Author Profiling Using Style-based Features
Notebook for PAN at CLEF 2013

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 present a method for profiling the author of an anonymous text. Our approach is based on learning the author profile with a focus on dimensions age and gender. Our system takes as input a document which is written in English or in Spanish and generates the age and the gender of its author. First, we computed a ranked list of words that occur in the corpus and we grouped them into classes according to their similarities. Then, we calculated the TF * IDF score of each class for each document in order to find the stylistic differences between men and women, on the one hand, and those between different age intervals on the other hand. After that, we applied the learning process on 66% of the English and the Spanish corpuses using decision trees through the J48 algorithm. In factwe got the second place in the competition for the English corpus;Our system has shown a high level of accuracy and effectiveness in treating the gender dimension and we got the best accuracy for the entire PAN 2013 competition.

Keywords. Machine learning, Author profiling, Style-based features, Decision trees.

1 Introduction

Over the past twenty years, the field of Information Retrieval (IR) has grown well beyond its primary objectives which are text indexing and the search for relevant documents in a collection. Today, IR includes modeling, classification and categorization of documents, plagiarism detection, data visualization, filtering languages, etc. Document classification allows answering the question: to which class does a text or a conversation pertain? However, many studies have focused on profiling the author of a particular text and more specifically on detecting the age of the writer of a given text, her gender, her native language, … [1, 2, 3, 4]. The aim of author profiling isto identify the stylistic differences in writing between a man and a woman and between authors from different age intervals. In this paper, we will answer the following questions: How to know if an anonymous text was written by a man or a woman? What is the author approximate age?

The training corpora are made up of an English corpus (236,000 documents) and a Spanish one (75,900 documents). For both corpuses, we used 66% of documents for training and 33% for test.

2 Related work

Text classification is based on text mining and statistical techniques that produce results from calculations of extracted terms frequency [5]. Text classification may also be based on machine learning approaches, such as Bayesian approaches and decision trees.

In [3], Koppel et al explored the possibility of automatically classifying documents according to author gender using an author profiling approach. Author profiling is the task of predicting features related to the text author [6]. It addresses several dimensions such as age, gender, native language, personality, level of education, etc.

According to Koppel et al [3], men who prefer to categorize things use more determiners (the/the, this/that, a/an,etc.) and quantifiers (two, more, little, etc.).Women are more interested in relationships and, therefore, use personal pronouns (I, you, me, her, etc.) more than men. Koppel algorithm therefore consists in quantifying the recurrence of 467 English keywords (a, too, yourself, their, etc.) in a text in order to calculate the gender of its author. Indeed, the program was trained and was conditioned on a corpus of texts from the Blog Authorship Corpus [7]. The works analyzed were within all styles of writing: fiction, textbooks, tests, etc. After this learning phase, the software was able to provide a correct answer four out of five times. Gaustad[6] work on an automatic classification of messages in Arabic. They received a rate of 81.5% of well classified documents in relation to the gender and of 72% in relation to age. Hariharan [8] and Kose [9] have also obtained promising results in their work. They worked on the detection of gender; in this category they managed to obtain an accuracy of 0.9.

In this paper, we focus on two-dimensional author profiles. The author profile attributes are age and gender. We consider English and Spanish texts.

3 Our approach

Our approach is purely statistical. It accepts input from any document written in English or Spanish. It is based on the calculation of numbers of term frequencies to identify the differences between men and women on the one hand and the differences between age intervals (10s, 20s and 30s) on the other hand. This method consists of four stages, namely calculating the number of occurrences of words, choosing classes and building ARFF[1] files, machine learning and presentation of results.

Our method provides one of the following six classes:10s Male,20s Male,30s Male,10s Female,20s Female and 30s Female.

3.1 Features

The step of classes selection was very important and had a great impact on the results. We computed the number of occurrences of all the words that occur in the

[1] An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato.

corpus and ranked them in descending order of their appearance. However, we made use of the top-200 attributes.

Figure 1 depicts some examples of occurrences of the English corpus:

the = 1039688	internet = 18067	company = 11565
to = 769817	money = 13080	down = 11551
and = 567952	love = 12342	she = 11547
time = 41661	better = 12320	…..
other = 40873	marketing = 12275	…..
people = 37332	high = 11932	offer = 11398
good = 31590	website = 11672	does = 11364
business = 19854	feel = 11578	

Figure 1.Examples of attributes occurrences of the English corpus for the class "male 20s"

We repeated this task six times for each language. Then we tried to group the attributes belonging to the same classes together.

