Ensemble Learning Approach for Author Profiling

Notebook for PAN at CLEF 2014

Gilad Gressel, Hrudya P, Surendran K, Thara S, Aravind A, Prabaharan Poornachandran
Amrita Center for Cyber Security, Amrita University, Kollam, India
giladgressel@gmail.com, hrudyap@am.amrita.edu, surendrank@am.amrita.edu,
tharas@am.amrita.edu, aravindashok@am.amrita.edu, praba@amrita.edu

Abstract. With the evolution of internet, author profiling has become a topic of great interest in the field of forensics, security, marketing, plagiarism detection etc. However the task of identifying the characteristics of the author just based on a text document has its own limitations and challenges. This paper reports on the design, techniques and learning models we adopted for the PAN-2014 Author Profiling challenge. To identify the age and gender of an author from a document we employed ensemble learning approach by training a Random Forest classifier with the training data provided by PAN organizers for English language only. Our work indicate that readability metrics, function words and structural features play a vital role in identifying the age and gender of an author.

1 Introduction

With more than 2 billion users, internet has provided a solid platform for people to share, communicate and express their ideas globally. Though online social media have brought people together, they are vulnerable to crimes like identity thefts, false information, identity masking etc. A lot of people fake their original identity either to remain anonymous or to perform different cybercrimes. Zheng et al. in [1] has showed that anonymity is a significant characteristic in online communities. The process of identifying the traits of an author like age, gender, country, religion etc from a document has become one of the hot topics for researches in the fields of security, forensics, marketing, etc.

In this paper, we present the working of our system which performs the Author Profiling task by PAN-2014. This task aims at identifying the age and gender of an author from four different datasets which are twitter tweets, blog data, social media posts and hotel reviews. The training data for this analysis is also provided by the PAN organizers. We employed different Natural Language Processing techniques to extract features from a text document and using the Random Forest classifier we determine the age and gender of the author. This paper presents the working model of our system along with our architecture diagram, the machine learning algorithms and techniques we used to complete the task.

The rest of the paper is organized in the following order. Section 2 discusses the related work by other researches we went through. In Section 3 we discuss our system details and in section 4 we explain in detail our methodology and implementation details. Finally in section 5 we draw conclusions and present future plans for extending and improvising our author profiling task.

2 Literature Survey

A considerable amount of research has already been done to identify the traits of an author from a document using different machine learning algorithms and statistical models. In [2] Peng et al. mention that each author has his/her own unique stylometry of writing and refer to this feature as author profile. Pennebraker and Stone in [3] used the LIWC dataset and showed a relationship between language used and the age of the author. The study performed by Vimala Balakrishnan and Paul H.P [4] prove that age and gender of a mobile phone users do influence their texting style. John D. Burger, John Henderson and co-writers presented a language independent classifier for gender prediction from the twitter micro-blogging site [5]. In [6] Dong Nyugen et al. presented a linear regression model to predict the age of an author from a given text document. Claudia Peersman performed an exploratory study for predicting age and gender form chat texts using Netlog corpus data [7]. R Chandramouli in [8] compared the working of SVM, AdaBoost and Logistic regression for gender identification. Shlomo Argamon shows the differences in the male and female writing using the British National Corpus in [9].

In [10] the authors propose a tool TAT to profiling the authors of Arabic emails. Calix et al. [11] used 55 different features to analyze the stylometry of authors for email author identification and authentication. All previous years PAN Author Profiling research papers could be found from [15], [16] and [17].

3 System Architecture

In this notebook we propose our solution for author profiling from a given set of text documents. We have used a combination of Semantic, Syntactic and Natural Language Processing (NLP) analysis for finding the same. The output of each analysis is fed into a trained ensemble classifier which determines the age and gender of the author. Figure 1 shows the detailed architectural diagram of our work.

