ameer*

*COMSATS University Islamabad, Lahore Campus, Pakistan
abdulsittar72@gmail.com
Iqraameer133@gmail.om

Abstract. Author profiling is the identification of an author's traits by examining the text written by an author. Author profiling has many useful applications in security, criminal, marketing, identification of false accounts on shared communication websites, etc. We submitted our system to the FIRE'18-MAPonSMS (Multi-lingual Author Profiling on SMS), a shared task to classify the attributes of an author like gender and age group from multilingual text specifically English +Roman Urdu. Roman Urdu is common language specifically in SMS messaging, Facebook posts/comments and chats blog of games etc. Our presented system is based on 29 different stylistic features. On the training dataset, we have achieved best Accuracy = **73.714**, for gender while using all 14-language-independent features together and Accuracy = **58.571** for age group by using all 29 features together. We obtained Accuracy = **0.55** and **0.37** on testing data for both gender and age respectively.

Keywords: Author profiling, Multi-lingual text, Machine Learning, Roman Urdu, Stylistic.

1 Introduction

Author profiling (AP) is the task of determining the writer's traits, like age, gender, profession, personality types and mother language by analyzing the written document. Due to the heap of information on social networks, it became essential to classify user's characteristics. This chore has diverse applications in forensics, security, and in marketing fields [1]. For example, by using forensics and terrorism applications, we can decrease the search space for the suspicious writer. In marketing point of view, these facts can be essential to predict and target specific customers and to form new strategies according to consumer's interest and preferences.

Recent tendencies in the domain include Multi-lingual AP [2], that is, the Multi-lingual document: "occurrence of more than two languages in a text document" [3]. Multi-lingual AP settings match the necessities of a real-life situation of security applications when the produced text by the authors can belong to a mixture of different languages from the texts under examination.

Following the emerging field, the FIRE'2018 shared task on Age and Gender identification in SMS based author profiles (MAPonSMS), provided the training and testing

corpora that were composed of multi-lingual (English +Roman Urdu) SMS based documents.

Since the provided training data was in multi-lingual, one of our primary objectives was to determine how our proposed technique is performing in the multi-lingual text. The author profiling is supervised document classification task. We carried out different experiments, i.e. (i) using all 29 stylistic features, (ii) using all 14-language independent stylistic features and (iii) using individual language independent stylistic features.

In recent period, deep-learning methods [4], such as word, character, document-embedding and word approaches [5], have been used for this specific problem; still, linear models perform well, as they seem to be more robust in picking up stylistic information in the author's writing. So, we applied frequently used linear machine-learning (ML) approaches.

The document is prepared as follows. Section 2 discusses related work has been done in this domain. Section 3 details about our approach for the FIRE'2018 shared the task. Results and their analysis in section 4. The final section is five which concludes the paper.

2 Related Work

The International PAN Competitions made remarkable progress in author profiling, especially for the gender and age identification tasks [1, 6, 7, 8, 9]. In PAN-2017, 22 teams participated, and traditional machine-learning algorithms were used by most of them [9], like Logistic Regression [3, 13] or SVM [10, 11, 12]. Some of them applied deep-learning techniques, especially word and character embedding [4, 13, 16], which are considering competing techniques, but still, results are not up to the mark for the Author Profiling task.

As researchers are attracted towards the multi-lingual settings, however, there is only one research study found in the literature that considering the same genre author profiling task for multi-lingual text. [2] worked on multi-lingual corpus based on Roman Urdu and English Facebook posts and comments for same genre author profiling. Content-based and stylistic features were explored in this study and 10-fold cross-validation was used for evaluation. They have achieved *Accuracy = 0.875* on multi-lingual corpus by word uni-gram, char 3-gram, and char 8-gram content-based approach. By using the word bigram content-based approach, they got *0.750 accuracy* for the age group classification task.

3 Proposed Approach

3.1 Stylistic Feature Set

There are three comprehensive categories have been used for automatic identification of an author's traits: (1) content-based methods – that aim to detect characteristic

of a writer by using content of the text, (2) stylistic-based methods – which try to predict a writer's demographics traits by analyzing writing style of the writer, and (3) topic-based methods – are applied to classify characteristics of an author by using debated topics in the text.

