Automatic Author Profiling Based on Linguistic and Stylistic Features

Notebook for PAN at CLEF 2013

Braja Gopal Patra[1], Somnath Banerjee[1], Dipankar Das[2], Tanik Saikh[1],
Sivaji Bandyopadhyay[1]

[1]Department of Computer Science & Engineering, Jadavpur University, India
[2]Department of Computer Science & Engineering, NIT Meghalaya, India
{brajagopal.cse, s.banerjee1980, dipankar.dipnil2005, tanik4u}@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract. The rapid expansion of blog and electronic data in Web 2.0 is abounding and thus it is becoming important to identify the author's profile also. The problems of automatic identification of author's gender and age based on linguistic and stylistic pattern have been a subject of increasingly research interest in the recent years. The research methodologies are also helpful for several other applications like criminal detection, security and author detection etc. We have used lexical, syntactic and structural features for identifying the gender and age group of the authors. We have employed the Decision tree classifier for classifying the author profile. We have achieved the accuracies of 56.83% and 28.95% for gender and age group classification, respectively.

Key words: Automatic author profiling, gender identification, age group identification, word class, decision tree.

1 Introduction

In the 21st century, popularity of the Internet is increasing abundantly. Internet media like emails, blogs/internet forum and websites have been identified as the ideal communication platform for the people. In the recent years, the researchers have been paid increasingly interest to analyze of authorship of emails, electronic messages [5] and plagiarism detection [7] etc. Thus, analyzing the web content has become more important to the intelligence and security agencies that monitor the author information as much as possible [1]. The task of automatically predicting the authorship from anonymous text by extracting the linguistic and stylistic features has a number of potential applications [5]. For example, if you have important textual information of an unknown author and want to know author's gender, age, demographic and cultural background etc., just by analyzing the text.

In this paper, we present the task of automatic authorship identification from anonymous data provided by the PAN-2013. PAN-2013 is the 9th evaluation lab on uncovering plagiarism, authorship and social software misuse. Author Profiling task is concerned with predicting an author's demographics from her writing. Besides being personally identifiable, an author's style may also reveal her age and gender**.** We have used linguistic and stylistic feature for identifying the authors' age and gender. Different word lists have been created for calculating the frequencies of each document. We have also created the stopword list, smiley list, positive and negative word lists etc. for creating the feature vector. A machine learning algorithm has been employed for classifying the authors' profile. The Decision tree of Weka[1] tool has been used for the classification task. The accuracy of the system has been calculated by the PAN organizers and we have achieved 56.83% and 28.95% for gender and age group classification.

The rest of the paper is organized in the following manner. Section 2 discusses briefly the related work available till date. Section 3 provides an overview of the data used in the experiments. Section 4 describes the feature selection for implementation of Machine learning algorithm and Section 5 gives the details of the system architecture. Section 6 presents the experiments with detail analysis of results. Finally, conclusions are drawn and future directions are presented in Section 7.

2 Related Studies

Automatic author profiling is the task of predicting author's traits automatically using any of the machine learning algorithms. It is more important to identify the authors' trait when any of that author's documents does not belong to the training data. In contrast, we can say that the greater accuracy can be achieved when the author's documents are present in the training data. Most of the research on author's profiling focuses on the less number of traits. For example, in addition to gender, age and the native language, level of education and country of residence are also identified in [8]. But, in our present work, efforts have been given to identify only the age and gender.

A considerable amount of research on automatic classification of texts into predefined categories has already been conducted by different research groups using several machine learning techniques [11]. Over the last few decades, a great variety of methods have been implemented for classifying the authorship attribution [1], [2], [3], [4], [5], [6]. Different machine learning algorithms previously tried to include different techniques such as Lazy learners (IBk) [5], [8], Support Vector Machine (SVM) [6] and SMO [8], LibSVM [5], RandForest [5], Information Gain [2], Baysian Regression [6], Exponential Gradient [7] etc.

