Author Profiling: Predicting Age and Gender from Blogs
Notebook for PAN at CLEF 2013

K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma

International Institute of Information Technology, Hyderabad
{santosh.kosgi, romil.bansal, mihir.shekhar}@research.iiit.ac.in, vv@iiit.ac.in

Abstract Author profiling is the task of determining age, gender, native language or personality type of author by studying their sociolect aspect, that is, how language is shared by people. In this paper, we propose a Machine Learning approach to determine unknown author's age and gender. The approach uses three types of features: content based, style based and topic based. We were able to achieve an accuracy of 64.08%, 64.30% for age and 56.53%, 64.73% for gender in English and Spanish respectively.

Keywords: Author Profiling, Topic Modelling, Text Categorization, Natural Language Processing

1 Introduction

The problem of identifying the user's profile from the text is always of importance as it helps in various fields like forensics and marketing. For example, in marketing, a manager might want to find the gender and age group of people who like or dislike their products from the public reviews. The increasing accessibility of public blogs offers an unprecedented opportunity to harvest information from texts authored by hundreds of thousands of different authors. In this paper, we tried to exploit these public blogs to find the relations between the author's profile and the language style used by them. The main idea behind this task is to analyse how everyday languages reflects basic social and personality traits. The profiling dimensions we considered are age and gender.

2 Approach

2.1 The Corpus

We used the blog corpus provided by PAN 2013[1]. The corpus consisted of blogs written in both English and Spanish and each blog is written by either male or female and belongs to one of three age groups(10s: 13-17, 20s: 23-27 and 30s: 33-47). The corpus is described in more details in Table 1.

	EN			ES		
	10s	20s	30s	10s	20s	30s
Male	8,600	42,900	66,800	1,250	21,300	15,400
Female	8,600	42,900	66,800	1,250	21,300	15,400

Table 1. Blogs Distribution for English and Spanish Dataset

2.2 Features and Experiments

Different people tend to write differently. These differences occur due to variations in the topics of interest and style of writing like word choices and grammar rules. For example, females tend to write more about wedding styles and male tends to write more about technology and politics. Further females use more adverbs and adjectives while writing compared to males[8]. We considered these differences in the writing styles and content of male and female bloggers of different ages. Overall we considered three different types of features that are useful for distinguishing between different categories; they are: content based features, style based features and topic based features. These features are described in details below.

Content Based Features Male and female authors tend to speak about different topics, so they will use different words. Thus content based features are important to distinguish between male and female bloggers[9]. 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 cricket may contain words like cricket, no ball, wide, world cup, icc world cup etc. Thus the occurrence of words like world cup, cricket will increase the chances of it being written by male rather than female blogger and occurrence of words or phrases like my husband, pink, boyfriend will increase the chances of it being written by female. The words which are used more frequently by one of the classes when compared to other can be used as features. We calculated the frequencies of different N-grams in the documents written by a particular gender. Then, for every N-gram, we calculated the ratio of its frequencies in the blogs written by male and female bloggers. We took the top k N-grams (We used k as 50000 and 40000 for English and Spanish gender analysis respectively) that differentiate males from females and females from males as features.

Similarly, teenagers tend to write more about their friends and mood swings, whereas people of 20's write more about college life and people of 30's 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. We used k as 40000 for both English and Spanish age analysis.

Style Based Features Style based features includes N-grams of POS tags in documents, punctuation symbols and number of href links[2,9]. For each of these features we calculated its frequency with which it appears in the corpus. We used their normalised count for creating numerical vector. This was the only language dependent feature.

Topic Based Features N-gram based approach models the top words used by both males and females. But many times same words are used in different contexts. For example, males usually use words like '*daily life*' to describe their work and whereas females use '*daily life*' to describe their love or spiritual life. Males use '*dresses*' in context with pants and coats whereas females use '*dresses*' with words like bridal wears and gowns etc. Topic based features consider the fact that different categories of people have different topic of interests. We tried to model these differences to predict age and gender of the person. We ran LDA[1] algorithm to find topics from the blog and created a machine learning model based on the probability distribution of the blog over different topics and the class it is in.

