Ensemble-based classification for author profiling using various features

Michał Meina1, Karolina Brodzińska2, Bartosz Celmer1, Maja Czoków1,   
Martyna Patera1, Jakub Pezacki1, Mateusz Wilk1

*1 Faculty of Mathematics and Comptuer Science. Nicolaus Copernicus University, Toruń, Poland   
2 Faculty of Physics, Astronomy and Informatics. Nicolaus Copernicus University, Toruń, Poland  
{mich, bcelmer, maja, patera, kpezacki, mtszwlk}@mat.umk.pl 1  
karola@fizyka.umk.pl 2*

This paper provides a summary of our approach to author profiling  
task with perform the best in English portion of data. We have developed a solution that is based on Ensemble-Based Classification using more than 400 features describing individual conversations. The features are organized into seven categories: (1) Structural Features, (2) Latent Semantic Analysis – statistical approach for topics modelling, (3) part-of-speech ratio and part-of-speech sequences, (4) Text Difficulty and Readability Measures, (5) Language Modelling – n-grams, (6) dictionary based features and different types of errors, and (7) unsupervised clustering analysis.

Exploration of feature set shows that the most important features for the English portion of the data was related to structure of the conversation. On the basis of this observation, we prepared solution that roughly estimate Author Behaviour Profile (by unsupervised learning over structural and topical features) and then we learn how to utilize this information in conjunction with the rest of the feature set. Concerning the Spanish portion of the data, the most significant features was constructed by n-gram modelling over part-of-speech Sequences.

Furthermore, a crucial sub-problem for author profiling was spam detection. We addressed this problem only partially by using spam detection software and example-to-centroid distance.

We will describe and discuss experimental results, which include the  
selected approaches to Ensemble-Based Classification, namely:  
Classifier Committees and Random Forest. Our experiment shows that the  
best result was achieved by the Random Forest classification.  
Moreover, we will discuss exploration of importance of the features.