KCE_DAlab@MAPonSMS-FIRE2018: Effective Word and Character-based Features for Multilingual Author Profiling

Sharmila Devi V[1], Kannimuthu S[1], Ravikumar G[2], and Anand Kumar M[3]

[1]Department of Information Technology, Karpagam College of Engineering, Coimbatore,
[2]Department of Computer Science and Engineering, CIET, Coimbatore
[3]Department of Information Technology, National Institute of Technology-Karnataka, Surathkal
sharmiladevi1002@gmail.com

Abstract. This paper illustrates the work on identification of gender and age-group in Multilingual Author Profiling on SMS messages (MAPonSMS) shared task conducted in the Forum for Information Retrieval and Evaluation (FIRE 2018). To develop the Multilingual Author profiling system, the organizers released the training corpus which includes multilingual (Roman Urdu and English) SMS messages and its corresponding profiles. In gender identification, a profile may be either male or female. The author's age-group fall into one of the three categories: 15-19, 20-24, 25-xx. We have developed the author profiling system[1] using the word and character-based Term Frequency & Inverse Document Frequency (TFIDF) features and classify with Support Vector Machine classifier. The proposed system achieved the State-of-Art performance in the multilingual author profiling on SMS task. The accuracy obtained for identification of age-group is 65% and for gender, it is 87%. The performance is also evaluated jointly where the accuracy gained is 57%. We also experimented with the system by changing different parameters and report the cross-validation accuracy.

Keywords: Author profiling, Support Vector Machine, TFIDF, Machine Learning, Word and Character-based features, Multilingual SMS.

1 Introduction

In our day-to-day life, social media has provided various ways to share the information through the faster growth of the electronic devices. Transferring information and sharing perceptions into the world wide web is becoming an inevitable scenario. The social networking sites like Facebook, Twitter, blogs, newsgroups, etc are growing in popularity because they connected various peoples who can share and express

[1] https://drive.google.com/drive/u/0/folders/1UIVUZfk98V_KvIITnncl856X66MVi0Z0

their ideas pertained to interesting topics around the world. Author profiling is generally said to be an identification of the demographic features of the author's traits such as gender, age-group, and nativity etc. The author profiling has various useful applications like forensics, security, politics, and marketing etc. For example, from the security point of view, author profiling competently examine the linguistic profile of a person who writes the aggressive messages which will lead to valued background information to evaluate the context of the thread. Similarly, in the online marketing, the companies want to know about the user's profile who absolutely interested or not interested in their product. This information helps to recommend the same product to the users who are all having the similar profile such as demographic, age-group and gender etc. In order to develop the system for author profiling, the main challenge is collecting the annotated corpora which contain different attributes of the user. In this paper, we described the straightforward approach to the multilingual author profiling on SMS messages (MAPonSMS) shared task conducted. The task involves the identification of the demographic appearance of the user traits such as gender and age-group in code-mixed Roman Urdu language. The rest of the paper is mentioned as follows: In section 2, we discuss the related work about the author profiling in various languages. Section 3 presents the corpus statistics and how it was preprocessed. In section 4, we explain the methodology used for author profiling and the final section describes the experiments conducted and the results obtained. In section 6, we conclude the paper and present the limitations and future work

2 Literature Review

The growth and popularity of social media platforms have generated a new social interaction environment between people and the internet thus a new collaboration and communication network among individuals. To establish the benchmark corpora for author profiling that the attempt has been made by the research community in newly years.

