Identification of Author Personality Traits using Stylistic Features
Notebook for PAN at CLEF 2015

Ifrah Pervaz[1,*], Iqra Ameer[2,*], Abdul Sittar[3,*], Rao Muhammad Adeel Nawab[4,*]

[*]COMSATS Instiute of Inforamtion Technology, Lahore
[1]ifrahpervaz23@gmail.com,
[2]iqraameer133@gmail.com,
[3]abdulsittar72@gmail.com,
[4]adeelnawab@ciitlahore.edu.pk

Abstract. Author profiling is the task of determining the age, gender or type of the author's personality by studying their sociolect aspect, that is, how the language is shared by people. This paper presents the COMSATS Institute of Information Technology, Lahore entry for the PAN 2015 competition on Author Profiling task. Our proposed system is based on stylometry features. We implemented 29 different stylistic features, many of which are language independent. Since the training data was available in multiple languages, one of our main objectives was to explore which language independent features are most effective. The problem of author profiling was casted as a supervised document classification task. Results showed that features (Percentage of Question Sentences, Average Sentence Length, Percentage of Punctuations, Percentage of Comma and Percentage of Full stops) were most effective multilingual features.

1 Introduction

Personality in Encyclopedia of Psychology is defined as "*individual differences in characteristic patterns of thinking, feeling and behaving."* [1] These differences can be reflected through one's speech, writing, images etc. Authorship analysis deals to find out the methods to identify these differences and use them to know profile of author such as gender, age, native language, education, profession or personality type. So author profiling can simply be defined as: given the set of texts, you need to identify age, gender, profession, education, native language and similar personality traits.

Authorship analysis has attracted much attention in recent years due to the rapid increase of electronic text and the need for expert systems able to handle this information. Like from a marketing viewpoint, companies may be interested in knowing about the trends regarding their products, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products, what is

required by the customer and which customer category/ class they can attract more. Similarly from a forensic viewpoint, determining the linguistic profile of a person i.e. who wrote a "suspicious text"' may provide valuable background information.

In this paper we explore how different stylistic features help in conveying about the personality of writer. Moreover, we also tried to find out how the use of different stylistic features affect multilingual results. We used the datasets provided by PAN organizers, and applied different machine learning practices for predicting writer's traits.

The rest of this paper is organized as follows: Section 2 describes related work. Section 3 presents the approach used to identify personality traits of author. Section 4 describes the results obtained on training data. Section 5 presents the results obtained on test data and section 6 concludes the paper.

2 Related Work

With the advancement in web technology, there is an abundant increase in use of electronic media. Along with useful information there is also heap of anonymous data available so the need to identify "who is who" has become important.

A lot of establishments have been done in this field so far, Pennebaker et al. [2] joined dialect utilization with identity qualities, examining how the variety of semantic attributes in a content can give data in regards to the gender and age of its author.

Argamon et al. [3] analyzed formal written texts extracted from the British National Corpus [4][1] combining function words with part-of-speech features. Koppel studied the problem of automatically determining an author's gender by proposing combinations of simple lexical and syntactic features.

Holmes and Meyerhoff [7], Burger and Henderson [6] have also investigated obtaining age and gender information from formal texts.

Seifeddine and Maher [8], focus is on author's discussions, they combine content based approach and statistic approach. They show a hemi- strategy that merges the analysis of information in writings with a machine learning technique. To get a higher organization of these statistics, they depended on the utilization of the "Decision table calculation".

3 Our Approach

3.1 Stylistic Features

The approaches commonly used for the automatic identification of an author's personality traits from text can be categorized into three broad categories:

[1] www.natcorp.ox.ac.uk

(1) Stylometry based approaches (which aim to identify an author's traits from his writing style), (2) Content based approaches (which identify author traits using features extracted from the content of the document) and (3) Topic based approaches (which try to predict an author's profile based on the topics used in the document).

Table 3. 1: List of all the stylistic features that are used in the development of the author profiling detection system.

