Datasets

|  |  |
| --- | --- |
| Name | Description |
| [**Personality Traits on Twitter(WASSA 2015:**](http://aclweb.org/anthology/W15-2913) | * 1.2M English tweets, 1500 users (type, text) * Myers-Briggs Type Indicators: MBTI * <https://bitbucket.org/bplank/wassa2015/downloads/> * uses four binary dimensions to classify users (INTROVERT–EXTROVERT, INTUITIVE–SENSING, THINKING–FEELING, JUDGING–PERCEIVING) * 16 different types * Although MBTI has some limited expressiveness and big5 has a preference, it is highly available in social media content. * I-E and T-F were predicted reliably while others were hard to predict * 100-2000 tweets/user * (19M) tokens, length of a tweet: 16.1 tokens |
| **Kaggle** | * 8600 instances/ (users?): type, text * MBTI * PersonalityCafe forum |
| **PAN 2015** | * Big 5 in range of (-5, +5) * Data from Twitter * Users (152 train), (42 val), (142 test) * Rangel F., Celli F., Rosso P., Potthast M., Stein B., Daelemans W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391 |
| [**Reddit: A Gold Mine for Personality Prediction**](http://www.aclweb.org/anthology/W18-1112) | * 9000 users * Myers-Briggs Type Indicators: MBTI * Reddit * <http://takelab.fer.hr/data/mbti> * Vocab size (3,774,588 ???) |

To read:

Penn projects on social media contents: <http://wwbp.org/publications.html>