Cross-Genre Author Profile Prediction Using Stylometry-Based Approach

Notebook for PAN at CLEF 2016

Shaina Ashraf, Hafiz Rizwan Iqbal, Rao Muhammad Adeel Nawab

Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan.
shainaashraf@ciitlahore.edu.pk, rizwan.iqbal@ciitlahore.edu.pk, adeelnawab@ciitlahore.edu.pk

Abstract. Author profiling task aims to identify different traits of an author by analyzing his/her written text. This study presents a Stylometry-based approach for detection of author traits (gender and age) for cross-genre author profiles. In our proposed approach, we used different types of stylistic features including 7 lexical features, 16 syntactic features, 26 character-based features and 6 vocabulary richness (total 56 stylistic features). On the training corpus, the proposed approach obtained promising results with an accuracy of 0.787 for gender, 0.983 for age and 0.780 for *both* (jointly detecting age and gender). On the test corpus, proposed system gave an accuracy of 0.576 for gender, 0.371 for age and 0.256 for *both*.

1 Introduction

The main concept behind author profiling is to determine the traits of a writer from his/her written text. We can predict different characteristics of an author by analyzing his/her written text, for example, age, gender, native language, qualification and personality traits etc.[1]. The writing style demonstrates the profile of an author and provides valuable information about his demographics. Identification of these author traits can be very helpful in different applications e.g. forensics analysis, security, intelligent marketing decisions, sentiment analysis and classification[2].

In this paper, we present an approach, based on different types of stylistic features. In total, we applied 56 stylistic features. These features are divided into four categories including lexical, syntactic, character-based and vocabulary richness measures. The reason for selecting this methodology is that the training and test datasets are on different genres i.e. the training has done using Twitter data and the evaluation performed on other genre different from Twitter tweets. We expect that capturing an author's writing style on different types of training and testing data will yield good results.

The problem of gender and age identification also treated as a supervised document classification task. Different machine learning algorithms including J48, Random Forest and LADTree were explored for classification task. Various *feature selection* methods including Best-First and Ranker etc. were also investigated to identify the subset of best features from the set of 56 features. Best results on the training data were obtained (using the LADTree machine learning algorithm), where all the 56 features were used for the gender and age identification task. The trained system deployed on TIRA [11] for final evaluation on test dataset(s). The comparison of our system with other participants has shown in [12].

Rest of this paper is organized as follows: Section 2 describes related work. Section 3 describes the proposed approach. Section 4 presents the experimental setup. Section 5 discusses results and their analysis. Finally, Section 6 concludes the paper and discusses future work directions.

2 Related work

Previous studies have commonly used Stylometry-based features to identify an author's traits from his/her writing style. For example, one of the pioneers in author profiling [3] explored some linguistic patterns in writing styles of authors which can be helpful in identifying different author traits like personality attributes, gender and age group. They carried out Part-Of-Speech tags analysis to get different stylistic features (i.e. function words, prepositions, pronouns, auxiliary verbs) for identification of gender of an author and the accuracy of 72% and 66% for gender and age identification respectively. Argamon [4] et. al identified the demographics of an author by combining different features i.e. (function words with POS tags and obtained an accuracy of 80% for gender identification, In [5, 6], authors presented a set of features like word unigrams, function words, non-dictionary words, hyperlinks for detection of age and gender of an author. Results showed 80% accuracy for gender identification and 75% accuracy for age identification. In the previous PAN Author Profiling Competitions, many submitted systems used stylistic features for predicting age, gender and personality types [7-9].

3 Proposed Approach

Our proposed approach is based on Stylometric (the study of linguistic style) features, which help us to capture a set of elements of writing. Since the writing style of one author is likely to be different from others, therefore, these Stylometric features can be useful in discrimination between an author's traits. The other reason for selecting various types of stylistic features is that the training data is in one genre and the test data is in another genre. Therefore, stylistic features were expected to accurately identify author traits even if they are trained and tested on different types of data.

Our proposed approach combines different types of stylistic features including lexical, syntactic, vocabulary richness and character based features. The next sections describe these feature types in more detail.

3.1 Lexical Features

Lexical features represent text as a sequence of tokens forming sentences, paragraphs and documents. A token can be numeric number, alphabetic word of a punctuation mark. These tokens are used to get statistics like average sentence length and average word length [5]. These features have the ability to get insights of a text in any language without special requirements. In our proposed system, we have implemented 7 lexical features: (1) average sentence length in characters, (2) average sentence length in words, (3) average word length, (4) percentage of question sentences, (5) total number of words, (6) total unique words and (7) words ratio of length 3.

