Author profiling using LDA and Maximum Entropy

Notebook for PAN at CLEF 2013

Aditya Pavan, Aditya Mogadala, Vasudeva Varma

Search and Information Extraction Lab,
International Institute of Information Technology , Hyderabad

aditya.pavanm@students.iiit.ac.in , aditya.m@research.iiit.ac.in , vv@iiit.ac.in

Abstract. This paper describes the traditional authorship attribution subtask of the PAN/CLEF 2013 workshop. In our attempt to classify the documents based on gender and age of an author, we have applied a traditional approach of topic modeling using Latent Dirichlet Allocation[LDA]. We used the content based features like topics and style based features like preposition-frequencies, which act as the efficient markers to demarcate the authorship attributes based on age and gender. We demonstrated tenfold cross validation and observed that our classification approach using Maxent and LDA gave an accuracy of 53.3% for English language and 52% for Spanish Language.

1 Introduction

Authorship Attribution or author profiling has been a standard problem addressed in the areas of Information Retrieval, Statistical Natural Language Processing and Machine Learning. With increase in the number of user blog-posts and micro-blogs in the massive internet domain, author profiling task serves as a pre-processing step to help augment the prospects in several areas of text processing like Opinion Mining, mood mining and Polarity extraction. Every user comment or blog post is directly or indirectly associated with several attributes of author like age, gender and other demographic features. Extracting these features on a given document is of paramount priority. As a part of PAN competition, we have applied a traditional approach for extracting features of a document and predict the gender and age of an author. We have considered the topics used by the authors in the article as standard features and built a topic model from the corpus using unsupervised learning techniques like [LDA] Latent Dirichlet Allocation [4]. From the generated topic model, we trained a discriminative model using Maxent classification to profile the documents based on gender and age of the author. The same discriminative model was used for inferring tenfold validation data set.

The paper is organized as follows. Section 2, provides a brief explanation on various features we have adopted to derive authorship attributes like age and gender. Section 3, explains our approach. Section 4 concludes our work.

2 Features

2.1 Explaining the features

Based on the variations in the expressions of authors, features used for author profiling can be categorized into two types: Content-based features and Style-based features [1]. In the earlier work, several markers like textual style, Vocabulary complexity, Orthographic errors and morphological mapping were used for capturing the authorship attributes. But, preponderance of evidence suggests that wide variety of features were captured by simple markers like function-words [2] and individual parts-of-speech. However, in this paper we focus on extracting age and gender of an author based on the topics used in the document and the distribution of the corresponding topics with in the corpus. Since characteristics of an author are directly dependent on the age [2] and gender [2,3] of the author, which in turn are contingent on the usage of the topics in the article, our work primarily is focused on building essential topic model that naturally subsumes simple markers like Noun-phrases in parts-of-speech and other complex markers.

In addition to content-based features like topics, we also considered style-based features like frequency of prepositions used by the author and the number of superlative adjectives used within a document.

2.2 Features for Age and gender

As mentioned earlier, topics play a significant role in predicting the age of an author. In our present work, we have observed that usage of the topics vary from one age group to the other. The corpus of author documents used in this task provide a substantial evidence that the articles of users ranking within the age groups of 10s (13-17) comprise of topics related to adolescence, school activities and immature crush. While users in an age group of 20s (23-27) write about their college life, favorite heroines/ heroes, Pre-marital affairs, etc. Whereas, users belonging to age group of 30s (33-47) post more about Corporate / Social activities, Post-marriage life, etc [2]. Similarly, male authors stress on topics related to sports, politics and technology whereas the female authors post on topics like beauty, shopping, kitty parties, etc. [3]

But we have observed from the data that although the topic-set used by an author abets in demarcating the age groups, there are considerable overlaps in the topics among the age groups and genders. In order to resolve these overlaps, we considered a topic distribution model rather than just a set of topics. We have used a generative model called Latent Dirichlet Allocation (LDA) [4] to get a probabilistic distribution

of the topics in the document. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Thus generating models using LDA has been an essential step in extraction of features in our experiment.

3 Approach

3.1 Processing of Corpus

We have used the corpus available in PAN website. Since the data was in the form of mark up, we generated a clean data by parsing the tags and eliminating the unnecessary duplications. In order to discriminate train and test data, we created ten-fold cross validation sets and within training sets we generated datasets for individual age groups and individual genders. Our working model is independent of the language. So for both the Spanish and English data sets, we have employed similar approach.

3.2 Calculating frequencies

Prior Works [2, 3] imply that male authors tend to use more prepositions in the articles or blog posts than the female authors. As a part of our style-based features we have generated the frequencies of prepositions of authors in each document and generated the tf-score. We have not considered the anomalies and other dialectic exceptions as it can lead to over fitting of the model. So we have used this generalized observation to demarcate the gender based authorship attributes.

3.3 Generating topic models

In order to implement the concept of topic modeling, we used a java-based package named Mallet [5]. Since the topic distribution disregards the usage of function words and stop words, we eliminate them from our individual data sets. We have also precluded the preprocessing steps like stemming and lemmatization on the datasets in order to retain the style based features of the authors. For example, an author posting an article on cricket would allude the term ‘bowling’ in the context of the game. If we run our preprocessing steps like lemmatization of stemming on this word, the result would be ‘bowl’, which can have multiple contexts to kitchenware or cricket. Though LDA takes care of these differences, in order to retain the author style and subsume the noise in the corpus, we precluded these steps.

The gender specific data sets and age specific data sets were subjected to topic modeling and we have generated five corresponding topic models. Each topic model was built with a distribution on 250 topics and 1000 iterations.

3.4 Classification using Maxent

Earlier, linear classifier like Winnow, which overcomes differences between the genres and dependencies between features or the generative model like Naïve Bayes, which considers bag of words were used by several teams for author profiling. But we chose to use a discriminative model like Maxent as it would suffice our goal of classifying the document based on gender as well as age groups. Since the input for the classification task is the distribution of topics, in order to improve the maximum likelihood during estimation, the maximum entropy was used. The model essentially eliminates the over fitting aspects as it can normalize the duplication and co-occurrences of same features. During classification, we merged the features like preposition frequencies with the topic vector and trained our Maxent Classifier. We imported the Maxent classifier provided by mallet and ran our experiments with default hyper parameters and nine-tenth of training portion.

4 Conclusion and Future work

In this task of author profiling, we have applied an unsupervised learning method to extract the distribution of topics. We used a topic size of 250 for 1000 iterations on the dataset. We used a Maxent classifier to classify the documents based on gender and age groups and observed that performance of these models are independent of the language.

In order to improve the performance of the system, one can use better stylometric features concomitant to the content-based features. Better markers like POS tagging, superlative adjective occurrence can be used to improve the performance of the gender specific profiling task.

5 References

1. S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009), Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2): 119–123.
2. J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006.
3. M.Koppel, S. Argamon and A. Shimoni (2003), Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17(4), November 2002, pp. 401-412.
4. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). "Latent Dirichlet allocation". In Lafferty, John. *Journal of Machine Learning Research* **3** (4–5): *pp.* 993–1022.
5. McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.