Below are examples of the classes found in the Spanish corpus:

Class	Attribute
Determiners	el, los, la, las, lo, unos, unas…
Prepositions	Ante,bajo,contra,hacia,hasta,excepto,para,contrario,junto…
Pronouns	Todos, otro, cualquier, cualquiercosa, ambos, cada, el uno al otro, todo el mundo, ni, ninguno, otro, otros, qué…
Amor	quiero, amor, amore, corazon, corazón.
Smiley	Smiley ☺
Teenager	Haha, escuela, lol, quiero, aburrido, windows ,mamá, padre, madre, sistema, internet, web, casa, entonces, mientras, manera , versión , cualquier , gomu.
Young adult	apartamento, oficina, trabajo, bar, gusta, ellos
…….	……

Table 1. Example of classes from the Spanish corpus

There are two basic types of features that can be used for authorship profiling: content-based features and style-based features [4]. Indeed, we looked for the similarities that can group a set of terms in the same class.

The corpus of the English text is much larger. Therefore, we identified many more classes (25 classes), which are***: 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.

3.2 The Learning Method

Once the classes were set, they are used to perform the training. We used the learning software and data mining "Weka"[12] to perform this task. We started the construction of ARFF files, a file for the gender dimension and one for the age dimension for each language. Indeed, we calculated TF * IDF for each class in order to assess the relevance of each class in a given body of the documents:

$$w_d = f_{w,d} * \log (|D|/f_{w,D})$$ [10]

D: a document collection,
w: word,
d: an individual document d ϵD,
fw,d : equals the number of times w appears in d, |D| is the size of the corpus,
fw, D : equals the number of documents in which w appears in D.

We were not satisfied with the calculation of TF (term frequency) only because IDF measures the importance of a term in the corpus and therefore it gives more importance to discriminatory terms (which are less frequent). Then we put the TF * IDF of each document in the corpus of each class in the two ARFF files and got our training base for both age and gender. Here comes the role of Weka. We developed a system based primarily on the notion of a conditional probability system. There are numerous statistical learning methods. After trying several learning algorithms (Naive-Bayes, SVM, Multilayer perception, DMNB text) we realized that the learning-based decision tree method differs from other statistical methods by its tree structure. This structure makes the learning method readable to humans, unlike in other approaches where the predictor is built, in a black box [11]. This favors the use of decision trees that have a role to determine, from a graphical representation of a set of rules, an instance of the class with a probabilistic model. We implemented and supervised the learning method of the decision tree based on a set of statistical techniques used to model problems, extract information from raw data and make decisions in a coherent and rational way.

4 Experiments

For the English language, we obtained the best results in the competition for the gender dimension. 58% of the documents were correctly classified [13]. However, for the age dimension, the results are encouraging, but not as good as those for the gender ones. Indeed, 58% of the documents were correctly classified (second place in competition).But, we encountered a problem at the level of the corpus for 10s (10-17 years), since the corpus is of a small size. Almost 90% of the documents for this corpus were misclassified.

The graph below shows the variation of precision depending on the number of classes:

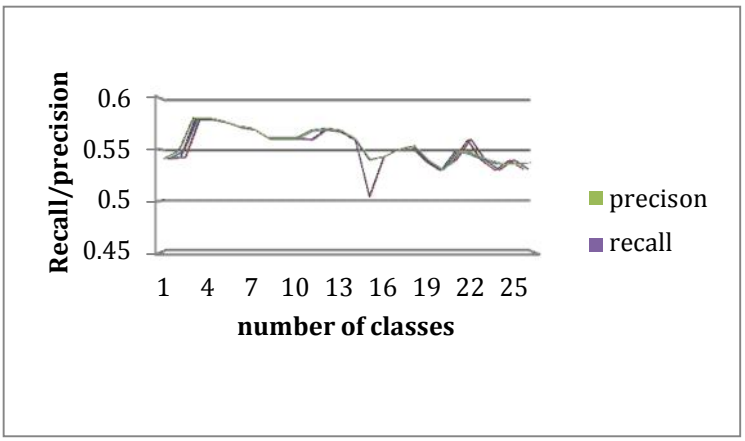

Figure2. The variation of precision depending on the number of classes

If the number of classes is high, that doesn't mean that precision will be high ; instead we had the best accuracy with only three classes.

