A statistical approach to gender and age range classification in multilingual corpus

Òscar Garibo i Orts[1][0000−0001−8089−1904],
Francisco Rangel[2][0000−0002−6583−3682]

[1] Optical Tech and Support, València 46025, SPAIN
[2] Autoritas Consulting, València 46011, SPAIN

Abstract. This paper describes a statistical approach to the task of gender and age range classification in SMS. Our approach started developing our own implementation of Low Dimensional Representation (LDR) method, with the idea to add some other statistics which had not been used in the original implementation, such as skewness, kurtosis and central moments. The proposed method calculates term frequencies and uses 3 statistics per class: mean, standard deviation and skewness.

Keywords: Gender identification · age identification · author profiling · statistical approach · LDR.

1 Introduction.

The Author profiling task is widely studied and some new ideas arise from time to time [4,3,1]. We have pursued a model to classify SMS that fits into Big Data environment. We could consider models based on Deep Learning, which are time and resources expensive to build and train, and to predict as well. Our goal, in the scenario on MAPonSMS Author Profiling task was to test our algorithm with a multilingual corpus for gender and age range classification. The presented approach implements a variation of the Low Dimensional Representation (LDR) method [2] with new statistics: skewness, kurtosis and central moments. While LDR uses 6 characteristics per class, the proposed algorithm reduces the number of characteristic to 3 per class, which is interesting in the context of a Big Data application. Whereas in BoW or similar approaches every word is a characteristic, our method reduces the number of characteristics to 3 per class by codifying the words in probabilities per class, thus calculating the average, standard deviation and skewness. This is implemented retrieving the associated values from previously built dictionaries, which is very efficient. In a Big Data environment, velocity tends to be critical, and our method speeds up the process.

2 Corpus.

The MAPonSMS Author Profiling Train corpus is composed by one single collection of SMS messages written in Roman Urdu and English languages mixed.

The goal of the task is to classify the user by gender (Male/Female) and by age range (15-19, 20-24, 25-xx). We were provided with SMS messages from 350 different users, being the number of SMS messages variable. So, we had not the same number of SMS messages for each user. In regard of gender, the classes were not balanced, being 60% of the SMS labeled as Male, thus 40% Female. Age range classes were again not balanced. 30.86% of the users belong to 15-19 class, 50.29% to 20-24 class and 18.85% to 25-xx class.

Table 1. Class distribution for Gender in Train Dataset.

	Male	Female
% of users	60%	40%

Table 2. Class distribution for Age Range in Train Dataset.

	15-19	20-24	25-xx
% of users	30.86%	50.29%	18.85%

3 Methodology.

Our goal was to develop a method that was not language dependent, and that required no prior knowledge of the language. We started implementing TF representation for each user in the dataset, counting how many times each word appears in each user and globally for all users. We decided to use TF since this way we could represent a priori class dependent probability for each term for each class simply by counting the number of times a term occurs for each class, and then dividing this amount by the total number of times this term shows for all classes. For the sake of understanding we wanted to be sure we were dealing with probabilities. In addition, calculating TF is less time and resource consuming than calculating Tf-Idf (as in the original LDR). We build a vocabulary set including each word we have seen in the training corpus. We decided to discard the words that appear less than 5 times in the corpus, which highly reduces the size of the resulting dictionaries. Then, we went over the training corpus, one user each time, checking for words for this user and writing down into an array the related a priori probability for each class. Finally, we got one vector per class per user (2 for gender, 3 for age range) with the a priori probability of each word to pertain to each class. Then, we can calculate the different statistics from these a priori probabilities arrays that represent the text used by each user. Once we had the average, standard deviation and skewness for every user in the training dataset we used these characteristics to train a LinearSVC machine learning algorithm from Python's Sklearn library. This is a Support Vector Machine with Linear kernel, where multiclass support is handled according to a

one-vs-the-rest scheme. It is important to notice that when we want the model to include more users or SMS messages to build the a priori probability vectors, we only have to run the procedure with new labeled SMS messages. This new vector is what we will use to predict new incoming SMS messages. The whole procedure is simple and fast, and can be done in parallel. Once the new vectors are built, we only need to point the algorithm to these new vectors. This is an easy way of keeping the vectors up to date, and more importantly, this is a clean, fast and reliable updating procedure. The code used in this task can be found at `https://github.com/OscarGariboiOrts/MAPonSMS18`

