Multilingual Author Profiling on Short Message Services (SMS)

[1]Hardik Joshi, [2]Dr. Hiren Joshi, [3]Aatithya Hitkar, [4]Shifa Khan

[1]Asst. Professor, [2]Professor, [3,4]M.Tech Students
[1]Department of Computer Science
[1]Gujarat University, Ahmedabad, Gujarat, India

Abstract : Author profiling is the task of identifying different personality traits of author by analyzing their written text. Different personality traits are author's gender, age, native language, native city, qualification, occupation etc. Author profiling has different applications such as it could be used in forensics to find the suspect, in marketing, for fake profile identification etc. We aim to predict the author's age and gender from the multilingual SMS. In this paper we have classify the author gender into male or female categories. Moreover we have classify the author age into three different groups: (i) 15-19, (ii) 20-24, (iii) above 25. Several machine learning classification models like Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes and Linear Regression are used for experiment. Experimental Results show that Support Vector Machine provide better accuracy when used with Count Vectorizer and TFIDF (Term Frequency Inverse Document Frequency).

***IndexTerms* - Multilingual Author Profiling, Natural Language Processing, TFIDF, Support Vector Machine**

1 INTRODUCTION

People use social media like WhatsApp, Facebook, YouTube, Instagram, and Twitter to communicate with each other and to share their thoughts, feedback, ideas and opinion. Using a single language like English in communication is Monolingual. Sometimes people uses more than one language in their messages which is Multilingual. As some people are not comfortable to express themselves properly in English they use their preferred language or local language to communicate. Social media allows people to use their preferred language in communication. According to CTIA, 6 billion SMS messages are sent each day in the US And according to Portio Research, SMS traffic was at peak in 2015 with 8.3 trillion messages in a year and they also forecasted that SMS traffic will start declining after 2015.

Author profiling is the process of finding author demographics features like gender, age, native city, native language, occupation, qualification etc. by analysing their text. Author profiling is used in many different application like Forensics, Marketing, and Security etc. For example in forensics it can be used to detect the fake profile, in marketing it can be used to know the interest of the people of different age group and gender.

Uncovering the exact identity of the person is very complicated and sometimes unsolvable task, whereas to reveal his/her Meta information (i.e., demographic features: age, gender, etc.) is easier, but still very useful. The revealed meta-information that, e.g., a 40-year-old man is impersonating a 15-year-old girl may encourage the police to dive more detailed into the data or even take decisive actions for the criminal offense. The manual Internet space monitoring and manual text analysis is hardly possible, because it requires enormous amounts of human resources. Thus, natural language processing technologies become the only solution for tacking similar problems.

In this paper we aimed to predict the author's traits like age and gender from the messages which are in multilingual language. Much of the work on author profiling is done on English Language and other languages like French, German, and Spanish etc. However very few work is done on South Asian language like Sanskrit, Urdu, Hindi, and Nepali etc. Like in Hindi we write "आप क्या कर रहे हो? ", in Roman Hindi it is written as "Aap kya kar rahe ho?" and it translates to "what are you doing?"

The rest of the paper is mentioned as follows: In section 2 we discuss the literature review that the related work on author profiling in different languages. Section 3 explain the corpus and pre-processing. Section 4 shows the different experiment and result. Section 5 concludes the paper and suggest the future work.

2 LITERATURE REVIEW

SMS is considered as one of the most widely used source of communication. People uses their preferred or local language while communicating through SMS. Extracting useful information from SMS data is very crucial task and it attracted many researches.

A lot of the work has been performed on author profiling with the dataset collected from social media sites and blogs in English, French, Dutch, Spanish Language whereas work on SMS in Roman Hindi it's rather a new subject and less work has been done on Multilingual author profiling using SMS.

In Fatima et al. [1] explored Multilingual author profiling on Facebook data for English-Urdu languages using content-based features and stylistic based features to identify age and gender. In Monika Briediene et al. [2] based on Lithuanian text, author profiling is done using the machine learning, they used Naive Bayes Multinomial method and determined the age, gender, educational, marital status and personality type. In Francisco Rangel et al. (2015) [3] developed a system to identify the age and gender based on the emotions.

