ABSTRACT

Today enormous amount of information is stored and handled in the electronic media. For each and everything in real life, information becomes very vital. This information is stored in the form of text. Internet is flooded with so much of textual information and supplemented with pictures and videos. But text becomes the basic representation of all these things. But the important question is how to manage this enormous amount of text information. The most fundamental aspect of managing the information is categorizing the text. Like the foundation of a building, text categorization is basis of text management. This work concentrates on finding an effective method to do this work apart from analysing the past text categorization methods.

In the past, so many people have contributed to finding a better solution to categorize text. Several algorithms have been developed to do text categorization in a better manner. Some of those algorithms are really good. In this research work, four of these algorithms which are found to be effective, are compared and analysed. They are Naive Bayes, NGrams, K-Nearest Neighbourhood with Simulated algorithm and Back Propagation network algorithms. These algorithms are thoroughly compared with each other and a discussion based on their results is also conducted in this work. Results of some of these algorithms are better. Some of them are very promising even now. Another important point that should be noted here is that the performance of all these algorithms depends on the pre-processing work that should be carried earlier on the text documents. Hence a discussion on pre-processing is also conducted during this work.

Though these algorithms produce better results in the overall context, the research cannot stop there. This is because Information Technology is moving in a
rapid pace today. Along with it, the users of IT are also moving faster. They expect faster and accurate systems. Hence there is always a search for an improved algorithm which produces result in faster time and in an effective and accurate manner. Hence based on the above comparison a new algorithm has been designed and tested. This algorithm is termed as NGramsSA algorithm. A detailed discussion on the results of this algorithm is conducted. The performance of this new algorithm seems to be good in terms of execution time and precision.

In this work, for testing these algorithms more than hundred documents have been used. These documents mainly represent Sports, Politics and International News. In the real world, numerous text categories are available. Hence this type of analysis should be conducted on these categories too. Apart from this aspect a lot more work has to be done in the text categorization field. Some suggestions are provided at the end of this thesis for the future work.