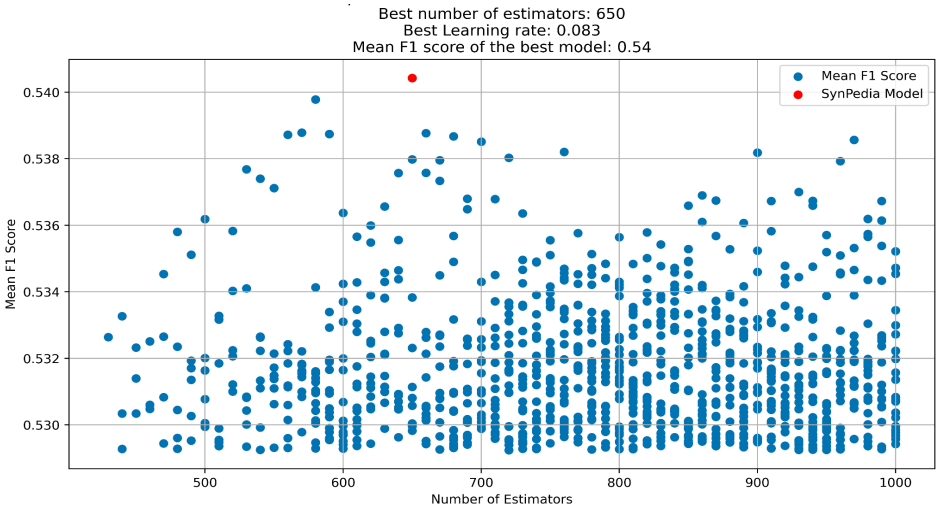
SynPedia Classifier: *Using Syntactic Structures and Named Entity Recognition for Effective Wikipedia Page Classification of Individuals*

Paschalis Agapitos [1], Luis A. Miccio [1,2], Juan Luis Suarez [3] and Gustavo A. Schwartz [1,4]

[1] Donostia International Physics Center, P. M. de Lardizábal 4, 20018 San Sebastián, Spain.
[2] Institute of Materials Science and Technology (INTEMA), National Research Council (CONICET), Colón 10850, 7600 Mar del Plata, Buenos Aires, Argentina.
[3] CulturePlex Lab, Western University, London, Ontario N6A 3K7, Canada.
[4] Centro de Física de Materiales (CSIC-UPV/EHU) - Material Physics Centre (MPC), P. M. de Lardizábal 5, 20018 San Sebastián, Spain.

Wikipedia serves as a vital repository of knowledge, accommodating diverse cultural analytics approaches. Unravelling the hidden connections and stories of individuals through their internal links in Wikipedia has garnered attention over the past few years and it is far from over. However, most of the studies about individuals are limited to the labels usually provided by Wikimedia or the information contained in the infobox. In this study, we leverage and recontextualise an idea from computational stylometry to develop the SynPedia classifier, a model based on the syntactic structure of Wikipedia pages. We utilise Spacy's Syntactic Dependency Parser (SDP) on the first sentence of each page, employing a simple frequency distribution as training and evaluation material. Our hypothesis posits that the frequency distribution of syntactic patterns in the first sentence differs between pages referring to individuals and non-individual entities (e.g., companies, organisations, concepts). This allows us to classify people with a precision score above 80% for the test set of our data. Additionally, we also experimented with other implicit structures, such as Named Entity Recognition (NER) labels. While SynPedia already demonstrates promising results, our ultimate goal is to develop a scalable model for classifying Wikipedia pages, regardless of whether they represent biographical content or not. Through this study, we showcase the effectiveness of the SynPedia classifier while remaining mindful of potential areas for refinement and future improvements. Our research contributes to the advancement of Wikipedia page classification and holds promise for broader applications in information retrieval and knowledge organisation.



The top 1000 models ranked based on their F1 score after conducting a grid search, testing 45,500 combinations of hyperparameters (due to class imbalance we chose to report the F1 score here)..