CHAPTER 1

INTRODUCTION

"Know more than others"
"Work more than others"
"Expect less than others"

---William Shakespeare

1.1 INTRODUCTION

Information is accumulated over a period in the process of human civilization. People realized the importance of archiving and pursuing information. With the advent of computers, storing large amounts of information in electronic form became an easy task. The field of Information Retrieval (IR) is derived out of this necessity. In the past fifty years this field found considerable amount of maturity. Several IR systems are put to use by a variety of users. In the recent past, content-based document management tasks otherwise known as Information Retrieval (IR) gained prominent status due to the increased availability of large categories of documents in digital form and the ensuing requirement to access them in flexible way from the user point of view.

Text Categorization (TC) otherwise named as Text Classification is the activity of labelling natural language texts with thematic categories from a predefined set. TC approaches are observed in early '60s, at a primitive stage and in the recent period a major subfield of the
information systems discipline, with the increased applicative interest and the availability of more powerful hardware. The most popular and oldest one is Knowledge Engineering (KE) task, where manually defining a set of rules on how to classify documents under the given category. Creation of automatic document classifiers by means of Knowledge Engineering techniques consists of manually building, an expert system capable of taking TC decisions. Such an expert system would typically consist of a set of manually defined logical rules, at least one per category.

This approach lost popularity over a period of time favouring Machine Learning (ML) paradigm, where a general inductive process builds an automatic text classifier by means of learning, from a set of pre-classified documents embodying the characteristics of the respective categories. Numerous advantages are found reflecting the accuracy and savings in terms of expert labour power. In Machine Learning terminology, the classification problem is an activity of supervised learning. This learning process is "supervised" by the knowledge of categories and the representative set of training instances. This effort is a continuous process towards the construction of a dynamic classifier rather than a static one.

Corpus-based machine learning approaches are predominant in language engineering over the knowledge-based approaches, which use explicit rules hand-crafted by domain experts. In machine learning approaches, a set of training data is given and the machine builds a model for performing the intended task. It is expected to be generic and it is understood that training is based only on the intrinsic properties of the data, as expressed through a set of features. Extraneous indicators such as clues from scripts or fonts used, header information or explicit
mark up tags in the document structure may not contribute significant knowledge in a generic system.

A learner is available off-the-shelf, to construct a classifier from a set of manually classified documents. The same happens if a classifier already exists with the original set of categories updated or if the classifier is ported to a completely different domain. The Machine Learning approach is more convenient than the Knowledge Engineering approach to manually classify a set of documents than to build and tune a set of rules, since it is easier to characterize a concept extensionally (i.e., to select instances of it) than intentionally (i.e., to describe the concept in words, or to describe a procedure for recognizing its instances).

1.2 TEXT CATEGORIZATION

Text categorization is the task of assigning a Boolean value for every document labelled into a category in a domain of documents. The value of true or false indicates a decision to a document using an appropriate target function otherwise called as a classifier. These classifiers or categories are symbolic labels without any additional knowledge. A more generic definition states [SEB 02] that classification is to be accomplished on the basis of endogenous knowledge without any exogenous knowledge. This general model is used as a basis for the present work.

Subjective notation based on the semantics of a document cannot be a deterministic decision for the purpose of the classification. The phenomenon of inter indexer inconsistency influences the subjective judgment of an expert in the real world scenario. This situation leads to the single label and multi label categories.
A single label category is a special case of text categorization based on the binary judgment, where as the multi label classification of the document results in overlapping categorization. From the theoretical point of view multi label classification can be visualized as independent problems of binary classification where the categories are treated as stochastically independent. Two approaches are generally found from the classifier point of view. Document Pivoted Categorization (DPC) is an approach in which for a given document all the categories related to it are mapped. The other approach, Category Pivoted Categorization (CPC) is aimed at mapping of all the documents related to the respective category. DPC is thus suitable when documents are available at different moments in time such as filtering of E-mails. CPC is instead suitable in a scenario where addition of a new category to the existing set after classifying a number of documents under a predefined set of categories and all the documents need to be reconsidered for new classification. A complete automation of the TC task requires a True or False decision for each pair of documents and categories in the real world scenario.

A system demands for ranking of categories according to their appropriateness to a document. This is known as hard categorization but such categorization is of great help to a domain expert in deciding final categorization by way of restricting the choice to the category (or categories) at the top of the list, rather than having to examine the entire set. Alternatively, given a category a system might simply rank the documents from a document set, estimating the appropriateness to a category. For such classification a domain expert examines the top-ranked documents instead of the entire document set. These two modals are sometimes called Category-Ranking TC and Document Ranking TC respectively, and are treated as counterparts of DPC and CPC.
Text categorization approaches are observed with a large number of applications in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application in the area of document organization or selective and adaptive document dispatching are influenced by Text categorization algorithms.

