Automatic Author Profiling Based on Linguistic and Stylistic Features

Braja Gopal Patra1, Somnath Banerjee1, Dipankar Das2, Tanik Saikh1,   
and Sivaji Bandyopadhyay1

*1Department of Computer Science and Engineering, Jadavpur University, Kolkata – 700032, India*

*2Department of Computer Science and Engineering, NIT Meghalaya, India*

*{ brajagopal.cse* *, s.banerjee1980, dipankar.dipnil2005, tanik4u }@gmail.com,*

*sivaji\_cse\_ju@yahoo.com*

The rapid expansion of blog and electronic data in Web 2.0 is abounding and thus it is becoming important to identify the author’s profile also. The problems of automatic identification of author’s gender and age based on linguistic and stylistic pattern have been a subject of increasingly research interest in the recent years. The research methodologies are also helpful for several other applications like criminal detection, security and author detection etc.

Author profiling task of PAN2013 is to determine author’s age and gender for a given document. The training data set are in English and Spanish languages and with regard to age, posts are of three classes: 10s (13-17), 20s (23-27), and 30s (33-47). We have worked on English language only. The English corpus incorporates a total of 236,000 authors (files) containing 413,564 conversations and 180,809,187 words.

An overview of our approach is as follows: we have used linguistic and stylistic feature for identifying the author’s age and gender. Different word lists have been created for calculating the frequencies of each document. We have also created the stopword list, smiley list, positive and negative word lists etc. for creating the feature vector. A machine learning algorithm has been employed for classifying the author’s profile. The Decision tree of Weka tool has been used for the classification task.

The accuracy of the system has been calculated by the PAN organizers and we have achieved 56.83% and 28.95% for gender and age group classification.