Text Analytics Assignment 2 Classification

This assignment will give you hands-on experience in building text classification models, using the application of email spam filtering. You will use Weka to convert the textual data (emails) into feature vectors and build text mining models for automatic spam filtering. The email messages you will use were delivered to a particular server between 8 Apr 2007 and 6 Jul 2007. The target variable represents whether an email is either spam or ham (non-spam). Follow the directions and answer any questions. Your report should not be verbose, but should present your results clearly and professionally.

. On the blackboard, you will find a file called *spam_data_Text.arff*, which contains all the emails in our dataset and is in a format ready-to-input by Weka. Follow the instructions in "Workshop_TextClassification_in_Weka.docx" and answer the following 5 questions. (Make sure you follow the instruction in the workshop)

Q1. ***Report the evaluation results of your model using 10-fold cross-validation***. Do not include the entire output. Only copy the last three sections in the result window: "Summary", "Detailed Accuracy By Class" and "Confusion Matrix." Think about how all these evaluation measures were calculated.

Q2. ***Report the 10-fold cross-validation results and compare with the occurrence-based results in the previous question.***

Q3. Calculate the total cost and expected cost (per email) based on the confusion matrix you obtained in question 1. ***Copy the confusion matrix and present the formulas you used to get the results.*** [Be careful with the dimensions of the confusion matrix: which are the "actuals" and which are the "predictions"?]

Q4. ***Calculate the total cost and expected cost. Compare the accuracy and the costs with those of the cost "insensitive" model you built earlier.***

Q5. ***Compare the results of evaluating this new model with the one you generated in question 1 using the full set of features. What is your observation?***

Extra credit:

Explore different ways to improve the classification performance (accuracy or expected cost). You can consider the following:

- *Feature representation: binary vs. frequency vs. tf-idf*
- *Feature selection: different feature/attribute selection methods or parameters (e.g., numToSelect)*
- *Classifier: compare classifiers such as decision trees, neural nets, etc.*