An empirical method for evaluation of author profiling framework

Bsir Bassem
LATICE Laboratory Research Department of Computer Science / Tunisia
bsir.bassem@yahoo.fr

Mounir Zrigui
LATICE Laboratory Research Department of Computer Science / Tunisia
mounir.zrigui@fsm.rnu.tn

Abstract

Author's profile analysis allows identifying some characteristics of the authors, such as gender, age, mother tongue and character. Indeed, the problem of author profiling has gained great importance, in the last few years, thanks to its new applications and utility in marketing and legal linguistics.

In this paper, we developed a novel approach for age and gender detection. We used different techniques for classification combined with linguistic, stylistic and structural features to determine the author's gender and age. The experimental results and the encouraging achieved accuracies approve that our approach is efficient in age and gender identification.

1 Introduction

Authorship Analysis aims at extracting information about the authorship of documents from features within those documents. It is based on combining three different techniques, namely Authorship Profiling, Authorship attribution or Identification and Plagiarism Detection. In fact, Author profiling is the task of determining demographic features of authors like native language, education, gender, age and personality traits of an author by understanding their writing styles.

With the development of social networks, Facebook has become a widely-used social network containing different types of documents (text, images, video, etc.). Indeed, people generally use it to express their opinions and exchange their ideas in different languages (French, English, Arabic, etc.), which makes it an important and rich source for the extraction of corpus to determine the author's profile (Markovikj and al., 2013).
The objective of our approach is to find the author's age and gender from text messages written in the forums, by analyzing the vocabulary used by each user.

This article is organized as follows: In the first section, we define author profiling. In the following section, we present the data set on which our tool was trained and tested. Section 4 presents the different types of the extracted features. Section 5 depicts the system architecture. Then, we illustrate the conducted experiments and the obtained results in section 6. Finally, we end the paper with brief conclusion and possible future works.

2 Related Work

The writing style of the words reflects the mental, social and even the physical as well as the psychological state of the authors. Indeed, statistical studies exploring the stylistic features of a text have begun one century ago with T. Mendenhall. Afterwards, several linguistic, sociological and natural language processing approaches were introduced.

The research works on profile detection can be divided into those proposed before the rise of the social media and those introduced after it. At the beginning, author profiling approaches, such as that of koppel et al., were based on formal written texts. Indeed, researchers, in (Koppel and al., 2002) represented 604 documents from the British National Corpus (BNC) in the form of trees whose roots are words sets or parts of the discourse. Then, they applied automated text categorization

techniques. They obtained 80 % accuracy to infer the gender of the author.

Researchers, in (Corney and al., 2002), described an investigation of authorship gender attribution mining from e-mail text documents. They used structural characteristics and gender-preferential language features together with a Support Vector Machine learning algorithm to identify the author's gender. The corpus of e-mail documents used in their experiments to evaluate the author's gender categorization method was sourced from the inboxes of academic organization members (more than 15,000 users). They obtained 70.2 % precision rate for gender detection.

Recently, thanks to the incompatible volume of data in social networks, such as Twitter and Facebook, more interest has been given to other kinds of writing that are more colloquial, less structured and formal. For instance, Koppel and Peneebaker, analyzed a corpus of 71,000 blogs incorporating almost 300 million words. They used the learning algorithm Multi-Class Real Winnow (MCRW) to learn models that classify blogs according to the author's gender and age. They obtained 43.8% and 86% for age and gender accuracy prediction, respectively (Schler and al., 2006).

We can also mention, for instance, the study carried out by Zhang and al. who examined 10,000 short blog messages; each of which contains 15 words. They got 72.1% accuracy for gender prediction. (Zhang and al., 2011)

To identify age and gender, authors in (Bayot and al., 2016) used an approach based on SVM algorithm and word embeddings. They employed a number of features equal to 100 in the test carried out on PAN 15 dataset. Their model achieved 44.8% and 68.2% for age and gender classification accuracy, respectively.