For the FIRE'18-MAPonSMS Author Profiling competition, our system[1] is based on different statistical features. As this year training data of FIRE'18-MAPonSMS is based on the multi-lingual SMS messages, i.e. Roman Urdu and English. This systematic investigation purposes to detect some language independent stylistic features, which are likely to perform in multi-lingual text. List of language-independent as follows: Avg. Word Length, Avg. Sentence Length, %age of Words with Six and More Letters, %age of Words with Two and Three Letters, %age of Question Sentences, %age of Semicolons, %age of Punctuations, %age of Comma, %age of Short Sentences, %age of Long Sentences, %age of Capitals, %age of Colons, %age of Digits, and %age of Full Stop. However, further are not language independent: Avg. Syllables Per Word, %age of Pronouns, %age of Prepositions, %age of Coordinating Conjunctions, %age of Articles, %age of Words with One Syllable, %age of Words with Three Plus Syllables, %age of Adjectives, %age of Determiners, %age of Interjections, %age of Modals, %age of Nouns, %age of Personal Pronouns, %age of Verbs, and %age of Adverbs.

We can observe that the list mentioned above of features purposes to catch different stylistic facts from a multi-lingual written text, which can be useful to uncover the age group and gender of an unknown author.

4 Experimental Setup

The focus of the FIRE'18-MAPonSMS shared task 2018 is on two author attributes (1) age and (2) gender identification on the multi-lingual text. The organizers provided us training dataset composed of multi-lingual (Roman Urdu and English) SMS text messages. There are two classes for gender (male, female) and three classes for the age group (15-19, 20-24 and 25-xx).

4.1 FIRE'18-MAPonSMS Training and Test Dataset for Author Profiling

In the training dataset, for the gender classification, we have 210 male while 140 female labeled text documents. On the other hand for age group classification 108 text documents are in the 15-19 age group, 176 are in the 20-24 category, and 66 are in 25-xx age group. However, 150 files were provided for the testing phase.

[1] The implementation (source code) details of our approach is provided in a repository at https://github.com/abdulsittar/Multilingual-Author-Profiling

4.2 Evaluation Methodology

Author profiling classification problem is handled as a supervised ML problem. For detection of the age group, there are multi–groups problem and objective is to classify the age amongst 3-groups: (1) 15-19, (2) 20–24 and (3) 25–xx. For the gender categorization, there are binary-groups and objective is to differentiate between 2 groups: (1) female as well as the (2) male.

10–fold cross-validation was used in experiments to evaluate the performance of our model. We conducted our experiments by using four different ML algorithms named Naive Bayes, J48, Random Forest and Logistic Regression. Implementation of WEKA was used for these algorithms. The scores produced using the stylistic features are manipulated as input to above-stated ML algorithms.

4.3 Evaluation Measure

As suggested by the organizing team of the FIRE'18-MAPonSMS shared the task, the performance of the submitted automatic system for the age and gender was measured using accuracy. Accuracy is described as the proportion of the correctly classified predictions $\boldsymbol{p_c}$ out of all the predictions $\boldsymbol{p_a}$ made.

$$Accuracy = \frac{p_c}{p_a}$$

5 Results and Analysis

For all the tables shown in this results and analysis section, next mentioned terminologies are used. The "Classifier" indicates the ML algorithm which we have applied to generate the numeric scores (NB → Naive Bayes, RF → Random Forest, LR → Logistic Regression). Best results are highlighted in bold typeface. We performed three sets of experiments: (1) performance on all 29 features, (2) performance on all 14 language independent features (see section 3.1) and (3) performance on individual language independent features (by using every single feature).Results on Training Dataset

Table 1 shows the scores using all 29 stylistic features for both groups, i.e., age and gender. Using all features together we obtained best results by using Random Forest, Accuracy = **73.714** for the gender and 53.142 for the age group classification. Table 2 depicts the results for all 14-language independent stylistic features collectively. We achieved the best results for gender (Accuracy = 72.000) using Random Forest classifier and for the age (Accuracy = **58.571**) by using Logistic Regression.

Table 1. Results using all 29 stylistic features

Classifiers	Age (Accuracy)	Gender (Accuracy)
NB	51.428	58.285
RF	56.571	**72.000**
LR	**58.571**	66.000
J48	49.142	70.571

Table 2. Results using all 14-language independent features

Classifiers	Age (Accuracy)	Gender (Accuracy)
NB	47.428	63.428
RF	**53.142**	**73.714**
LR	52.857	70.857
J48	46.285	69.428

Table 3. Results using language independent features individually by Random Forest

Features	Age (Accuracy)	Gender (Accuracy)
Avg. Word Length	37.714	51.714
Avg. Sentence Length	38.857	44.000
%age of Words with Six and More Letters	42.285	52.000
%age of Words with Two and Three Letters	37.142	51.142
%age of Question Sentences	41.714	56.285
%age of Semicolons	50.185	64.285
%age of Punctuations	35.428	**64.857**
%age of Comma	47.142	52.857
%age of Short Sentences	50.085	60.000
%age of Long Sentences	**50.285**	60.000
%age of Capitals	41.714	48.000
%age of Colons	40.857	56.000
%age of Digits	43.142	56.571
%age of Full Stop	43.142	58.285

Table 3 displays the results using 14-language independent stylistic features individually (for every single feature). RF was performing better on both age and gender identification; therefore we are only reporting the results on single language-independent features by using RF.

