Experiments on Individual datasets

This document presents the results of predicting the Big5 traits for each of the datasets separately. All the features are used for this set of experiments. Each of these experiments are conducted using a randomly selected training and test set in one setting and using 5 fold cross validation for the other.

1 Facebook Data

Figure 1 shows the data skew for Facebook data. Extraversion, Neuroticism and Agreeableness are highly skewed, while the data skew is considerably less for Conscientiousness and Openness. Table 1 shows the results of using all the features for a randomly selected training and test set for Facebook data. For most of the traits, SVM works better than Naive Bayes. In order to get an unbiased evaluation, a 5-fold cross validation has been performed. The results indicate that for Extraversion, Neuroticism and Agreeableness, SVM works better (or at par in case of Neuroticism) than Naive Bayes. Since there is a data skew, the majority classifier tends to work better or at par with the other classifiers in case of Extraversion, Neuroticism and Agreeableness.

Table 3 the effect of emotion features (with a 5-fold cross validation) on predicting personality from Facebook data. It can be seen that the emotion features does a better job predicting personalities than using the entire feature set, especially for Agreeableness, Conscientiousness and Openness. This is because all the features may not be equally informative when it comes to predicting personalities.

2 Essay Data

Figure 1 shows the data skew for Essays data. Agreeableness and Extraversion show the most data skew.

Table 4 shows the results of 5 fold cross validation on the Essays dataset. The results for all traits except for openness are at par with the majority classifier.

Table 5 shows the results of experiments using all the features for essays data. Overall, random forest performs best for 3 labels out of 5. This can be attributed to the size of the essays dataset (2417 datapoints while Facebook and Twitter have 250 and 152 datapoints respectively).

3 Twitter Data

Figure 3 shows the data skew for Twitter data. The PAN dataset contains scores (-0.5 to +0.5) instead of “yes/no" labels. For this project, anything greater than 0 has been labeled as “yes", while any score below 0 has been labeled as “no". However, this has yielded an extremely imbalanced dataset. Openness has the worst imbalance among the five traits. Table 6 shows the result of 5-fold cross validation on Twitter dataset. Due to the class imbalance for Twitter data, the majority classifier yields close results to that of any classifier for any particular personality trait. The cross validation has not worked for openness as the number of members for “no" is only 3.

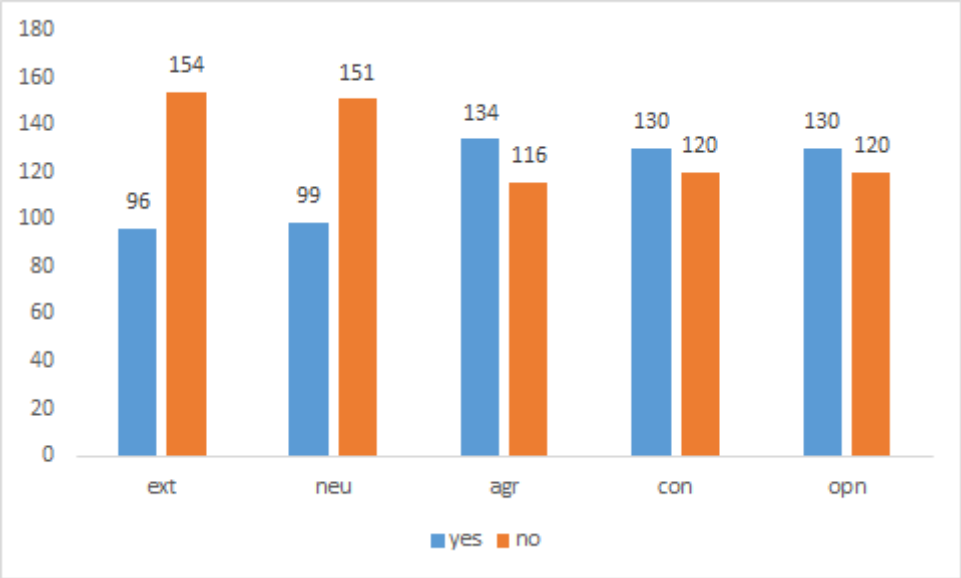

Figure 1: Data skew for Facebook status

	Majority	SVM (Linear)	SVM (RBF)	Naive Bayes
Extraversion	0.66	**0.68**	0.66	0.56
Neuroticism	**0.72**	**0.72**	**0.72**	0.5
Agreeableness	**0.52**	0.44	**0.52**	0.44
Conscientiousness	**0.5**	0.48	**0.5**	0.54
Openness	0.5	0.48	0.5	**0.54**

Table 1: Experiments on Facebook dataset with randomly chosen train and test sets

	Majority	SVM (Linear)	SVM (RBF)	Naive Bayes	Random Forest
Extraversion	**0.62**	0.55	**0.62**	0.57	0.58
Neuroticism	**0.6**	0.5	**0.6**	**0.6**	0.49
Agreeableness	**0.54**	0.52	**0.54**	0.51	0.52
Conscientiousness	0.52	0.49	0.52	**0.53**	0.48
Openness	0.52	0.49	0.52	**0.53**	**0.53**

Table 2: Experiments on Facebook dataset using 5-fold cross validation for an 80/20 split

	Majority	SVM (Linear)	SVM (RBF)	Naive Bayes	Random Forest
Extraversion	**0.62**	0.61	0.61	0.6	**0.62**
Neuroticism	**0.6**	0.58	**0.6**	0.54	0.55
Agreeableness	0.54	0.52	**0.6**	0.47	0.57
Conscientiousness	0.52	0.52	0.44	**0.56**	0.49
Openness	0.52	0.52	0.44	**0.56**	0.49