Oren Halvani et.al (2017) [1] proposed an inherent author verification method which produced aggressive result compared to a number of state-of-the-art approaches, situated on support vector machines or neural networks. Michael Tschuggnall and Gunther Specht (2015) [2] explained the entire grammar trees of the sentences of a document and the substructure of the documents must be extracted by using pq-grams. The high effectiveness of grammar analysis for automatic author profiling. In Francisco Rangel et al. (2015) [3], developed a system to identify the age and gender based on the impact of emotions. They used the emograph and emotion labeled graph method to attain the better performance. In Monika Briedienė et al. [4] based on the Lithuanian texts the author profiling is performed by machine learning, they used Naive Bayes Multinomial method which gives the best accuracy in gender, age, education, marital status and personality type. In Fatima et.al [5] explored the multilingual author profiling on Facebook for English-Urdu languages using content-based features and 64 different stylistic based features to identify the age and gender on multilingual and translated corpora. Seifeddine Mechti. et al. (2013) [6] identified the

age and gender of an anonymous author text using the J48 algorithm in the learning process of the English and the Spanish corpora. Ben Verhoeven et al. [7] discussed the gender profiling system for multiple languages to generate the categorized discourse lexicons for three languages, English, Dutch and German and the method used is Rhetorical structure theory (RST) discourse parser. Francisco Rangel et al. (2013) [8] used the method for automatically identifying the emotions in Spanish written texts of Facebook media.

Nandhini et al. (2015) [9] explained a method to detect the cyberbullying activities on social media. Vineetha et al. (2018) [10] explored the Malayalam gender identification for WhatsApp data using conventional features and SVM. Maarten Sap et al. (2014) [11] developed the lexica for predicting age and gender in social media, they also tried the regression and classification models for Facebook, blogs, and Twitter data. Francisco Rangel et al. (2017) [12] explored to identify the age and gender of an author and they used four different languages such as Arabic, English, Portuguese and Spanish. Vasiliki Simaki et al. (2016) [13] automatically identifying the age and gender based on the sociolinguistics and achieved better results. These results indicate that the model based on the knowledge features and the linguistic choices that they are preferred for social media users. Mohammed AliAl-garadi et al. (2016) [14] detecting the cyberbullying in the Twitter using the supervised machine learning approach. Anand Kumar et.al [15] conducted the shared task on Indian native language identification for six Indian languages in FIRE-2017.

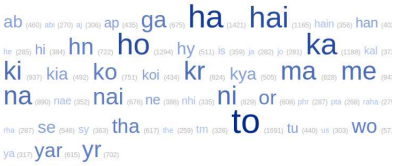
3 Dataset Description

In order to develop a multilingual author profiling system for Roman-Urdu, the MAPonSMS shared task organizers provided the training and testing corpus which includes the multilingual (Roman Urdu and English) SMS messages grouped as documents In the training corpus, there were 350 documents with annotations in which 210 documents were from male and the remaining 140 were from female. The training and testing documents are given in the ".txt" format. In testing, the dataset contains 150 documents in the same format. For gender, an author profile may belong to the male or female class and regarding the age-group, there are three categories: 15-19, 20-24 and 25-xx. The detailed statistics of the document counts in age-group and gender are given in Table 1.

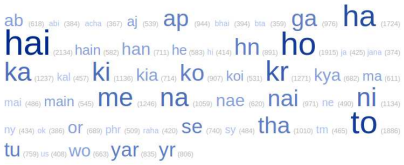
Table 1.Training dataset details (in documents)

Age-Groups						Total
15-19		20-24		25-xx		
108		176		66		350
Male	Female	Male	Female	Male	Female	
70	38	112	64	28	38	350

In order to discriminate the word usage, identify the features, for the different classes we categorized the data into five classes and plot it using the word cloud. We have visualized the frequency top 50 words of each category. The top 50 words and its corresponding count for different age-group are given in Fig. 1. The same for different gender is shown in Fig. 2. Since we have not removed the stop words, the top 50 words contain most of the stop words but the proportion of usage differs between classes. From the word cloud in figure 1 and 2, the most of the top 50 words are the same for the different classes. On the other hand, in the Fig.1, the word "main" is not in the top 50 words of 15-19 age group. The word "mein" only occurs in the 25-xx age group. The word "to" is also not in the top 50-word list of 25-xx age-group. Interestingly, in Fig. 2, even though most of the words are overlapped between the gender, the proportion shows some discrimination in the gender. For example, the words "yar", "ap","bhai" and "gi" may discriminate the gender.



