Author Profiling: Predicting Age and Gender from Blogs

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*

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. 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. We propose a Machine Learning based approach to determine unknown author's age (10s, 20s or 30s) and gender (male or female). The approach uses three types of features: content based, style based and topic based. We analyzed the problem of determining age and gender independently. We used PAN 2013 author profiling corpus to analyze the same.

Male, female and authors of different age groups tend to speak about different topics, so they use different words. Content based features exploit this fact and try to model each class based on the N-grams of words used by them.

Style based features model the fact that female bloggers use more pronouns and adjectives words while male bloggers use more articles and prepositions. Style based features include N-grams of POS tags in documents, punctuation symbols and number of href links.

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. Similarly, 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, so same word can mean different to different class of people. We tried to model these differences to predict age and gender of the person. Topic based features used LDA algorithm to find topics from the blog, and created the probability distribution of the blog over different topics as a feature to predict the class of the blog.

Finally, we trained the decision tree from the output of above classifiers to predict the final class of the blog. We used standard machine learning libraries to implement the algorithm.

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 achieved when the context information is used along with the content and style of the blog. We obtained the accuracy of 64.08%, 64.30% for age and 56.53%, 64.73% for gender in English and Spanish respectively. The overall accuracy is 35.08% and 42.08% for English and Spanish respectively.