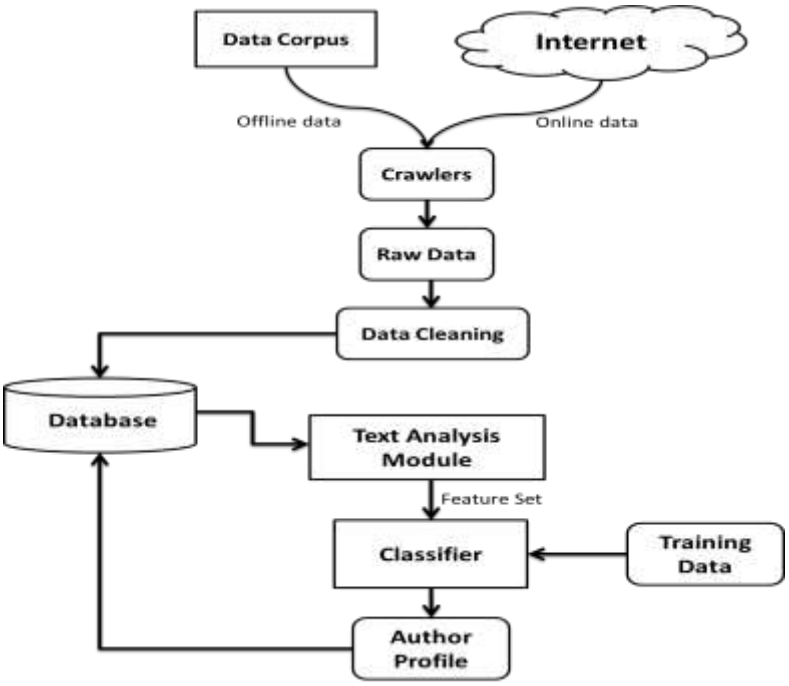

Figure 1 System Architecture Diagram

For performing the Author Profiling task we used the data corpus provided by PAN-2014. The corpus consisted of various xml documents which had to be handled in an offline (Hotel reviews and Social media) and online (Twitter and Blog) mode. The Twitter training data had to be downloaded due to Twitter terms of service. The cleaned blog corpus from RSS feeds was provided by PAN organizers, but it only contained partial data. To obtain full data we decided to crawl it online and then use it for analysis. The data crawled from both these modes is then cleaned for the removal of xml contents, urls, user mentions (twitter), etc. The cleaned data is then pushed into a database.

After crawling and data cleaning we had the following data entries in the database for training. Table 1 and 2 shows the number of posts we extracted overall.

Table 1 Male/ Female data set count:

SR. NO	CORPUS TYPE	MALE/FEMALE	TOTAL COUNT
1	Blogs	980/913	1893
2	Hotel Review	2823/2629	5452
3	Social Media	85086/83643	168729
4	Twitter	96980/69959	166939
Total Posts:		**185869/157144**	**343013**

Table 2 Age group data set count:

AGE GROUP	BLOGS	HOTEL REVIEW	SOCIAL MEDIA	TWITTER	TOTAL COUNT
18-24	102	436	34694	12498	47730
25-34	643	1351	47599	55843	105436
35-49	732	1366	48385	66844	117327
50-64	334	1285	37143	2229	65991
65 or above	82	1014	908	4525	5515
Total Posts:					**343013**

The data set is then retrieved from the database and sent for text analysis which involves Natural Language Processing, Readability Analysis, Syntactic analysis and Structural Analysis. We extract a total of 22 features from each data set and send it to Random Forest Classifier for training purposes. This trained ensemble classifier is then used later for training purposes. In the next module we will explain our implementation process in detail.

4 Implementation Details

As shown in Figure 1 our model consists of four main modules:

a. Crawlers: The training corpus provided by PAN-2014 contains documents in the XML format. However for some data sets like Twitter and Blogs, data had to be taken from HTML links in the XML file. Hence we had two modes of crawlers; one for offline data sets (Social Media and Hotel Review) and another for online data sets (Twitter and Blogs).

b. Data Cleaning: The raw text obtained from the crawlers has to be cleaned to remove noisy data like '\ufffd', XML tags, urls, twitter user mentions, 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.