Figure 3 depicts a comparison between stylistic features and content features:

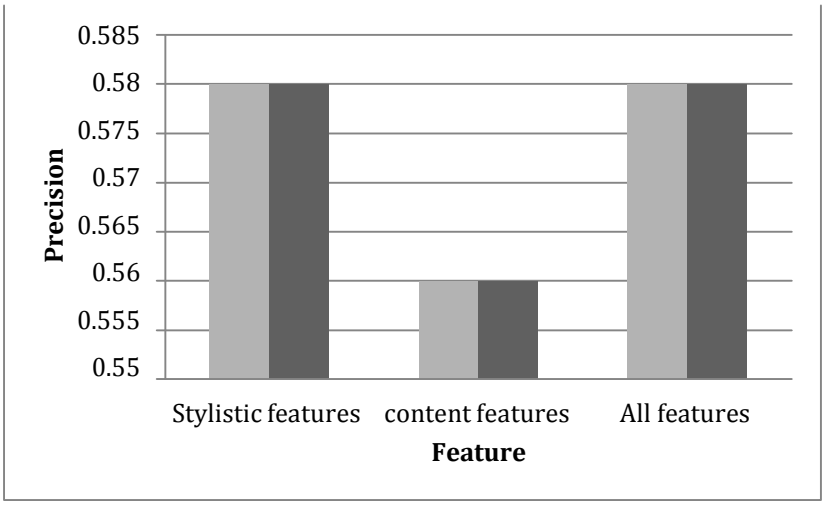

Figure 3. percentage split(66%) results for the gender classifiers

For the gender dimension:(Styles features) prepositions, pronouns and verbs were highly effective. This is due to the fact that these classes are the most represented in all documents in the corpus. The advantage of our approach compared with previous approaches [4] is that it does not use content features. This way, we will no longer need 25 classes.

Focusing on the Spanish corpus, we also had good results both for the gender and age dimension (55,6% and 51% correctly classified documents respectively).

As shown in table 2 we present a detailed accuracy by class for the spanish corpus:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	**Class**
	0	0	0	0	0	0.572	**10s**
	0.966	0.935	0.569	0.966	0.716	0.555	**20s**
	0.068	0.033	0.583	0.068	0.122	0.564	**30s**
Weighted Avg.			**0.556**	**0.57**	**0.451**	**0.56**	

Table 2. Detailed Accuracy by Class for the gender dimension of the Spanish corpus

5 Conclusion

We performed the classification of the documents to personalize the author of a given text. The results are encouraging, especially for the gender dimension. It turned out that the use of the lexical classes alone is not enough. However, we intend to integrate other aspects such as the syntactic aspect, morphology, semantics, etc ... On the other hand, in order to well personalize the author, we intend to explore other dimensions apart from age and sex We will also address the detection of the native language, the geographical data of the author, etc ...

Currently, our approach is based on a statistic learning model where the corpus is not updated and the test documents already predicted are not included in the training for the next test. Therefore, in future work, we plan to address this issue to be more convenient for real-time author profiling scenarios...

References

[1] Argamon S., Koppel M., Pennebaker J. and Schler J. Automatically profiling the author of an anonymous text, Communications of the ACM, pages 119–123. New York,USA, 2009.

[2] Schler J., Koppel M.,Argamon S. and Pennebaker J. Effects of Age and Gender on Blogging, Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, England, 2006.

[3] Koppel M. S. Argamon and A. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing, pages 401-412, 2003.

[4] Pennebaker J. The secret life of pronouns: What our words say about us. New York,USA, 2011.

[5] Salton G. and McGill M.J. Introduction to modern information retrieval.1983.
[6] Gaustad T., Estival D. and Hutchinson B. TAT: an author profiling tool with application to Arabic emails. Proceedings of the Australasian Language Technology Workshop, pages 21-30, Melbourne, Australia, 2007.
[7] TheBlogAuthorshipCorpus: http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
[8] Hariharan S., Gender Prediction in Chat based Medium's Using Text Mining, in: International Journal of Research and Reviews in Information Sciences (IJRRIS),Kohat, Pakistan, 2011.
[9] Köse C., Özyurt Ö. andAmanmyradov G.Mining Chat Conversations for Sex Identification, Emerging Technologies in Knowledge Discovery and Data Mining (PAKKD),Nanjing,China, 2007.
[10] Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval, information processing and management, pages 513--523,1988.
[11] Caron S, decision tree learning for python. http://scaron.info/pydtl.2011.
[12] University of Waikato.Weka. http://www.cs. waikato.ac.nz / ~ ml/weka.
[13] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13 web/pan13-ap-final-results.pdf