Another important application is automatic document indexing for IR systems that relies on a controlled dictionary where each document is assigned with one or more key words or key phrases describing its content.

The global statistics on Internet usage based on linguistic population [WEB 1] reveals the fact that non English users are dominating in numbers. At the same time the web content based on the linguistic information found with a phenomenal shift towards the specificity of the language. European languages and East Asian languages are in the fore front of this growth. Unicode standard is acting as a catalyst in this process. The research in the filed of Information Retrieval and Text mining is also influenced by this phenomenal change.

In general the choice of representation of text depends on defining the meaningful units of text embodying the meaningful natural language rules evolving into a set of compositional semantics. Decomposition of the text document into meaningful units and model it into any computational form is a major task. In any Machine Learning approach considering all decomposed units will be a Herculean task due to a large number of language rules leading to a dimensionality problem. In this context a trade off between the meaningful units and computation of the associative rules will be considered to evolve an effective representative model while performing the task of Text categorization.
1.3 MOTIVATION

Document representation based on words and/or phrases are the most popular models which are mainly dependent on language. N grams are used as alternatives to the above models. N grams are modelled as language independent units. The strategies used for information retrieval or document classification based on N-grams are found [CAV 95, COH 95] to be effective on languages other than English. N grams are found to be more effective units while distinguishing documents of different languages in multilingual collections. Gauging the topical similarities between the documents is another interesting area of N gram applicability. In the present work a novel model is attempted to identify the topical similarity between documents of Telugu Script using N grams.

N grams are consecutive n character sequences derived from an input stream. Decomposition of an individual character from the input stream is a simple task in case of languages possessing one to one correspondence between the characters and machine representation. This process is complex phenomenon in the case of Indic scripts due to the very nature of canonical structure defined by an individual syllable (generally called as Akshara). The canonical structure ((C)C)CV is a common structure for all Indic scripts resembling the phonetic nature derived from Brahmi Script. At the same time multiple variations are found in the respective scripts.

In this context the present work is aimed at evolving a specific model for decomposition of individual Aksharas from the input stream of Telugu script documents. The decomposed units are the representative set of the respective document. The information content of that document
will be reflected in those units. Although N grams are treated as language independent units, the set of N grams derived from a specific document is to be treated as the information content of that document. The Shannon’s proposal of entropy is used as a parameter to evaluate the information content of the N gram model in the present work. The proposal made in this work is to validate the efficiency of N gram model in terms of information content derived from Telugu Documents. Most of the recent approaches are attempted towards retrieval efficiency for various values of N. The present work emphasises that the amount of information represented by a set of N grams is an influencing factor over the end result. This principle is treated as a basis in regard to the applicability of N grams on Telugu script documents.

The individual Akshara derived from the canonical structure is one of the unit among a possibility of more than $5 \times 10^5$ possible units. An independent bigram is an element in a large set of $25 \times 10^{10}$ possible elements. The Machine Learning algorithm has to handle a large search space as n increases in N grams. The necessity of dimensionality reduction motivated the present work to identify the possible solutions and evolve a model without a compromise in terms of efficiency. The present work is aimed at handling the dimensionality problem in manifold and also reducing the computational complexity.

1.4 ORGANIZATION OF THE THESIS.

This thesis is organized into six chapters where the first chapter introduces basic introduction to the area of text classification with an emphasis on motivation of the present work. In the last chapter all the references are listed.

In Chapter 2, various proposed methods in the literature related to Information Retrieval in terms of text categorization and classification are reviewed. A brief discussion on machine
learning approaches is presented in terms of supervised and unsupervised strategies, which were applied on Text categorization methods. Applicability of N grams on Text Categorization methods is explored. A brief presentation is made on N gram model on different languages. The complexity of Indic scripts is discussed and in continuation, problem definition as well as methodology is described in this chapter.

Chapter 3 deals with the formulation of the classifier using Bayesian model and the proposed method of handling dimensionality problem. Representation of N grams in Telugu script and the complexity involved while tokenizing the Telugu word into N grams are presented in this chapter. The proposed model uses posterior probability for classification purpose. Dimensionality reduction problem is addressed using supervised corpus. The training and testing phases are described in this chapter. Evaluation of the proposed classifier is also described.

In Chapter 4, Results of the proposed model are discussed. Bigram and Trigram evaluation along with entropy values are presented. The evaluation of classifier performance on bigrams, and combination of bigrams plus trigrams are discussed.

The fifth chapter (Chapter 5) presents a brief summary of the work followed by a discussion on the future scope of the work.