c. Text Analysis: From the database the cleaned data is retrieved and we employ Natural Language Processing (NLP) techniques on the text data for its analysis. We used the NLTK [12] platform for performing NLP techniques like Stemming, Tokenizing, Parts of Speech (POS) tagging etc. Using these techniques we extract features which is further divided into 3 subsets listed below:

i. **Readability Metrics:** Though readability metrics were created to find the factors that help in making the text easy to read, it also plays a vital role in identifying the characteristics of the author [13]. The readability score is a statistical technique that computes readability based on the structure and

semantics of the sentence [14]. We use the following readability metrics for our analysis:

a. ARI:

$$4.71\left(\frac{no.of\ letters}{no.of\ words}\right)+0.5\left(\frac{no.of\ words}{no.of\ sentences}\right)-21.43$$

b. Flesch Reading Ease:

$$206.835-1.015\left(\frac{total\ words}{total\ sentences}\right)-84.6\left(\frac{total\ syllables}{total\ words}\right)$$

c. Flesch Kincaid Grade Level:

$$0.39\left(\frac{total\ words}{total\ sentences}\right)+11.8\left(\frac{total\ syllables}{total\ words}\right)-15.59$$

d. SMOG Index:

$$1.0430\sqrt{complex\ words\times\frac{30}{sentences}}+3.1291$$

e. Gunning Fog Index:

$$0.4\times\left[\frac{words}{sentences}+\left(100\times\frac{complex\ words}{words}\right)\right]$$

f. Coleman Liau Index:

$$5.879851\left(\frac{no.of\ letters}{no.of\ words}\right)+29.587280\left(\frac{no.of\ sentences}{no.of\ words}\right)-15.8$$

g. LIX:

$$\left(\frac{no.of\ words}{no.of\ sentences}\right)+100\left(\frac{no.of\ long\ words}{no.of\ words}\right)$$

h. RIX:

$$\left(\frac{no.of\ long\ words}{no.of\ sentences}\right)$$

ii. **Function Words:** For extracting the function words features we used the POS tagger of NLTK toolkit. We extract a total of 7 features from the text which include the number of nouns, adjectives, verbs, pronouns, determiners, adverbs and foreign words. Foreign words are those words which are mostly slangs used in internet like "Helloooo", "Whaaaat", "yipeee", "ROFL" etc.

iii. **Syntactic Features:** We extract the syntactic features from the text using the NLTK Tokenizer. We extract a total of 7 features which include number of single quotes, commas, periods, colons, semi-colons, question marks, exclamation marks etc.

d. Classifier: All of the 22 features collected are then fed into an ensemble classifier. For our analysis we used Random Forest classifier due to speed and accuracy. The classifier is trained with the whole data corpus and used later for testing purposes. The working of a cleaned test corpus is shown in Figure 2 given below.

