MAPonSMS - Overview of the Multilingual SMS-based Author Profiling Task at FIRE'18

Muhammad Sharjeel, Mehwish Fatima, Saba Anwar, and Rao Muhammad Adeel Nawab

COMSATS University Islamabad, Lahore Campus, Pakistan.
{muhammadsharjeel,adeelnawab}@cuilahore.edu.pk
{mehwishfatima.raja,sabaanwar}@gmail.com

Abstract. This paper presents the overview of 1st International shared task of *Multilingual Author Profiling on SMS (MAPonSMS)* at *Forum for Information Retrieval Evaluation (FIRE'18)*. The aim of the *MAPonSMS* task is to identify the author's gender and age for a given multilingual (Roman Urdu and English) SMS messages profile, where each profile consists of an aggregation of SMS messages from a single author. This paper provides the details of the dataset and its distribution, overview of the submitted approaches and the evaluation framework used for measuring the performance of the submitted multilingual author profiling systems.

Keywords: Natural Language Processing, Multilingual Author Profiling, SMS Corpus, Roman Urdu

1 Introduction

Authorship profiling is a task where the objective is the identification of author's demographic traits (age, gender, native language, etc.) by analyzing the author's written text. The subject of author profiling is beneficial in many domains such as digital forensic analysis [16], marketing intelligence for business [11], sentiment analysis and classification for social and physiological behaviors [3]. The profile of an author can be either: (1) monolingual, or (2) multilingual. In the former case, the entire author profile is written in one language, while in the latter case, a single author profile will contain text in two or more languages. Author profiling on profile containing text in two or more languages is known as Multilingual Author Profiling [8, 7].

With the advancement of technology and the Internet, people can interact globally via different mediums (text messaging, social media, blogs, etc.). The phenomenon of multilingualism emerged due to the communications among various nationalities having different native languages. Because, people usually use a common language (like English) for global communication, but somehow, they have an inclination to their native language(s). Thus, these global connections

have influenced not only how the languages are being used among different communities, but also morphing the vocabularies of different languages. Moreover, multilingualism has also affected the texting trends (SMS messaging, chatting applications) in the past few years. It might be because a multilingual person tends to opt vocabulary from multiple known languages during a spontaneous speech or typing process. So, the research on multilingual text is attracting the attention of the research community due to the rapid growth of multilingual text.

The development and evaluation of automatic author profiling techniques demand standard evaluation resources in various languages and genres. Because the selection of language and genre influences the structure, style and content of the document. The example of structure based attributes is document length, sentence length, etc., While the example of style and content based attributes is vocabulary/ word construction, grammatical forms, punctuation choices, use of special characters/ emojis, etc. The impact of language and genre can be comprehended with the following cases. The SMS messages are considered short, having informal language, including slang and emojis. While, Facebook posts are regarded as a different genre (length can be short to long), also having informal language containing emojis. On the other hand, a book chapter or scientific article is classified as a different genre usually having long to very long document length where language is formal having dense vocabulary with proper grammar form. Due to this, the feature extraction process from different genres and languages would be very important for the training of the author profiling systems. Therefore, the selection of language and genre is very important because it can affect the robustness of the author profiling system.

In previous research studies, different genres (Twitter, Facebook, blogs, web forums) have been considered for mostly English and other European languages in monolingual setting [4, 24, 38, 32]. The SMS genre has been neglected for author profiling regardless of its global popularity, ease of use and access. The most probable reason of this negligence is its challenging and time consuming collection as a standard resource. In short, the problem of author profiling has not been thoroughly explored neither for South Asian languages (particularly Urdu and Roman Urdu) nor for SMS genre. Therefore, this competition focuses on multilingual (English and Roman Urdu) SMS-based author profiling.

The aim of MAPonSMS-Fire'18 (Multilingual Author Profiling on SMS) shared task is to identify the author's gender and age for a given multilingual (Roman Urdu and English) SMS messages profile, where each profile consists of an aggregation of SMS messages from a single author.

The rest of the paper is organized as follows. Section 2 discusses the existing work that has been done on SMS corpora and author profiling. Section 3 gives the details of train and test datasets, evaluation measure used to evaluate the performance of submitted systems, and system submission process. Section 4

describes the overview of submitted systems. Section 5 presents the results and analysis of submitted approaches. Finally, Section 6 concludes the paper.

