Using Bag-of-Words and Psycho-Linguistic Features For MAPonSMS*

Asmara Safdar, Osama Akhter, Osama Inayat, and Abdullah Khalid

COMSATS University Islamabad, Lahore Campus, Pakistan
asmarasafdar@cuilahore.edu.pk , osama.ocreed@gmail.com ,
rajaosamainayat@gmail.com , sagittarius.here@gmail.com

Abstract. This paper presents the use of Bag-of-Words(BoW) and Psycho-Linguistic(P-L) approaches based upon the demographic trends in modeling multilingual(Roman-Urdu and English) SMS text(Short Message Service) for gender and age prediction. The data set[1] was provided as a standard source to work for the multilingual author profiling task in the contest FIRE'18-MAPonSMS[2]. The proposed approaches, as compared to the baseline results, adequately classify the test set to age and gender separately.

Keywords: Author profiling · Multilingual · Bag-of-Words · Psycho-Linguistic.

1 Introduction

Authorprofiling is a task in automatic authorship identification that finds characteristics, particularly: demographic, of the author of a document. Having known the profile of an author can help in resolving many issues, such as, crime investigation(e.g., by identifying the linguistic profile of a suspected message), developing a recommendation system to recommend different products to different users(by finding the demographic features of authors through his reviews on a product) etc.

MAPonSMS task is about the prediction of gender and age of the authors based on the multilingual i.e. English and Roman-Urdu SMS data set of 350 documents.

In this paper, we present our proposed approaches and their contribution towards the classification of gender and age for the contest. We have applied some stylometric(lexical, syntactic, structural features) and content based(based upon the textual content rather than the metadata) approaches to perform the task. In stylometric approaches, features based upon Psycho-Linguistic have proposed as taken from [3] and we refer such features as P-L in this write-up. As

* https://lahore.comsats.edu.pk/cs/MAPonSMS/index.html
[1] https://lahore.comsats.edu.pk/cs/MAPonSMS/de.html
[2] "Forum for Information Retrieval Evaluation-Multilingual Author Profiling on SMS" https://lahore.comsats.edu.pk/cs/MAPonSMS/index.html

content based features, we have devised some sets of Psycho-Linguistic content based words as a representation of the document. We call these features as Psycho-Linguistic Bag-of-Words features and write as P-BoW in the whole discussion that follows. The proposed features outperformed the baseline accuracy results both for age and gender prediction. The software submitted for the contest can be downloaded from https://github.com/Osama081/MapOnSMS.

In the sections those follow, we present the related work in authorprofiling in section 2. In section 3 we describe the approaches we used to generate features of the given dataset. Section 3.4 provides an overview of the results for both prediction tasks separately. Section 4 concludes the paper and suggests potential improvements.

2 Literature Survey

In the literature-world of automatic authorprofiling, a lot of work has performed on datasets mainly collected from social media sites and blogs while SMS, as data-source, remains neglected. Besides, the multilingual datasets are in the languages which are spoken in developed countries while a little work(such as given by Fatima et al. in [5]) has done on multilingual datasets with Roman-Urdu as one of the languages.

In [2], Chen et al. did their studies based on a dataset of gendered usage of emojis containing 134,419 Android-smartphone users across 183 countries, in 58 languages to analyze various aspects of emoji usage and find out that the people of different genders tend to use emojis of slightly different categories.

Fatima et al. in [5] provide a standard multilingual resource of 810 SMS based user profiles annotated with 7 demographic traits including age and gender. They have applied stylometric and content based features for the gender identification task. However, the classification of other demographic features has not yet explored for the corpus.

Cheng et al. propose Psycho-Linguistic and gender preferential cues along with stylometric features for gender prediction in [3]. They performed the experiments for short length, multi-genre, content-free English text. The empirical studies show an accuracy up to 85.1%.