		Languages			
No.	**Feature**	**English**	**Dutch**	**Spanish**	**Italian**
1.	Percentage of Question Sentences	Yes	Yes	Yes	Yes
2.	Percentage of Short Sentences	Yes	Yes	Yes	Yes
3.	Percentage of Long Sentences	Yes	Yes	Yes	Yes
4.	Average Sentence Length	Yes	Yes	Yes	Yes
5.	Average Word Length	Yes	Yes	Yes	Yes
6.	Percentage of Words with Six and More Letters	Yes	Yes	Yes	Yes
7.	Percentage of Words with Two and Three Letters	Yes	Yes	Yes	Yes
8.	Percentage of Semicolons	Yes	Yes	Yes	Yes
9.	Percentage of Punctuations	Yes	Yes	Yes	Yes
10.	Percentage of Pronouns	Yes	----	----	----
11.	Percentage of Prepositions	Yes	----	----	----
12.	Percentage of Coordinating Conjunctions	Yes	----	----	----
13.	Percentage of Comma	Yes	Yes	Yes	Yes
14.	Percentage of Articles	Yes	----	----	----
15.	Percentage of Words with One Syllable	Yes	----	----	----
16.	Percentage of Words with Three Plus Syllables	Yes	----	----	----
17.	Average Syllables per Word	Yes	----	----	----
18.	Percentage of Adjectives	Yes	----	----	----
19.	Percentage of Adverbs	Yes	----	----	----
20.	Percentage of Capitals	Yes	Yes	Yes	Yes
21.	Percentage of Colons	Yes	Yes	Yes	Yes
22.	Percentage of Determiners	Yes	----	----	----
23.	Percentage of Digits.	Yes	Yes	Yes	Yes
24.	Percentage of Full stop	Yes	Yes	Yes	Yes
25.	Percentage of Interjections	Yes	----	----	----
26.	Percentage of Modals	Yes	----	----	----
27.	Percentage of Nouns	Yes	----	----	----
28.	Percentage of Personal Pronouns	Yes	----	----	----
29.	Percentage of Verbs	Yes	----	----	----

The system presented in this PAN Author Profiling Competition is based on stylometry. Table 3.1 shows the list of 29 stylistic features used for the development of our proposed author profiling detection system. As can be noted that these features aim to extract different stylistic information from a single document, which can be helpful in identifying the age, gender, personality type i.e. stable, open, extroverted, agreeable and conscientious .In this year's training data, the tweets are available for four different languages: (1) English, (2) Dutch, (3) Spanish and (4) Italian. This study aims to identify some language independent stylistic features which are probable to perform on multiple languages as well. As can be noted from Table 3.1 that features numbered 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 20, 21, 23 and 24 are language independent and can be used for extracting stylistic information from document in any of the four languages i.e. English, Dutch, Spanish and Italian, while the remaining ones can only be used for English language.

3.2 PAN 2015 Author Profiling Training Datasets

Our proposed system was trained using the PAN 2015 Author Profiling training data, which consists of Twitter tweets in four different languages: (1) English, (2) Dutch, (3) Spanish and (4) Italian. In the training dataset there are 152, 34, 100 and 38 author profiles for English, Dutch, Spanish and Italian languages respectively.

In the PAN 2015 Author Profiling training dataset, there are two classes for gender (male and female), four classes for age (18-24, 25-34, 35-49 and 50-xx) for the remaining personality traits open, stable, agreeable, extroverted and conscientious there are two classes: (yes or no).

3.3 Evaluation Methodology

The task of identifying an author's profile from text was treated as a supervised machine learning problem. N-fold-cross-validation was used. Due to difference in the sizes of training data for different languages, we used 5-fold cross validation for English language corpus, 4-fold cross validation for Dutch and 3-fold cross validation for the Italian and Spanish corpora respectively.

We applied a range of machine learning algorithms on the training data including *Navies Bayes, Support Vector Machine, Random Forest, J48 and Logistics.* The WEKA's[2] [9] implementation of these algorithms was used. The scores generated using the stylistic features (see Table 3.1) were used as "input features" to the machine learning algorithms.

We used WEKA's "attribute selection" approach for selecting "best features" from the complete feature set. For that purpose we explored *CFsSubSetEval, Filtered Attribute,*

[2] mailto:https://weka.wikispaces.com/

SVM Attribute and Classifier Subset as evaluators, and *Best first and ranker* as search methods.

3.4 Evaluation Measure

As recommended by PAN 2015 organizers, the performance of the proposed system for *age* and *gender* personality traits was measured using *accuracy,* whereas for other personality traits (stable, open, extroverted, agreeable and conscientious) *average Root Mean Squared Error* was used.

4 Results on Training Data

We carried out three sets of experiments: (1) performance on individual features, (2) performance on combined features (mean all 29 features are used as input) and (3) performance on "selected features". The best performance was obtained using "selected features", therefore, we are only reporting the results for them.

Table 4.1: shows the results for training data on all four language, for age and gender accuracy scores are reported, while for other personality traits average root mean square error scores are presented

Language	Gender	Age	Stable	Open	Agreeable	Extroverted	Conscientious
English	0.75	0.65	0.45	0.14	0.43	0.41	0.42
Dutch	0.64	---	0.36	0.00	0.39	0.20	0.47
Spanish	0.73	0.53	0.51	0.37	0.47	0.00	0.34
Italian	0.73	---	0.14	0.16	0.33	0.43	0.29

Table 4.2: shows the features selected for training data on all for language.

