Evaluation Metrics for Inferring Personality from Text

David N. Chin
University of Hawai'i at Mānoa
Dept. of Information and Computer Sciences
1680 East-West Road, POST 317
Honolulu, HI 96822 USA
chin@hawaii.edu

William R. Wright
University of Hawai'i at Mānoa
Dept. of Information and Computer Sciences
1680 East-West Road, POST 317
Honolulu, HI 96822 USA
wrightwr@hawaii.edu

1. INTRODUCTION

There have been a rich variety of studies of the relationship between speaker or author language usage and human personality. With each additional effort, the hope is that we will better identify text features consistently predictive of personality across domains, and identify the appropriate modeling techniques to convert those features into predictions about personality. However because each study uses different data from often very different domains, it is impossible to directly compare personality prediction algorithms.

1.1 Features

The community has examined a broad variety of text features: LIWC categories [5, 9], MRC (which places words in emotion, perception, cognition, and communication categories), POS n-grams [1], proper noun marking [8], word frequency (bag-of-words), word n-grams, and various hybrid features that combine these to form meaningful structures, not to speak of the variety of personality tests employed, from a basic 10-item questionnaire to those with numerous items, as well as observer reports of personality.

What remains is, (I) how to identify relevant features predictive of personality across the many different contexts, including different localities, time periods, and writing purposes and (II) how to evaluate predictive models built on some combination of such features. Progress on these two tasks will promote models of increasing utility to practitioners even when their subjects differ from the typical research study participant. A community corpora will be quite helpful to allow researchers to compare how well their choices of language features allow their algorithms to predict personality.

2. CORPORA

The ideal corpora for evaluating different techniques for inferring personality from text would include large amounts of text from many different contexts including different localities, time periods, and writing purposes (e.g., emails, text messages, blogs, essays, tweets, fiction, technical writing, etc.). These texts would be associated with personality profiles, preferably with scores from the prevailing Five Factor Model, which are useful for comparing individuals, but also when available with the Myers-Briggs Type Indicator, which is useful for other purposes. Text from multiple different contexts is important because text from a single context would likely have some coincidental correlations of personality influenced by current events that would not be found in text from other contexts. For example, Pennebaker's student essays [9] show strong correlations of the word "hurricane": positively with Conscientiousness and Agreeableness and inversely with Openness that are likely an artifact of the fact that the essays were written soon after hurricane Katrina had hit nearby and would likely not have similar correlations in other contexts. Scores should be standardized to eliminate units, or else our evaluation metrics will differ wildly between researchers depending on the personality test used.

2.1 Features of interest

Some studies focus on building a classifier, but not on identifying which features were useful for classification. They run the model building tool like a black box, but what is really interesting is what is inside. Announcing which language features are most predictive of personality for their dataset would be more interesting than, say, the classification accuracy they obtain.

For the good of the broader community, care should be taken to identify and announce the features f believed to be associated with personality, accompanied by the frequency mean m, Pearson's correlation coefficient ρ_x, the p-value p_x expressing h, the probability of the null hypothesis in the presence of the current feature. Of course given the large number of features that serve as candidates for personality prediction, and the oft witnessed sparsity of text data, some features that initially look promising will just turn out to be noise, regardless one's filtering method. This is just the right moment for comparison with prior research. By looking up or computing the aforementioned statistics from preexisting corpora, researchers can adjust h to account for prior appearances of the features f. Authors should address what to infer, if anything, from a feature's absence in any of the corpora.

This task is distinct from feature selection, which some modeling techniques require as a preprocessing step. The feature selection process often arbitrarily selects a single instance of a group of collinear features, discarding the rest. In that way relevant features are excluded from consideration by an arbitrary ordering imposed by a selection algorithm random: perhaps as a result of a randomizer seed, computer hardware differences or idiosyncrasies of implementation. Should researchers report the resulting feature list without explicit allowances made for these issues, some confusion may result as to why some features are present, and why others (perhaps present in other studies) are missing.