The authors, in (Bilan and al., 2016), built a Crossgenre Author Profiling System (CAPS). This classification system considered part-of-speech, collocations, connective words and various other stylometric features to differentiate between the writing styles of male and female authors as well as between different age groups. Their system attained 74.36% accuracy for gender identification on the test set for English corpus collected from PAN 2015. Besides, for age classification, they reported accuracy rate equal to 44.87%.

In (Dichiu and al., 2016), researchers applied SVM classifier and neural network on TF-IDF and verbosity features. Results showed that SVM classifiers are better for English datasets. They proved that neural networks performed better for Dutch and Spanish datasets. For English, the best findings were obtained using a tf-idf at character level combined with the verbosity feature. Their results are almost similar to those provided in (Bayot and al., 2016). They got 0.6154 gender accuracy and 0.4103 age accuracy.

In (Nguyen and al., 2011), authors showed how the linguistic style of a Twitter text may give hints about the author's age. They employed the logistic regression approach to predict the author's age in Twitter Dutch messages. In the constructed corpus, the texts were very short with an average less than 10 words per message. They obtained 86 % accuracy for gender classification.

In 2015, (Grivas and al., 2015) proposed a grouping of features combined with appropriate reprocessing steps for each group. Their method consists in creating a framework to test various features and preprocessing combinations. They focused mainly on the gender and age subtasks. Their system performed less optimally in the case of age classification where more features were considered. In fact, authors got80.78 % accuracy for gender detection.

Thanks to the incompatible volume of data in social networks, personality recognition was developed.

Although Arabic is spoken by almost 400 million people, research works of author profiling performed on Arabic are not numerous. For example, the study of (Abbasi and al., 2005) focused on author identification. They constructed a corpus of 20 authors and 20 messages written by each one. The second work investigating the multilingual messages was performed by (Estival and al., 1998). In this research, authors collected Arabic and English e-mails written by 1033 English people and other e-mails written by 1030 Egyptian Arabic speaker. They studied several demographic and psychometric features for author profiling.

3 Data

Our method is based on the Web texts, especially Facebook, to form textual sources. The studied

corpus includes 4444 documents written in standard Arabic and labeled both by both gender and age.

When examining a given corpus for author profiling, researchers are faced with the problem of assigning the writer's age and gender to each text, which makes corpus construction a hard task.

The proposed method permits designing public groups and pages in order to make easier the communication between web users (individuals, freelancers, companies, etc.). The latter can do businesses, publish advertisement, sell products and even organize meetings. In order to extract the content of the comments on each posts written on Facebook page, we used API of Facebook.

Our corpus extraction procedure is based on two important stages:

In the first one, we entered Aljazeera Channel Facebook page in order to limit our corpus to Arabic text. We obtained ultimately a list containing the user's profiles (id accompanied by comments). After that, we labeled each id profile with age and gender.

To model statistically the language, we pre-processed the comments in the extracted corpus and define four regular expressions. The first one permits identifying Arabic texts and omitting those written in foreign languages. However, in the second expression, two or more successive spaces were replaced by one space. The third expression omitted diacritization; whereas the forth deletes comments in form of hypertexts (advertisements links, images ...).

The published posts and the comments were respectively classified manually and automatically. At the end of this process, each comment was classified under a given post and inherited its domain.

The obtained corpus includes 10819 user's profiles. We deleted 5244 of them to have, at the end, 5575 profiles. Obviously, the corpus was balanced in terms of gender, but imbalanced in terms of age. The total number of the validated profiles was 4444 containing 70121 words with almost 15 words per profile.

As can be seen, the text-length varies according to the author. There are three authors with average text-length shorter than15words. Figure 3 presents the distribution of the corpus according to the text-length. Approximately 50% of the texts have a text-length shorter than 10 words.