In Murat Karabatak et al. [4] identifies gender from SMS text messages. He has used naïve bayes, J48 and Multi-layered Perceptron algorithms to find out gender from the SMS message. The dataset was taken from National University of Singapore's website. As a result they got accuracy of 69.81% with naïve bayes, 71.93% accuracy with J48 algorithm, and 69.16% accuracy with Multi-layered Perceptron. In Jahna Otterbacher et al. [5] finds out gender of the movie reviewer with help of writing style, content and metadata. The author gets accuracy of 73.7% with logistic regression. Same way K Santosh et al. [6] finds age and gender with content based, style based and topic based features and achieves accuracy of 56.53%, 64.8% for gender and age respectively.

In Jonathan Schler et al. [7] has done research for author profiling on blogs with stylistic and content based features with support vector machine and got 67% accuracy for gender. In Hernandez et al. [8] proposed linguistic markers, slangs, emotions and semantic similarity for the identification of author's age and gender from the multilingual text.

Francisco Rangel et al. [9] used the method for automatically identifying the emotions in Spanish written texts of Facebook media. Gonzalez et al. [10] have used TF-IDF for feature extraction and used that with SVM to obtain the best result on a dataset containing both Spanish and English text.

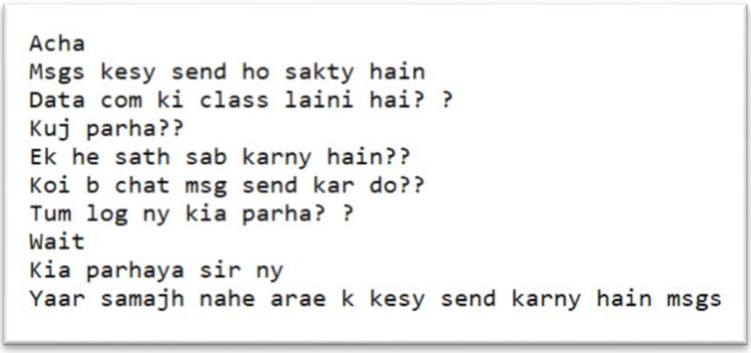
Maharajan et al. [11] have proved that simple n-gram or using word n-grams have been observed to be effective in getting better result for feature extraction than character n-grams. Argamon et al. [12] has predicted author gender for formal text, they found differences in the use of pronouns. Females' uses more pronouns than male and they uses more first person pronouns. In Ankush et al. [13] has predicted the author's gender in code-mixed content and used the dataset of English-Hindi collected from Twitter.

Schler et al. [14] and Goswami et al. [15] used combinations of simple lexical and syntactic features to determine the gender and age of the authors of anonymous blog posts. Nguyen et al. [16] studied the use of language and determined the age in Twitter dataset. Francisco Rangel et al. [17] has identified age and gender of authors based on their use of language and used stylistic features and obtained results with SVM-based approach on the PAN-AP-13 dataset. Raghunadha et al. [18] has predicted the age and gender of the author by analyzing their writing styles on hotel review dataset. They had used the pivoted unique term normalization measure and calculated the weight of the terms specific to each profile group.

In Shlomo et al. [19] used style-based and content-based method and determined the gender on the blog dataset, they had used Bayesian Multinomial Regression. In Malcolm et al. [20] has identified author's gender using SVM approach. They found out that women more punctuates like assertions, apologies, questions etc. while men uses strong assertions, aggressive, self-promotion, challenges, humor etc.

