Semantic-based Features for Author Profiling Identification: First insights

Delia-Irazu Hernandez1, Rafael Guzmán-Cabrera2, Antonio Reyes3 and Martha-Alicia Rocha4

*1Universidad Politécnica de Valencia, España   
2 Universidad de Guanajuato, México*

*3Instituto Superior de Intérpretes y Traductores, México   
4 Instituto Tecnológico de León, México*

*dirazuherfa@hotmail.com,* [*guzmanc@ugto.com*](mailto:guzmanc@ugto.com)*,* [*antonioreyes@isit.edu.mx*](mailto:antonioreyes@isit.edu.mx)*,* [*mrocha@dsic.upv.es*](mailto:mrocha@dsic.upv.es)

In this article we present an approach for identifying two main characteristics regarding the way in which Internet users interact: age and gender. The approach is grounded on detecting textual features considering different types of information: from to textual markers to semantics, they are: 1) *Signatures*, intend to identify textual markers that are used to throw focus onto certain aspects of a text. 2)*Chatslang*, a set of words that are often used by internet users as a subcode to communicate their messages more accurately. 3) *Context*, common elements across the different classes of the corpus, we employed a cluster algorithm to obtain a set of descriptive and discriminating words to represent each class. 4)*Emotionality*, is a feature to integrate information related to the communication of subjective matters through the selection of particular words, we used a dictionary to represent this type of information. 5) *Semantic similarity*, used to measure the semantic relatedness of the words. In addition to the features described, a list of most frequent words (BOW) in the corpus was used, finally the Jaccard similarity was applied over the text in order to focus on informative words. For our experiments we use a subset of conversations from the PAN 13 Training Corpus for Author Profiling Task and were considered all the six classes included (female 10s-30s and male 10s-30s). Each conversation is represented as a numerical vector. We make different combinations of the features proposed and we classified the conversations using various learning algorithms. As from the results obtained from each one of combinations, we defined a final model integrated for semantic similarity and emotionality measures + Jaccard Distance + BOW, the Naive Bayes classifier was applied to participate in the Author Profiling Task in PAN 2012 competition. The final results obtained in the competition reached a 0.2816 of accuracy in English and 0.1757 in Spanish.