Two methodologies applied to the author profiling task

Yuridiana Aleman, Nahun Loya, Darnes Vilariño, David Pinto

*Benemérita Universidad Autónoma de Puebla, Faculty of Computer Science, Puebla, 72000, Mexico*

*{candy.aleman, nahun.loya, darnes, dpinto}@cs.buap.mx*

This paper describes two methodologies applied to the author profiling task submitted to the PAN 2013 competition of the CLEF 2013 conference. For the English corpus, we applied a methodology based in classical techniques of machine learning. The set of features were extracted in order to feed a Random Forest classifier. The methodology used for this corpus is twofold: pre-processing and classification.

In the pre-processing step, we attempt to normalize terminology by replacing unrecognizable terms, smiles, and weird symbols (e.g. URLs, pictures) from the dialogues by their corresponding normalized term. In order to apply this normalization procedure, we used three lexical resources that we have constructed for this purpose.

In the classification process we used the frequencies of the following sets of features with further were used for feeding a Random Forest classifier:

– Emoticons

– Contractions

– Conversation length (in words)

– Conversation length (in characters)

– Mispelled words

– Average length of words in the dialogues

– Words capitalized

– Words in uppercase

– URLs

– Each different POS tag

– Each different suffix

– Each different punctuation symbol

– Each stopword

The methodology proposed for the Spanish corpus focuses on the use of graphs as a strategy for feature extraction. Moreover, this methodology uses the extracted features with the purpose of feeding a supervised classification algorithm which allows to determine the gender and age of the authors. As carried out with the English corpus, we performed a pre-processing step in order to normalize the input texts. Afterwards, the texts are represented by means of graphs which will be further used for extracting relevant features. Graphs are mined using the SUBDUE tool. The obtained results of the mining graph phase are used as features in machine learning algorithms with the aim to obtain a classification model.

We succeed in the first approach obtaining the 7th place in the competition, but the second one was not able to capture regularities or patterns from the graphs. The results obtained by the first approach indicates that the features selected allow to discriminate gender and age of a given author with an Fmeasure of 0.59.We are interesting in evaluating the first methodology presented in the Spanish language.