3.2 Syntactic Features

Syntactic features consist of function words and parts-of-speech tags. Syntactic pattern varies significantly from one author to another. These features were extracted using more accurate and robust text analysis tools i.e. Part-of-speech taggers, chunckers and lemmatizers. In our proposed system, for the extraction of syntactic features, we have used Stanford Log-linear Part-Of-Speech Tagger[1]. The proposed approach contains 16 syntactic features: (1) number of adjectives, (2) number of nouns, (3) number of adverbs, (4) number of verbs, (5) number of cardinal number, (6) number of preposition, (7) number of particle, (8) number of symbol, (9) number of conjunction, (10) number of determiner, (11) number of Interrogative, (12) number of foreign words, (13) number of pronoun, (14) POS unigram density (see Equation 1), (15) POS bigram density (see Equation 2), (16) POS trigram density (see Equation 3).

$$POS\ Unigram\ Density = \left(\frac{number\ of\ different\ POS\ Unigram}{total\ number\ of\ \ POS\ Unigram}\right) \times 100 \qquad (1)$$

$$POS\ Bigram\ Density\ = \left(\frac{number\ of\ different\ POS\ Bigram}{total\ number\ of\ \ POS\ Bigram}\right) \times 100 \qquad (2)$$

$$POS\ Trigram\ Density = \left(\frac{number\ of\ different\ POS\ Trigram}{total\ number\ of\ \ POS\ Trigram}\right) \times 100 \qquad (3)$$

3.3 Vocabulary Richness

Every piece of text is composed of a set of unique words called its vocabulary. Vocabulary richness functions try to measure the diversity of vocabulary in a given

[1] http://nlp.stanford.edu/software/tagger.shtml Last visited: 25-05-2016

text i.e. how rich is the vocabulary [10]. Easiest and common example to understand vocabulary richness is hapax-legomena (number of words occurring exactly once) and type-token ratio i.e. *V/N* - where *V* is number of unique words in the text and *N* is the total number of words in the same text. Size of text/document directly affects the vocabulary size i.e. smaller documents will have less number of unique words while the larger ones will have higher number of unique words. To cater the influence of text size of vocabulary richness measures, a number of formulas have been used. In our proposed system, we have implemented 6 vocabulary richness measures (see Table 3.1).

Table 3.1 Vocabulary richness measures

Sr. No	**Feature Name**	**Formula**
1	Brunet W Measure	$W = N^{v^{-.165}}$
2	Hapax Legomena	$V_1 = (number\ of\ words\ appear\ exactly\ once)$
3	Honore R Measure	$\mathrm{R} = \frac{100 log \mathrm{N}}{1 - (\frac{V_1}{V})}$
4	Sichel S Measure	$S = \frac{V_2}{\mathrm{V}}$
5	Simpson D Measure	$\mathrm{D} = \frac{\sum n\,(\mathrm{n}-1)}{\mathrm{N}(\mathrm{N}-1)}$
6	Yule K Measure	$\mathrm{K} = \frac{10^4 \sum_{i=1}^{\infty} t^2 V_i - N}{N^2}$

3.4 Character Based Features

Character based features consider text as a sequence of characters. Thus, a number of character based measurements are defined including punctuation count, digit count, character count, colon count, comma count, question mark count etc. [5]. Such information is easily available in any language and corpora. Our proposed system contains 26 character-based features: (1) character count, (2) percentage of punctuation characters, (3) character count without spaces, (4) percentage of semi colons, (5) ratio of digits, (6) percentage Of commas, (7) ratio of letters,(8) apostrophe count, (9) ratio of upper case letters, (10) brackets count, (11) ratio of white-spaces to N (total no of characters in an author profile), (12) colon count, (13)

ratio of tabs to N, (14) comma count, (15) ratio of special character to N, (16) dash count, (17) number of upper case characters, (18) ellipsis count, (19) digit count, (20) exclamation count, (21) number of white-spaces, (22) full-stop count, (23) number of tabs, (24) question-mark count, (25) semicolon count, (26) slash count

4 Experimental Setup:

4.1 Training Corpus

We have used *pan16-training-dataset-english* to train our proposed system (we did not attempt author-profiling task for other languages i.e. Dutch and Spanish). The training corpus for English language is composed of Twitter tweets and contains 436 author profiles (see Table 4.1 for detailed statistics). The goal is to identify two author traits: (1) gender and (2) age. Gender identification task aims to discriminate between two classes: (1) male and (2) female, whereas age identification task aims to discriminate between five classes: (1) 18-24, (2) 25-34, (3) 35-49, (4) 50-64 and 65-xx.

Table 4.1 Distribution of data for age and gender attributes in the PAN16 training corpus

Total Author Profiles: 436						
Gender		**Age-Group**				
Male	**Female**	**18-24**	**25-34**	**35-49**	**50-64**	**65-xx**
218	218	28	140	182	80	6

We pre-processed both training and test datasets by removing xml tags, html tags etc. and only used plain text for experimentation.

