Author Profiling: Age Prediction Based on Advanced Bayesian Networks

Seifeddine Mechti [1], Maher Jaoua[2], Rim Faiz[1,3], Heni Bouhamed[2] and Lamia Hadrich Belguith[2]

[1] LARODEC Laboratory, ISG of Tunis B.P.1088, 2000 Le Bardo, Tunisia
[2] ANLP Group, MIRACL Laboratory, University of Sfax, BP 1088, 3018, Sfax
[3]IHEC of Carthage, 2016 Carthage Présidence, Tunisia

mechtiseif@gmail.com,Rim.faiz@ihec.rnu.tn
{maher.jaoua,l.belguith}@fsegs.rnu.tn,
heni_bouhamed@yahoo.fr

Abstract. In this study, we present a new method for profiling the author of an anonymous English text. The aim of author profiling is to determine demographic (age, gender, region, education level) and psychological (personality, mental health) properties of the authors of a text, especially authors of user generated content in social media. To obtain the best classification, authors resort to machine learning methods. Focusing on the works which use the Bayesian networks, all those methods rather apply the Bayesian naïve classifiers which do not yield the best results. Therfore we propose a method based on advanced Bayesian networks for age prediction to over come the mentioned detail problem. We obtained promising results by relying on an English PAN@CLEF 2013 corpus. The obtained results are comparable to the ones obtained by the best state of the art methods. The software and data can be publicly downloaded from www.cicling.org/2016/data/248/CICLING_248.zip.

Keywords: Author profiling, advanced Bayesian networks, age prediction

1 Introduction

There is no doubt that social networks are experiencing significant growth. Social networks require profiling from their users. These users provide false information about themselves. In 2012 Facebook estimate that there were 83 million false profiles[1]. The detection of user profiles in a discussion is an important piece of information for the providers of certain services. This is specifically to study the way in which certain linguistic characteristics vary depending on the profile of the author of a text. Author profiling can be used in other circumstances, for example, in forensic linguistics; the detection of the linguistic profile of the author of a text could be ex-

[1] http://edition.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/

tremely valuable for evaluating the suspects. Similarly, in the marketing perspective, companies may be interested in determining what types of people prefer their products. In the literature, many works have focused on the classification of a conversation or a given text and more precisely on the detection of the age, gender, native language and personality of the author [1].

In this paper we present our method for predicting the age of an author based on his/her linguistic attributes. We resorted to the use of advanced Bayesian networks. The paper is organised as follows. Section 2 summarises related research regarding author profiling. Section 3 presents our method of age prediction based on advanced Bayesian networks. Section 4 reports experiments and evaluation carried out using the advanced Bayesians networks. Finally, conclusions are stated and future lines of research are analysed in Section 5.

2 Related work

The detection of the author's profile is the study of how linguistic features vary depending on the authors profiles [2]. The study achieved by the pioneers Koppel et al. has shown that there are linguistic differences between men and women. Indeed, men who prefer to categorize things, use more determiners (the/ this / that, a, etc.) and quantifiers (two, more, a few, etc.). Women, more than men, resort to personal pronouns (I, you, me, etc.) [3]. Argamon [1] worked on the British National Corpus. They used part of speech features. They were able to get 80% accuracy for the prediction of gender. In another study [4] the authors worked on segments of blogs using features such as punctuation, average words/ sentence length, part of speech and word factor analysis. They achieved a gender prediction rate of 72.2%. Peersman [5] used a corpus of Netlog trying unigrams, bigrams, trigrams, and tetagrams. They obtained an average accuracy of 88.88% for the prediction of the authors' age and gender.

In [6], the authors worked on the automatic classification of emails; they got a rate of 81.5% of well classified documents for the gender dimension and 72% for the age dimension. The works of [7]; [8] showed promising results regarding the detection of the author's gender in chats. Recently, [9] tried to perform the prediction of age in conversations among dutch Twitter users. Although the documents are very short (an average length of less than 10 words), 74% of the discussions were highly ranked. In fact, the authors were able to find a mean absolute error between 4.1 and 6.8 years. Pennebaker [10] relies on the change of language features for the prediction of some personality traits of authors in discussions [11]. The author considered unsupervised learning to detect the personality traits of the authors in texts. Besides personality, [12] used the logistic regression method or the binomial model for the detection of the author's native language.

