

AUTOMATED ADVERTISEMENTS CLASSIFICATION SYSTEM USING STATISTICAL METHODS¹

W.N. Wickremasinghe*

Abstract

Newspaper advertising is one of the most popular and commonly used mode of promoting products and services among the public. Even in Sri Lanka, some weekend newspapers have even gone into the extent of devoting a complete section to this purpose. In these complete sections, advertisements are classified into several categories for convenience. Currently, domain experts or editors manually assign advertisements to these categories. This study describes how machine learning and automatic document classification techniques can be used for managing a large number of advertisements, lightening the load on domain experts or editors. Classified advertisements from the *Sunday Observer* are used for this task and classified into predefined categories using three well known classification algorithms namely, k-Nearest Neighbor (*KNN*), Centroid Based (*K-Means*), and Naive Bayes. This study also discusses the different parameters and design decisions such as stemming, stopwords, dictionary (*searchable item*) size, and amount of data for improving classifier performance that normally appear when building automatic classifiers.