HWxPI: A Multimodal Spanish Corpus for Personality Identification

Gabriela Ramírez-de-la-Rosa, Esaú Villatoro-Tello, and Héctor Jiménez-Salazar

Language and Reasoning Research Group
Universidad Autónoma Metropolitana, Cuajimalpa, México
{gramirez, evillatoro, hjimenez}@correo.cua.uam.mx

Abstract. Historically, a large number of resources regarding a broad number of problems are available mostly in English. One of such problems is known as Personality Identification where based on a psychological model (e.g. The Big Five Model), the goal is to find subject's personality traits given, for instance, a text written by that subject. We present a corpus of handwritten essays for Personality Identification: HWxPI. Our corpus contains information of 836 undergraduate Mexican students. We provide two modalities of each handwritten text: the manually transcribed essay and the scanned image of such essay.

Keywords: Language resource · Personality identification · Handwritten recognition · Text classification

1 Introduction

There is a growing interest on studying subjects' personality, specially among the natural language processing (NLP) community. This is because through techniques developed by psychologists, identification of one's personality has been proved efficient for predicting thought patterns, emotions and behaviour [1].

In order to study this area, the NLP community needs to have resources, i.e. labelled corpora. While there is a large number of resources in English about a great number of problems, very few resources exists for Spanish, and even less for the Personality Identification task in Spanish.

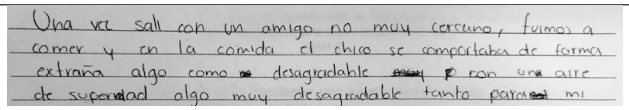
To tackle this problem we have been collecting, during 2 years, a corpus of handwritten short essays of undergraduates Mexican students. The personality information of each subject was obtained using a psychological instrument called TIPI (Ten Item Personality Inventory) [2]. Our corpus, called HWxPI (Handwritten text for Personality Identification), contains information from 836 subjects. Recently was used in the Multimedia Information Processing for Personality & Social Networks Analysis Challenge at ICPR (International Conference on Pattern Recognition)[1].

[1] http://chalearnlap.cvc.uab.es/challenge/27/description/

2 HWxPI corpus

The corpus consists of handwritten Spanish essays from undergraduate Mexican students.[2] For each handwritten essay we have two sources of information: the manual transcription and the scanned image of the handwritten essay. The corpus is available at `https://competitions.codalab.org/competitions/18362`. An example of these two modalities can be seen in Table 1.

Table 1. Example of a scanned image of a handwritten essay' fragment and its manual transcription with added tags.



Una vez sali <FO:salí> con un amigo no muy cercano, fuimos a comer y en la comida el chico se comportaba de forma extraña algo como <DL> desagradable <DL> <DL> con un <MD> aire de superioridad <MD> algo muy desagradable tanto para <DL> mi <FO:mí> ...

Ground truth. During the gathering process we asked each subject to answer a psychological instrument called TIPI to identify its personality according to the Big Five Model (i.e., Extroversion, Emotional stability, Agreeableness, Conscientiousness, and Openness to experience traits). The TIPI allows to divide each trait into four classes: high, medium high, medium low, and low. For HWxPI corpus we binarized the personality information of each trait, such as, high and medium-high classes are converted into 1 and low and medium-low are converted into 0.

Manual transcriptions and annotations. An important aspect of this corpus, beside its manual transcription, is a set of seven tags used to labelled handwriting phenomena: insertion of drawings or emojis <D:desc.>, insertions of a letter into a word <IN>, modification of a word <MD>, elimination of a word <DL>, two words written together <NS>, syllabification <SB> and misspelling <FO:word>. To the best of our knowledge there is no other corpus for personality identification with this kind of information. Preliminary analysis suggests that some tags might be positively correlated with a personality trait.

We keep working on gathering more subjects to participate on this research project. Therefore, eventually we can add more instances to our corpus.

References

1. Funder, D.C.: Personality. Annual Review of Psychology **52**(1), 197–221 (2001). https://doi.org/10.1146/annurev.psych.52.1.197
2. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality **37**(6), 504 – 528 (2003). https://doi.org/10.1016/S0092-6566(03)00046-1

[2] A subset of this corpus and the complete gathering methodology is described in [3].

3. Ramírez-de-la-Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H.: TxPI-u: A resource for personality identification of undergraduates. Journal of Intelligent & Fuzzy Systems **34**(5), 2991–3001 (2018). https://doi.org/10.3233/JIFS-169484