To ensure on effective prediction, authors resort to preprocessing. Indeed, in their work, [13] resort to HTML Cleaning to obtain plain text and discrimination between human-like posts and spam-like posts , while Ashok [14] use the deletion of URLs, hashtags and user entries in Twitter. On the other hand, [15] uses case conversion, invalid characters, multiple white spaces and tokenization and the selection of subcorpus. The study of [3] distinguished two types of attributes: style based features and content-based features. To determine the age or gender of the author of a document, it

is important to consider the function words. Prepositions, pronouns and determiners have shown their effectiveness in an author's profile detection process [16]. In other works, the authors resort to the frequency of punctuation, of capital letters and of citations [17]. HTML attributes such as the URL of an image or the links of a Web page have been used by [18]. In the works of [19], the authors relied on specific vocabulary items (foreign words) to distinguish between authors. These terms are tags in the Stanford Core NLP tagger such as meeee, yessss, thy, u, sisters, etc. Unlike other authors, [20] resort to calculating the frequency of emoticons as one of the discriminating attributes to predict the author's profile. In [21] for instance, the authors resort to Automated Readability Measures such as the readability index, the Coleman-Liau Index, the Rix Readability Index, the Gunning Fog Index and the Flesch-Kinkaid Index. [22] use stylistic features: frequencies of punctuation marks, size of sentences, words that appear once and twice, use of deflections, number of characters, words and sentences. Ashok [14] uses Lexical Analysis such part of speech, proper nouns and character flooding in this choice of attributes and even attributes which are rarely used like those of emoticons have been considered in the work of [2].

In addition to the style used, the content of documents can be of great help in the classification process. What differentiates several age classes, for example, might be the content of their discussions. Indeed [1] distinguished several classes to categorize the authors. For the English language, they identified classes like home, smartphone, games, sports, Job, Marketing, etc. Then, they choose the first k attributes providing the best discrimination. [23] uses content features (n-grams, bag-of-words) while Ashok uses Dictionaries per subcorpus and class, lexical errors, foreign words and specific phrases like : 'my husband', 'my wife',' my son', etc. Finally, [24] uses second order representation based on relationships among terms, documents, profiles and sub-profiles.

However, the major drawback of content based attributes is that they depend on the psychological and mental state of the author (negative emotions, positive emotions) when writing, which might distort the classification results. In order to obtain a prediction for different output classes, several methods and machine learning algorithms were used like Logic Boost, Rotation Forest, Multi-Class Classifier, Multilayer Perceptron, Single Logistic, Logistic Regression, Multinomial Naive Bayes, Random Forest, Support Vector Machines [14]. In another work [22] uses his own frequency-based prediction function. To our knowledge, focusing on the works that use the Bayesian classifiers, we found that all those methods rather apply the Bayesian naive classifiers which do not yield good results for author profiling. In fact, in the work of [23] the author reached 39% accuracy for Blogs, 31% for hotel reviews, and 35% for social media. These results are relatively poor and reflect the ineffectiveness of the naive Bayesian classifiers. As a solution, we resort to advanced Bayesian networks to ameliorate the process of profiling anonymous authors (section 4).

We note that function words serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Function words might be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles, all of which belong to the group of closed-class words.

3 Proposed Method

As shown in Figure 1, our method is composed of four steps:
Preprocessing: The raw text obtained from the crawlers has to be cleaned to re- move noisy data, tags, urls, hashtags etc. The presence of this noisy data could affect and reduce the accuracy of the entire analysis. The cleaned data is then pushed into a database.

Text analysis: We started by calculating the number of occurrences of all words found in the corpus ranking them in order of their appearances. However, we focused onto the first 200 attributes only. We calculated CF (the class frequency) for each class of attributes in order to measure the frequency of occurrence of each class of attributes in each document of the corpus.

Feature set generation: The most common approaches in the literature distinguish two main types of attributes that can be used to detect the author's profile: the stylistic and the content based ones [10]. We manually grouped the terms belonging to the same class of attributes. We identified 15 classes, namely: Prepositions, Pronouns, Determiners, Adverbs, Verbs, No, Of, I, Medicine, Music, Sport, Phone, Beer, Love, Money. For the 'gender' output class, we realized that the purely stylistic attributes yield good results (based on style). Indeed, we selected three attributes: prepositions, pronouns and verbs. These attributes give good performance with decision trees. In addition, for the age output class, we used both of the content based and stylistic based. For this dimension, each age class discusses well-defined topics.

