Author's Traits Prediction on Twitter Data using Content Based Approach

Notebook for PAN at CLEF 2015

Fahad Najib, Waqas Arshad Cheema, Rao Muhammad Adeel Nawab

Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan.
choudharyfahad@gmail.com, waqascheema06@gmail.com, adeelnawab@ciitlahore.edu.pk

Abstract This paper describes the methods we have employed to solve the author profiling task at PAN-2015. The proposed system is based on simple content based features to identify an author's age, gender and other personality traits. The problem of author profiling was treated as a supervised machine learning task. First content based features were extracted from the text and then different machine learning algorithms were applied to train the models. Results showed that content based features approach can be very useful in predicting the author's traits from his/her text.

1 Introduction

Authorship attribution concerns with the classification of documents into the classes to be predicted based on the writing style of their authors. In the case of author verification and author identification tasks, the style of individual authors is examined. Whereas author profiling mean to distinguish between classes of authors studying their sociolect aspect, that is, how language is shared between people. This helps in predicting profiling aspects such as age, gender or personality type. Author profiling is a problem of increasing importance in several applications like forensics, security and marketing. E.g., from a forensic linguistics prospect, the linguistic profile of the sender of a harassing SMS message can be identified. Similarly, from a marketing perceptive, companies would like to know the demographics of the people that like or dislike their products on the basis of the text analysis of online product reviews and blogs.

In recent years, automatic detection of an author's profile from his/her text has become an emerging and popular research area (Rangel et al., 2013). Automatically predicting the identity of authors from their texts has a lot of future applications. for e.g., forensics analysis (Corney et al., 2002; Abbasi and Chen, 2005), marketing intelligence (Glance et al., 2005) and classification and sentiment analysis (Oberlander and Nowson, 2006).

2 Related work

A significant amount of research in automatic classification of texts into the classes to be predicted, has already been done by different researchers and linguists using several different machine learning techniques (Sebastiani, 2002). Over the past few years, a large

variety of techniques have been devised for predicting the text based on its author's traits (Abbasi and Chen, 2005; Houvardas and Stamatatos, 2006; Schler et al., 2006; Argamon et al., 2009; Estival et al., 2008; Koppel et al., 2009). Previously different machine learning classifiers tried to include variety of techniques such as Lazy learners (IBk)(Estival et al., 2008, 2007), Support Vector Machine (SVM) (Koppel et al., 2009; Estival et al., 2007), LibSVM (Estival et al., 2008), RandForest (Estival et al., 2008), Information Gain (Houvardas and Stamatatos, 2006), Baysian Regression (Koppel et al., 2009), Exponential Gradient (Koppel et al., 2002) etc.

Several approaches have been implemented and experiments conducted for selecting the best possible features set for the most accurate classification. Houvardas and Stamatatos (2006) showed the usefulness of n-gram, whereas Koppel et al. (2009) shown the effect of gender and age in blogging sites by considering different word classes and showing the relation of the word classes with the author's age and gender. Koppel et al. (2009); Estival et al. (2007) have identified that the Part-of-speech is also an commendable linguistic feature and (Calix et al., 2008) achieved the accuracy of 76.72% using 55 different features.

3 Experimental setup

The data used in our experiments is the training dataset of PAN-2015 [1]. The corpus consists tweets on different topics, grouped by author and labeled with his/her language, gender, age group and 5 personality traits (extroverted (Ex), stable (St), agreeable (Ag), conscientious (Co) and open (Op)). The documents are categorized as in languages (English, Dutch, Italian and Spanish), two genders (male and female), and four groups (18-24, 25-34, 35-49 and 50-XX). With regard to personality traits, for each trait the scores lies between -0.5 and 0.5. Documents in the corpus consist of a collection of posts made by a single user.

Language				**Gender**		**Age-Group**			
English	Dutch	Italian	Spanish	Male	Female	18-24	25-34	35-49	50-XX
152	34	38	100	162	162	80	106	44	94

Table 1. Distribution of data in language, gender and age.

The corpus was balanced gender wise within each age group but imbalanced in terms of age representation and the five personality traits scores distribution (-0.5 to 0.5). The proportion of languages, gender and each age group in the corpus within the training dataset is presented in Table 1, whereas the personality traits distribution in table 2.

Prior to any model training or testing, we apply some pre processing steps to all documents. We eliminated all the data contents that were not determined to be the text written from the user like XML tags, as our primary source of features is the text written

[1] `http://pan.webis.de/`

Class	-5	-4	-3	-2	-1	0	1	2	3	4	5
Agreeable	0	0	5	7	30	33	81	105	34	13	16
Conscientious	0	0	0	3	6	58	78	59	55	41	24
Extroverted	0	0	4	4	15	33	87	89	36	31	25
Open	0	0	0	0	9	12	102	74	38	52	37
Stable	0	0	13	17	56	24	42	64	47	41	20

Table 2. Distribution of data in personality traits.

by an author. Now as all the user posts lie within the unparsed data tags of the source's xml file, we disregard any text not within these tags and HTML tags also.