Several experiments have been conducted for selecting the features for classification. Houvardas and Stamatatos [2] showed that n-gram is the best feature for authors' profiling classification whereas Schler and his group [6] have shown the effect of age and gender in Blogging sites. The authors have considered different word classes and shown the relation of each word class with the authors' gender and age. Some works [6], [8] have identified that the POS is also an admirable linguistic

[1] http://weka.wikispaces.com/Use+WEKA+in+your+Java+code

feature in this type of classification. Calix and his group [10] have used 55 different features and achieved the highest accuracy of 76.72%.

3 Experimental Data

The data used in our experiments are provided by the PAN-2013. The corpus consists of XML documents containing the conversations in HTML format. Many different topics are grouped by author and labeled with his/her language, gender and age group. The documents are of two languages (English and Spanish), two genders (Male and Female), and three groups of age (10s: 13-17 yrs, 20s: 23-27 yrs and 30s: 33-47yrs). Basically, we have developed the system for English language only. Each of the authors is presented in a separate XML file, the name of which provides information about the language, gender and age group in order to facilitate file tasks, and grouped by language in two separate folders, EN and ES.

The English corpus incorporates a total of 236,000 authors (files) containing 413,564 conversations and 180,809,187 words. The detailed distribution of data is given in Table. 1. In our present task, a total of 1200 XML files are used for the development purpose and rest of the data is used for training whereas the test data is provided by the PAN organizers.

Age Group	Gender	Number of authors
10s	MALE	8,600
	FEMALE	8,600
20s	MALE	42,900
	FEMALE	42,900
30s	MALE	66,800
	FEMALE	66,800

Table 1. Statistics of experimental data for English

4 Feature Selection

Feature selection plays an important role in machine learning framework and depends upon the data set used for the experiments. Thus, we have considered different combination of features to get the best results in the classification task. Initially, we experimented with the features like simple Unigram, Bigram and Trigram [2]. But, it has been observed that the features are proved not effective while trying to find out any similarity for the automatic identification process. Thus, we have incorporated the knowledge of word class frequencies. The features are as follows:

Word class Frequency: The word class frequency feature plays an important role in author profiling as in [3]. Each word class contains a set of stemmed words related to synonyms and hypernyms. We have manually created a small number of seed list for each of the classes.

Then, we have used the hypernym and synonym relations of RiTaWordNet[2] to increase the seed list. The increased word lists are checked manually. There are 9 classes, namely *money, job, sports, television, sleep, eat, sex, family* and *friend*. Each of the lists contains an average of 1400 unique words. The statistics of each class is given in the Table 2.

Positive and Negative word class: It is found that the positive and negative word classes are also key features for automatic author profiling [3]. Thus, these two classes contain the words which do not appear in our existing 9 word classes. After getting all possible POS from RiTaWordNet, the sentiment scores of the words have been calculated using the SentiWordNet 3:0[3] lexicon. Then, the words having sentiment score greater than 0.1 and less than -0.1 (threshold value: |0.1|) have been considered as the positive and negative sentiment word classes. The Positive and Negative word classes contain a total of 9627 and 10383 words respectively.

Class	**Number of Words**
Money	881
Job	1145
Friends	508
Family	302
Eating	3120
TV	261
Sports	591
Sleep	19
Sex	1008
Positive	9627
Negative	10383

Table 2. Statistics of Word class.

Stop words frequency: Stop words have been found as one of the important features. We have observed that the age group 20 has used more number of stop words in their text. A total of 329 stop words have been prepared manually.

Smiley words list: Frequency of Smileys in each file has been calculated using the handcrafted Smiley List. We have listed 55 smileys prepared manually by us. After calculating the frequency of smileys in each of the files, smileys are replaced by full stop words.

List of Foreign Words (FW): These are the words, which are tagged as FW by the StanfordCoreNLP[4] POS tagger. These are basically *meee, yesss, thy, u* and *urs* etc.

List of Punctuations: 10 types of punctuations are prepared manually.