Table 3: Experiments on Facebook data using emotion features only

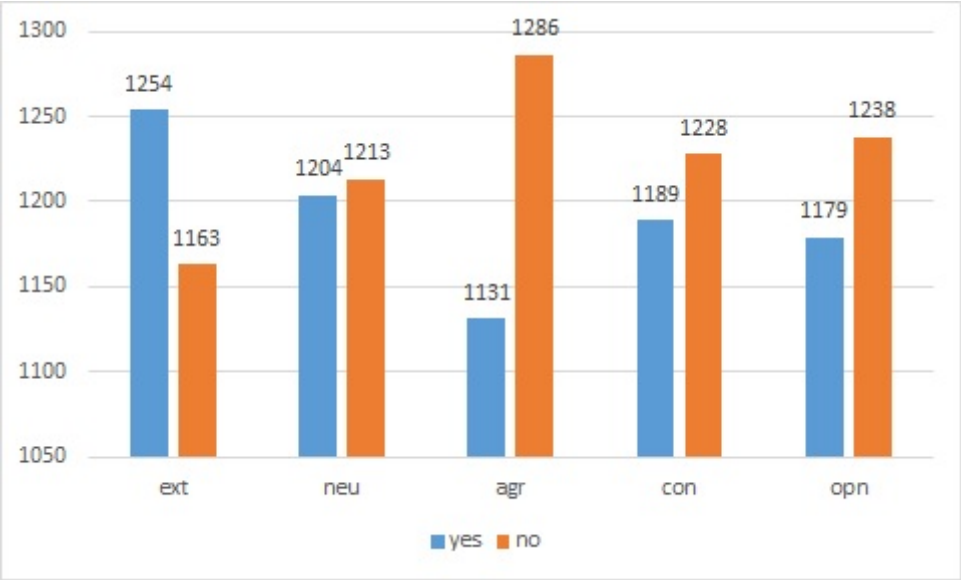

Figure 2: Data skew for Essay data

	Majority	SVM (Linear)	SVM (RBF)	Naive Bayes	Random Forest
Extraversion	**0.52**	0.51	**0.52**	**0.52**	0.49
Neuroticism	**0.50**	**0.50**	**0.50**	**0.50**	**0.50**
Agreeableness	**0.53**	0.50	**0.53**	0.49	0.49
Conscientiousness	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**
Openness	0.51	**0.53**	0.51	0.51	0.52

Table 4: Experiments on Essays dataset using 5-fold cross validation for an 80/20 split

	Majority	SVM (RBF)	Multinomial Naive Bayes	Random Forest	SVM (Linear)	Gaussian Naive Bayes
Extraversion	0.529	0.529	0.527	**0.556**	0.517	0.469
Neuroticism	0.498	0.498	0.457	0.490	**0.525**	0.500
Agreeableness	0.514	0.514	0.492	0.517	**0.523**	0.502
Conscientiousness	0.496	0.496	0.461	**0.541**	0.498	0.500
Openness	0.504	0.504	0.457	**0.525**	0.488	0.479

Table 5: Experiments on Essays dataset with random train and test

	Majority	SVM (Linear)	SVM (RBF)	Naive Bayes	Random Forest
Extraversion	0.79	0.67	0.79	0.49	**0.80**
Neuroticism	**0.69**	0.60	**0.69**	0.53	0.64
Agreeableness	0.75	0.74	0.75	0.52	**0.78**
Conscientiousness	**0.77**	0.70	**0.77**	0.55	0.74
Openness	**0.98**	0.95	**0.98**	0.56	**0.98**

Table 6: Experiments on Twitter dataset using 5-fold cross validation for an 80/20 split

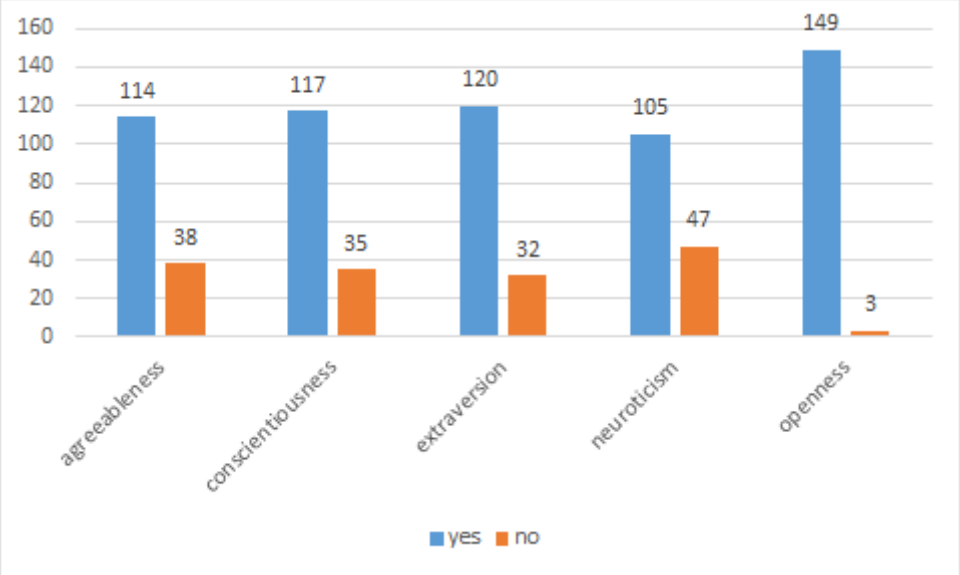

Figure 3: Data skew for Twitter data

References

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 627–634, Stroudsburg, PA, USA. Association for Computational Linguistics.