Classification: It is possible to construct an effective classifier using Bayesian networks [25]; [26]; [27]. A Bayesian classifier has n + 1 nodes for a model with n variables. In the classification models, there is necessarily a discrete multinomial central node which has k modalities corresponding to the class; it can be called "class node" and is added to other nodes of descriptive variables. Descriptive variables are denoted by Xi (i from 1 to n). The simple Bayesian classifier structure is that of the naive Bayesian network classifier also called Naive Bayesian (CBN) [28]. For these CBNs, the inter-variable descriptive correlations are not shown and all descriptive variables contribute equally to the classifier. The class node uses the information from each attribute independently of information from other attributes, which is very limited and not optimal for a classification problem. Accordingly, there have been several CBN structure enrichment proposals considering the possible correlations between the descriptive nodes. In [26], the authors proposed the Tree Augmented Naive Bayes method (TAN) in order to enrich the network structure using the shaft structure [29]. [30]. The construction of this structure is not greedy in computational complexity, but the restriction of the number of parents of a node to 2 (1 + class node) represents a real gap and risks taking the model away from reality. The resulting structure represents neither cases where a variable is correlated with several other descriptive variables nor the case where a variable is independent of all the others (in this case the node representing this variable only needs the class node as parent and the addition of another parent node only increases the complexity of the learning of the settings). Consequently, other authors proposed the use of the Augmented Naive Bayesian (BAN)

networks [26]; [31] where the addition of arcs between descriptive variable nodes is carried out with algorithms which do not impose any restrictions. Other authors simply proposed the use of general methods for the learning of Bayesian network structures (GBN) [26]; [31]) where the class node is regarded as an ordinary node and is not automatically connected with all other nodes [32]; [30]; [33]; [34]; [35]; [36]. It is true that thanks to the last we obtain closer to reality Bayesian classifier structures and therefore, the possibility of have more efficient classifiers. Hence, for the building and operation of the Bayesian network, we will use the Bayesian network toolbox (BNT) [37] running with the matlab software (Version 2010). Specifically, we will use the "Greedy Search" (GS) [38] for the learning structure and the 'Click-tree propagation algorithm [39] for the inference. We used a portion of 30000 examples for the learning phase and another portion of 30000 examples for the test phase of the classifier.

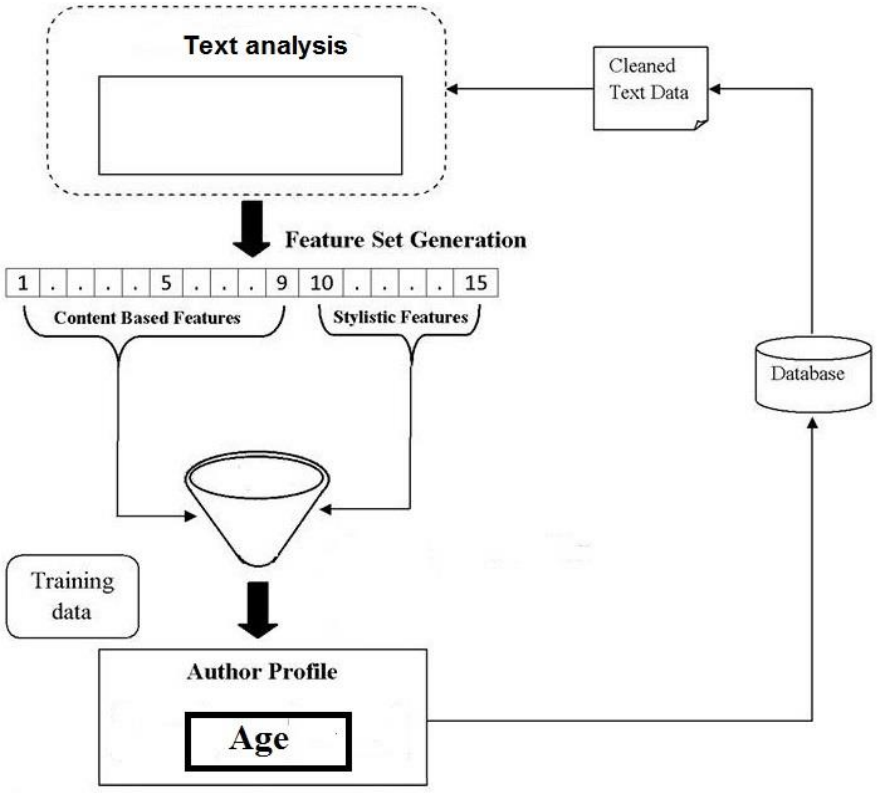

Fig. 1. System architecture diagram

4 Experimentation and Evaluation

4.1 Dataset Description

In our data, the adopted documents are blog posts written in English. The variety of themes provides a wide spectrum of topics, making the task of determining age and

gender more realistic. The age groups were defined according to [40]: the 10s is the class of individuals between 13 and 17 years old, the 20s are those between 20 and 33 year olds, and finally the 30s are those between 33 and 47 years old. Table 1 summarizes the contents of the corpus. We note that each file has a different author and more files cannot have the same author.