In third authorprofiling task at PAN 2015 [11], Rangel et al. organized the tasks of age, gender, and personality recognition. The given dataset was collected from Twitter and consisted of English, Spanish, Dutch and Italian languages. The participants used content-based features(including bag of words and n-grams) and style-based features (frequencies, punctuations and some Twitter specific such as hash-tags).

Rangel et al. in the author profiling task at PAN 2013 contest [10] describe the identification of age and gender using multilingual dataset, consisting of English and Spanish, collected from social media sites. The approaches used by the participants include content-based, stylistic-based, n-grams based, IR-based and collocations-based features. Empirical studies show the difficulty of the task especially in gender prediction and collective prediction of gender and age.

Researchers have been providing pieces of evidence for the last few decades that a person's physical and mental health is strongly correlated with the words he/she uses. Gottschalk et al. [6] and Rosenberg et al. [12] discuss the different factors and theoretical bases of psychological states. We have presented our work using some existing and rest by modification of the existing approaches that have applied to various types and genre of the dataset in the past. Our proposed approaches are mainly related to psycholinguistic features on the given SMS based multilingual dataset in English and Roman-Urdu.

3 Authorprofile Experiments

We conducted feature extraction on the dataset provided for age and gender prediction separately. The results show that both of the proposed approaches have proved better than the baseline approaches.

3.1 Dataset

The given dataset was a training set to work for gender and age classification task for the contest FIRE18'-MAPonSMS. It is an SMS based multilingual corpus containing 350 total documents each from a different user. Each document contains multiple text messages and annotated with age groups and the gender such that the instances of all groups are balanced. For gender classification, it has 60% and 40% documents written by male and female authors respectively. For age classification, there are 31% documents categorized in age group *15-19*, 50% in age group *20-25* while 19% in the group *25-xx*.

3.2 Approaches Used

We applied 1) Stylometry and 2) Content-based approaches for gender and age prediction tasks among the renowned methods for authorprofiling i.e. stylometry, content-based and topic-based [1, 3].

Feature Extraction For Gender Classification For gender classification task, we used 67 stylometric features in groups of three namely: character based(Table1), vocabulary richness(Table 2) and word based(Table3). Under the group of word based features are introduced some P-L as well. In content based(Table 4) both features are P-BoW.

For character based approaches, there are 43 features in total(as shown in Table1) most of which have been employed by [3, 5] for prediction of demographic features of the author.

For vocabulary richness, we used total 9 features as given in Table2). These vocabulary richness features have used by many researchers for age and gender prediction problem such as in [14, 8, 3, 5].

Table 1. Character based features for gender classification task.

Feature	Description	Feature	Description
F1	Total number of all characters(C)	F26	Count of underscore
F2-F12	Punctuation marks(. , ? etc less m-dsh[3])	F27-34	Count of @, &, *, $, =, /, %, + sign
F13	Percentage of punctuation marks to C	F35	Count of all sorts of brackets
F14-F15	Opening and closing curly braces	F36	Count of white spaces
F16-F17	Opening and closing square brackets	F37	Percentage of of white spaces to C
F18-F19	Opening and closing parenthesis	F38	Percentage of letters to C
F20-F21	Opening and closing angle brackets	F39	Percentage of upper case letters to C
F22	Count of white spaces	F39	Percentage of upper case letters to C
F23	Count of vertical lines	F41	Percentage of white spaces to C
F24	Count of uppercase letters	F42	Percentage of digits to C
F25	Count of digits	F43	Percentage of tabs to C

Table 2. Vocabulary richness features for gender classification task.