Traits	English	Dutch	Spanish	Italian
Age	Ranked in order: 28,7,19,11,12,10,29,21,16,14,6,17,9,18,27,22,24,23,15,26,4,13,1,5,25,20,8,3,2	----	Ranked in order: 4,13,9,1,6,21,7,24,23,8,5,20,3,2	----
Gender	Ranked in order 23,10,22,20,18,9,1,5,28,11,21,14,25,24,12,27,13,7,6,26,8,29,15,16,17,4,19,3,2	Ranked in order 21,9,8,13,6,1,20,23,4,14,7,5,3,2	Ranked in order 1,9,23,7,8,11,4,21,5,6,14,13,3,2	Ranked in order 24,1,4,8,20,5,9,13,23,6,7,21,3,2
Open	1	1	1	1
Stable	6,27,29	Ranked in order 6,8,5,7,1,23,21,20,13,24,9,4,3,22	24	1
Agreeable	1	Ranked in order 13,20,1,6,13,4,8,24,5,7,9,21,3,2 Ranked in order 21,24,5,8,6,1,13,9,4,20,7,23,3,2	24 Ranked in order 24,23,21,20,13,9,8,7,6,5,4,3,2,1	1 Ranked in order 8,21,9,7,6,24,20,13,5,23,1,4,3,2
Conscientious	7,21	4,6,7,21,23,24	1	1

Table 4.1 demonstrates the accuracy for age and gender trait and average root mean square error for personality types, for tweets in all four languages. Best features for age and gender traits in English, gender, stable, agreeable and extroverted traits in Dutch, age, gender and extroverted traits in Spanish and gender and extroverted traits in Italian are obtained by using "ranker" search method while remaining are obtained through "best-first" search method, the features chosen by these methods are shown in table 4.2

5 Results on Test Data

Table 5.1 shows the accuracy for age and gender trait and average root mean square error for personality types, for test data in all four languages i.e. (1) English (2) Dutch (3) Spanish (4) Italian. For evaluation of test data we train our modal using all features as discussed in table 3.1., and applying "chi squared" evaluator and "ranker" search method. It can be concluded from table 5.1 that overall high accuracy for age and gender is achieved for English language, while for other traits i.e. open, stable, agreeable, extroverted and conscientious overall best results are achieved for Dutch language.

Table 5.1: shows the results for test data on all four language, for age and gender accuracy scores are reported, while for other personality traits average root mean square error scores are presented.

Language	Gender	Age	Stable	Open	Agreeable	Extrovert	Conscientious
English	0.690	0.718	0.317	0.215	0.215	0.213	0.196
Dutch	0.594	------	0.168	0.087	0.144	0.168	0.142
Spanish	0.693	0.534	0.281	0.214	0.143	0.279	0.141
Italian	0.583	-------	0.230	0.245	0.146	0.107	0.133

6 Conclusion

In this paper we have discussed our participation in the Author Profiling task. We have covered the role of stylistic features in identification of author personality traits. For that purpose we figured out 29 features and perform different experiments on these, like comparing accuracy by using all features, then checking accuracy for single feature and finally using subsets of them and come to conclusion that best results are achieved by feature selection techniques.

References

1. Alan E. Kazdin, PhD, Editor-in-Chief Encyclopedia of Psychology: 8 Volume Set

2. Pennebaker,J.M., Mehl,M.R., and Niederhoffer, K.G.,Psychological aspects of natural language use: Our words, our selves. Annual review of psychology, 54(1):547–577, 2003.

3. Argamon,S., Koppel,M., Fine,J., and Shimoni,A.R., Gender, genre, and writing style in formal written texts. TEXT, 23:321–346, 2003.

5. Argamon,S., Koppel,M., Pennebaker,J.W., and Schler,J., Automatically profiling the author of an anonymous text. Commun. ACM, 52(2):119–123, February 2009.

6. Burger, J.D., Henderson, J., Kim,G., and Zarrella,G., Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

7. Holmes, J., and Meyerhoff,M., The Handbook of Language and Gender. Blackwell Handbooks in Linguistics. Wiley, 2003. ISBN 9780631225027.

8. Mechti,S., Jaoua,M., Belguith,L., and Faiz1,R., Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media Notebook for PAN at CLEF 2014.

10. Jason Brownlee Feature Selection to Improve Accuracy and Decrease Training Time, March 12, 2014 in Uncategorized

11. Ian H. Witten, Eibe Frank and Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques

12. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inche, G., Navigli,R., and Tufis,D., (ed), Overview of the Author Profiling Task at PAN 2013 Working Notes Papers of the CLEF's. 2013 Evaluation Labs, September 2013. ISBN 978-88-904810-3-1.