Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media

Notebook for PAN at CLEF 2014

Seifeddine Mechti, Maher Jaoua, Lamia Hadrich Belguith

ANLP Research Group MIRACL Laboratory –University of Sfax Tunisia
mechtiseif@gmail.com,maher.jaoua@fsegs.rnu.tn,l.belghith@fsegs.rnu.tn

Abstract. In this paper, we focus on detecting the profile of authors (age, gender) through their discussions. The 2014 Pan@Clef corpus consists of 4 sub-corpuses: tweets, blogs, social media and reviews. The proposed method is based on automatic classification, which uses some data extracted statistically from a source corpus. We present a hybrid method that combines the analysis of data in texts with a machine learning method. In order to obtain a better management of these data, we relied on the use of the "Decision table algorithm".

Keywords: Author profiling, Author attribution, Decision table.

1 Introduction

Author attribution seeks to determine the author of an anonymous piece of writing or one whose attribution is still uncertain [1]; it is used as a text categorization. The idea is then to predict the author of a text and in whose drafting he is suspected to have participated. In addition, several studies of stylistic and statistical nature, but also taken from machine learning, allow further analysis of a text which is to be attributed, and provide useful information for its attribution.

Author profiling is the study of how certain linguistic features vary according to the profile of their authors [2]. On Twitter or on Facebook most of the users enter only 20% of their profiles. In the literature, a lot of studies have focused on the classification of a given conversation or text and specifically the detection of the age of the author, the genre, his personality, his native language, etc. [2,3,4]

Our method of author attribution is based on author identification and author profiling.

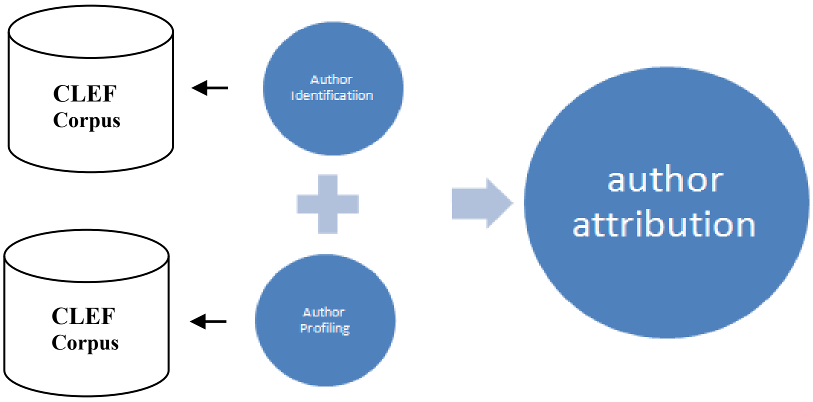

Fig.1- Author attribution based on author identification and author profiling

The remainder of this paper is organized as follows: In Section 2 we present our approach to representation of authors. Then, in Section 3, we focus on the author profiling part, basing our study on the attributes selected in this phase. In section 4 we present our method of author profiling. The final section presents our experimental study.

2 Stylometric approach for the representation of authors

Our approach uses stylometry to represent authors. The stylistic cues forming this type of model come in the first level: "function words”. In fact, many studies [4] have shown the effectiveness of function words for author attribution. In the second level, we follow a lexical analysis which is rendered by the cue of the lexical frequency of proper names, verbs, adjectives. In addition, we rely on punctuation to distinguish between authors. Finally, we use some statistic aspects linked to the author such as the average size of a sentence, the number of words that occur only once in a text, the number of words that occur twice, the number of deflections in a sentence compared with an average sentence, etc.

In the following section we discuss the various attributes used in author profiling.

3 Attributes for the detection of the author's profile

Koppel [4] distinguished two types of attributes: style-based features and content features.

3.1 style based features

To determine the age or gender of the author of a piece of writing, it is important to consider the function words. Prepositions, pronouns and determinants have shown their effectiveness in the process of detecting an author's profile. [5] In other studies, the authors use the frequency of punctuation, the frequency of capital letters and the frequency of citations [6,7]. Also, HTML attributes such as the URL of an image or link on a Web page have been used by [8]. In the work of [9], the authors rely on specific terms (foreign words) to distinguish between authors. These terms are tags in theStanfordCoreNLPPos tagger like meeee, yessss, thy, u, urs, etc. Unlike other authors, [10] resort to the calculation of the frequency of emoticons to predict the authors.

3.2 content features

What differentiates several age classes is the content of their discussions. Indeed, Koppel [4] distinguished several classes to categorize authors. We identified classes like money, home, smartphone, games, sports, Job, marketing, etc. We chose the first 200 attributes providing the best discrimination. The major drawback of content-based attributes is that they depend on the mental state of the author (negative emotions, positive emotions), which can distort the results.