2 Related Work

Although, collecting SMS messages for creating a standard evaluation resource is a very challenging task, however, few efforts have been made in developing datasets by using SMS messages for various tasks including SMS text normalization [22], linguistic[36], machine translation systems [34] and spam detection [10]. Among the existing SMS-based corpora, NUS SMS corpus[1] is the largest and most widely used SMS based dataset, which was initially developed to improve the predictive text in mobile devices [5]. Its first version was released in 2004 having English SMS messages and the second version came out in 2010 with an increase in the size of corpus as compared to the first release. The final corpus released in 2013, consisted of two sub-corpora for English and Chinese [5]. However, not all the profiles were associated with demographic information because many people shared only messages. The NUS SMS corpus has also been used for forensic authorship analysis task [13–15], authorship detection [25] and author identification [19].

To date, the PAN competitions provide a major contribution of benchmark monolingual corpora for identifying different author traits, particularly age and gender, in various languages and genres. In the 2013 PAN competition, English and Spanish blog posts were collected for monolingual age and gender prediction tasks [26]. In the 2014 PAN competition, four genres (hotel reviews, tweets, social media and blogs) in English and Spanish were considered for monolingual age and gender prediction tasks [29]. In the 2015 PAN competition, tweets were collected in four different languages, including English, Spanish, Italian and Dutch for monolingual personality trait detection, age and gender prediction [28]. In 2016, PAN competition task shifted from same genre author profiling to cross genre author profiling in monolingual setting. [30]. The train and test datasets of PAN 2014 were merged for this year competition. The training was carried out on tweets, and the test dataset constituted of blogs, social media and hotel reviews for monolingual age and gender prediction [30]. In PAN 2017, the task was gender and language identification for tweets considering four languages (Arabic, English, Spanish and Portuguese) [27]. In PAN competitions from 2014 to 2017, it can be noted that one out of four sub-corpora consisted of tweets.

Apart from PAN competitions, some research studies also explored tweet based datasets for the authorship analysis task, such as author identification [20, 17], gender identification [2, 37]. Few researchers carried out experiments on the combined datasets of SMS messages and tweets for sentiment analysis [18,

[1] http://www.comp.nus.edu.sg/ rpnlpir/downloads/corpora/sms/ Last visited: 22-09-2018

1]. Although, the construction of a tweet based corpus is quite easy due to having less privacy concerns and its readily availability, but tweets cannot be an alternative of SMS genre. It is because, SMS messages are purposely built for private conversations while tweets are meant for public conversations [7].

To summarize, the above mentioned corpora are predominantly monolingual (for English and other European languages) and are not suitable for South Asian languages such as Roman Urdu. Moreover, existing SMS and tweet based corpora are not suitable for the multilingual author profiling task. Therefore, this competition addressed the problem by providing a dataset of multilingual SMS based author profiles for gender and age prediction. We believe that this competition will foster research on multilingual text (in general) and Roman Urdu (an under-resourced language) more specifically.

3 Evaluation Framework

This section describes the characteristics of the train and test datasets, performance measure used to evaluate the performance of submitted systems, baseline approach and the procedure of submissions by the participants.

3.1 Corpus

A subset of SMS-AP-18 corpus [7] is used for the first shared task on *Multilingual Author Profiling on SMS*. The original SMS-AP-18 corpus consists of 810 author profiles. For the MAPonSMS-FIRE'18 shared task, a subset of 500 author profiles was selected from the SMS-AP-18 corpus. The reason for selecting a subset of original corpus is to have a balanced train/test dataset.

Train Dataset The train dataset consists of 350 multilingual (Roman Urdu and English) SMS based author profiles (see table 1 for detailed statistics). For gender, a multilingual author profile may belong to either Male or Female class. With regard to age, a multilingual author profile may fall into one of the three categories: 15–19, 20–24, 25–xx.

The gender and age information associated with each multilingual author profile were stored in a separate truth file which was provided with the train dataset. All author profiles in the train dataset were stored in the "*.txt*" format.

Test Dataset The dataset consists of 150 multilingual (Roman Urdu and English) SMS based author profiles (see table 1 for detailed statistics). The associated information (age and gender) was unknown for participants. All author profiles in the test dataset were also stored in the "*.txt*" format.

Table 1. Distribution of author profiles in train and test datasets for MAPonSMS-FIRE'18 task.

Age	Gender	Train Dataset	Test Dataset
15-19	Male	70	30
	Female	38	16
20-24	Male	112	48
	Female	64	28
25-xx	Male	28	12
	Female	38	16
	$\sum$	**350**	**150**

3.2 Performance Measure

The performance of submitted author profiling systems was computed using Accuracy measure. Accuracy is defined as the proportion of correctly classified author profiles.

$$Accuracy = \frac{No.\ of\ Correctly\ Predicted\ Author\ Profiles}{Total\ No.\ of\ Author\ Profiles}$$

We computed Accuracy in two ways: (1) Individual Accuracy of gender and age traits, and (2) Joint Accuracy of gender and age traits. The submitted systems were ranked based on Joint Accuracy score.