The best results are obtained when "%age of Punctuations" single stylistic feature is used for gender (Accuracy = 64.857) and "%age of Long Sentences" for age (Accuracy = 50.285). This indicates that in SMS multi-lingual messages, one of the gender prefers Punctuations than the other, while one of the age groups prefers longer messages than others.

Overall concerning algorithms, for age group identification, the best scores are obtained using two classification algorithms named as RF and LG. For gender estimation, best sores are achieved by RF algorithm. This shows that the RF algorithm is suitable if we give a collection of attributes as an input to the algorithm in the classification problem.

5.1 Results on Test Dataset

We obtained *Accuracy* = ***0.55*** and ***0.37*** on testing data for both gender and age respectively, which is below the baseline (baseline for gender = 0.60 and age = 0.51). Joint estimation of our model for both age and gender is *0.23*.

6 Conclusion

Correct profiling of an unknown author is getting a reputation for security point of view, investigation of criminal activities and the market research opinion. In this paper, we have participated in our approach in the FIRE'18-MAPonSMS author profiling shared task on age and gender identification in multi-lingual text. We have shown how the stylistic features and machine learning techniques enable an automatic system to determine different characteristics of an unknown author efficiently.

We have considered the stylistic features to uncover the traits of an author on the multi-lingual corpus. We implemented 29 stylistic features and performed three different set of experiments, i.e., compared the results by using all 29 features, analyzed the scores for 14-language independent features altogether and at the end using single language-independent features. We observed that best results are achieved when we used all 29 features together for gender (Accuracy = **73.714**) identification by Random Forest and for the age (Accuracy = **58.571**) group when used all 14-language independent features by Logistic Regression classifier.

References

1. Rangel, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. In CLEF http://fire.irsi.res.in/fire/2018/home, last accessed 2018/07/27.
2. Mehwish Fatima, Komal Hasan, Saba Anwar, Rao Muhammad Adeel Nawab (2017), "Multilingual author profiling on Facebook", Information Processing & Management, Elsevier, pp: 886 - 904, Vol: 53, Issue: 4, Standard: 0306-4573.
3. Meylaerts, R., 2010. Multilingualism and translation. Handbook of translation studies 1, 227{230.
4. Sebastian Sierra, Manuel Montes-y-Gómez, Thamar Solorio, and Fabio A. González. 2017. Convolutional Neural Networks for Author Profling. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
5. Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and

Alexander Gelbukh. 2017. Author Profiling with Doc2vec Neural Network Based Docment Embeddings. In Proceedings of the 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Vol. 10062. Part II, LNAI, Springer, Cancún, Mexico, 117–131.

6. Rangel, F., Rosso, P., 2013. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of Their Writing Style. In: Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013). Vol. 1096. CEUR Workshop Proceedings, Turin, Italy, pp. 34{46.
7. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G., 2013. Overview of the Author Profiling Task at PAN 2013. In: CLEF 2013 Evaluation Labs and Workshop Working Notes Papers. Valencia, Spain.
8. Rangel, F., Rosso, P., Potthast, M., Stein, B., 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings. CLEF and CEUR-WS.org.
9. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B., 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum. CEUR-WS.org, vora - Portugal, pp. 750{784.
10. A. Pastor Lopez-Monroy, Manuel Montes-y-Go ´ mez, Hugo Jair-Escalante, Luis ´ Villasenõr Pineda, and Thamar Solorio. 2017. Social-Media Users can be Profled by their Similarity with other Users. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
11. Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. 2017. Language- and Subtask-Dependent Feature Selection and Classifer Parameter Tuning for Author Profling. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
12. Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, and Daniela Moctezuma. 2017. Gender and Language-Variety Identifcation with microTC. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
13. Andrey Ignatov, Liliya Akhtyamova, and John Cardiff. 2017. Twitter Author Profling Using Word Embeddings and Logistic Regression. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
14. Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. 2017. PAN 2017: Author Profling - Gender and Language Variety Prediction. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
15. Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba. 2017. Subword-based Deep Averaging Networks for Author Profling in Social Media. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.
16. Sebastian Sierra, Manuel Montes-y-Gómez, Thamar Solorio, and Fabio A. González. 2017. Convolutional Neural Networks for Author Profling. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.