[2] http://www.rednoise.org/rita/wordnet/documentation/

[3] http://sentiwordnet.isti.cnr.it/

[4] http://www-nlp.stanford.edu/software/corenlp.shtml

List of Pronouns: The frequencies of the pronouns are also computed. Pronouns are tagged as PRP by StafordCoreNLP POS tagger.

Average Length of Sentence: We have considered the average sentence length in documents. The sentence boundary is detected by the StanfordCoreNLP tool.

It has been found that the size of each document varies, i.e., some documents contain more number of words and some documents contain less words. So, we have normalized each bag of word feature by dividing the total number of words in a document.

5 System Architecture

Figure 1shows different processing modules and step by step flow of processing the documents. Initially, the XML documents have been cleaned. The cleaning process involves the removal of garbage words (e.g. '\ufffd' and '\u007f' etc.) and XML tags. Then, the frequency of smiley or emoticons has been calculated and the smileys have been replaced by full stop because it has been observed from the corpus data that smileys are generally occurred at the end of the sentence and sentence end marker i.e., full stop is not present for those sentences.

The Stanford CoreNLP package has been used to detect the sentence boundary and then, the average word in a sentence has been calculated from the sentence boundary detected text. Each word of the XML document has also been stemmed by the Stanford CoreNLP package. The punctuation list has been used to calculate frequencies from the text.

The Stanford CoreNLP POS tagger has been used to tag parts of speech (POS) of each word. Pronouns and Foreign words frequencies have been calculated from the tagged text. Pronouns and Foreign Words are tagged as PRP and FW by the Stanford CoreNLP POS tagger.

Word class frequencies have also been calculated by using the manually prepared lists. The Positive and Negative word frequencies have also been determined by using the RiTaWordNet. The stop word frequency has been determined by using stop words list. The frequencies of word class, positive word, negative words, pronoun, punctuation and foreign words have been normalized by dividing the total number of words present in the document. The extracted features are also used to prepare our test templates.

We have used the API of Weka 3.7.7.5 to accomplish our classification experiments. Weka is an open source data mining tool. It presents a collection of machine learning algorithms for data mining tasks. We employed the Decision tree (J48) for classifying the documents. The decision tree model has been trained by training template and the model has been used to classify the test template.

