MAPonSMS'18: Multilingual Author Profiling using Combination of Features

Ramsha Imran[1] and Muntaha Iqbal[2]

[1] COMSATS University Islamabad, Lahore Campus, Pakistan.
ramsha.rabbiya@gmail.com
[2] Virtual University of Pakistan, Lahore, Pakistan.
muntaha.iqbal@vu.edu.pk

Abstract. Author profiling is used to identify different personality traits of an author from the given text. Different applications of author profiling are marketing, security, forensics, spam detection etc. In this paper, we describe the methodology used in author profiling for the FIRE'18-MAPonSMS shared task. Our main objective is to identify age and gender of an author from the multilingual (English and Roman Urdu) text. We incorporated different style, vocabulary and emoticon based features to build our proposed system using the training data provided by the task organizers. Furthermore, we used a combination of different extracted features and investigated different ML algorithms (i.e. Random Forest, Naïve Bayes and Support Vector Machines) for the classification problem. Results show that our proposed system achieved 73% accuracy in gender identification and 53% accuracy in age prediction tasks.

Keywords: Gender Identification · Age Prediction · Author Profiling

1 Introduction

Author profiling task is to identify different personality traits of an author i.e. age, gender, personality type, native language etc. from their social dialect. In other words, author profiling is to detect different attributes of an author's personality from the written text. The task can be further categorized into monolingual and multilingual author profiling. In monolingual, the profile text is in one language while in multilingual, the profile text is in more than one language. Author profiling has vast applications in marketing, security and for the detection of fake profiles.

In this research work, we used style, vocabulary and emoticon based features to detect age and gender of a person from a multilingual author profiling corpus provided for the MAPonSMS'18 task. Our major contribution are as follows:

- We used 83 distinct features from the multilingual author profiling corpus, grouped into three main categories (1) style based features (2) vocabulary based features and (3) emoticon based features.
- We used an ensemble approach to combine different features for the classification task.

- For classification, we investigated three Machine Learning (ML) algorithms i.e. Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB).
- We performed better than the baseline and achieved 73% accuracy in gender identification, 53% accuracy in age prediction and an overall 38% joint accuracy.

The rest of paper is organized as follows. Section 2 summarizes existing work on author profiling. Section 3 explains the multilingual corpus and its statistics. We then present the proposed system and details of the three types of features in Section 4. Section 5 discusses results and their analysis and Section 6 concludes the paper.

2 Related Work

With the evolution of mobile technology, the Short Message Service (SMS) is considered as one of the most widely used source of communication. Extracting useful information from SMS data is very crucial task and it attracted many researches.

In recent past, a lot of work has been done on author profiling using data from social media i.e. Twitter and Facebook data, but less work has been done on multilingual author profiling using SMS. In [1], authors proposed stylistic and content based features for multilingual author profiling using Facebook data.

Argamon et al. in [2] also investigated the task of author profiling and categorized the features into two main classes (1) style based features and (2) content based features.

In [3], authors used dictionary of emotions (consists of 344 entries), dictionary of contractions (consists of 65 entries) and dictionary of unrecognized or misspelled words for the task of author profiling.

Hernández et al. in [4] proposed linguistic markers, slangs, emotions and semantic similarity for the identification of author's age and gender from the multilingual text.

Cruz et al. in [5], De-Arteaga et al. in [6], Flekova and Gurevych in [7], Lim et al. in [8], Meina et al. in [9], Patra et al. in [10] and Santosh et al. in [11] make use of different stylistic features for the task of multilingual author profiling.

As can be seen from the above discussion, researchers have used Facebook and Twitter data for the author profiling task but less work has been done for multilingual author profiling using SMS data. Fatima et al. in [12], recently developed a corpus for multilingual author profiling using SMS data called SMS-AP-18. SMS-AP-18 corpus consist of a total 810 profiles of authors with different demographics i.e age, gender, native language native city, personality type, qualification and occupation. From these 810, 610 are male profiles (with 64,968 messages) and 200 female profiles (with 19,726 messages). These profiles were categorized in three age groups i.e. 15-19, 20-24 and 25-xx. They used 64 stylistics features and 12 content based features and reported 97.5% joint accuracy for gender identification task.

3 Corpus

For the task of FIRE'18-MAPonSMS organizers provided a total of 500 author profiles from the SMS-AP-18 corpus [12]. Out of these 500 profiles, 350 were for training and 150 for testing. The training data was distributed into two gender groups i.e. male and female, and three age groups i.e. 15-19, 20-24 and 25-xx. The training data consists of 141 female and 209 male profiles. Moreover, 108 profiles were from age group of 15-19, 176 from age group of 20-24 and 66 were from age group of 25-xx. The test data contains 150 profiles, which was provided by the organizers to test our system and report accuracies.