Feature	Description	Feature	Description
F1	Sichel's Measure	F5	Hepax Lgumena
F2	HonoureR Measure	F6-F7	number of Unique character 5-gram and 7-gram
F3	Brunet Measure	F8-F9	number of unique word 1-gram and 2-gram
F4	Yule K measure		

Intuition for proposing word ending with *i/I*(F7 in Table3) and word ending with *a/A*(F6 in Table3) is the fact that in Roman-Urdu, many words are gender specific(that discriminate a masculine noun[4] from a feminine noun[5]). The ending of a word(things, abstract nouns, participles), if a vowel, usually helps in this gender classification. Words, ending with *a* are usually masculine whereas if a word ends with *i* or *ii*, it is usually a feminine[6] word. For example, *"Answer ne kr saki mein"* is written in Roman-Urdu that means "I couldn't answer"(as written by a female author). The word "saki" means *could* that is a feminine version of participle while the same word is written as "saka" if referred by a male. There are many such words in Roman-Urdu those are used with the slight change of *i* and *a* letter, in the end, to refer to female and male author respectively. Other examples for such words are "khata-khati"(*eat* in English), "ata-ati"(*come* in English), "karta-karti"(*do* in English), "sota-soti"(*sleep* in English) and so on. Limitation of this approach is the fact that there are many neutral words(with no gender) that might have been counted as a masculine or a feminine. Additionally, a male author may refer to many feminine words and vice versa.

Two features(F9 and F10 in Table3) are related to the use of emojis and smilies in the SMS document by each user. Emojies are combinations of different characters to express emotions(emojis are also called emoticons, winks or smileys). Chen et al. claim in [2] that women are more likely to use emojis than men. We got a resource for a number of text-emojis from[7]. One feature is the

[4] all male human beings,animals and plants those are considered "masculine" are masculine in Roman-Urdu

[5] all female human beings,animals and plants those are considered "feminine" are feminine in Roman-Urdu

[6] https://en.wikibooks.org/wiki/Urdu/Nouns
Last Visited: 05, 08, 2018

[7] http://cool-smileys.com/text-emoticons
Last Visited: 05, 08, 2018

count of emojis(F9 in 3) and the other is the average number of emojis per message(F10 in 3).

As the given multilingual dataset has been generated having collected from different mobile users in Pakistan, another approach for proposing P-L features is to see the tendency of authors to use English words in the multilingual dataset. Urdu is Pakistan's official language yet English is used equally in offices especially for writing many official documents. Moreover, many text displays on different banners, billboards, organization's name boards and many other activities use English. Besides English is the medium of education in almost all of the educational setups and institutes in Pakistan[8]. Studies show that some demographic features affect the language one uses [4, 9]. Keeping this in view, the proportions of English to Roman-Urdu contents sounds a potential feature to contribute substantially in predicting the demographic features like age and gender of authors. To see its effect on the given classification tasks, we proposed 4 such features(F11-F14 in Table3) for gender classification. We used the standard English word dictionary, used in Linux, as a resource to match the English words in the given dataset.

Table 3. Word based features for gender classification task.

Feature	Description	Feature	Description
F1	Count of Multiple "?"	F9	Count of emojis(P-L)
F2	Count of Multiple "!"	F10	Average emojis per message(P-L)
F3-F4	Percentage of words with length 3, 4	F11	Count of English(P-L)
F5	Total number of sentences	F12	Count of Roman-Urdu words(P-L)
F6	Count of words ending with a/A	F13	Ratio of English to Roman-Urdu(P-L)
F7	Count of words ending with i/I	F14	Ratio of Roman-Urdu to English(P-L)
F8	Percentage of questioned sentences		

Table 4. Content based features for gender classification task.

Feature	Description
F1	Percentage of Assent words to total words(P-BoW)
F2	Percentage of negation words to total words(P-BoW)

We added some P-BoW features as well: 1) Percentage of Assent words to total words, and 2) Percentage of Negation words to total words given in Table 10. Such categories of P-BoW are to see the effect of count-based representation of the document based on the correlation of the linguistic factors and psychological aspects of an author. Cheng et al. [3] propose several P-L features to build the feature space for gender prediction. In our case, where the data set provided is multilingual, we identified the group of some psycholinguistic words as given by [3] and added some Roman-Urdu words in the selected categories. One feature related to P-BoW words is the percentage of assent words to total words. English assent words we selected are *ok, agree, alright, right, yes, yup, yeah.* The same category also included Roman-Urdu words as *sai*(ok or alright in English), *k*[9] *ok*[9], *han, haan, sae, h*[9].

