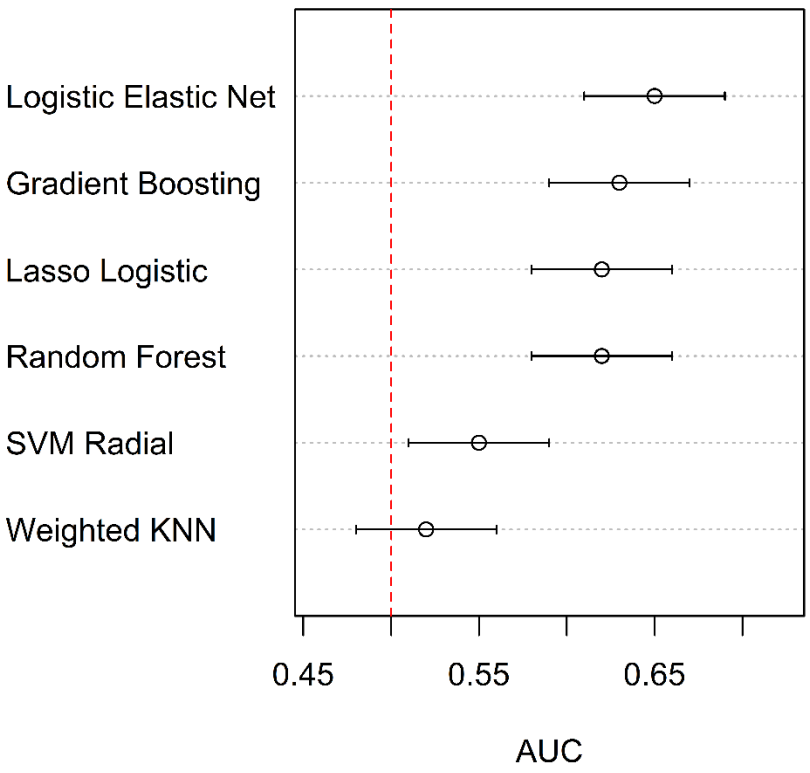
Supplemental Table 1. The list of hyperparameters for each of the 6 trained classifiers.