4 System Description

We propose an author profiling system to detect age and gender from the multilingual text for the MAPonSMS'18 task. Our system has three main steps 1) feature extraction, 2) combination of features and 3) classification.

4.1 Feature Extraction

In the feature extraction step, we extracted three types of features i.e. style based features, vocabulary based features and emoticon based features. Using these three types, a total of 83 features were extracted [3]. In the following sections, we describe each of the feature types in detail.

Style based Features We used a total of 24 style based features which were extracted from the training corpus. This set of 24 features has 17 special characters features and 7 count of occurrences features. The details are given below.

- number of tokens
- number of distinct words
- average words per line
- number of lines
- number of capital letters
- number of small letters
- number of digits used

- number of commas [,]
- number of full stops [.]
- number of commercial at [@]
- number of start braces [(]
- number of end braces [)]
- number of exclamation marks [!]
- number of dashes [-]

[3] Before extracting these features, the text was preprocessed by converting it into lower case.

- number of question marks [?]
- number of percentages [%]
- number of ampersands [&]
- number of underscores [_]
- number of hashes [#]
- number of equal-tos [=]
- number of colons [;]
- number of semicolons [;]
- number of spaces []]
- number of forward slashes [/]

Vocabulary Based Features We also used a number of vocabulary based features i.e. abbreviations, academic terms, contractions and slang words. Each of these are explained in the next sections.

Table 1. List of abbreviations

Word	Language	Actual Word	English Translation
aoa	Roman Urdu	Asslam-o-alikum	Hello
w.s	Roman Urdu	Walaikum assalam	Peace be with you
w8	English	Wait	Stand by
ty	English	Thank you	Gratitude
tc	English	Take care	Lookafter
msg	English	Message	Inform
y	English	Why	How
ia	Roman Urdu	In sha Allah	By the will of God
os	Roman Urdu	Os	He/she
lol	English	Laugh out loud	Laughing
wow	English	Wow	Wonderful
pls	English	Please	Requesting
btw	English	By the way	Incidentally
hmmm	Roman Urdu	Hm	Ok

Abbreviations Abbreviations are defined as shortened form of a word. In SMS conversations, people prefer to use short form of words. Considering this fact, we used a list of Roman Urdu and English abbreviations to extract 14 features from the training corpus. As the corpus data comes from the SMS messages, people often use different forms of a single word (abbreviations) in SMS communication. Therefore, we further normalized these word variations to their single form. For

example, “aoa” is abbreviated in many ways i.e. “a.o.a”, “a.a.w.w” and “a.w”. We replaced all of these variants with a single abbreviation “aoa”.

Table 1 lists the abbreviations along with their English translations that we extracted from the training corpus.

Academic terms We also used the academia related terms as features for the classifier. Academic terms such as mam (madam), sir, assignment, class, quiz, office, teacher, test, uni (university) and clg (college) were extracted from the training corpus. Similar to the abbreviations (See Section 4.1) different spelling variants of a word were normalized to their single word form e.g. mam was used for ma’am, maam and ma’m.

Contractions Contractions are shorten form of words excluding some internal letters. We used the most commonly used 15 contractions as features from the training corpus. The list of contractions is shown below.

- I’m – I am
- you’re – You are
- we’re – We are
- they’re – They are
- Who’re – Who are
- I’ll – I shall
- She’ll – She will
- It’ll – It will
- Can’t –Can not
- Don’t – Do not
- Isn’t – Is not
- Aren’t – Are not
- Doesn’t – Does not
- Wasn’t – Was not
- Didn’t – Did not

Slang Words Slangs are those informal words that people use in their conversation especially on the Web and different social media forums. We developed a small dictionary of these commonly used slang words in Roman Urdu language and their English translations. Some of the examples from the dictionary are shown in below.

- chuss – Unpleasant
- scene – Situation
- chill – Enjoyment
- pagal – Mad
- miss – Remembering
- mood – Temper
- burger – Impersonate
- bhai – Brother

- drama – Acting
- fit – Healthy
- chawal – Stupid
- yar – Friend
- dfa – Get lost

It should be noted that our dictionary also converts different variations of these Roman Urdu words into their normalized forms.

Emoticon based Features: Emoticons are small images or icons that are used to express moods, expressions and feelings in day-to-day communication. We collected seven main types of emoticons. The following list shows all types of emoticons extracted from the training corpus.

