

PANDORA Talks: Personality and Demographics on Reddit

Matej Gjurković  Mladen Karan Iva Vukojević Mihaela Bošnjak Jan Šnajder

Text Analysis and Knowledge Engineering Lab

Faculty of Electrical Engineering and Computing, University of Zagreb

Unska 3, 10000 Zagreb, Croatia

name.surname@fer.hr

Abstract

Personality and demographics are important variables in social sciences and computational sociolinguistics. However, datasets with both personality and demographic labels are scarce. To address this, we present PANDORA, the first dataset of Reddit comments of 10k users partially labeled with three personality models and demographics (age, gender, and location), including 1.6k users labeled with the well-established Big 5 personality model. We showcase the usefulness of this dataset on three experiments, where we leverage the more readily available data from other personality models to predict the Big 5 traits, analyze gender classification biases arising from psychodemographic variables, and carry out a confirmatory and exploratory analysis based on psychological theories. Finally, we present benchmark prediction models for all personality and demographic variables.

1 Introduction

Personality and demographics describe differences between people at the individual and group level. This makes them important for much of social sciences research, where they may be used as either target or control variables. One field that can greatly benefit from textual datasets with personality and demographic data is computational sociolinguistics (Nguyen et al., 2016), which uses NLP methods to study language use in society.

Conversely, personality and demographic data can be useful in the development of NLP systems. Recent advances in machine learning have brought significant improvements in NLP systems' performance across many tasks, but these typically come at the cost of more complex and less interpretable models, often susceptible to biases (Chang et al., 2019). Biases are commonly caused by societal biases present in data, and eliminating them requires a thorough understanding of the data used to train

the model. One way to do this is to consider demographic and personality variables, as language use and interpretation is affected by both. Incorporating these variables into the design and analysis of NLP models can help interpret model's decisions, avoid societal biases, and control for confounders.

The demographic variables of age, gender, and location have been widely used in computational sociolinguistics (Bamman et al., 2014; Peersman et al., 2011; Eisenstein et al., 2010), while in NLP there is ample work on predicting these variables or using them in other NLP tasks. In contrast, advances in text-based personality research are lagging behind. This can be traced to the fact that (1) personality-labeled datasets are scarce and (2) personality labels are much harder to infer from text than demographic variables such as age and gender. In addition, the few existing datasets have serious limitations: a small number of authors or comments, comments of limited length, non-anonymity, or topic bias. While most of these limitations have been addressed by the recently published MBTI9k Reddit dataset (Gjurković and Šnajder, 2018), this dataset still has two deficiencies. Firstly, it uses the Myers-Briggs Type Indicator (MBTI) model (Myers et al., 1990), which – while popular among the general public and in business – is discredited by most personality psychologists (Barbuto Jr, 1997). The alternative is the well-known Five Factor Model (or Big 5) (McCrae and John, 1992), which, however, is less popular, and thus labels for it are harder to obtain. Another deficiency of MBTI9k is the lack of demographics, limiting model interpretability and use in sociolinguistics.

Our work seeks to address these problems by introducing a new dataset – *Personality AND Demographics Of Reddit Authors* (PANDORA) – the first dataset from Reddit labeled with personality and demographic data. PANDORA comprises over 17M comments written by more than 10k Reddit users, labeled with Big 5 and/or two other person-

