Predicting an author's demographics from text using Topic Modeling approach

Notebook for PAN at CLEF 2015

Hafiz Rizwan Iqbal, Muhammad Adnan Ashraf, Rao Muhammad Adeel Nawab

COMSATS Institute of Information Technology, Lahore

rizwan.iqbal@ciitlahore.edu.pk,adnan.ashraf@ciitlahore.edu.pk,
adeelnawab@ciitlahore.edu.pk

Abstract. The paper presents an approach to predict personality traits of a writer for the author profiling task of the PAN CLEF 2015. The task aimed at predicting authors' demographics based on the written tweets of an author. These demographics included traditional authorship attributes of age, gender and various personality traits of an author. We applied topic modeling using LDA as baseline approach and used the generated topic to get hierarchical probabilities of the topics. J48 decision tree was used for training classification model. The trained models were then used to successfully predict the demographics of training and test datasets

1 Introduction

Identifying various demographic traits such as, age, gender, native language and other personality aspects, from the authors writing style is termed as Author profiling [2]. Due to its high implication in the computer forensics, marketing and content recommendations over the internet, it has become a hot research area in Natural Language Processing.

Twitter has been the field of quantitative study on a number of aspects and characteristics recently. The primary interest of researchers has been to process the user tweets to interpret users' interests and to correlate social and global happenings [1] whereas this research focus on predicting the author profiling attributes. Twitter dataset has been used in this research for author profiling.

PAN 15 is the competition held as a part of CLEF Conference. The PAN 15' competition is designed for three different tasks namely, Plagiarism Detection, Author Verification and Author Profiling. Each task required to develop a composite software and submission on the TIRA, an evaluation engine.

The PAN 15' Author profiling task was designed to evaluate seven demographic constraints of the author from his/her tweets. These demographics include identifying authors' age, gender and five personality traits which include extroverted, stable, agreeable, conscientious and open. The training corpus was provided by PAN in four

different languages, English, Spanish, Italian and Dutch. The target was to achieve the highest ranking rating, which included ratio for accurately identifying the authors age and gender and the average Root Mean Squared Error for the personality constraints.

To predict a given author's attributes, we generated LDA based topic models using mallet and used J48 decision tree in Weka for training and evaluation of our model. LDA identifies latent topic associations in multi-document collection where each topic is assigned a probability with respect to all other topics in a document and also each topic is assigned a probability with respect to number of words [1]. Topic modeling using standard LDA has gained attention recently and work has been conducted in community detection using LDA [11] and author profiling. Topic modeling using LDA has also provided encouraging results in microblogging and its application [12].MALLET [9], a famous topic modeling and inferring toolkit, uses LDA to build the topic models for given text.

This paper focuses on the English tweets of the PAN 15' provided dataset for both training and testing phases [7]. The detail of the methodology is explained in the Section 2 while results of training phase and testing phase are discussed in the Section 3 and Section 4, respectively. Section 5 provides conclusion and future work.

2 Proposed Approach

We used topic modeling [3] as the baseline approach to predict an author's profile on the basis of his/her tweets. Why topic modeling as baseline approach? It has been analyzed that different categories of people have different topics of interests [6] e.g. women mostly talk about fashion, dresses and cooking etc. whereas men like to discuss politics, cricket and technology etc. This natural phenomenon leads us to predict a person's age, gender and other personality traits on the basis of his/her written text. There are the three stages in our proposed approach (1) Dataset Pre-processing, (2) Fabrication of Topic and Classification Models (3) Prediction of author traits.

2.1 Pre-processing:

The English Language training dataset provided by PAN 15' was selected for the author profiling task. The training dataset consisted of 152 users' tweets. Each user's data was placed separately in an xml file. The classifications of all xml files were placed in a single text file.

During pre-processing phase only tweets were extracted from each xml file and were stored in a separate text file for each user. There was no further pre-processing performed on the dataset, such as stop word removal, stemming, removal of punctuation marks, lemmatization, as the topic model disregards it and to retain the author's original style based features [4].

2.2 Fabrication: Topic and Classification Models

The provided dataset consisted of three main demographic traits of users, i.e. gender, age and personality constraints. Age and gender had accuracy values in classification

whereas the five personality constraints had root mean error as the classification values.

A directory structure was created with subdirectories for two demographics (age and gender) and five personality traits (extroverted, stable, agreeable, conscientious and open). Table 1 enlists the classification details of the dataset provided in PAN 15'. The text files extracted in pre-processing stage were placed in their classification based subdirectory structure. The dataset contained equally distributed profiles for the male and female authors. By analyzing the dataset, it was found that the majority of the profiles' authors were from the first two age groups (i.e. 18-24 and 25-30) whereas the profiles from age group 34-50 and 50+ were relatively lower. Each personality identifier was further classified based on provided root mean square error value ranging between -0.5 and 0.5 [7].