Baseline Approach For baseline approach, we used MCC (Majority Common Category) which is computed by assigning the most common category to all the instances in the dataset. The MCC of test dataset for: (1) Gender = 0.60, (2) Age = 0.51 and (3) Joint = 0.32.

3.3 Submission

The participants were asked to submit: (1) Executable multilingual author profiling system, (2) Output of the system (predictions) for the test dataset in "*.csv*" format for age and gender.

For multilingual author profiling system[2], some guidelines were provided: (i) It should be executable generically by commands for both age and gender so that it can be re-trained on demand for maximizing the sustainability. (ii) It should predict for each case found in the test corpus and write the output in .CSV file(s) for both age and gender. The results of multiple runs were not allowed for the submission.

[2] The participants retain the full copyrights of their submitted systems.

4 Overview of Submitted Systems

For the first MAPonSMS competition, a total of 9 submission were received, however, one of the participating teams did not submit the notebook paper. We now present the detailed analysis of the 8 approaches we received.

4.1 Preprocessing

Four of the total eight participants that submitted their systems used shallow text preprocessing before applying methods to extract features from the multilingual corpus. The authors in [6] cleaned the text by removing multiple space characters, tabs and garbage characters. In [35] only punctuation marks were removed while in [12] only case conversion (lowercasing) was applied during text preprocessing. The authors of [9] discarded stop words, punctuation marks and then lowercased the text in the preprocessing step. Four participating systems [21, 33, 31, 23] did not use any text preprocessing method.

4.2 Feature Extraction

In terms of methods used to extract features from the multilingual corpus, majority of the submitting systems [6, 23, 9, 31, 21, 35] opted for content based methods using BoW (Bag of Words) and *Tf-Idf* (Term frequency - Inverse document frequency). One of the participating team [33] used language dependent and independent style based methods. Another team [12] utilized style, vocabulary and emoticon based methods for feature extraction.

The authors in [6] used both word and character-based *Tf-Idf* whereas [21, 35, 9] used only word based *Tf-Idf*. Moreover, before applying *Tf-Idf*, [9] first normalised the text using a dictionary to translate Roman Urdu words to English. Another participant, [23] used *Tf* only and did not consider words with less than 5 occurrences. Furthermore, the authors in [35] applied a statistical approach to select the best features out of a large set of generated features.

Different style based (e.g. punctuation marks and other symbols, count of distinct words, words per line, number of lines etc.), vocabulary based (e.g. abbreviations, academic terms, contractions and slang words) and emoticons based (i.e. happy, sad, cry, unsure, squint, kiss and wink) features were extracted by [12]. The authors also experimented with different combinations of these three set of features. A set of stylistic features which are language independent (i.e. avg. word and sentence length, number long short words and sentences, number of different punctuation marks) and language dependent (POS-based e.g. number of adjectives, interjections, nouns etc.) were used by [33] during the feature extraction step.

4.3 Classification

All the participating systems employed supervised ML to identify age and gender from the multilingual text. Most of them used multiple ML classifiers and

reported the results using the best one(s). In some cases, age was reported with one classifier while gender with a different one. All the systems submitted for the task used Support Vector Machines as one of the classifier, however, the authors in [6, 23] used only Support Vector Machines. Apart from [21], all the other approaches used Random Forest too. Logistic Regression was another favorite used by 3 [9, 33, 21] submitted systems.

In [35], the authors experimented with 11 different classifiers i.e., Multinomial Naïve Bayes, Gaussian Naïve Bayes, Decision Tree, Random Forest, Extra Trees, Ada Boost, Gradient Boosting, Support Vector Machines, Stochastic Gradient Descent, Multi Layer Perceptron and Multinomial Naïve Bayes. They reported best results using Multi Layer Perceptron and Multinomial Naïve Bayes. In [9], Random Forest, Support Vector Machines, Logistic Regression and Naïve Bayes were used, Naïve Bayes outperformed others. The authors of [12], tried 3 different classifiers i.e. Random Forest, Naïve Bayes and Support Vector Machines. They showed that Random Forest for gender and Support Vector Machines for Age performed best. In [31], the authors went for Random Forest and Meta Bagging by Decision Tree as its component classifier. In [33], Naïve Bayes, J48, Random Forest and Logistic Regression were used for the classification task. The authors reported that Random Forest performed best for gender and Logistic Regression for age. In [21], Logistic Regression, Naïve Base, Multi-layer Perceptron and Gradient Boosting were used. Furthermore, the authors ensemble all four classifiers to report the best result.

