CS Masters' Thesis Defense

Title: *Text Classification using Machine Learning*
Speaker: Vinay Polisetty
Date: Wednesday, May 2, 2012
Time: 10:00 a.m.
Location: GMCS 418
Thesis advisor: Dr Joseph Lewis

Abstract:
Automatic Text Classification has always been given importance in the field of computer since the beginning of digital documents. Considering the large amounts of documents online and the speed with which the digital information is being produced, automating the task of text classification has a great practical use. Given the task of automation, the documents can be classified based on the genre of the articles, for instance: politics, sports, religion etc. The digital documents are available in the form of news feeds, online news article, journal papers etc.

Text Classification is a task of classifying a document into a predefined category. If we have a document d in a set of document D, and we have predefined classes c1,c2,c3…..cN, the document d will be classified and be associated with a class ci, based on what it contains. Text Classification is done based on the readily available statistical algorithms, these algorithms need to be trained with a set of labeled documents and a set of test document are classified with the these algorithms. The accuracy with which the test documents are classifies gives us a measure of how well the algorithm can perform and thus can be used to categorize unlabeled documents.
I aim to develop the Bayesian Classifier in java and train the algorithm with a certain test data and calculate the accuracy of the classifier and how well it fairs when applied to a testing data which is already labeled. Bayes Hypothesis :
Description: Description: \begin{displaymath}c_{map} = \arg