The Linguistic Status of Predictions and Feature Ranks from SVM Text Classifiers

Jonathan Dunn and Shlomo Argamon[1]

LSA Annual Meeting, Minneapolis, January 2–5, 2014

1. Introduction. Text classification systems are capable of predicting certain characteristics of a text's author using only linguistic properties. Such characteristics include, for example, the author's gender (Mukherjee & Liu, 2010; Sarawgi, et al., 2011) and age (Nguyen, et al., 2011; Rosenthal & McKeown, 2011). This paper asks why such predictions are possible and how they can be interpreted. There are three important factors: (1) the nature of the features used by the system; (2) the robustness of the predictions across time and genres; (3) the amount of data required for training and testing the system (e.g., the danger of over-fitting to a particular dataset).

The linguistic status of the predictions and feature ranks from a particular implementation depend upon these three factors. This paper provides an empirical case-study of how these factors interact. There are three possibilities: (1) the predictions may be based on over-fitting a particular dataset and thus spurious; (2) the predictions may be based on topic-dependent information (e.g., content) and thus not linguistic in nature, even though derived from language; (3) the predictions may generalize across topics and content and thus reflect strictly linguistic patterns.

This issue is important because machine learning algorithms, such as SVMs, together with the very large datasets now available, allow the investigation of linguistic patterns (e.g., the relation between thousands of variables and speaker characteristics) which are not visible to individual analysts or to previous statistical methods. Thus, this methodology has the potential to improve the study of linguistic variations with large-scale quantitative evidence at the morphological and syntactic level. However, given the difficulty of interpreting individual predictions (i.e., error analysis is impossible because it is never intuitively clear why a given prediction is made for a given instance), these methods first require theoretical justification.

2. Datasets and Features. As a case study, this paper implements an SVM classifier using two types of features: first, word-based features (word-forms and lemmas); second, grammatical features (part of speech). Both features are implemented with contextual information (varying n-gram windows) and calculated using their Relative Frequency. A third feature set combines both the word-form and part of speech information. This classifier is used on political text consisting of speeches from the U.S. Congress contained in the *Congressional Quarterly* from 1995 to 2007, a total of 500,000 words. To avoid over-fitting, the system is trained on the 108th and 109th congresses and tested on the 105th congress. Genre-dependence is tested using the distinction between speeches from the House and those from the Senate. Classification is conducted along the following variables: Sex, Age, Race, Previous Military Service, Political Party, and Ideology (operationalized both as scalar ratings by special interest groups and as scalar ratings derived from voting patterns; Poole & Rosenthal, 2007). Further tests are conducted using a corpus of blog posts, representing a more informal genre than congressional speeches (Schler, et al., 2006).

[1] Authors: Jonathan Dunn, Illinois Institute of Technology (jonathan.edwin.dunn@gmail.com) and Shlomo Argamon, Illinois Institute of Technology (argamon@iit.edu).

Descriptive statistics for these two datasets are given in Table 1. The breakdown of textual features for the *Congressional Record* dataset is given in Table 2.

Table 1. Datasets

Dataset	***Years***	***Texts***	***Words***	***Meta-Data***
Congressional Record	1995-2006	500k	200 mil.	Gender, Age, Geo., Party, Mil. Serv.
Blog Posts	2004	680k	140 mil.	Gender

Table 2. Feature Extraction for Congressional Record Corpus

N-Gram	***Word-Form***	***Part-of-Speech***
Unigram	25.18%	0.52%
Bigram	24.47%	7.88%
Trigram	7.00%	34.90%
Total	**56.59%**	**43.33%**

The first issue for evaluating text classifiers beyond pure performance (as traditionally measured by accuracy or F-Measure) is to determine how topic-dependent the predictive power of a particular model is. In other words, what features or linguistic information enables the classifier to predict author attributes? We start with a linear SVM because it produces coefficients for each feature (the same method can be used with Naïve Bayes or Logistic Regression classifiers, although the coefficients have different interpretations). Coefficients range from 1 to -1, with 0 representing little predictive power and 1 and -1 representing high predictive power for one of two class labels. We take the absolute value of the coefficient as representing its predictive power abstracted away from a particular class value. A simple measure for determining the topic dependence of a given model is given in Table 3, with the results for the *Congressional Record* corpus shown in Table 4.