5 Evaluation of the Submitted Systems

In this section, we discuss the results of the 9 teams that submitted their systems for the MAPonSMS task. Table 2 shows the age, gender and joint accuracies obtained by the submitting systems. As can be seen, majority of the systems performed better than the baseline accuracies. The highest reported accuracies are with *sharmila-18*, as they performed best in age, gender as well as joint accuracy. *abdul-18* secured the lowest results and is the only approach that is below the baseline. Expectedly, as the gender prediction was binary classification task whereas age was multi classification, all the team have performed better in the former. Moreover, the low scores obtained on age classification has effected the joint accuracies as well.

The approach used by *sharmila-18* [6] outperformed others and achieved the highest accuracy for the MAPonSMS task. Their team used both word and character based *Tf-Idf* features which resulted in its overall best performance. On the other hand, *thenmozhi-18* [35] and *ali-18* [21] achieved results very close to the *sharmila-18*, and they are among the top three. It can be observed that the top 3 ranked teams have used *Tf-Idf* for feature extraction from the multilingual corpus. Contrarily, the teams that utilized stylistic features are ranked last and 3rd last.

Table 2. Results of submitted approaches

Teams	**Gender**	**Age**	**Joint**
basline	*0.60*	*0.51*	*0.32*
sharmila-18	**0.87**	**0.65**	**0.57**
thenmozhi-18	0.85	0.63	0.52
ali-18	0.83	0.60	0.49
deepanshu-18	0.75	0.64	0.47
dijana	0.74	0.59	0.43
òscar-18	0.77	0.57	0.43
ramsha-18	0.73	0.53	0.38
asmara-18	0.69	0.53	0.35
abdul-18	0.55	0.37	0.23

The top two teams have used shallow text preprocessing methods before feature engineering which indicates that text preprocessing have shown positive impact on the results of the task.

6 Conclusion

In this paper, we present [6, 35, 21, 31, 33, 9, 23, 12] the results of the 1st International shared task of *Multilingual Author Profiling on SMS (MAPonSMS)* at FIRE'18. Given a reasonable and realistic collection of SMS messages for age and gender identification with multilingual setting was a challenging task and 9 teams participated in the competition.

Participants used several different methods for solving the task such as BoW (Bag of Words) based *Tf-Idf*, stylistic, vocabulary and emoticon based features. Majority of the participating teams performed better than the baseline accuracies. The highest age, gender, and joint accuracy (0.87, 0.65, and 0.57) was achieved by *sharmila-18* [6] by using word and character based *Tf-Idf* method and Support Vector Machines.

References

1. Aboluwarin, O., Andriotis, P., Takasu, A., Tryfonas, T.: Optimizing short message text sentiment analysis for mobile device forensics. In: 12th IFIP WG 11.9 International Conference on Advances in Digital Forensics XII. pp. 69–87. Springer, New Delhi, India (2016)
2. Alowibdi, J.S., Buy, U.A., Yu, P.: Language independent gender classification on Twitter. In: ASONAM '13: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 739–743. ACM, Niagara, Ontario, Canada (2013)
3. Anstead, N., O'Loughlin, B.: Social Media Analysis and Public Opinion: The 2010 UK General Election. Journal of Computer-Mediated Communication **20**(2), 204–220 (2015)

4. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1301–1309. Association for Computational Linguistics, Edinburgh, United Kingdom (2011)
5. Chen, T., Kan, M.Y.: Creating a live, public short message service corpus: the NUS SMS corpus. Language Resources and Evaluation **47**(2), 299–335 (2013)
6. Devi V, S., Kannimuthu, S., Safeeq, G., Kumar M, A.: KCE_DAlab@MAPonSMS-FIRE2018: Effective Word and Character-based Features for Multilingual Author Profiling. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
7. 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 **24**(5), 695–724 (2018)
8. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on Facebook. Information Processing & Management **53**(4), 886–904 (2017)
9. Gaur, D., Ayyar, M., Kumar Singh, A., Shah, R.R.: Multilingual Author Profiling from SMS. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
10. Giannella, C.R., Winder, R., Wilson, B.: (Un/Semi-)supervised SMS text message SPAM detection. Natural Language Engineering **21**(4), 553–567 (2015)
11. Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., Tomokiyo, T.: Deriving Marketing Intelligence from Online Discussion. In: KDD '05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. pp. 419–428. Chicago, Illinois, USA (2005)
12. Imran, R., Iqbal, M.: MAPonSMS'18: Multilingual Author Profiling using Combination of Features. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
13. Ishihara, S.: A forensic authorship classification in sms messages: A likelihood ratio based approach using n-gram. In: Proceedings of the Australasian Language Technology Association Workshop 2011. pp. 47–56. Canberra, Australia (2011)
14. Ishihara, S.: A Forensic Text Comparison in SMS Messages: A Likelihood Ratio Approach with Lexical Features. In: WDFIA 2012 : Seventh International Workshop on Digital Forensics & Incident Analysis. pp. 55–65. Crete, Greece (2012)
15. Ishihara, S.: A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. International Journal of Speech, Language & the Law **21**(1) (2014)
16. Juola, P.: Industrial Uses for Authorship Analysis. In: Mathematics and Computers in Sciences and Industry, pp. 21–25. INASE (2015)
17. Kebede, A.M., Tefrie, K.G., Sohn, K.A.: Anonymous Author Similarity Identification. In: 2015 5th International Conference on IT Convergence and Security (ICITCS). pp. 1–5. Kuala Lumpur, Malaysia (2015)
18. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research **50**, 723–762 (2014)
19. Kretchmar, M., Zhao, Y.: Text Message Authorship Classification Using Kernel Support Vector Machines. In: CSCI 2014: International Conference on Computational Science and Computational Intelligence. vol. 2, pp. 215–218. IEEE, Las Vegas, Nevada, USA (2014)

20. Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second. pp. 1–8. IEEE (2010)
21. Nemati, A.: Gender and Age Prediction Multilingual Author Profiles Based on Comments. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
22. Oliva, J., Serrano, J.I., Del Castillo, M.D., Igesias, A.: A SMS normalization system integrating multiple grammatical resources. Natural Language Engineering **19**(01), 121–141 (2013)
23. Orts, O.G.i., Rangel, F.: A statistical approach to gender and age range classification in multilingual corpus. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
24. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting Age and Gender in Online Social Networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC'11). pp. 37–44. ACM, Glasgow, Scotland, UK (2011)
25. Ragel, R., Herath, P., Senanayake, U.: Authorship detection of SMS messages using unigrams. In: 2013 IEEE 8th International Conference on Industrial and Information Systems (ICIIS 2013). pp. 387–392. IEEE, Sri Lanka (2013)
26. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers. Valencia, Spain (2013)
27. Rangel, F., Rosso, P., Potthast, M., Stein, B.: 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 (2017)
28. Rangel, 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. CEUR-WS.org, Toulouse, France (2015)
29. Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daeleman, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org, Sheffield, UK (2014)
30. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: 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. pp. 750–784. vora - Portugal (2016)
31. Safdar, A., Akhter, O., Inayat, O., Khalid, A.: Using Bag-of-Words and Psycho-Linguistic Features For MAPonSMS. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
32. Shrestha, P., Rey-Villamizar, N., Sadeque, F., Pedersen, T., Bethard, S., Solorio, T.: Age and gender prediction on health forum data. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA) (2016)
33. Sittar, A., Ameer, I.: Multi-lingual Author Profiling Using Stylistic Features. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th In-

ternational Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
34. Song, Z., Strassel, S., Lee, H., Walker, K., Wright, J., Garland, J., Fore, D., Gainor, B., Cabe, P., Thomas, T., Callahan, B., Sawyer, A.: Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
35. Thenmozhi, D., Kalaivani, A., Chandrabose, A.: Multi-lingual Author Profiling on SMS Messages using Machine Learning Approach with Statistical Feature Selection. In: Working Notes for MAPonSMS at FIRE'18 - Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation (FIRE'18). CEUR-WS.org, CEUR, DAIICT, Gujarat, India (2018)
36. Treurniet, M., De Clercq, O., Van Den Heuvel, H., Oostdijk, N.: Collecting a corpus of Dutch SMS. In: 8th International conference on Language Resources and Evaluation Conference (LREC 2012). pp. 2268–2273. European Language Resources Association (ELRA), Istanbul, Turkey (2012)
37. Vicente, M., Batista, F., Carvalho, J.P.: Improving Twitter Gender Classification using Multiple Classifiers. In: ESCIM 2016 : 8th European Symposium on Computational Intelligence and Mathematics 2016. pp. 121–127. Sofia, Bulgaria (2016)
38. Wanner, L.: Multiple Language Gender Identification for Blog Posts. In: Proceedings of the 37th Annual Meeting of the Cognitive Science Society. pp. 2248–2251 (2015)