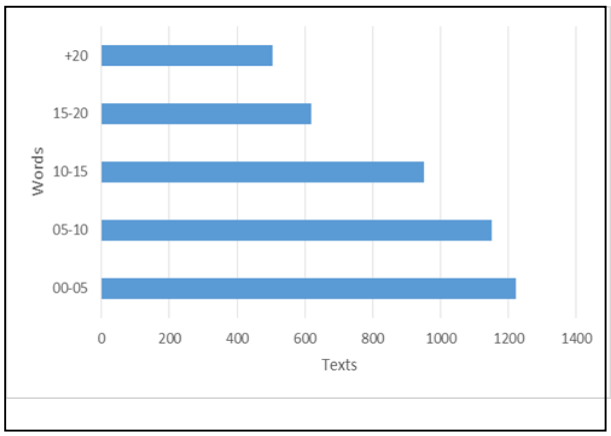

Fig 1 .Text-length distribution in the used corpus

4 Features selection

We considered the comments as a vector in a multi-feature space. Then, we used the obtained labeled vectors to construct our classification model. In fact, six types of style-related features were determined and different combinations of features were considered for authorship profile detection. We tried to coherently group features so that we could perform the adequate pre-processing steps for each group and facilitate the performance of the experimentation with various combinations of features groups.

Features were divided into several sub-groups (Arabic Lexical feature, Arabic n-gram, Arabic Syntax feature Arabic character frequencies, bag of smileys and Arabic stop words):

- Arabic Lexical feature: As for lexical features, the text was regarded as a set of tokens. These features were obtained by frequency calculations. We distinguish the number of words appearing once and those appearing twice, the average length of sentences, the number of sentences, the number of verbal sentences and the number of negative sentences starting with negation. (ل ل،ام، سي ل،لا، م ن).

- Arabic n-gram: The size of this n-gram representation considerably increased for the compression of electronic texts and for the identification of languages and authors. In fact, we used bigrams and trigrams to detect Arabic authors' profiles.

- Arabic syntax features: It consists in labeling comments and giving each word in the extracted corpus its syntactic category (proper name, common noun, verb, adjective, etc). Words are tagged by Penn Treebank Project Tagger.
- Arabic character frequencies: In order to extract the character features, we considered the text as a simple sequence of characters. However, we calculated the frequency of letters, punctuation marks (number of colons, exclamation marks, question marks and commas), upper-case characters, lower-case characters, numerical characters, alphabetical characters , numbers and symbols like (@, #, and, %, *).
- Bag of smileys: The use of emoticons has been grown thanks to the tremendous widespread of social networks. With the development of information and communication technology, emoticons have been employed in everyday online communications. For instance, Facebook and twitter contain more than 400 and 1100 emoticons, respectively. In our study, frequency of Smileys in each file was calculated using the handcrafted Smiley List. We listed 57manually - created emoticons (:-), , : (, ; -), ;) , ;-P, ;P , :P , :-p, :p...) which express different sentiments (happiness, sadness, anger, etc.)
- Arabic stop words: We defined 4 groups of stop words, such as personal pronouns ("he", وه),("she", يه),("they", مه), demonstratives (this اذه , these ءالؤه , there ni , ىلإ ot , من morf) snoitisoperp ,(أولـ ئك ekil sevitagorretni dna (عن tuoba , في ("where", أينـ,"who", نم).

5 Architecture of the system

This architecture is based on a set of modules which allow combining the different studied features obtained through the preliminary analysis in order to associate them with their authors. Indeed, during the training process, classifiers were created by selecting the sets of frequencies for each trait. A variety of machine learning algorithms, including decision trees (J48), naïve bayes, k-nearest neighbors and Support Vector Machines, were tested.

Our system constructed a prediction model that received, as input, a comment of author. As output, it gave the comment several labels to identify the age and the gender.

Figure 2 shows different processing modules and steps to detect the author's age and gender in an Arabic document.

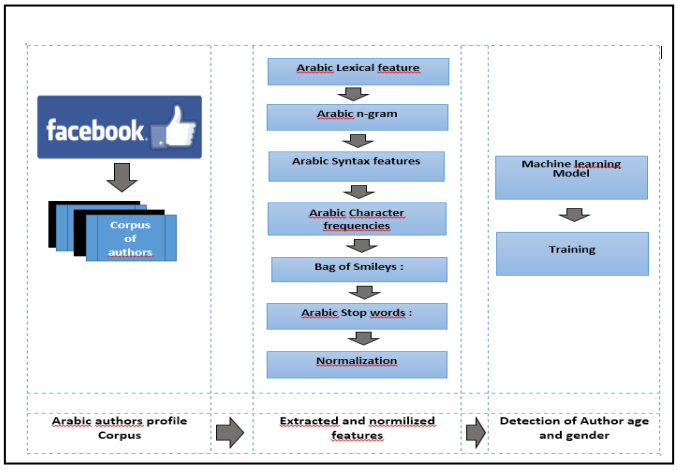

Fig. 2. Architecture of the system.