Table 3. Measure of Topic Dependence

$\sum \mid F_{d1} \ldots F_{d2} \ldots \mid / (\sum \mid F_{d1} \ldots F_{d2} \ldots \mid + (\sum \mid F_{i1} \ldots F_{i2} \ldots \mid)$
F_d = Topic Dependent Feature, Coefficient
F_i = Topic Independent Feature, Coefficient

Table 4. Topic Independence By Congressional Record Corpus Type

Profile	***Individual***	***Aggregated***
Age	45.41%	41.19%
Military Service	46.32%	43.57%
Party Membership	38.42%	38.22%

A further issues is that models which share similar performance, as measured by F-Measure, may have different predictive features. Cross-validation techniques tend to cover up this situation in that each of the folds shares most of the same training data. We divide the blog dataset in half and train / test twice, with no overlapping training data, with the results in Table 5.

Table 5. Stability of Predictive Power of Features, Blog Corpus

Train on Set 1, Test on Set 2	0.764 F-Measure
Train on Set 2, Test on Set 1	0.760 F-Measure
Pearson Correlation Between Coefficients	0.303

A final problem is confounds between social and political characteristics. In other words, we need to determine which author attributes are actually being predicted. If we train models for multiple attributes using the same features on the same dataset we can test the correlation between the predictive power of features across all attributes and look for these confounds. This analysis is performed in Table 6 for the *Congressional Record* dataset.

Table 6. Pearson Correlations Between Feature Weights for Classification, All Above 0.2

	Age	***Chmbr***	***Mil.***	***Party***	***Race***	***Sex***	***SIG 1***	***SIG 2***	***M-W***	***N-S***
Age	---	0.206	0.350	0.210			0.232	0.211	0.253	0.207
Chamber	0.206	---		0.210			0.246	0.216		
Military	0.350		---	0.261		0.215	0.266	0.262	0.221	0.214
Party	0.210	0.210	0.261	---	0.276		0.628	0.971	0.233	0.263
Race				0.276	----		0.250	0.279		
Sex			0.215			----				
SIG 1	0.232	0.246	0.266	0.628	0.250		---	0.650	0.254	0.416
SIG 2	0.211	0.216	0.262	0.971	0.279		0.650	---	0.235	0.279
Midwest-West	0.253		0.221	0.233			0.254	0.235	---	0.221
North-South	0.207		0.214	0.263			0.416	0.279	0.221	---

Finally, we now evaluate the performance of the models using traditional methods (e.g., the F-Measure, based on proportions between true and false positives and negatives). The results are shown in Figure 1 and Figure 2 for the *Congressional Record* dataset. This measure of performance is important, but needs to be supplemented by information about (1) topic dependence, (2) feature stability, and (3) attribute confounds. The above measures are simple ways to determine these properties. The idea is that well-performing models, using F-Measure, also need to be well-performing on these other measures before interpreted linguistically.