3 Experimental Dataset

Our Dataset is taken from FIRE-2018 which consist of 500 author profiles. We aim to find out age and gender for each author profile. Each profile consist of multiple sms messages of the same author in a text file. These sms messages are written in Roman Hindi. Out of 500 profiles 350 were for training and 150 for testing. Sample dataset is shown in Figure 1.



```
Acha
Msgs kesy send ho sakty hain
Data com ki class laini hai? ?
Kuj parha??
Ek he sath sab karny hain??
Koi b chat msg send kar do??
Tum log ny kia parha? ?
Wait
Kia parhaya sir ny
Yaar samajh nahe arae k kesy send karny hain msgs
```

Figure 1. Sample Dataset

Training dataset consist of 350 profiles. From each of them we aim to find gender and age. For each file labels are given in separate file called 'truth.csv'. The 'truth.csv' file consist of 3 columns. First column contains the name of the filename, Second column consist Gender of the filename given in the first column and in the third column Age-group is given for same. We will find age and gender individually and combined. Age falls in three groups 1) 15-19. 2) 20-24. 3) 25-XX. And gender as Male and Female. Testing dataset consist of 150 profiles for which associated labels/answers are unknown. Table 1 summarizes the test dataset.

Table 1: Summary of Training & Test Dataset

Age	Gender	Training	Testing
15-19	Male Female	70 38	30 16
20-24	Male Female	112 64	48 28
25-XX	Male Female	28 38	12 16
	Total	350	150

4 Experiments and Results

In this section we have presented workflow and accuracy for the dataset. We have used word base n-gram (upto trigram) approach to find out the accuracy of the author profiles. Since our author profiles consist Roman Hindi, we have not experimented language specific features. Our results shows that Support vector machine has performed better than other Models.

4.1 Dataset

Our dataset is taken from FIRE-2018. We have multiple text messages of an author in Roman Hindi in a single text file from that file we aim to find some demographic traits like age, gender and combine (age and gender both). We have 500 author profiles from which we have taken 350 for training purpose and 150 for testing.

4.2 Pre-processing

In pre-processing of data we have removed unnecessary words and characters which can decrease prediction accuracy of our model while training and testing.

- List of removed words and Characters are as follows:
- Extra spaces and Extra lines
- Non-Ascii Characters
- English stop-words
- Punctuations

4.3 Model Training and Testing

In this section we have trained and tested our dataset. For training and testing of our models we have used n-gram approach with Support vector machine, Naïve bayes, Decision Tree and Random forest classifiers for prediction. We have find out features with n-gram approach. We have tried with unigram, bigram and trigram. Where unigram stands for a single item of a text. Bi-gram stands for a pair of consecutive written characters or words and Tri-gram stands for three consecutive letter or words of text. We have used these unigram, bigram, trigram features for training and testing of out models. Out of all Unigram with Support vector machine (SVM) outperformed. We got 91.43% accuracy from SVM with unigram for gender, for age we got 61.43% and for combine we got 50.00% accuracy.

4.4 Accuracy and results

We have evaluated and trained all the models with different approaches like unigram, bigram and trigram. We got best accuracy with Support vector machine in all approaches which we tried. After that Random forest performed very well with very near accuracy results to SVM. Decision Tree performed poor for predicting combine accuracy. We have evaluated accuracy on the basis of equation 1.

$$Accuracy = \frac{Number\ Of\ Correctly\ Identified\ Author\ Profiles}{Total\ Number\ of\ Author\ Profiles} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(1)$$

The results obtained by our experiments are shown in Table 2. We evaluated different models for prediction of Gender and Age separately and with the combined approach. Support Vector Machine showed better accuracy when compared to other models like Naïve Bayes, Decision Tree and Random Forest.

Table 2: Results Obtained from Experiments

Model	Gender			Age			Combine		
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
Support Vector machine	**91.43**	**85.71**	**78.57**	**61.43**	**59.14**	**57.14**	**50.00**	**45.71**	**42.86**
Naïve Bayes	80.00	65.71	60	57.14	50	45.71	34.28	27.14	27.14
Decision Tree	74.29	64.29	65.71	54.29	54.29	52.86	34.28	35.86	32.86
Random Forest	78.57	72.85	68.57	58.57	57.14	55.71	41.43	37.14	31.43

5 Conclusion and Future Work

In this paper, we presented our on identification of gender and age-group in Multilingual Author Profiling on SMS on Roman Hindi and English language. Using the training dataset, we have developed the system using word based Term Frequency & Inverse Document Frequency (TFIDF) features and then classified with different ML classifiers i.e. Random Forest, Support Vector Machine, Naive Bayes, Decision Tree, and Logistic Regression. We have done the pre-processing by removing the stop words from the dataset. 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. We have also determined the results using Word Unigram, Bigram, and Trigram. We obtained 85% accuracy for gender identification and 67% accuracy for age identification. The joint accuracy gained is 47%.