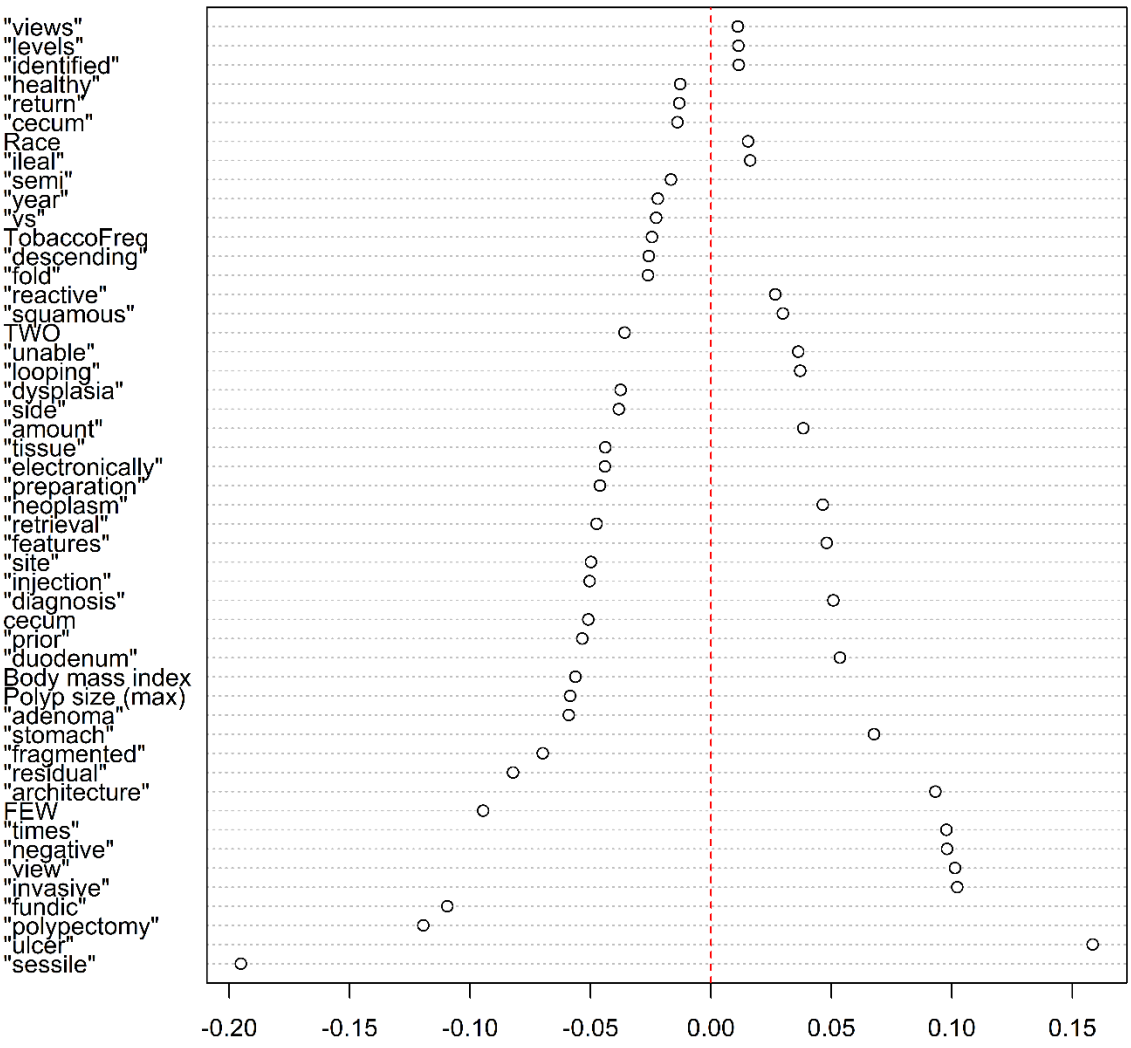
Model	**Hyperparameters and their optimal values in our training**
Random Forest	mtry = 31
SVM Radial	sigma = 0.0026; C = 4
Weighted KNN	kmax = 13; distance = 2; kernel = optimal
Logistic Elastic Net	alpha = 0.55; lambda = 0.0314
Gradient Boosting	nrounds = 50; max_depth = 1; eta = 0.4; gamma = 0; colsample_bytree = 0.6; min_child_weight = 1; subsample = 0.75
Logistic	cost = 0.25; loss = L1; epsilon = 1

Supplemental Table 2. Comparison of models with and without addition of unigrams. For all models, addition of unigrams to the model either improved the model's AUC score or left it without significant change. Note that all models experienced improvement with unigrams with the exception of weighted KNN (whose performance is poor regardless). Confidence intervals for unigram-containing models were calculated with pROC with 2000 bootstrap iterations.

Model	**AUC with no Unigram Feature**	**AUC and 95% Confidence Interval with Unigram Features**	**Comparison**
Logistic	0.63	0.65 ± 0.04	Within range
Logistic Elastic Net	0.62	0.63 ± 0.04	Within range
Gradient Boosting	0.61	0.62 ± 0.04	Within range
Random Forest	0.55	0.62 ± 0.04	Better
SVM Radial	0.53	0.55 ± 0.04	Within range
Weighted KNN	0.53	0.52 ± 0.04	Within range



Supplemental Figure 1. A comparison between the six trained classifiers by their areas under the curve (AUC). Error bars represent the 2000-iteration 95% bootstrap confidence intervals on the ROC curve computed with pROC. The red dotted line indicates an AUC value of 0.5, which is equivalent to a model with random predictions. Because the confidence interval for all models except weighted KNN fall to the right of the red line, these models performed better than the random chance.



Supplemental Figure 2. Top 50 coefficients selected by the logistic elastic net (glmnet) model, ranked by their absolute values. The red dotted line corresponds to the coefficient at 0 (no effect on the outcome variable); the further a feature is from the red line, the more weight that feature is ascribed in this model and hence it is more important as a predictor.