The corpus consists of 236600 files for training and 25440 for testing. For machine-learning, the class of 30s includes 133508 authors unlike the class of 10s which includes only 17200 files. The corpus is balanced in terms of gender but imbalanced in terms of age.

Table 1. Dataset description

		Number of authors	
Age	Gender	Training	Test
10S	Male	8600	888
	Female	8600	888
20S	Male	42828	4576
	Female	42875	4598
30S	Male	66708	7184
	Female	66800	7224
Total		236600	25440

4.2 Baseline Method

For comparison purposes a baseline was used so as to evaluate one's own results. We rely on the results of [41] in PAN@CLEF2013[2] as a baseline method. They ranked 3rd in this competitive conference. Using the free learning software Weka[3], this method started with the construction of ARFF (Attribute Relationship File Format) age dimension. The features are collected and then fed into an ensemble classifier. For categorization, authors used decision trees classifier (J48) due to speed and accuracy. The classifier is trained with the whole data corpus and used later for testing purposes. They got a good classification rate of 0.58.

4.3 Results Based on Advanced Bayesian Networks

Based on the advanced Bayesian networks, the proposed method has good performance. According to the confusion matrix, for the age prediction we got a good classification rate of 0,6175. Compared to the results reached with the decision trees, we notice the added values brought about by the Bayesian networks in this classification. Also, a good classifier is expected to yield the best recall measure. Indeed, the classifier retrieves 74,5% of the relevant documents against 55% with decision trees.

[2] http://pan.webis.de/clef13/pan13-web/author-profiling.html

Table 2. Confusion matrix for age prediction

	10s	**20s**	**30s**	**Total**	**Accuracy**
1	3124	1782	5094	10000	0.3124
2	0	5537	4463	10000	0.5537
3	0	132	9868	10000	0.9868
Total	3124	7451	19425	30000	**0.6176**

5 Conclusion

In this study, we have performed a document categorization so as to provide an author profile classification according to his/her text's characteristics. Content based attributes could be discriminative elements in the documents partitioning among age classes. Such a deduction can be predicted since children, the middle aged adults and elderly people never discuss the same topics. The improvements of our performance are mainly due to the proposition of a new method based on advanced Bayesian networks for classification. The performances of these networks prove their effectiveness in terms of accuracy and recall. It can be concluded that the use of the lexical classes is not enough. That is why, and as a perspective, we intend to integrate other aspects like syntax, morphology and semantics. Furthermore, to allow a better author detection we think of going beyond the age dimension and consider the detection of the native language and geographical data of the author and above all the detection of his/her personality. The software and data can be publicly downloaded from www.cicling.org/2016/data/248/CICLING_248.zip.

References

1. Argamon, S., Koppel M., Pennebaker J., and Schler J. Automatically detection the author of an anonymous text. Communications of the ACM, p 119-123. 2009.
2. Maharjan S., Shrestha P, and Solorio T. A Simple Approach to Author Profiling in MapReduce. England. CLEF 2014.
3. Koppel M. Argamon S. and Shimoni A., Automatically categorizing written texts by author, gender, Literary and Linguistic Computing, pages 401-412, 2003.
4. Zhang., C and Zhang.,P .Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010.
5. Peersman C., Daelemans., W, and 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, pages 37-44, New York, NY, USA, ACM.2011.
6. Gaustad T., Estival D. and Hutchinson B. TAT: an author profiling tool with application to Arabic emails. Proceedings of the Australsian Language Technology Workshop, pages 21-30, Melbourne, Australia, 2007.
7. Hariharan, S.,Gender Prediction in Chat based Medium s Using Text Mining, in: International Journal of Research and Reviews in Information Sciences (IJRRIS), Pakistan, 2011.
8. Kose C., Ozyurt O, and Amanmyradov G. Mining Chat Conversations for Sex Identification, Emerging Technologies in Knowledge Discovery and Data Mining (PAKKD), Nanjing, China, 2007.