6 Experimentation and Evaluation

6.1 Data normalization

To avoid the dependence on the choice of measurement units, the data should be normalized. In order to remove the impact of different scale spaces, we normalized the value of each feature using AMZD normalization. (Patro and al., 2015).

$$\Psi = (|X|) - (10^{n-1}) * (|A|) / 10^{n-1} \quad (1)$$

where the new value is scaled one value between 0 and 1. X is the particular value of feature, A denotes the first digit of data feature X, N corresponds to the number of digits in feature X.

6.2 Cross-validation

To conduct performance tests, we applied the cross-validation technique. It is a statistical method which involves applying a training model on large

portion of data set. It has a good ratio testing data points and it allows changing the training and test data set distribution. It splits dataset into k"folds" randomly-selected groups. This method also calculates an error rate for each group. The average of the k recorded errors rate served as performance metric for the model. In our case, the number K was equal to 10.

6.3 Baseline method

We used the best results of [41] in PAN@CLEF2016 as a baseline method to assess our method and show its efficiency.

In the 4th Author Profiling Task at PAN 2016, 22 approaches of participants were evaluated. The participants used several different feature types and approaches to identify age and gender in a cross-genre framework in English, Spanish and Dutch.

When we analyzed the best results per language, we observed that results for gender identification in English and Spanish were quite similar.

They were also much better than those obtained for gender identification in Dutch. The best participant got 75.64% of precision in gender identification and 58.97% of precision in age identification for English language. We obtained 73.21% and 61.80% for Spanish and Dutch gender accuracy prediction, respectively.

6.4 Results

The obtained results were computed on the Arabic data set described in Section 3 using the different classifiers.

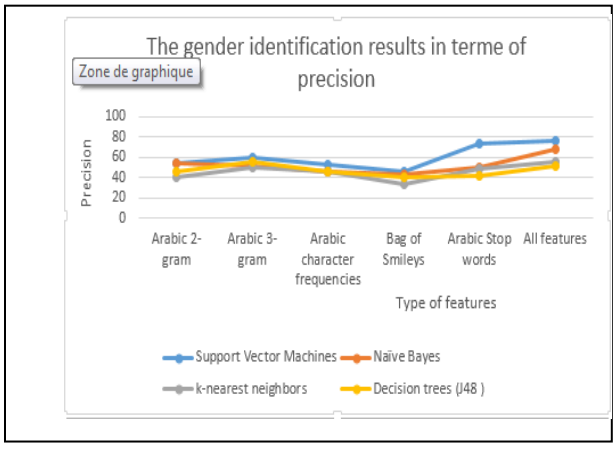

Fig. 3. The precision rates obtained for gender detection according to the type of the studied feature and the applied classifiers.

We examined separately each of the six categories of features. The findings presented in Figure 3 demonstrate that the Naïve Bayes classifier achieved a precision equal to 68.15%, but its performance was not as good as that provided by the Support Vector Machines (SVM) which is the most effective approach with different features. It reached 71.52% precision rate for gender identification applying "cross-validation".

Comparing our results with those achied by the best participant in PAN@CLEF2016 for gender prediction, we got a good classification rate greater than those obtained by the participants for the two other languages (Spanish and Dutch).

The distribution of age can be seen in Figure 4. Our results were not so significant. We obtained an average precision of 53.08% with SVM classifier. Comparing our results with those achieved by the best participant in PAN@CLEF2016, for the gender prediction, we got a good classification rate greater than those obtained by the participants for the two other languages (Spanish and Dutch).