4 Feature selection

As male and females like to write about different topics, they use different words accordingly. This leads to the fact that content based features can be an important tool to distinguish between texts of males and females (Schler et al., 2006). For example, a tweet concerned to sports will be more likely to be written by a male author rather by a female. That tweet may contain words like goal, score, world cup etc. So the occurrence of words like these will increase the chances of it being written by a male author. Similarly occurrence of words or phrases like my husband, shopping, nailpolish etc will increase the chances of it being written by a female author. In a similar fashion, people in their teen age like to write more about their school life, and friends. Whereas people in their 20's like to write more about their college life and people of 30's write more about jobs, marriage and politics. So the content based features can be an important tool to distinguish between texts written by people belonging to different profiles.

Word	Gender		Age-Group			
	Male	Female	18-24	25-34	35-49	50-XX
fuck	.22	.14	.29	.07	0	0
love	.16	.37	.28	.17	.05	.03
peopl	.19	.14	.21	.08	.04	0
feel	.13	.07	.14	.06	0	0
data	.13	.01	0	.02	.01	.02
life	.13	.13	.16	.04	.06	0
time	.12	.19	.02	.01	0	1
job	.09	.01	.01	.09	0	0
girl	.09	.02	.1	.01	0	0
day	.21	.14	.2	.09	.03	.03

Table 3. Frequencies comparison of unigrams of English language for age groups and gender.

We calculated the frequencies of different unigrams in the texts written by a particular profile. Then, for every unigram, we calculated the ratio of its frequencies in

the tweets of different classes like male and female, different age groups and different personality traits. Finally we selected the features on the basis of two combined factors. First the unigrams with the highest frequencies in the corpus, and second the difference of frequencies in different classes which are to be classify from one another. The frequencies of some of the most frequently used unigrams for English language in the corpus and their frequency comparison along gender and age groups is given in table 3. Similarly this routine has carried out for all the four languages individually and four sets of content based features selected for each language each. Then for each language, two different sets of features has been used, one for gender and age group prediction and one for the personality traits classification.

5 Models and Evaluation

For each of the four languages, we trained different models. Then for each language, we trained two models, one for age and gender and one for the personality traits. So there were total training models build all on the content based features. We ran the experiments on four machine learning classifiers: J48, Random Forest, Support Vector Machines (SMO), and Naive Bayes. The evaluation measures used as instructed by PAN 2015 [2], accuracy for age and gender and root mean squared error for the five personality traits.

6 Results and Analysis

Language	Gender	Age	Both	Ex	St	Ag	Co	Op	RMSE	Global
English	0.914	0.967	0.894	0.076	0.093	0.088	0.087	0.077	0.084	0.905
Italian	1.000	NA	NA	0.028	0.051	0.048	0.032	0.038	0.039	0.980
Spanish	0.940	0.990	0.930	0.085	0.110	0.084	0.102	0.083	0.093	0.918
Dutch	1	NA	NA	0	0	0	0	0	0	1

Table 4. Results on training data.

Based on the performance of the four classifiers (see section 5) on the training data, we choose only one single classifier (SVM) for all the classes of age, gender and personality in our final software that we have submitted in the competition. We are repoting the results that we have achieved on the training data in table 4 and finally on the testing data in table 5. Here the combined accuracy for age and gender (open), the combined root mean squared error for all five personality traits (RMSE) have been reported as well. Again here the age group results for Italian and Dutch are missing in both training and testing data suggesting there is only 1 class.

In comparison of performance in different languages, our system performed better for English and Dutch as compared to Spanish and Italian. Traits wise the performance

[2] http://pan.webis.de/

Language	Gender	Age	Both	Ex	St	Ag	Co	Op	RMSE	Global
English	0.591	0.669	0.422	0.187	0.261	0.176	0.161	0.195	0.196	0.613
Italian	0.527	NA	NA	0.160	0.220	0.157	0.136	0.190	0.173	0.667
Spanish	0.840	0.568	0.454	0.159	0.247	0.188	0.152	0.171	0.183	0.635
Dutch	0.468	NA	NA	0.136	0.176	0.091	0.123	0.091	0.124	0.672

Table 5. Results on testing data.

is reasonably good for Spanish gender, English age and Dutch personality. Overall, our system couldn't perform as well on the testing data as it performed on training data. The possible reasons for this are the variation of language in training and testing data and the nature of the content based approach.

7 Conclusion

In this paper, we have presented a content based technique for the automatic classification of the author's gender, age and personality from their writing. This work has a number of potential applications like marketing, forensics, and security. We have performed our experiments on the training data provided by the PAN-2015 organizers. We applied some simple content based techniques and the results we achieved are highly motivational showing the usefulness of content based features in predicting the author's profile from text. In future work, the results can be further improved by incorporating and finding more suitable features.

Bibliography

1. Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to arabic web content. In *Intelligence and Security Informatics*, pages 183–197. Springer, 2005.

2. Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

3. K Calix, M Connors, D Levy, H Manzar, G MCabe, and S Westcott. Stylometry for e-mail author identification and a uthentication. *Proceedings of CSIS Research Day, Pace University*, 2008.

4. Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, pages 282–289. IEEE, 2002.

5. Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLINGâĂŹ07)*, pages 263–272. PACLING, 2007.

6. Dominique Estival, Tanja Gaustad, Ben Hutchinson, Son Bao Pham, and Will Radford.Author profiling for english and arabic emails. 2008.

7. Natalie Glance, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428. Association for Computing Machinery, 2005.

8. John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer, 2006.

9. Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4): 401–412, 2002.

10. Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

11. Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.

12. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, pages 23–26, 2013.

13. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

14. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.