We aim to extend the model by making it more efficient by using different

Techniques, we did not explore the deep neural network. Since our author profiles consist Roman Hindi, we have not tried language specific features to find out author profiles. So far the text contained two languages but in the future, it would be beneficial to include more South Asian languages as they are relatively less explored and contain potential to be very useful. We would like to demonstrate other author traits like native language, native area, personality, type, qualification and occupation.

6. Acknowledgment

We would like to extend gratitude to the FIRE community for providing us with the dataset. The equipment used for this research were funded by MEITY under the project no. DIT/R&D/TDC/13(2)2015.

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.] Monika Briediene, Jurgita Kapociute Dzikiene. An Automatic author profiling from Non-Normative Lithuanbian Texts.
[3.] Fransisco Rangel, Paolo. On the impact of emotions on author profiling. Information Processing and Management 52(2016)73-92.
[4.] Murat Karabatak, Shannon Silessi, Cihan Varol : Identifying Gender From SMS Text Messages at 2016 15th IEEE International Conference on Machine Learning and Applications.
[5]. Jahna Otterbacher: Inferring Gender of Movie Reviewers at 19th ACM international conference.
[6.] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma : Author Profiling: Predicting Age and Gender from Blogs at CLEF 2013.
[7.] Jonathan Schler, Moshe Koppel, Shlomo Argamon, James Pennebaker : Effects of Age and Gender on Blogging at Conference of Computational Approaches to Analysing Weblogs – Stanford, California, USA – 2006.
[8.] Hernandez, D., Guzman-Cabrera, R., Reyes, A., Rocha, M.-A.: Semantic-based Features for Author Profiling Identification: First insights. In: Proceedings of CLEF, (2013).
[9.] Francisco Rangel , Paolo Rosso .On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style.http://www.uniweimar.de/medien/webis/research/events/pan-13/pan13- web/author-profiling.html.
[10.] Bayot, R., Gonçalves, T.: Multilingual author profiling using word embedding averages and svm. In: Software, Knowledge, Information Management & Applications (SKIMA), 10th International Conference on. pp. 382–386. IEEE (2016).
[11.] Maharjan, S., Shrestha, P., Solorio, T.: A simple approach to author profiling in mapreduce. In: CLEF (Working Notes). pp. 1121–1128 (2014).
[12.] Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R.: Gender, genre, and writing style in formal written texts. Text, 23, August 2003.
[13.] Ankush Khandelwal, Sahil Swami, Syed Sarfaraz Akhtar and Manish Shrivastava: Gender Prediction in English-Hindi Code-Mixed Social Media Content at Cornell University Library Arxiv (2018).
[14.] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In American Association for Artificial Intelligence (2006).

[15.] Goswami, S.; Sarkar, S.; and Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: International AAAI Conference on Weblogs and Social Media ICWSM (2009).
[16.] Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How Old Do You Think I Am?" A Study of Language and Age in Twitter in 7th International AAAI Conference on Weblogs and Social Media ICWSM (2013).
[17.] Fransisco Rangel, Paolo: Use of Language and Author Profiling: Identification of Gender and Age in Natural Language Processing and Cognitive (2013).
[18.] T. Raghunadha Reddy, B. Vishnu Vardhan and P. Vijayapal Reddy: Author Profile Prediction using Pivoted Unique Term Normalization in Indian Journal of Science and Technology (2016).
[19.] Shlomo Argamon, Moshe Kopple, James W. Pennebaker, and Jonathan Schler: Automatically Profiling the Author of an Anonymous Text in Communication of ACM (2011).
[20.] Malcolm Corney, Alison Anderson, George Mohay, Olivier De Vel: Language and Gender Author Cohort Analysis of E-mail for Computer Forensics in Digital Forensic Research Conference (2017).