4.2 Evaluation Methodology

The task of identifying an author's gender and age from his/her text is casted as a supervised document classification task. For gender identification, we have performed binary classification task i.e. goal is to distinguish between two classes: (1) male and (2) female. For age identification, we have performed multi-classification task i.e. goal is to categorize age among five classes: (1) 18-24 (2) 25-34 (3) 35-49 (4) 50-64 (5) 65-xx. We have used 10-fold cross validation for experiments. We explored multiple classifiers including J48, Random Forest, LADTree, to train and test our proposed system. The numeric values generated by 56 different Stylometry features (see Section 3) were used as input to these classifiers. Evaluation is carried out using *accuracy measure* for both age and gender identification tasks.

5 Results and Analysis

Table 5.1 shows the results for the proposed approach on both training and testing datasets. The results show that our proposed approach obtained promising results on the training data (0.983 accuracy for age trait, 0.787 accuracy for gender trait and 0.780 for *both* (joint identification of age and gender). This demonstrates that combination of different types of Stylometric features, which capture different types of information from a given text, is helpful in identifying age and gender of an author from his/her written text[2]. Overall, results of the proposed approach are low for both testing datasets (pan16-test-dataset1-english and pan16-test-dataset2-english), particularly for the early bird's evaluation corpus (pan16-test-dataset1-english).

Table 5.1 Results for age and gender on training and test data sets

Corpus	**Age**	**Gender**	**Both**
pan16-training-dataset-english	0.983	0.787	0.780
pan16-test-dataset1-english	0.290	0.497	0.149
pan16-test-dataset2-english	0.371	0.576	0.256

On the final evaluation corpus (pan16-test-dataset2-english), our proposed approach obtained an accuracy of 0.371 for age, 0.576 for gender and 0.256 for *both*. It can be noticed that these results are very low compared to the training corpus. The possible reason for this is that proposed system is trained on one genre (tweets) and it is tested on another genre (blogs, reviews, social media etc.). Also the effect of evaluation on a test dataset with different genre as that of training dataset is reflected in the difference of accuracy scores for training and test datasets. The proposed system gives very high accuracy on age (0.983) and it drops to 0.371 on test dataset. On the other hand, the accuracy for *gender* on training dataset is low as compared to *age,* but it is high on the test dataset. This clearly shows that models trained on one genre may not give the same pattern of performance if they are evaluated on a data set, which contains author profiles from a different genre.

[2] Note that we also applied feature selection on the set of 56 features but it did not improve performance. Best results were obtained when all the 56 features were used for age and gender identification

6 Conclusion and Future Work

In this paper, we presented an approach based on different types of stylistic features for identifying two author traits i.e. gender and age. The proposed system contains total 56 stylistic features including 7 lexical features, 16 syntactic features, 26 character-based features and 6 vocabulary richness measures. The system was trained using all the 56 features and different machine learning algorithms were explored including Random Forest, J48 and *LADTree*. Using the proposed approach, promising results were obtained on the training dataset (0.983 for age, 0.787 for gender and 0.780 for *both* (jointly identifying age and gender)). On the test data set, the proposed approach obtained accuracy of 0.371 for age, 0.576 for gender and 0.256 for *both*.

In future, we plan to combine other features, for example, content based, topic based etc., with stylistic features for cross-genre author profiling task.

References

1. Guthrie, D., Guthrie, L., Wilks, Y.: An Unsupervised Approach for the Detection of Outliers in Corpora. LREC (2008)
2. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. In: ACM Transactions on Information Systems (TOIS). 26(2): p.7 (2008)
3. Argamon, S., Koppel, M., Pennebaker, J. W., Schler, J.: Automatically profiling the author of an anonymous text. In: Communications of the ACM. 52(2): p.119-123 (2009)
4. Argamon, S., Koppel, M., Fine, J., Shimoni, A. R.: Gender, genre, and writing style in formal written texts. In: Text-The Hague Then Amsterdam Then Berlin. 23(3): p. 321-346 (2003)
5. Stamatatos, E.: A survey of modern authorship attribution methods. In: Journal of the American Society for information Science and Technology. 60(3): p. 538-556 (2009)
6. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. In: Computers and the Humanities. 35(2): p. 193-214 (2001)
7. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Working Notes. CEUR (2015)
8. Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daeleman, W.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF Evaluation Labs and Workshop (2014)
9. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation CELCT (2013)
10. Toutanova, K., Manning, C.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint

SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. Association for Computational Linguistics (2000)

11. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Tjoa, A., Liddle, S., Schewe, K.D., Zhou, X. (eds.) 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. IEEE, Los Alamitos, California (2012)

12. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Evaluations Concerning Cross-genre Author Profiling. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2016)