2.2 Predictive modeling

The corpora should be pre-divided into multiple training and test sets to make it easier to compare different classification or score prediction algorithms. The purpose of these sets is as follows:

- **Training set.** Feature selection, data for algorithms for training regression or classification models, checking their accuracy and tuning the algorithms, tuning parameters, and re-checking. How the training set might be further subdivided into feature selection, training and validation subsets is left to the discretion of individual researchers.
- **Test set.** For a last and final test of the model created from the training set. None of the activities mentioned above should take place after this event. Nothing from the test set should be used for training. Most importantly, no feature selection should be performed using the testing set.[1]

The corpora should be pre-divided into training and test sets to make it easier to compare different classification algorithms. The division should not be purely random, but should also take care that common text-analysis features are approximately evenly distributed across the two sets and the relative distribution of personality traits are also approximately evenly distributed across the two sets. That is the test set should be representative of the corpora in distribution of language features, personality, and any other demographic markers like gender, age, and location. Also, multiple such pairs of training and test sets should be prepared and ordered so that researchers without enough time or resources to repeat their analyses for all training/test partitions can compare results for the designated first partition with all other researchers. We would recommend 5 to 10 different divisions of training and test sets. There is also a question about the relative sizes of the training vs. test sets. With a large enough corpora, we believe a 75% training and 25% test size would be a good division.

Needless to say, the corpora should be anonymized to protect the privacy of the writers. It may be useful to include gender, time period and broad geographical location tags with the corpora. Further steps to protect against de-anonymization attacks might include replacing all names with generic markers like NAME1, PLACE2, etc. via named entity recognition.

3. METRICS

For personality prediction, some applications will require classifying users as high/low (above/below the mean) in the five personality dimensions. For other applications, it may be more useful to predict 3 classes, high (one standard deviation above mean), low (one standard deviation below mean) and medium (between high and low). Finally, regression models are useful for applications that require more fine-grained prediction of personality values.

Binary and 3-class classification algorithms should report percentage accuracy for each personality trait. Also for each class within each personality trait, report precision and recall rates. When the researchers are able to analyze multiple training/test partitions, the average and standard deviation for all partitions should be reported in addition to the individual partition results. Regression models should report both root mean squared error (RMSE) and mean absolute error (MAE) since MAE is often less sensitive to infrequent occurrences of very large errors than RMSE.

For binary and 3-class classification algorithms, [11] recommend testing the significance of improvements in classification accuracy with the Binomial test. They argue that a t-test is simply the wrong test for comparing classifiers because the t-test assumes that the test sets for each "treatment" (each algorithm) are independent and when two algorithms are compared on the same data set, the test sets are obviously not independent. Instead they recommend using the Binomial test to compare the number of examples that algorithm A got right and algorithm B got wrong versus the number of examples that algorithm A got wrong and algorithm B got right, ignoring examples that both got right or both got wrong. To apply the Binomial test, researchers should report in an online form which entries in each test set were correct/incorrect to allow for proper significance comparisons with future/past classification algorithms.

An alternative approach to statistical significance of classification algorithm differences is given by [2]. They recommend averaging the t-values from a paired Student t-test of each training/test partition and converting this to a significance value. This allows discounting of the effects of a single partition versus multiple partitions. To allow comparisons with past/future classification algorithms, researchers should report t-values for each training and test partition of the corpus.

For regression models, [3] recommend pairwise comparisons of RMSE using a test proposed by [6]. To allow comparisons with past/future regression models, researchers should report RMSE for each entry in their test set(s).

4. CONCLUSION

An established corpora with detailed reporting requirements will allow researchers to much more easily compare their algorithms for inferring personality from text. However, there will always be a need to extend the corpora to increase the coverage of different types of writing, time periods, and localities. If the shared corpus is viewed merely as a benchmark set, we risk overfitting the benchmark. Therefore we recommend a series of corpora, perhaps one every few years to keep adding new data to the community.

5. REFERENCES

[1] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

[2] Jeffrey P Bradford and Carla E Brodley. The effect of instance-space partition on significance. *Machine Learning*, 42(3):269–286, 2001.

[3] A. Feelders and W. Verkooijen. On the statistical comparison of inductive learning methods. In *In D. Fisher and H.-J. Lenz (Eds.), Learning from Data:*

[1]Those who disregard this and choose to perform automated feature selection and train classifiers on the same observations should consider that massive overfitting will likely occur. A classifier trained on the "best" 300 features drawn from 600,000 may be overfitting most of those features, undermining external validity.

Artificial and Intelligence V, pages 271–279. Springer-Verlag, 1996.

[4] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[5] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE, 2011.

[6] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[7] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[8] J. Oberlander and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.

[9] J.W. Pennebaker and L.A. King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[10] A. Roshchina, J. Cardiff, and P. Rosso. User profile construction in the twin personality-based recommender system. *Sentiment Analysis where AI meets Psychology (SAAIP)*, page 73, 2011.

[11] Steven L. Salzberg and Usama Fayyad. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, pages 317–328, 1997.

[12] Robert E Schapire. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406, 1999.