International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 12, December-2014

Author Profiling: Predicting Gender and Age from Blogs, Reviews & Social media

Satya Sri Yatam
Department of Computer Science & Engineering,
Swarnandhra Institute of Engineering & Technology,
Seetharampuram, Narsapur, W.G.Dt., A.P.

T Raghunadha Reddy,
Assoc. Prof. & Head,
Department of Computer Science & Engineering,
Swarnandhra Institute of Engineering & Technology,
Seetharampuram, Narsapur, W.G.Dt., A.P.

Abstract: **Author profiling aims to determine the gender, age, and mother language, level of education or socio-economic categories of authors by analyzing their published texts. In the recent times, several solutions are being proposed by different researchers focusing primarily on Age and Gender prediction of the authors. In this paper, we propose a Machine Learning approach to determine authors Age and Gender. Our approach uses two types of features: Content based and Style based. In content based features, we considered the words used by authors, in style based features we have considered parts of speech taggers of the words. We evaluated our system using PAN 2014 author profiling dataset on Blogs, Reviews and Social Media data.**

Keywords: Author Profiling, Classification, Age and Gender Prediction.

I. INTRODUCTION

Author profiling [3] [6] has received a growing importance due to the enormous impact of the social media on our daily life. Several applications like forensics, internet security, and commercial recommendation systems require information specific to the creator of the content. For example, author profiling can help police identify characteristics of the perpetrator of a crime when there are too few (or too many) specific suspects to consider. Similarly, in the online marketing settings, companies want to accurately predict recommendations that suits to the user interests. For this purpose, they analyze the user activity on social forums like blogs and online product reviews, to mine the demographic information of people.

In social media, we are mainly interested in everyday language and how it reflects basic social and personality processes. The increasing accessibility of public blogs, reviews and social media offers new ways to harvest information from texts authored by hundreds of thousands of different authors. In such scenarios, author profiling can be used to study the sociolect aspect, that is, how language is shared by people.

The aim of this work is to contribute to the topic of author profiling by experimenting with two features based and a popular machine learning classification algorithm, Support Vector Machine [8]. In this paper, we attempt to exploit these blogs, reviews and social media to find the correlation between the author‘s of various profiles to the language styles used by them. We believe that the ideas used in this work can help to analyze how everyday language reflects basic social and personality traits. In this work, we consider the popular profiling dimensions: Age and Gender.

The rest of the paper is organized in the following order: Section II introduces to the two features used in our approach. Section III describes our experimental settings and evaluation metrics. Section IV discusses our results on PAN 2014 dataset for the author profiling task, and section V concludes this work.

II. IDENTIFYING THE CHARACTERSTICS OF THE AUTHOR

In this section, we detail the features used in our investigation as well as the classification approach that we adopted. People of different ages write differently due to the variations in the topics of interest and experience gained over several years of practice which might change the writing styles like word choices and grammar rules. For example, females tend to write more about shopping, design and wedding events while males typically tend to write more about sports, finance, technology and politics. Further, studies [5] have shown that females use more adverbs and adjectives while writing compared to males. Therefore, it is good to use features that can differentiate between various writing styles and content of male and female bloggers of different ages. We considered two different types of features that are useful to distinguish between different categories of authors: Content based features and Style based features.

CONTENT BASED FEATURES

The content based features are important to distinguish between male and female writers [7]. For example, a blog related to cricket is more likely to be written by a male author rather than a female. A blog related to the sports, cricket, typically contains the words like cricket, ODI, test match, innings, six, BCCI, world cup, IPL etc. Thus the occurrence of words like cricket, world cup will increase the chances of it being written by a male rather than a female blogger. Similarly, the occurrence of words or phrases like my husband, flowers, pink, boyfriend etc will

increase the chances of it being written by female. This shows that the most frequent class of words used by male and female writers are widely different and can be exploited to accurately discriminate between them. Therefore, we calculate the frequencies of different N-grams (we used tri-grams, i.e., N=3) in the documents written by a particular gender. For every N-gram, we compute the ratio of its frequencies in the blogs written by male and female bloggers. We took the top k tri-grams that can differentiate males from females and females from males as features. In our work, we fixed k to 50000 for gender analysis.

Similarly, teenagers typically write more about their friends and mood swings, college life etc whereas middle age group people write more about marriage, jobs and politics. Thus content based features are important to distinguish between bloggers belonging to different age groups. Again, the words with most skewed ratios are used as features. In our work, we fixed k to 40000 for age analysis.

STYLE BASED FEATURES

Style based features includes N-grams of POS (Parts of Speech) tags in documents, punctuation symbols and number of href links [2] [7]. For each of these features, we calculated its frequency with which it appears in the corpus. We used their normalized count for creating numerical vector. We have converted every sentence of training and testing documents to its equivalent POS sentence using Stanford part of speech tagger. After the conversion, we have divide the text into N-grams (we used tri-grams, i.e., N=3) and identified the top k tri-grams to build the features for the age and gender prediction. In our work, we fixed k to 3000 for both age and gender analysis.