Our results were almost similar to the best finding provided for English language.

The distribution of age can be seen in Figure 4, our result were not so significant. We obtained an average precision of 53.08% with SVM classifier.

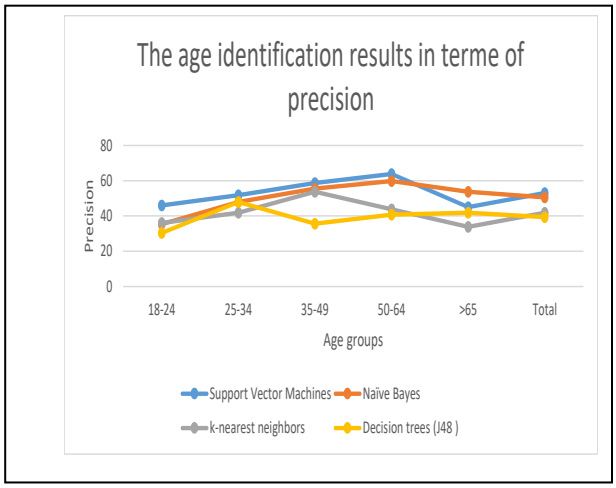

Fig. 4. The precision rates obtained age detection according to the type of classifiers applied.

7 Conclusion and future works

Detecting automatically some attributes, such as gender, age, background and personality, etc. from the author's writing style is a task of growing

importance for many application domains such as medico-legal, social networks analysis, legal linguistics and e-commerce.

In this paper, based on Arabic language from social networks, we showed how the combination of six types of features and machine learning methods can be effectively used to detect automatically the Arabic author's gender and age. The obtained results were encouraging, especially for the gender dimension.

In our future works, we will focus on personality features and extend our detection tool by applying it on texts written in other languages.

References

Cvijikj I. P. and Michahelles F. 2013. Understanding the user generated content and interactions on a Facebook brand page. International Journal of Social and Humanistic Computing 14, 2(1-2), 118-140.

Estival D., Gaustad T., Hutchinson B., Pham S. B., and Radford W. 1998. Author Profiling for English and Arabic Emails. Natural Language Engineering, Cambridge University Press.

Grivas A. Krithara A. & Giannakopoulos G. 2015. Author Profiling using Stylometric and Structural Feature Groupings. In CLEF (Working Notes).

Bayot, R., & Gonçalves, T. 2016. Author Profiling using SVMs and Word Embedding Averages. Notebook for PAN at CLEF.

Dichiu, D., & Rancea I. 2016. Using machine learning algorithms for author profiling in social media. Notebook for PAN at CLEF.

Koppel. M., Argamon S., & Shimoni A. R. 2002. Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

Bilan, I., & Zhekova D. 2016. CAPS: A Cross-genre Author Profiling System. Notebook for PAN at CLEF 2016.
Portugal.

Markovikj D., Gievska S., Kosinski M. & Stillwell D. 2013. Mining facebook data for predictive personality modeling. In Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA.

Nguyen. D. N. Smith A. & Rosé C.P. 2011. Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123. LaTeCH '11, Association for Computational Linguistics, Stroudsburg, USA.

Abbasi. A., & Chen H.. 2005. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 67-75.

Schler. J. Koppel M., Argamon S. & Pennebaker J. W. Effects of Age and Gender on Blogging. 2006. In AAAI spring symposium: Computational approaches to analyzing weblogs (Vol. 6, pp. 199-205).

Corney M., De Vel O., Anderson A. & Mohay G. 2002. Genderpreferential text mining of e-mail discourse. In Computer Security Applications Conference, 2002. Proceedings. 18th Annual (pp. 282-289). IEEE.

Rangel F. Rosso P., Verhoeven B., W. Daelemans, Potthast M., & Stein B. 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. Working Notes Papers of the CLEF.

Zhang C. Zhang P. 2011. Predicting gender from blog posts. Tech. rep., Technical P. Zikopoulos, & C. Eaton, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.

Patro S. & Sahu K. K. 2015. Normalization: A Preprocessing Stage. arXiv preprint arXiv:1503.06462.