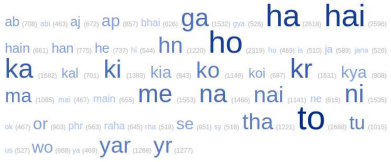
a) 15-19 age-group



b) 20-24 age-group

c) 25-xx age-group

Fig.1. Top 50 words of the different age-group class



a) Male

b) Female

Fig.2. Top 50 words of the male and female

4 Author Profiling using word and char features

We have developed the author profiling system using word-based and character-based Term Frequency and Inverse Document Frequency (TFIDF) features. Since we don't aware of Urdu language and its stop words, especially in Roman Urdu, we have not tried the language specific features. Before choosing the hybrid features (word and char) with the set of parameters, we experimented the author profiling for word based and char based features separately. But, the cross-validation results are not promising in the above-mentioned methods. The detailed cross-validation results for various feature sets are explained in section 5.

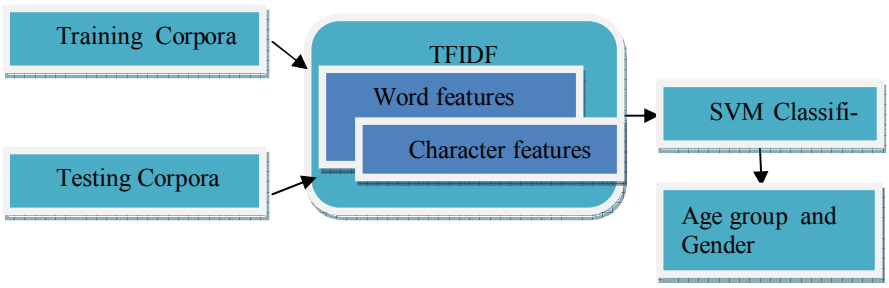


Fig.3. Methodology for Multilingual Author profiling

Fig.3 briefly explains the methodology used for the Multilingual Author profiling SMS messages system developed for the shared task. We cleaned the corpora in the preprocessing stage by removing multiple spaces, tabs and unknown characters. The set of cleaned documents are given to the TFIDF feature extraction module. We have considered word level and document level features in each document. The motivation behind the selection of word and character features are; a) we assumed that the word and phrase usage differs from a male, female and different age-groups in the SMS messages. b) We believed that since it is code-mixed multilingual corpora, the transliteration style, spelling, and usage of capital and small Roman letters consists of author traits. We varied the word features from unigram, bigram to trigram and character features from bigram to 5-grams. We have also tried the combination of features like unigram + bigram, bigram-trigram etc. in both word and character features. Finally, we combined the feature matrix of words and characters and cross-validated the performance of the author profiling for the given training corpora. The features are finally classified using the well-known Support Vector Machine linear classifier with default parameter settings. The testing corpora also converted to word and char based TFIDF features and given to the classifier. The classifier's output, as well as the de-

veloped system, is submitted to the organizer's for evaluation. Since only one submission is accepted by the organizers, we have submitted the system which gives high cross-validation accuracy.

Due to the time limit, we have not explored the preprocessing techniques and the features pertained to the author profiling task. Another reason for not using well-known preprocessing steps like case folding, stemming and stop word removal is that we believed the originally written style of the text holds the personality and behavior traits of the user.

4.1 Word and Char based TFIDF Features

We have used the conventional word and character-based features for developing the document based author profiling system. The experiments and the cross-validation results for different n-gram combinations are briefly given in section.5.

TFIDF is said to be Term Frequency-Inverse Document Frequency which is most often used in the application like document categorization in Information Retrieval and Text Mining. The number of times that word occurs in the document is proportional increases in the importance of the collection of word or corpus. Normally, TFIDF weight is computed by following steps:

STEP 1 :

We first compute the Term Frequency (TF) i.e, the number of times the term or word that occurs in the document that should be divided by the total no of a document in order normalize.

$$TF(t) = \frac{\text{Number of times term t occurs in a document}}{\text{Total no of terms in the document}}$$

STEP 2:

The Inverse Document Frequency (IDF) is computed by taking the logarithm of a total number of the document in the corpus divided by the number of the document where the particular term occurs. Thus we want to weigh down the repeated terms which scale up the different one by computing the following.

$$IDF(t) = \log(\text{Total number of documents} / \text{Number of documents with term t}).$$

STEP 3 :

To calculate the TFIDF is combing both Term frequency is multiplied with Inverse Document Frequency is given as

$$TFIDF = TF(t) * IDF(t)$$

5 Experiments and Results

In this section, we present the cross-validation accuracy for the Multilingual author profiling on SMS messages. The 10-fold cross-validation is used to find out the best model for different feature sets and parameters. The experimented models identified the age-group, gender and both jointly. The results of the 10-fold cross-validation are computed and shown in Table 2.

Gender identification accuracy of the age-group is less in the proposed model. It shows that identifying the age-group is a complicated task compared with the gender detection. For identifying gender the word-unigram model outperforms the word-bigrams, word trigram, and character-based features. For age-group identification, character-based model outperforms the word based models. As expected, the accuracy of the age-group, gender joint identification is less accurate. We combined the word features and character features and interestingly it gives the highest cross-validation accuracy compared to the models tested. The joint model accuracy is significantly increased in the word and character-based model. We combined the best features of word and character models for the hybrid models.

Table 2. Cross-Validation Results for Multilingual Author Profiling

Features		Gender	Age	Jointly
Word	Unigram	0.851	0.638	0.363
	Unigram+Bigram	0.829	0.603	-
	Unigram+Bigram+Trigram	0.738	0.588	-
Char	Bigram+Trigram	0.837	0.638	-
	Bigram+Trigram+4-grams	0.843	0.643	-
	Bigram+Trigram+4-grams +5-grams	0.840	0.640	-
Word and Char	Word{Unigram}+ Char{Bigram+Trigram+4-grams}	0.857	0.643	0.536

After the extensive cross-validation analysis, we fixed to use the hybrid model as the final submission. Before that in order to find out the discriminative features for the gender identification, we plot the SVM discriminative feature model for male and female. Fig.4. explains the discriminative features where red color indicates the female and blue indicates the male. We can easily understand that the word "ga" and "bhai" are used mostly by the male and "rai" is mostly used by the female. We have also seen that from Fig.2 the "ga" is used mostly in the male class.

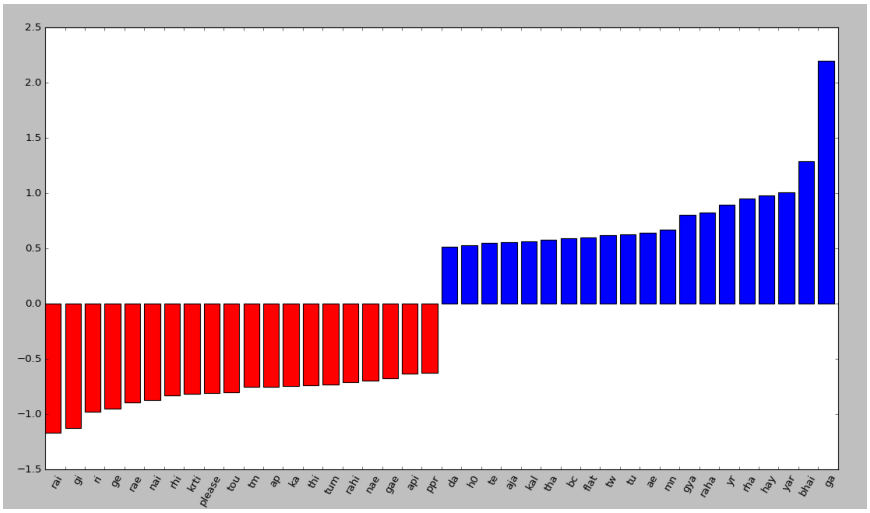


Fig. 4. Discriminative features in gender identification

Table. 3 describes the top-3 team's performance given by the organizers. We have used the unigram word features and character features of bigram, trigram, and 4grams.

