Our original approach was a simple bag of words, with no document preprocessing. Our current leading approach implements several different strategies. The document is first preformatted by replacing all numbers with a placeholder, and then it is tokenized, treating punctuation and single-letter words as valid tokens. We use a modified English stop words list that doesn't treat 'I' as a stop word. We reduce the feature space by applying chi-squared Univariate Feature Selection. For the SVM we apply L1 regularization with a penalty parameter of 0.5 to reduce overfitting. Note that the generalization error improved for every label, even though the feature space has been significantly reduced.

Model	Label	Train: F1	Test: F1
SVM	knowledgeable	0.986599	0.389445
SVM	neutral	0.996483	0.864509
SVM	confused	0.993741	0.35207
Naive Bayes	knowledgeable	0.737699	0.441833
Naive Bayes	neutral	0.928654	0.888782
Naive Bayes	confused	0.711252	0.367184
Logistic Reg	knowledgeable	0.91259	0.420034
Logistic Reg	neutral	0.974787	0.882031
Logistic Reg	confused	0.911703	0.378374

Table 1: Original Bag-of-Words Approach

Model	Label	Train: F1	Test: F1
SVM	knowledgeable	0.672791	0.396975
SVM	neutral	0.922132	0.867997
SVM	confused	0.701643	0.35299
Naive Bayes	knowledgeable	0.653983	0.487236
Naive Bayes	neutral	0.908834	0.879653
Naive Bayes	confused	0.648954	0.466741
Logistic Reg	knowledgeable	0.664368	0.421321
Logistic Reg	neutral	0.918744	0.873467
Logistic Reg	confused	0.675589	0.382527

Table 2: Full Approach