# Executive summary of Minor Research Project

**Name of the Principal Investigator: Sangeetha N**
**Title of research project: Text Classification using symbolic data analysis**
**UGC Reference No: MRP(S)-0129/12-13/KAMA002/UGC-SWRO dated 29-03-2013**

An important issue in text categorization is how documents are represented, and how features can be extracted from them which can be used for categorization. As the volume of information available on the internet and corporate intranets continues to increase, there is a growing need for tools helping people better find, filter and manage these resources.

The input to these methods is a set of documents (i.e., training data), the classes which these documents belong to, and a set of variables describing different characteristics of the documents. An important issue in text categorization is how documents are represented, and how features can be extracted from them which can be used for categorization.

Text classifiers are basically used for free flowing text documents that are basically an unstructured text. Text classification for such unstructured text is done with a statistical feature weighting method. To get the better result, text documents are pre-processed. Under pre-processing, high dimensionality of text are reduced by eliminating digits, punctuations, hyphens, stop words and high/low frequency words and by applying stemming. It also covers reduction technique used for elimination of synonyms due to which we get the effective result. This strategy of text classification cannot be applied to the domain of unstructured texts describing the advertisements.

When observations in large data sets are aggregated into smaller more manageable data sizes, the resulting descriptions of the new units invariably involve ``symbolic data". By symbolic data, we mean that rather than a specific categorical or numerical value, an observed value can be a set of categories or numbers, an interval or a probability distribution or any kind or more complex information than the usual one. Hence, Symbolic Data Analysis generalizes classical

methods of exploratory, statistical and graphical data analysis to more complex data issued from huge Conventional Data Bases. A Pseudo code is generated and to practically implement the pseudo code, a model construction is done for the matrimonial database which includes:

1. Construction of a symbolic database

2. Classification based on giving a query to the database

*****