[8] https://en.wikipedia.org/wiki/Pakistani_English
Last Visited: 05,08,2018

[9] one or more characters

Second group in P-BoW is the *Negation Words*, from [3]. It contains *no, never, not, na ,ni , nae, niii, nahi.* We proposed some Roman-Urdu words mostly used for negation in this category. Note that all non-English(Roman-Urdu) words in this group are variants of *no* in English.

As Roman-Urdu lacks standard lexicon, many spelling variations exist for a given word most of the times. For example, *nahi, ni, nae, nii* are all variations of a single word in Roman-Urdu that means *no* in English. So, it is important to mention here that any group of these P-BoW is not exhaustive because of the inherent inconsistency in the representation of the Roman-Urdu text.

Feature Extraction For Age Classification For age classification task, total 75 features were generated out of which 70 are stylometric and 5 are content based.

In stylometric features, we introduced total 44 character based features as listed in Table 5 and 8 vocabulary richness features as given by Table 6.

Table 5. Character based features for age classification task.

Feature	Description	Feature	Description
F1	Total number of all characters(C)	F25	Percentage of digits
F2-F12	Punctuation marks(. , ? etc less m-dsh[10])	F26-36	Count of @, &, *, $, =, /, %, +, _
F13	Percentage of punctuation marks to C	F37	Count of all sorts of brackets
F14-F15	Opening and closing curly braces	F38	Count of white spaces
F16-F17	Opening and closing square brackets	F39	Percentage of letters to C
F18-F19	Opening and closing parenthesis	F40	Count of upper case letters to C
F20-F21	Opening and closing angle brackets	F41	Percentage of digits to C
F22	Percentage of white spaces	F42	Percentage of tabs to C
F23	Count of vertical lines	F43	Count of tabs
F24	Percentage of uppercase letters to C	F44	Percentage of special characters to C

Table 6. Vocabulary richness features for age classification.

Feature	Description	Feature	Description
F1	HonoureR Measure	F4	Hepax Legumena
F2	Brunet Measure	F5-F6	number of Unique character 5-gram and 7-gram
F3	Yule K measure	F7-F8	number of unique word 1-gram and 2-gram

Rest of the stylometric features are word based. Some more P-L features have proposed for age classification[11]. These new P-L features are 1) Percentage of English words to Total words(F13) and 2) Percentage of Roman-Urdu words to total words(F14) in Table 7.

Table 7. Word based features for age classification.

Feature	Description	Feature	Description
F1	Count of Multiple "?"	F11	Ratio of English to Roman-Urdu(P-L)
F2	Count of Multiple "!"	F12	Percent English to total words(P-L)
F3-F4	Percentage of words with length 3, 4	F13	Percent Roman-Urdu to total words(P-L)
F5	Total number of sentences	F14	Count of I ending
F6	Percent questioned sentences	F15	Count of A ending
F7	Average word length	F16	Ratio of A ending to I ending
F8	Total number of words	F17	Count of emojis(P-L)
F9	Count of English words(P-L)	F18	Average emojis per message(P-L)
F10	Count of Roman-Urdu words(P-L)		

A few more P-BoW features in the content based are also proposed. Such P-BoW features are: 1) Count of slang(F3), 2) Percentage of slang(F4), and 3)

[11] Note that they didn't contribute well for gender classification so we did not select them for that

Percentage of certainty(F5) given in Table 8. As studies show that age strongly affects the use of language [7], knowing this, we proposed the feature of slang(F3 and F4 in Table8). We categorized slang words as the words those don't relate either to English or Roman-Urdu and are not used in formal speaking or writing. 16 such words were identified from the dataset. Some of the slang words we selected are *lol, plz, btw, k, idk, jigar, oye, oyee, yar, yr.* We identified a few words of certainty(F5 in Table8) having taken idea from [3]. Words of certainty[12], that we selected, include *always, hamesha, hmesha, never, ever, kabi, kabhi, kbhi, kbi, forever.*[13]

Table 8. Content based features for age classification .