Figure 1. Performance of Social Attributes by F-Measure, Congressional Record

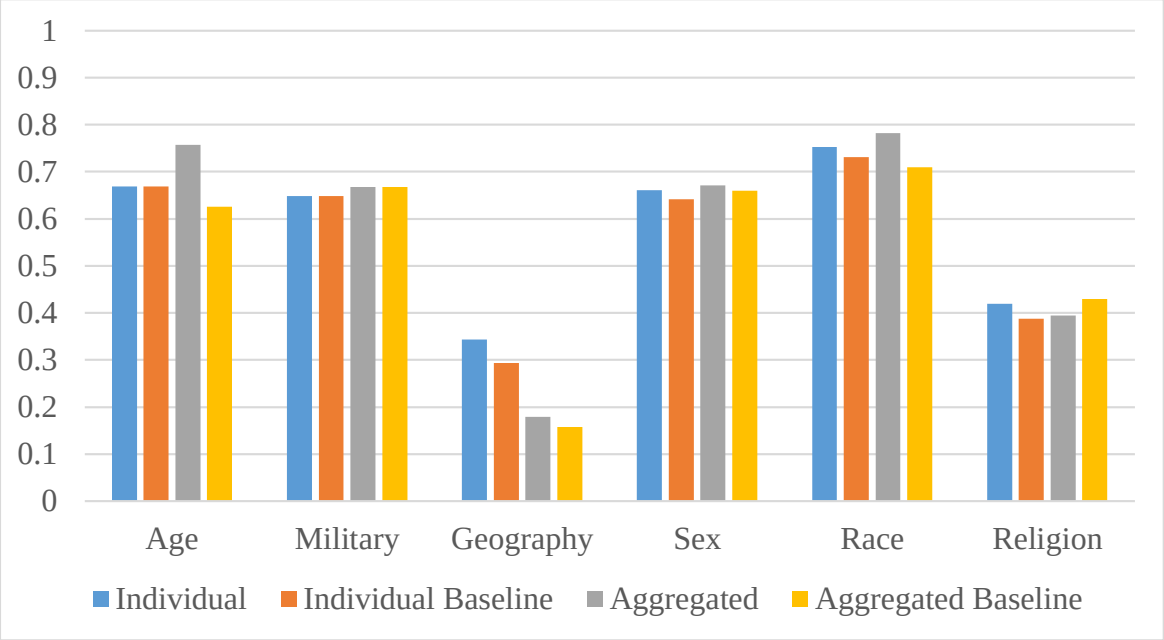
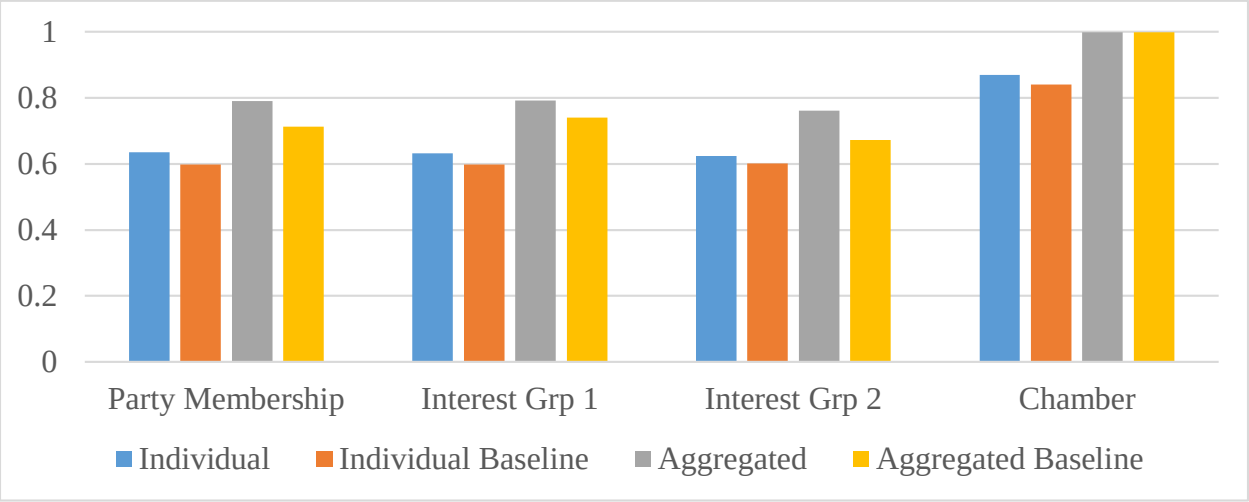


Figure 2. Performance of Political Attributes by F-Measure, Congressional Record



This case-study shows that some classification predictions, such as gender, are based on non-content linguistic material that generalizes across time, genre, and topics. These classifications are characterized by stable performance and feature ranks, and permit linguistic interpretation. Others, such as ideology, are content-based and topic-dependent, and do not permit linguistic interpretation.

Works Cited

Mukherjee, A., & Liu, B. (2010). "Improving Gender Classification of Blog Authors." In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 207–217. Stroudsburg, PA: Association for Computational Linguistics.

Nguyen, D., Smith, N. A., & Ros, C. P. (2011). "Author Age Prediction from Text using Linear Regression." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 115–123. Stroudsburg, PA: Association for Computational Linguistics.

Poole, K., & Rosenthal, H. (2007). *Ideology and Congress*. Edison, NJ: Transaction Publishers.

Rosenthal, S., & Mckeown, K. (2011). "Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 763–772. Stroudsburg, PA: Association for Computational Linguistics.

Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011). "Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre." In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 78–86. Stroudsburg, PA: Association for Computational Linguistics.

Schler, J., Koppel, M., Argamon, S, & Pennebaker J. (2006). "Effects of age and gender on blogging." In *Proceedings of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.