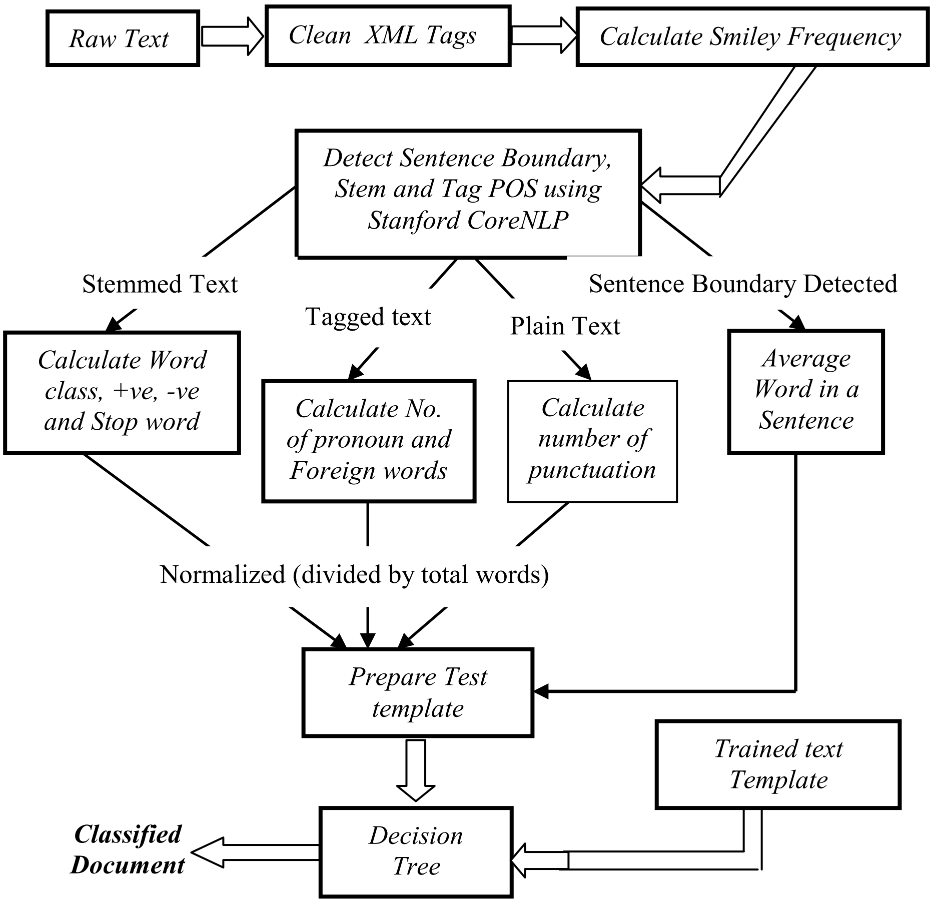

Raw Text
Clean XML Tags
Calculate Smiley Frequency
Detect Sentence Boundary, Stem and Tag POS using Stanford CoreNLP
Stemmed Text
Tagged text
Plain Text
Sentence Boundary Detected
Calculate Word class, +ve, -ve and Stop word
Calculate No. of pronoun and Foreign words
Calculate number of punctuation
Average Word in a Sentence
Normalized (divided by total words)
Prepare Test template
Trained text Template
Classified Document
Decision Tree

Figure 1: System architecture

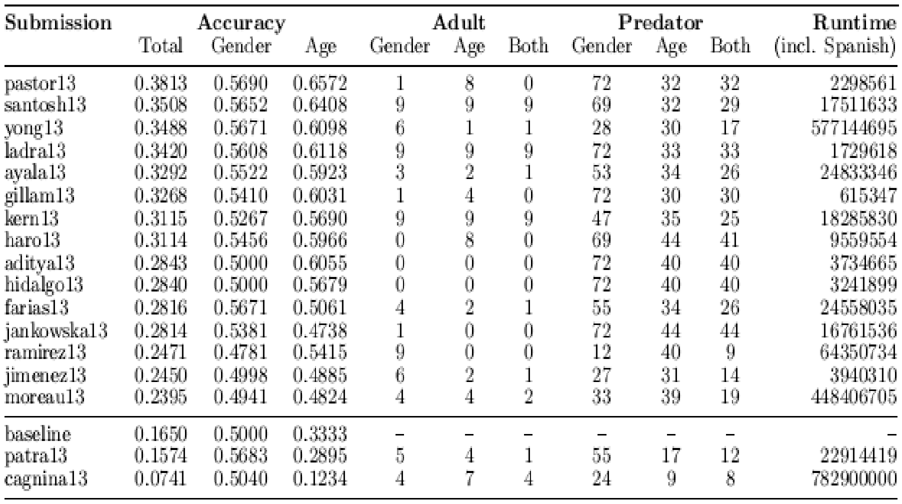
6 Experiments and Results

The detailed results of the PAN author profiling task have been shown in Figure 2. The result of our system has been evaluated by the PAN organizers. Though the result shows that the overall performance of our approach (patra13) is below the base line, but if we consider the accuracy of Gender identification, our approach is in the second best position in the results. Our gender identification accuracy achieved 0.5683 compared to the best score of 0.5690. But, our approach does not perform well to classify age and achieved accuracy of 28.95%. Though large corpus training data has been supplied by PAN-2013 organizers, but due to processing overhead we have considered small development data for training age identification. This may be one of the causes of the performance degradation. We have used same template for both gender and age classification and this may be another reason for degradation in age classification.