9. Nguyen D, Gravel R, Trieschnigg D, and Meder T. "how old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.
10. Pennebaker., J. The secret life of pronouns: What our words say about us. pp. 401-412. 2011.
11. Celi F. Unsupervised Personality Recognition from Text: Possible Applications. In proceeding of PAN at CLEF, England. 2014.
12. Sze-Meng Jojo., W and Dras.,M. Contrastive analysis and native language identification. In Proceedings of the Australasian Language Technology Association Workshop, pages 53-61, Sydney, Australia. 2009.
13. Marquardt., J, Fanardi,G., Vasudevan, G., Moens., M, Davalos., S, Teredesai, D and De Cock. M. Age and Gender Identification in Social Media.PAN at CLEF 2014.
14. Rangel F, Rosso P, Koppel M, Stamatatos E, and Inches G. Overview of the Author Profiling Task. PAN@CLEF. Valencia, spain .2013.
15. Edson R. D. Weren, Viviane P. Moreira, and Jose P. M. de Oliveira. Using Simple Content Features for Author Profiling PAN@CLEF.Valencia,Spain.2013.
16. Fermin, L., Cruz., Rafa Haro R., and Javier Ortega, F. ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling. Notebook for PAN at CLEF .Valencia. 2013.
17. Aleman Y., Loya N, Vilarino D, and Pinto D. Two methodologies applied to the author profiling task. Notebook for PAN at CLEF. Spain. 2014.
18. Sapkota ,U., Solorio, T., Montes-y-Gomez, M, and De-la-Rosa, G. Author Profiling for English and Spanish Text .PAN at CLEF. Spain. 2013.
19. Gopal Patra, B., Banerjee, S, Das., D, Tanik., S, and Sivaji Bandy.,O . Automatic Author Profiling Based on Linguistic and Stylistic Features. PAN at CLEF 2013.
20. Irazu D., Farias H., Guzman-Cabrera R.,Reyes A and Rocha M. Semantic-based Features for Author Profiling Identification: First insights- Notebook for PAN at CLEF. Spain. 2013.
21. Gressel., G, Hrudya P, Surendran K, Thara S, Aravind A, and Prabaharan P. Ensemble Learning Approach for Author Profiling.PAN at CLEF.England. 2014.
22. Baker ., C. Proof of Concept Framework for Prediction.Pan@CLEF. England. 2014.
23. Villena-Roman,J and Gonzalez-Cristobal,J. DAEDALUS: Guessing Tweet Author's Gender and Age. PAN@CLEF.England.2014.
24. Pastor L, Montes-Y-Gomez M, Escalante H, Villasenor-Pineda L and Villatoro-Tello E. INAOE's Participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013.
25. Langley, P., Selection of relevant features in machine learning. Acts of AAAI Fall Symposium on Relevance, p. 140-144.1994.
26. Friedman, N., Geiger, D., Goldszmid, M., Bayesian Network classifiers, Machine Learning, p. 131-163.1997.
27. Pernkopf, F., Bayesian network classifiers versus selective k-NN classifier, Pattern Recognition, p. 1-10.2005.
28. Domingos, P., Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss, MachineLearning, 1997, p. 103-130.
29. Chow, C., Liu, C., Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory, 14 (3), p. 462-467.1968.
30. Madden, M. G., A New Bayesian Network Structure for Classification Tasks , Actes de la 13th Irish Conference on Artificial Intelligence and Cognitive Science, 2002, p.203-208.
31. Cheng, J., Greiner, R., Learning Bayesian belief network classifiers: algorithms and system , Actes de la 14th Canadian Conference on Artificial Intelligence, 2001, p. 141-151.
32. Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P.,Optimization by simulated annealing , Science, , 1983, p. 671-681.

33. Stuart, M., Yulan, H., Kecheng, L.,Choosing the best Bayesian classifier : An empirical study, IAENG International Journal of Computer Science, 2009, p. 1-10.
34. Carta, J. A., Velazquez, S., Matias, J. M., Use of Bayesian networks classifiers for long term mean wind turbine energy output estimation at a potential wind energy conversion site , Energy Conversion and Management,p.1137-1149. 2011.
35. Lerner, B., Malka, R., Investigation of the K2 algorithm in learning Bayesian Network Classifiers, Applied Artificial Intelligence, p.74-96.2011.
36. Bouhamed, H., Masmoudi, A., Lecroq, T., Rebai A., Reducing the structure space of Bayesian classifiers using some general algorithms, Journal of Mathematical Modelling and Algorithms in Operations Research(Springer), Volume 14, Issue 2, pages 197-237.2015.
37. Murphy, K, The BayesNet Toolbox for Matlab , Computing Science and Statistics : Interface .33,http://www.ai.mit.edu/murphyk/Software/BNT/bnt.html.2001.
38. Chickering, D., A Transformational Characterization of Equivalent Bayesian Network Structures, Actes de la 11th Conference on Uncertainty in Artificial Intelligence (UAI'95), San Francisco, CA, USA, Morgan Kaufmann Publishers, , p. 87-98.1995.
39. Lauritzen, S., Spiegelhalter, D., Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society, 50, p. 157-224.1988.
40. Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. Effects of age and gender on blogging. AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 2006.
41. Mechti S., Jaoua M. Belguith L and Faiz R., Author profiling using style based features. In proceeding of PAN at CLEF, Spain. 2013.