EXECUTIVE SUMMARY

Recently, relentless optimization of information retrieval effectiveness has driven Web search engines to new quality levels. Web search has become a standard and often preferred source of information finding. The World Wide Web has become a principal driver of innovation for information retrieval due to the explosion of published material. However, this explosion of published information would be of no use if the information could not be categorized so that a user may quickly find information that is both relevant and comprehensive [14].

With the rapid growth of the World Wide Web, the task of classifying natural language documents into a predefined set of semantic categories has become one of the key methods for organizing online information. This task is commonly referred to as text classification [12]. Due to the volume of documents a machine learning approach may become necessary as a manual classification approach is not practical. Text classifiers create new challenges in machine learning which include large input space, little training data, noise, complex learning tasks, and computational efficiency [12].

In 2004 Domboulas [9] investigated whether nonlinear kernel–based classifiers may improve overall classification rates over those obtained with linear classification schemes for infrared imaging face recognition applications. The specific nonlinear kernel–based classifier considered in that study was the Generalized Discriminant Analysis (GDA) approach. Results showed that the GDA approach lead to better classification performances than those obtained with the linear classifiers considered on the image database selected. Alexandropoulos later investigated a GDA approximation which is based on a Feature Vector Selection (FVS) data selection process [5], [7]. Results showed that the FVS scheme followed by the Linear Discriminant Analysis (LDA) scheme can achieve performances similar to those obtained with the GDA method at a much reduced computational cost. This study applies the FVS-LDA approach to the field of text categorization and compares results to those obtained using the Latent

Semantic Analysis (LSA) Approach commonly used in text classification/categorization applications.

The text database considered in this study was collected from the IEEE Xplore database website [2]. The documents collected were limited to Electrical engineering journal article abstracts and titles from IEEE periodicals with publications dates between 1990 and 1999. Ten categories were developed; some were specifically chosen to lead to texts with similar topics while others were selected to lead to very distinct subjects. A total of 1026 unique documents containing both article title and abstract were collected, however three of these documents were found in two classes. Note the documents were relatively short, less than one page, and contained on average around 151 words.

One of the first steps in the text categorization process is the creation of the term-document matrix (TDM) which contains features used for the categorization task. The collection of documents that makes up the text database gets converted to the TDM where each column j corresponded to a specific document and each row i corresponds to a term found in the collection of documents. Thus, each TDM element α_{ij}, where i and j are the row and column index respectively, represents the relevance of a specific word i in document j. Sixty different TDMs were explored in this study.

Common metrics used in text categorization evaluation studies include precision, recall, accuracy, error rate, and the F1 measure. In multi-label classification, the simplest method for computing an aggregate score across classes is to average the scores of all binary tasks. The resulting scores are called macro-averaged metrics. Another approach to averaging is first to sum over true positive (TP), true negative (TN), false positive (FP), and false negative (FN) over all classes, and then to compute each of the metrics. The resulting scores are called micro-averaged. These two approaches are both informative and complementary to each other by measuring performance differently. Macro-averaging gives an equal weight to each category, and is often dominated by the system's performance on smaller classes. Micro-averaging gives an equal weight to each document and is often dominated by the system's performance on larger classes. [3].

The classifier algorithms considered in the study are; Latent Symantec Analysis (LSA), GDA with a FVS processing step and one of two kernels. The two kernels considered are Gaussian and polynomial of various degrees and either a constant one added or no constant added. Polynomial kernels of degree greater than one were considered but gave worse results, and are not included in this document.

Classifier performances were compared by selecting eight different TDM types among the 60 considered in this study. The selected TDMs were those leading to both best and worse macro-averaged and micro-averaged *F1* performances obtained with the LSA classifier and 300 feature vectors (FVs).

Results showed most classification errors are directly linked to class similarities. The manifestation of these errors was very distinct when comparing LSA and FVS-LDA classifiers. This difference may be directly linked to the different criterion used in each classifier; LDA designed to extract the most discriminating features, while LSA selects the most representative ones. Note that it may be possible to increase classification performance for a well selected TDM type with a hybrid classifier based both on LSA and FVS-LDA methods. Further, different vector distance measures may also contribute to increased performances. Timing results indicate the computational loads associated with the FVS-LDA with a polynomial kernel with an added constant and the FVS-LDA with a Gaussian kernel are significantly lower than those for the other configurations considered.

Overall, taking into account both classification performance and timing issues, results showed the FVS-LDA with a polynomial kernel of degree 1 and an added constant equal to 1 is the best classifier for the database considered.