4 Evaluation results.

In Table 3 we present our results vs Baseline. Baseline was built by controlling class. Our results improve Baseline for gender by almost 30% and age range by almost 12%. We have stated that we were removing from the vocabulary all words appearing in less than 5 users. This is a parameter that can be adjusted depending on the dataset characteristics and on the task involved. We would have liked to submit more than 1 single submission, as was stated in the task, since we have seen that removing more or less words might have a big influence in the algorithm accuracy. Even more, this is important in such a small dataset as the one used in this task. Our classification method ends up with the accuracies shown in Table 3.

Table 3. Classification accuracies.

	Gender	Age range	Joint
% Accuracy	0.77	0.57	0.43
Baseline	0.60	0.51	0.32

The method is using statistical parameters calculated from a priori probabilities of each word to belong to one of the classes. The average, standard deviation and skewness for these distributions should reflect the fact that different classes use language in different ways. At the end, we are relying on the mathematical concept the more similar 2 distributions are, the more similar their statistics will be. Hence, feeding the Support Vector Machine (SVM) with these statistics SVM should be able to find an hyperplane to linearly separate the related classes. In [1] we showed the importance of skewness in a 2 class classification problem. Here we have used skewness together with average and standard deviation to calibrate the accuracy of the method.

5 Literature review.

Works on Multilingual Author Profiling are rare. "Multilingual author profiling on Facebook" [5] refers to the use of state-of-the-art author profiling techniques,

such as content based features (word and character Ngrams) and 64 different stylistic based features (11 lexical word based features, 47 lexical character based features and 6 vocabulary richness measures) for age and gender identification on multilingual corpora. These techniques rely on lexical and stylistic features, whereas the method we presented relies on the fact that different classes use language in different ways. Men and women use language in different manner [6]. Age range classification is usually approached based on language features [7], which basically mean different ranges will use language in a different way. This is what we try to capture in the proposed method, which relies in the mathematical distribution of the words used by every different class.

6 Conclusions and future work.

In this paper we presented a method to classify Gender and Age rank in the MAPonSMS International Author Profiling Shared Task at Forum for Information Retrieval Evaluation (FIRE'18). We have shown that codifying the text with the statistical features extracted from a priori class dependent arrays is an easy, cheap and language independent method to classify texts with different features. We are always relying on the fact that different classes will use the language in different ways that our algorithm can extract information from. A bigger dataset should give us a clearer idea of how good the performance can be. We need to check how critical in the performance eliminating words can be. For age range we could consider not removing any word. In fact, during task works we saw that eliminating words that appeared less than 3 times made us get a test set prediction with the same % of Male and Female labels as we got in the training set. But as 1 single submission was finally allowed, we decided to rest in the number we had already made some tests in-house. We could try to introduce some other statistics as central moments, minimums, maximums, etc, to see whether these new features affect to the algorithm performance.

References

1. Garibo-Orts, O. A Big Data approach to gender classification in Twitter. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). In: Patrice Bellot and Chiraz Trabelsi and Josiane Mothe and Fionn Murtagh and Jian Yun Nie and Laure Soulier and Eric Sanjuan and Linda Cappellato and Nicola Ferro Editors (2018)
2. Rangel, F., Rosso, P., Franco-Salvador, M. A low dimensionality representation for language variety identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing, Springer-Verlag, LNCS (2016)
3. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Linda Cappellato and Nicola Ferro and Jian-Yun Nie and Laure Soulier (Eds.) CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2018)

4. 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: Cappellato L., Ferro N., Goeuriot L, Mandl T. (Eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866 (2017)
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. Xiufang Xia (2013), "Gender Differences in Using Language". ISSN 1799-2591. Theory and Practice in Language Studies, Vol. 3, No. 8, pp. 1485-1489, August 2013
7. Morgan-Lopez AA, Kim AE, Chew RF, Ruddle P (2017) Predicting age groups of Twitter users based on language and metadata features. PLoS ONE 12(8): e0183537. https://doi.org/10.1371/journal.pone.0183537