Table. 3. Results of top three teams given by organizers

Team	Gender	Age	Joint
KCE_DALab	0.87	0.65	0.57
SSN, India.	0.85	0.63	0.52
The University of Washington Tacoma, USA.	0.83	0.60	0.49

6 Conclusion and Future Work

In this paper, we illustrate the work on identification of gender and age-group in Multilingual Author Profiling on SMS messages (MAPonSMS) shared task on Roman Urdu and English language. Using the training dataset, we have developed the system using word and char based Term Frequency & Inverse Document Frequency (TFIDF) features and classified with Support Vector Machine classifier. We have discussed the dataset descriptions and experiments used for the Multilingual author profiling task. We experimented with the 10-fold cross-validation with different feature sets and the results were reported. We have also presented the results given by the organizers of the shared task. The submitted system with Word{Unigram}+ Char{Bigram+Trigram+4-grams} achieved the state-of-art best performance in terms of accuracy 87% for identification of age and 65% for gender. The joint accuracy

gained is 57%. Interestingly, our cross-validation results are very close to the results provided by the organizer. This shows the consistency of the dataset and the method used in the shared task. Detailed error analysis can be considered in near future to improve the accuracy further for age -group detection. The preprocessing pertained to the author profiling can be incorporated. The linguistic, behavioral features and the statistical test features can be incorporated to improve the performance of age-group identification. The role of stop words and the identification of specific words pertained to gender and age-group can be studied. Finally, the character based embeddings with deep neural networks can also be proposed for large-scale author profiling.

Acknowledgment

We would like to thank MAPonSMS organizers and Forum for Information Retrieval Evaluation-FIRE 2018 for organizing the Author profiling task.

References

1. Oren Halvani, Christian Winter, Lukas Graner.Authorship Verification based on Compression-Models arXIV:1706.00516 [cs.ir] 1 June 2017

2. Michael T.Schuggnall and Gunther Specht.Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors , 2015

3. Fransisco Rangel, Paolo. On the impact of emotions on author profiling. Information Processing and Management 52(2016) 73-92.

4. Monika Briediene, Jurgita Kapociute Dzikiene. An Automatic author profiling from Non-Normative Lithuanbian Texts.CEUR-WS.org\vol-2145\p18.

5. 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

6. Seifeddine Mechti , Maher Jaoua , Lamia Hadrich Belguith , and Rim Faiz .Author Profiling Using Style-based Features Notebook for PAN at CLEF 2013.

7. Ben Verhoeven, Walter Daelemans.Discourse lexicon induction for multiple languages and its use for gender profiling

8. Francisco Rangel , Paolo Rosso .On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style.http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13- web/author-profiling.html.

9. B.Sri Nandhini, J.I.Sheeba.Online Social Network Bullying Detection Using Intelligence Techniques.International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).Procedia Computer Science 45 (2015) 485 – 492

10. Vineetha Rebecca Chacko, Anand Kumar M, Soman K P, "Experimental Study Of Gender And Language Variety Identificaion in Social Media" In: Proceedings of the Second International Conference on Big Data and Cloud Computing, (Springer) Advances in Intelligent Systems and Computing (AISC), 2018.

11. Maarten Sap et.al Developing Age and Gender Predictive Lexica over Socia Media.Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1146–1151.

12. Francisco Rangel , Paolo Rosso , Martin Potthast , Benno Stein .Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter, http://webis.de/research/events/pan-13/pan13-web/author-profiling.html

13. Vasiliki Simaki, Iosif Mporas, Vasileios Megalooikonomou.Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification. American Journal of Engineering and Applied Sciences 2016, 9 (4): 868.876 DOI: 10.3844/ajeassp.2016.868.876.

14. MohammedAliAl-garadi,KasturiDewiVarathansri, Deviravana.Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network.Computer in Human Behaviour . Volume 63,October 2016,pages 433-443.

15. Anand Kumar M, Barathi Ganesh HB, Shivkaran Singh, Soman KP and Paolo Rosso. Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification. In Proc. of Forum for Information Retrieval Evaluation 2017.