Feature	Description	Feature	Description
F1	Percentage of assent(P-BoW)	F4	Percentage of slang(P-BoW)
F2	Percentage of negation(P-BoW)	F5	Percentage of certainty(P-BoW)
F3	Count of slang(P-BoW)		

3.3 Classifiers Used

As the training dataset was annotated, the gender and age prediction tasks are supervised machine learning problems with *Gender prediction* a binary classification task(class attributes as *male* or *female*) whereas *Age prediction* a multiple classification task(class attributes as *15-19*, *20-24*, or *25-xx*). We used two classifiers: Random Forest(RF) and Meta Bagging(MB) from Meta class. Both of these algorithms are ensemble machine learning algorithms and are closely related. Note that Bagging was used with its default settings and REP Tree as its component classifier[14].

We used 10-fold cross validation to evaluate the prediction models and reported accuracy as a measure to evaluate the performance because the dataset is balanced. Accuracy is the percentage ratio of correctly classified instances to incorrectly classified instances.

3.4 Results and Analysis

Gender Classification Table 5 shows the accuracy measure of different groups of features as reported by RF and MB. MB and RF gave accuracies of 60.8% and 52.85% for All P-BoW features respectively. All P-L and P-BoW features combined gave 66.85% accuracy for MB and 65.7% for RF. All word based features collectively gave an accuracy of 70.2% by MB and 71.4% by RF. All character based features combined resulted in 78.28% accuracy for MB and 77.14% for RF. Then the fifth set of features i.e. combination of all features gave the highest accuracy of 80.29% by MB.

[12] "Certainty is something that is certain or sure" https://www.merriam-webster.com/dictionary/certainty Last Visited: 05, 08, 2018

[13] P-BoW features of certainty and use of slangs proved more useful to discriminate age groups than the gender.

[14] Although we also selected other Decision Tree algorithms as component classifiers for Bagging but REP Tree gave the best results.

It is evident that overall results were best reported by MB as 80.29% for combination of all features. The best group of features, if analyzed group-wise, was *character based* that gave an accuracy of 78.28% for MB. There is also an interesting fact about using P-L and P-BoW features combined. There were total 8 such features for gender classification which collectively gave an accuracy of 66.85% with MB. Of these 8 features, 4 were related to use of English or Roman-Urdu words for which RF gave 65.7% accuracy. This shows that all P-BoW and P-L approaches contributed substantially to train gender classifier.

Table 9. Gender classification results using different groups of features.

Feature	Classifier	Accuracy	Feature	Classifier	Accuracy
All P-BoW	**MB**	**60.8**%	All character based	**MB**	**78.28**%
	RF	52.85%		RF	77.14%
All P-L and P-BoW	**MB**	**66.85**%	**All features combined**	**MB**	**80.29**%
	RF	65.7%		RF	78.57%
All word based [15]	MB	70.2%			
	RF	**71.4**%			

Age Classification The accuracy scores for different sets of features with the names of classifiers are given in Table10 for Age classification. MB and RF gave accuracies of 49.14% and 44.85% for All P-BoW features respectively. All P-BoW and P-L is the group of 10 psycholinguistic features. This combined group showed an accuracy of 55.7% for RF and 53.7% by MB. Then the combination of all word based features gave an accuracy result of 58% and 53.7% for RF and MB respectively. Group of all character based features when combined gave accuracy results of 55.14% by RF and 50.57% for MB. The combination of all stylometric and content based features gave the maximum accuracy of 60% for RF and 52% by MB

For age classification, results show that RF performed better than MB. All P-L and P-BoW features gave a combined accuracy of 55.7% for RF that is greater than 55.14% - the accuracy reported by RF for all character based(44 in total) features combined. This shows a strong contribution of the P-L and P-BoW approaches(total 12 such features). We can infer from the results that best results(i.e., 60%) are reported by the set of all features combined using RF Decision Tree algorithm.