III. EXPERIMENTS

We used the data set provided by PAN 2014 [1] for our experiments. The corpus consisted of blogs, reviews and social media documents written in English. Each document is written by either male or female and belongs to one of the below mentioned five age groups: 10s:13-17, 20s: 23-27, 30s: 33-47, 50s: 50-64 and 60s: 65-xx. The data set statistics are described in more detail in Tables 1,2 & 3.

Table 1: Blogs Distribution over Age and Gender.

Blogs	10s	20s	30s	50s	60s
Male	3	30	27	12	2
Female	3	30	27	12	2

Table 2: Reviews Distribution over Age and Gender.

Reviews	10s	20s	30s	50s	60s
Male	180	500	500	500	400
Female	180	500	500	500	400

Table 3: Social Media Distribution over Age and Gender.

Social Media	10s	20s	30s	50s	60s
Male	775	1049	1123	919	7
Female	775	1049	1123	919	7

We have divided the entire data set into two parts. One for training the system and another for testing the system. We use 70% of the data set for training, 30% of the data set for testing.

We use the training data to find the most frequently used tri-grams and then, we have picked top k tri-grams. Using the selected tri-grams, we calculated the feature vector of every document. Likewise, we have built the feature vectors for training data set and testing data set. We have used these feature vectors from the training set to learn a Support Vector Machine (SVM) [7] model. We train the SVM model using content based and style based features separately using the training data set. We used evaluated the testing set on the learned SVM model and reported the performance of our approach. The Table 4 shows the machine learning method and features used to build classifiers for age and gender prediction tasks.

Table 4: Features and Machine Learning Algorithms used.

Feature	Feature Description	ML Library Used
Content Based Features	N-grams of plain words	SVM Light [3]
Style Based Features	N-grams of POS tags	SVM Light [3]

We have considered the precision as the scoring metrics to evaluate the effectiveness of our system. Precision in our context is the ratio of number of correct age or gender predictions to the total number of age/gender predictions.

$$\text{Precision} = \frac{\text{No. of correct gender/age predictions}}{\text{No. of total examples considered}}$$

IV. RESULTS AND DISCUSSION

Experiments have been carried out as 2 tasks, one is considering content based features and another is considering style based features. As described in section III, the systems are evaluated using the described metrics. In content based features, we took top k N-grams to build the feature vectors. We have experimented with different values of k. We got relatively better results for the value of 50000 for k. Details of the results for content based features are described in Table 5

Table 5: Results of author profiling using Content based features.

Corpus	Age	Gender
Blogs	37.78%	57.8%
Reviews	25.56%	60.26%
Social Media	33.91%	52.75%

In Style based features, we took top k N grams of POS tags to build the feature vectors, we have experimented with different values of k, and we got relatively better results for the value of 3000 for k. Details of the results for style based feature are described in Table 6.

Table 6: Results of author profiling using Style based features.

Corpus	Age	Gender
Blogs	31.11%	62.22%
Reviews	28.05%	58.17%
Social Media	39.33%	52.71%

V. CONCLUSION

In our work, we have shown that a combination of linguistic features and machine learning methods enables us to accurately predict the age and gender of an unknown author. We believe that other important author profile components like mother language, level of education can also be extracted using such techniques, given appropriate training data. Future efforts can be put into inducing sentiment analysis to discover more differences in text written by authors representing different classes. However, it is an important question, to see the extent to which variation in genre and language might affect the nature of the machine learning models that can be used to solve various aspects of the profiling problem.

ACKNOWLEDGMENTS

The authors wish to thank PAN 2014 organizers for providing document collection dataset for Training and Testing our system.

REFERENCES

1. PAN author profiling task (2014) http://www.uni weimar.de/medien/webis/research/events/ pan-14/pan14-web/author-identification.html
2. S.Argamon, M. Koppel, J.W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text Communications of the ACM, 52(2):119–123, 2009.
3. D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender in twitter: Styles, stances, and social networks. arXiv preprint arXiv:1210.4567, 2012.
4. A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.
5. J. Nerbonne. The secret life of pronouns. What our words say about us. Literary and Linguistic Computing, page fqt006, 2013.
6. C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pages 37–44. ACM , 2011.
7. J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, volume 6, pages 199–205, 2006.
8. B. Schölkopf, C. J. Burges, and A. J. Smola. Advances in kernel methods: support vector learning. MIT press, 1999.

BIOGRAPHY

Ms. Satya Sri Yatam completed her MCA from Andhra University in 2009. At present, She is pursuing M. Tech (CSE) in Swarnandhra Institute of Engg. & Tech, Narsapur. Her areas of interest are Data Mining and Software Engineering.

Mr. T. Raghunadha Reddy completed his B. Tech (CSE) from JNTUH in 2003 and M. Tech from JNTUH in 2005. Now he is pursuing Ph. D at JNTUH. At present, he is working as Associate Professor & Head in the Dept. of CSE at Swarnandhra Institute of Engg. & Tech., Narsapur. His areas of interest are Data Mining, Web Mining & Author Profiling.