4 Proposed method

The proposed method is based on the classification of discussions by gender and age. The gender dimension is represented by the class man and woman. We started by calculating the number of occurrences of all terms found in the corpus classifying them in descending order of their appearance, however, we have only the first 200 attributes. We calculated the CF (class frequency) for each attribute class in the context of measuring the frequency of occurrence of each class attribute in each document of the corpus.

We grouped manually the terms belonging to the same class of attributes. we have identified 25 classes namely: Prepositions, Pronouns, Determiners, Adverbs, Verbs, He, She, No, Of, I, Me, Medicine, Chemistry, Music, Sport, TV, Phone, Beer, Sleeping, Eating, Sex, Love, Money, Internet Marketing. We used the most discriminating classes of attributes. Once the classes are determined, it is to perform the training. We used the free learning "software Weka and we started by building ARFF file (Attribute Relation File Format), a file for gender and one for the age dimension.

5 Experimental setup

We conducted our experiments with extracts from the training corpora. The CLEF 2014 corpora that were discerned represent tweets, reviews and blogs. In fact, for the gender dimension (male, female), which has a 0.5 precision baseline, we obtained good results; 56% of the documents were correctly classified. For the age dimension, which has a 0.2 precision baseline, we distinguish five classes (18-24 | 25-34 | 35-49 | 50-64 | 65-xx "). The results are promising and demonstrate the effectiveness of the method. Indeed, 34% of the documents were correctly classified. We found that the learning method which is based on the "Decision table" algorithm gives better results for the gender dimension in the English language.

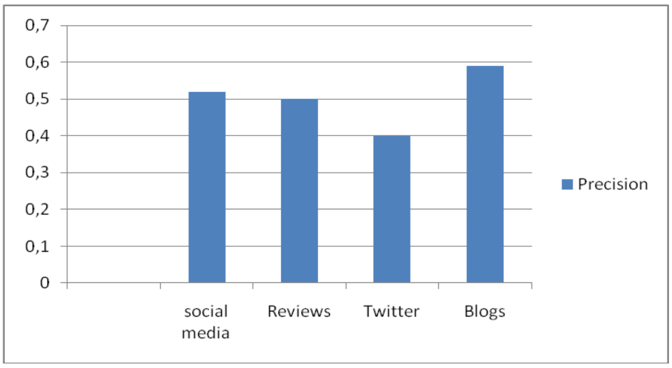

Fig. 2- Precisions obtained for the gender dimension

It turned that our approach gives better precisions for blogs but poorer results for tweets. This proves that our method is not effective on short texts.

6 Conclusion

We conducted the categorization of documents in order to provide a classification of the author of a given text according to its characteristics. The results are encouraging, especially for the gender dimension. The manual selection of the content of the classes has shown its limitations regarding language corpora which are not well known by the researcher. Automating this task turns out to be of great use, and the use of bilingual or multilingual dictionaries can cope with linguistic shortcomings.

It turned out that the use of lexical classes alone is not enough; we intend to integrate other aspects such as the syntactic, morphological and semantic aspects etc. On the other hand, to better perform the detection of the author's profile we intend to open up on other dimensions; apart from age and gender; we will also address the detection of the native language, the detection of the linguistic level, etc.

References

1. Patrick Juola. Authorship Attribution. In Foundations and Trends in Information Retrieval, Volume 1, Issue 3, March 2008.
2. Pennebaker, J. The secret life of pronouns: What our words say about us. pp. 401–412. 2011.
3. Schler, J., M. Koppel, S. Argamon, et J. Pennebaker. Effects of age and gender on blogging. AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. 2006.

4. M.Koppel, S. Argamon, J. Pennebaker and J. Schler, Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2): 119–123. 2009.
5. Fermín L. Cruz, Rafa Haro R., and F. Javier Ortega. ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling - Notebook for PAN at CLEF 2013.
6. Wee-Yong Lim, Jonathan Goh, and Vrizlynn L. L. Thing. Content-centric age and gender profiling - Notebook for PAN at CLEF 2013.
7. Yuridiana Aleman, Nahun Loya, Darnes Vilariño, and David Pinto.Two methodologies applied to the author profiling task. Notebook for PAN at CLEF 2013.
8. Upendra Sapkota, Thamar Solorio, Manuel Montes-y-Gómez, and Gabriela Ramírez-de-la-Rosa. Author Profiling for English and Spanish Text - Notebook for PAN at CLEF 2013.
9. Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent Trends in Digital Text Forensics and its Evaluation. 2013.
10. Delia Irazu Hernandez Farias, Rafael Guzmán-Cabrera, Antonio Reyes and Martha Alicia Rocha Semantic-based Features for Author Profiling Identification: First insights- Notebook for PAN at CLEF 2013.