The accuracy values given by the classifiers for any set of features could not go beyond 60% for age classification, unfortunately. One reason of the proposed features for not being able to classify the age groups adequately can be due to the fact that the division of age groups is so closely related(15-19, 20-24, 25-xx) in terms of many demographic traits such as education level, income background and even the type of educational institute(university) that the authors have many overlapping traits.

Table 10. Age classification results using different groups of features.

Feature	Classifier	Accuracy	Feature	Classifier	Accuracy
All P-BoW	RF	44.85%	All character based	**RF**	**55.14**%
	MB	**49.14**%		MB	50.57%
All P-L and P-BoW	**RF**	**55.7**%	All word based[16]	**RF**	**58**%
	MB	53.7 %		MB	53.7%
All features combined	**RF**	**60**%			
	MB	52%			

4 Conclusion And Future Work

In this paper, we presented our approaches for gender and age prediction tasks on the training data that is SMS based multilingual dataset containing English and Roman-Urdu text of 350 documents each from a different user. We used stylometric and content based approaches to extract the features and reported 80.29% accuracy for gender and 60% accuracy for age prediction task for MAPonSMS contest. The trained prediction models, when used to predict the test set containing 150 multilingual documents, outperform the baseline approaches. The improvement in accuracy for gender prediction goes from 0.60%(baseline) to 0.69% and for age prediction from 0.51%(baseline) to 0.53%. The joint result accuracy improvement is from 0.32%(baseline) to 0.35%. [17]

The task of authorprofiling for multilingual text displays a great cushion for improvement, particularly for the gender classification task. To improve the results some approaches, such as, preprocessing the dataset to normalize the Roman-Urdu text(as discussed by [13]), introducing topic based features [1], and devising methods for word-sense dis-ambiguity to differentiate English and Roman-Urdu text can be implied.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. pp. 151–162. Springer (2016)
2. Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., Liu, X.: Through a gender lens: Learning usage patterns of emojis from large-scale android users. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web. pp. 763–772. International World Wide Web Conferences Steering Committee (2018)
3. Cheng, N., Chandramouli, R., Subbalakshmi, K.: Author gender identification from text. Digital Investigation **8**(1), 78–88 (2011)
4. Collier, V.P.: Age and rate of acquisition of second language for academic purposes. TESOL quarterly **21**(4), 617–641 (1987)
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. Gottschalk, L.A., Gleser, G.C.: The measurement of psychological states through the content analysis of verbal behavior. Univ of California Press (1969)
7. Huffaker, D.A., Calvert, S.L.: Gender, identity, and language use in teenage blogs. Journal of computer-mediated communication **10**(2), JCMC10211 (2005)
8. Kubát, M., Milička, J.: Vocabulary richness measure in genres. Journal of Quantitative Linguistics **20**(4), 339–349 (2013)
9. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: " how old do you think i am?" a study of language and age in twitter. In: ICWSM (2013)

[17] https://lahore.comsats.edu.pk/cs/MAPonSMS/de.html

10. 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)
11. Rangel Pardo, F.M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers. pp. 1–8 (2015)
12. Rosenberg, S.D., Tucker, G.J.: Verbal behavior and schizophrenia: The semantic dimension. Archives of General Psychiatry **36**(12), 1331–1337 (1979)
13. Sharf, Z., Rahman, S.U.: Lexical normalization of roman urdu text. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY **17**(12), 213–221 (2017)
14. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational linguistics **26**(4), 471–495 (2000)