Each subdirectory was imported into MALLET, ran the topic modeling routine with setting of 20 topics for each subdirectory and inference file. As an output of this routine, list of extracted topics, topic composition file (file which contains the probability of participation of other topics into a single topic), trained topic model and topic inference file [9] was generated in a sequential order with respect to each trait directory.

ARFF (Attribute Relation File Format) [10] file was created from the topic composition file. Each topic was considered as one attribute and its probability taken as value of that attribute. Classification attribute was created for each arff file with respect to each personality trait. Each author arff file was sourced to WEKA and J48 tree classifier algorithm [10] was applied for construction of classification model for the respective personality attribute.

Table 1. -Classification of English dataset

Gender	**Male**			**Female**					
	76			76					
Age	**18-24**			**25-34**		**34-49**		**50+**	
	58			60		25		12	
Extroverted	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	1	4	10	17	41	37	20	13	9
Stable	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	11	5	22	9	19	37	19	18	12
Agreeable	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	5	2	12	19	44	46	13	7	4
Conscientious	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	0	1	4	30	38	27	33	12	7
Open	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	0	0	2	1	47	39	12	19	21

2.3 Prediction of Author Traits

To predict files in test data set, first two steps of the proposed approach with little variation in step 2, were applied on each test file to get the topics list, topic composition file and finally arff file. The test file was then compared with trained classification model to predict each personality trait value. The predicated results were then output in an xml file as per the task requirement.

3 Results for Training Phase

The final submission consisted of java based composite software which required an input directory consisting of xml files and an output directory to place the resultant xml files. The submitted software was first run on training dataset. Table 2 shows the results obtained on the PAN 15' training dataset with accuracy as evaluation measure for age and gender attributes whereas the personality traits' results based on Root Mean Square Error [RMSE] are presented in Table 3. The results show that our software was able to predict 54% correct classification for the age and 81.5% for the Gender whereas 44.7% correct predictions were made for both correct age and gender for the users. Similarly the results on personality traits are also encouraging.

Table 2. - Results on Age and Gender

Age	Gender	Both
0.540	0.815	0.447

Table 3. - Results on Personality Traits

Extroverted	Stable	Agreeable	conscientious	Open	RMSE	Global
0.150	0.200	0.154	0.149	0.100	0.151	0.648

4 Results for Testing Phase

The trained models were then run on the English test dataset 2 provided by PAN 15'. The evaluated test results are manipulated in the Table 4 and Table 5. The Test results on age and gender were different from the training dataset results. We were able to predict the age more accurately (69.7%) than the age on training dataset (54%) but gender prediction was poor (55.6 %) with respect to the gender on the training dataset (81.5%). Similarly the results of the personality traits on the test dataset were also encouraging with respect to the training dataset.

Table 4. - Test Results on Age and Gender

Age	Gender	Both
0.697	0.556	0.394

Table 5. - Test Results on Personality Traits

Extroverted	Stable	Agreeable	conscientious	open	RMSE	Global
0.208	0.315	0.191	0.190	0.214	0.224	0.585

5 Conclusion and Future Work

Author profiling requires an efficient and effective system for analyzing data for security and commercial purposes. In our approach, we developed a java based software that implied LDA for topic model and J48 classification algorithm to predict writers' demographics from the twitter dataset provided by PAN 15'. The results obtained are very encouraging especially the accuracy measures.

Future efforts can be focused on applying the different variations of topic modeling algorithm such as hierarchical LDA and implying supervised classification models to predict the demographic traits more accurately and precisely. The code will be optimized and effort can be put to minimize the total runtime of the software.

6 References

1. Liangjie Hong and Brian D. Davison.: Empirical Study of Topic Modeling in Twitter, 1st Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA (2010)
2. M. Suraj, S. Prasha and S. Thamar.: A Simple Approach to Author Profiling in MapReduce, Notebook for PAN, CLEF (2014)
3. Blei, David M.; Ng, Andrew Y.: Jordan, Michael I.: Latent Dirichlet allocation. In Lafferty, John. Journal of Machine Learning Research 3 (4–5): pp. 993–1022. (2003)
4. Pavan A., Mogadala A., Varma V.: Author profiling using LDA and Maximum Entropy, Notebook for PAN at CLEF (2013)
5. Caruana, R. and Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms". In Proceedings of the International Conference on Machine Learning. Pittsburgh, Pennsylvania, pp. 161–168 (2006)
6. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org, (2015)
7. K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma.: Author Profiling: Predicting Age and Gender from Blogs, Notebook for PAN at CLEF (2013)
8. D. Ramage, S. Dumais, and D. Liebling.: Characterizing microblogs with topic models. In International AAAI Conference on Weblogs and Social Media, (2010)
9. McCallum, Andrew Kachites.: MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu` (2002)
10. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. (2009)
11. H. Zhang, C. L. Giles, H. C. Foley, and J. Yen. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In AAAI'07: Proceedings of the 22nd National Conference on Artificial Intelligence, pages 663–668, (2007).