7 Conclusion

In this work, we have presented the task of automatic classifying the author's gender and age from their writing. This work is of interest for a number of potential applications like forensics, security and marketing etc. We have performed our experiment on the 23600 training data provided by the PAN-2013 organizers. The results provided in this work were calculated by PAN organizers. We have acquired the accuracies of 56.83% and 28.95% in gender and age classifications respectively.

In our future work, the accuracy of the classification can be improved by finding and incorporating more suitable features like POS, number of Ellipsis, average word length and number of paragraphs etc. It would also be interesting to perform deeper features engineering for finding demographic and psychometric author traits more correctly.



Submission	Accuracy			Adult			Predator			Runtime
	Total	Gender	Age	Gender	Age	Both	Gender	Age	Both	(incl. Spanish)
pastor13	0.3813	0.5690	0.6572	1	8	0	72	32	32	2298561
santosh13	0.3508	0.5652	0.6408	9	9	9	69	32	29	17511633
yong13	0.3488	0.5671	0.6098	6	1	1	28	30	17	577144695
ladra13	0.3420	0.5608	0.6118	9	9	9	72	33	33	1729618
ayala13	0.3292	0.5522	0.5923	3	2	1	53	34	26	24833346
gillam13	0.3268	0.5410	0.6031	1	4	0	72	30	30	615347
kern13	0.3115	0.5267	0.5690	9	9	9	47	35	25	18285830
haro13	0.3114	0.5456	0.5966	0	8	0	69	44	41	9559554
aditya13	0.2843	0.5000	0.6055	0	0	0	72	40	40	3734665
hidalgo13	0.2840	0.5000	0.5679	0	0	0	72	40	40	3241899
farias13	0.2816	0.5671	0.5061	4	2	1	55	34	26	24558035
jankowska13	0.2814	0.5381	0.4738	1	0	0	72	44	44	16761536
ramirez13	0.2471	0.4781	0.5415	9	0	0	12	40	9	64350734
jimenez13	0.2450	0.4998	0.4885	6	2	1	27	31	14	3940310
moreau13	0.2395	0.4941	0.4824	4	4	2	33	39	19	448406705
baseline	0.1650	0.5000	0.3333	–	–	–	–	–	–	–
patra13	0.1574	0.5683	0.2895	5	4	1	55	17	12	22914419
cagnina13	0.0741	0.5040	0.1234	4	7	4	24	9	8	782900000

Figure 2: Results of Author Profiling Task-2013

Acknowledgement

The work reported in this paper is supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

References

1. Abbasi, A., and Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE, 20(5), 67-75 (2005).
2. Houvardas, J., and Stamatatos, E.: N-gram feature selection for authorship identification. Artificial Intelligence: Methodology, Systems, and Applications. Springer Berlin Heidelberg, 77-86 (2006).
3. Schler, J., Koppel, M., Argamon, S., and Pennebaker, J.: Effects of age and gender on blogging. In Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 199-205 (2006).
4. Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123 (2009).
5. Estival, D., Gaustad, T., Hutchinson, B., Pham, S. B., and Radford, W.: Author Profiling for English and Arabic Emails. Natural Language Engineering, Cambridge University Press (1998).
6. Koppel, M., Schler, J., and Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1), 9-26 (2009).
7. Koppel, M., Argamon, S., and Shimoni, A. R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412 (2002).
8. Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B.: Author profiling for English emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (pp. 263-272) (2007).
9. Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L.: Author identification on the large scale. In Proc. of the Meeting of the Classification Society of North America (2005).
10. Calix, K., Connors, M., Levy, D., Manzar, H., MCabe, G., & Westcott, S.: Stylometry for e-mail author identification and authentication. Proceedings of CSIS Research Day, Pace University (2008).
11. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47 (2002).
12. Patra, B. G., Kundu, A., Das, D. and Bandyopadhyay S.: Classification of Interviews–A Case Study on Cancer Patients. In Proceedings of 2nd Workshop on Sentiment Analysis where AI meets Psychology (COLING-2012). IIT Bombay, Mumbai, India, pp. 27-36 (2012).