- Happy – :)
- Sad – :(
- Cry – :'(
- Unsure – :/
- Squint – -_-
- Kiss – :*
- Wink – ;)

4.2 Combination of Features

- **All style features:** In this combination set, we combined all style based features together (See Section 4.1). The set consists of 24 features which were used together to train the classifier for prediction of age and gender.
- **All vocabulary features:** For the second set, we combined all vocabulary based features (See Section 4.1). The set consists of 14 abbreviations, 10 academic terms, 13 slang words and 15 contractions.
- **All emoticon features:** The third set is the combination of all emoticon features which are 7 in number (See Section 4.1).
- **All features combined:** In this combination set, we combined all style, vocabulary and emoticon features together for the classification task. The combination set contains all 83 features.

4.3 Classification:

For the classification task of predicting age and gender from the multilingual text, we explored different ML classification algorithms including Random Forest (RF), Support Vector Machine (SVM) and Naïve Bayes (NB). The feature sets from Section 4.2 were provided as input to train the classifier.

Table 2. Accuracy on Training Data of MAPonSMS Subtask 2018

Combination	Gender			Age		
	RF	SVM	NB	RF	SVM	NB
Style	75.71%	65.42%	67.14%	52.28%	54.00%	53.42%
Vocabulary	64.85%	63.42%	56.28%	51.42%	51.42%	38.28%
Emoticon	67.14%	62.57%	66.28%	48.85%	51.71%	32.28%
All features	**78.57%**	71.71%	69.42%	52.57%	**55.42%**	46.00%

5 Results and Discussion

Based on the feature sets (See Section 4.2) and ML algorithms (See Section 4.3), we trained separate models for both gender identification and age prediction tasks. We applied 10-fold cross validation and evaluated the performance on the basis of accuracy.

From Table 2 it can be seen that RF classifier has outperformed in gender identification with 'All features' set with the highest accuracy of 78.57%. This shows that combining all distinct features has resulted in better predicting gender of an author from the multilingual text. The gender prediction also highlights that the max contribution to the best accuracy is by 'Style' based features (accuracy = 75%). On the other hand, for age prediction, SVM performs overall best with the highest accuracy of 55.42% again using 'All features'. Furthermore, all the three individual sets of features used in the age prediction task has reported almost the same results. Conclusively, we can say that the ensemble approach that we have used to combine the different individual features together has resulted in the better performance of different classifiers.

We submitted our system to the MAPonSMS'18 task and it was evaluated on the test dataset. Our system performed better than the baseline with the achieved results are 73% accuracy in gender identification task while 53% accuracy in age prediction task.

6 Conclusion

In this paper we used style, vocabulary and emoticon based features for gender identification and age prediction tasks on the multilingual author profiling corpus provided by the MAPonSMS'18 task. We trained a number of ML classifiers i.e. RF, SVM and NB using different combinations of the above mentioned sets of features. Though our system (gender = 73.57%, age = 53.42% and joint 38%) beat the baseline accuracies (gender = 0.60, age = 0.51 and joint 0.32), in

comparison to the top rank system, there is still much room for improvement. If the task is offered in the future, we will explore more sophisticated methods to improve our results.

References

1. Fatima, M., Hasan, K., Anwar, S., Nawab, R.-M.-A.: Multilingual author profiling on Facebook.Information Processing & Management **53**(4), 886–904 (2017)
2. 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)
3. Aleman, Y., Loya, N., Ayala, D.-V, Pinto, D.: Two Methodologies Applied to the Author Profiling Task. In: CLEF (Working Notes), Citeseer, (2013)
4. Hernández, D., Guzmán-Cabrera, R., Reyes, A., Rocha, M.-A.: Semantic-based Features for Author Profiling Identification: First insights. In: Proceedings of CLEF, (2013)
5. Cruz, F.-L., Rafa H,-R., Ortega, F. J.: ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling. In: Proceedings of CLEF, (2013)
6. De-Arteaga, M., Jimenez, S., Mancera, S., Baquero, J.: Author profiling using corpus statistics, lexicons and stylistic features. In: Proceedings of CLEF, (2013)
7. Flekova, L., Gurevych, I.: Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In: Proceedings of CLEF, (2013)
8. Lim, W.-Y., Goh, J., Thing, V.-L.: Content-centric age and gender profiling. In: Proceedings of CLEF, (2013)
9. Meina, M., Brodzinska, K., Celmer, B., Czokow, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based classification for author profiling using various features. In: Proceedings of CLEF, (2013)
10. Patra, B.-G., Banerjee, S., Das, D., Saikh, T., Bandyopadhyay, S.: Automatic author profiling based on linguistic and stylistic features. In: Proceedings of CLEF, (2013)
11. Santosh, K., Bansal, R., Shekhar, M., Varma, V.: Author profiling: Predicting age and gender from blogs. In: Proceedings of CLEF, (2013)
12. 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, 1–30 (2018)