For extracting the topic based features we divided the training data created in ratio 60%and 40%. The 40% of the data is used to train the MaxEnt model to predict the class based on the topic distribution of the blog. The rest 60% of the data we used for extracting relevant topics from the blogs. The topics were extracted as follows.

Overall Topics We gave the complete 60% of the data to generate topics from the blogs. The intuition was that the different category of people tends to write on completely different topics. So modelling the users based on the topics would tell us the class of the people the author belongs. Using this approach we achieved 52.3%(using 200 topics[2]) accuracy for gender classification. We analysed the topics of the blogs that are getting misclassified by method. We analysed that although few topics completely distinguish between males and females but most of the topics are written by both males and females. For example, the topic corresponding to 'dresses and shopping' was thought to be written by mostly females but males were also blogging about the topic. This causes the algorithm to find topics distribution vector that could distinguish between males and females completely. Similar case was observed with the different age groups.

Individual Topics Even if males and females write on the same topic, the words or context used by them to describe the topic is different. This could be seen from the above example as males are talking about pants and coats in the blogs for topic 'dressing and shopping' whereas females are talking about bridal wear and dresses in the similar topic. The method of 'Overall Topics' classified both in the same topic, thus making the topic noisy. So to improve the creation of topics, we trained the topics separately for individual classes and predicted the distribution over all of them. This helps us to model the context in which the topic was spoken about. Using this approach we obtained the overall accuracy of 54% for gender classification. We created 200 topics for each gender and 100 topics for each age group while creating the model.

Hybrid Topics The above method gave better results, but some of the overall topics are good enough to distinguish between different classes. So we created feature vector as probability distribution over both individual as well as overall topics. We took 200 topics from each gender and 100 topics from each age group along with 200 overall topics. Using this approach, we obtained the overall accuracy of 54.8% accuracy for gender classification.

[1] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[2] We experimented using different number of topics and found 200 topics to perform the best.

2.3 Learning Methods

We used the decision tree of classifiers to predict the class. We divided our corpus into three parts. We trained the ML algorithm using content based, style based and topic based features separately using the first part. We tested these models on the second part and the output is used to train the final decision tree classifier. The third part is used as a testing set. The table 2 shows Machine learning methods used to build classifiers.

Feature Name	Feature Description	ML Algorithm Used	ML Library Used
Content Based Features	Ngrams	SVM	SVM light[5]
Style Based Features	Ngrams of POS tags	SVM	SVM light
Topic Based Features	LDA Topic Model	MaxEnt	Mallet[6]
Merged Features	Scores of classes from different models	Decision Tree	Mallet

Table 2. Features used while training the models.

3 Conclusion and Future Work

A good system for author profiling is required in various domains ranging from analysing sensitive text for national security to commercially important data from various comments and product reviews. In our approach, we tried to model the author's profile using the writing style and content of the blog. We have shown that best results were acchieved when the context information is used along with the content and style of the blog.

Future efforts can be put into inducing sentiment analysis to discover more differences in text written by authors representing different classes. With further developments, we can expect much better accuracy rates in identifying the author's profile.

References

1. Pan author profiling task (2013), http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (Feb 2009), http://doi.acm.org/10.1145/1461928.1461959
3. Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17, 401–412 (2003)
4. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender in twitter: Styles, stances, and social networks. CoRR abs/1210.4567 (2012)
5. Joachims, T.: Advances in kernel methods. chap. Making large-scale support vector machine learning practical, pp. 169–184. MIT Press, Cambridge, MA, USA (1999), http://dl.acm.org/citation.cfm?id=299094.299104

6. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://www.cs.umass.edu/ mccallum/mallet
7. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents. pp. 37–44. SMUC '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2065023.2065035
8. Pennebaker, J.: The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury USA (2013), http://books.google.co.in/books?id=mJ4tLwEACAAJ
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (Mar 2006)