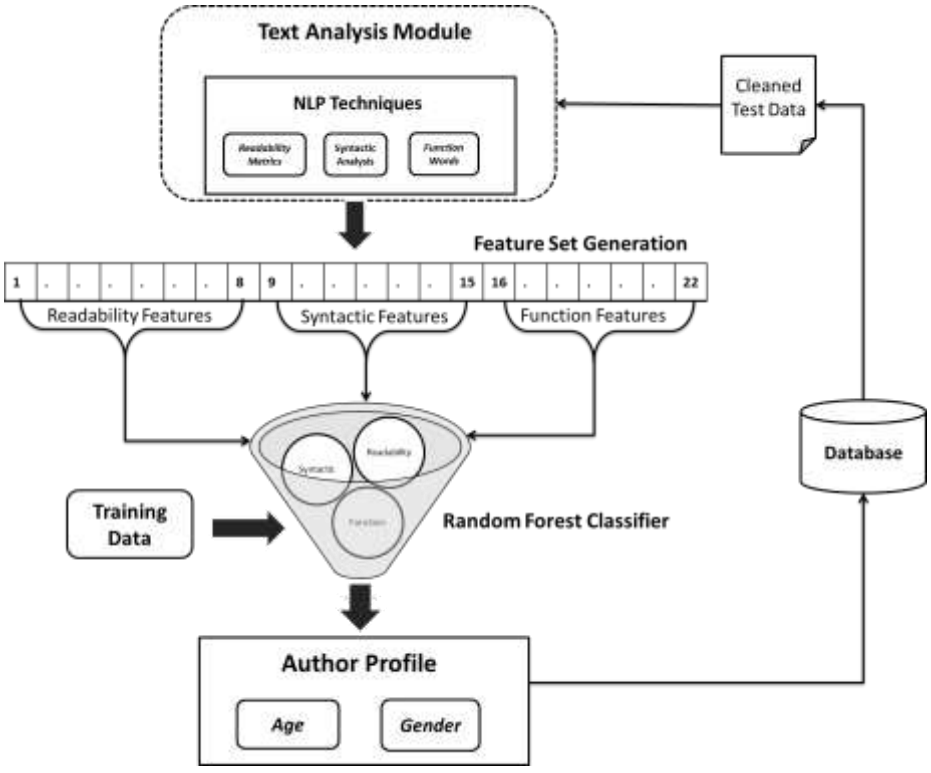

Figure 2 Working of cleaned test corpus

5 Results

In this section we present the results we obtained from our PAN Author Profiling Task. We evaluated our analysis on the test corpuses provided by PAN organizers using TIRA software. Table 3 and 4 show the age and gender predictions for English corpus 1 and 2 respectively.

Table 3 Results of Author Profiling - English Corpus-1:

DATA TYPE	AGE	GENDER	BOTH	RUNTIME (HH:MM:SS)
Blogs	0.1667	0.4583	0.0833	00:01:57
Hotel Review	0.2317	0. 5854	0.1220	00:04:30
Social Media	0.2558	0.5072	0.1272	03:54:55
Twitter	0.4667	0.5000	0.2333	00:39:28

Table 4 Results of Author Profiling - English Corpus-2:

DATA TYPE	AGE	GENDER	BOTH	RUNTIME (HH:MM:SS)
Blogs	0.2564	0.4231	0.1282	00:05:57
Hotel Review	0.2454	0.5189	0.1291	00:19:03
Social Media	0.2515	0.5198	0.1318	18:26:49
Twitter	0.3896	0.5000	0.1948	03:23:36

6 Conclusion

In this work, we present our system which identifies the age and gender of an author from a given document. We employed supervised Random Forest ensemble classifier for the Author Profiling task. We have performed our analysis on the 343013 training data for the English language corpus provided by the PAN-2014 organizers.

In our future work, we would like to perform a deeper analysis on the different features and traits and techniques that would help to improve the efficiency of our current system.

References

1. Zheng, Rong, et al. "A framework for authorship identification of online messages: Writing-style features and classification techniques." *Journal of the American Society for Information Science and Technology* 57.3 (2006): 378-393.
2. Peng, Fuchun, et al. "Language independent authorship attribution using character level language models." *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003.
3. Pennebaker, James W., and Lori D. Stone. "Words of wisdom: language use over the life span." *Journal of personality and social psychology* 85.2 (2003): 291.
4. Balakrishnan, Vimala, and Paul HP Yeow. "Texting satisfaction: Does age and gender make a difference." *International Journal of Computer Science and Security* 1.1 (2007): 85-96.
5. Burger, John D., et al. "Discriminating gender on Twitter." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.
6. Nguyen, Dong, Noah A. Smith, and Carolyn P. Rosé. "Author age prediction from text using linear regression." *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 2011.
7. Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh. "Predicting age and gender in online social networks." *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011.
8. Cheng, Na, Rajarathnam Chandramouli, and K. P. Subbalakshmi. "Author gender identification from text." *Digital Investigation* 8.1 (2011): 78-88.
9. Argamon, Shlomo, et al. "Gender, genre, and writing style in formal written texts."*TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN*- 23.3 (2003): 321-346.
10. Estival, Dominique, et al. "TAT: an author profiling tool with application to Arabic emails." *Proceedings of the Australasian Language Technology Workshop*. 2007.
11. Calix, K., et al. "Stylometry for e-mail author identification and authentication. "*Proceedings of CSIS Research Day, Pace University* (2008).
12. Bird, Steven. "NLTK: the natural language toolkit." *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.
13. Pitler, Emily, and Ani Nenkova. "Revisiting readability: A unified framework for predicting text quality." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.
14. Luyckx, Kim, and Walter Daelemans. "Shallow text analysis and machine learning for authorship attribution." (2005).

15. 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. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13), September 2013. Springer. ISBN 978-3-642-40801-4.
16. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos and Giacomo Inches. Overview of the Author Profiling Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013. ISBN 978-88-904810-3-1.
17. Rangel, Francisco, and Paolo Rosso. "Use of Language and Author Profiling: Identification of Gender and Age." Natural Language Processing and Cognitive Science (2013): 177.