Chapter 1 Introduction

Document Classification (DC) is the process of analyzing a set of documents and labeling each one of them with an appropriate category according to its relevance towards one of a pre-defined set of categories. The field of DC is rife with potential for several modern services document centric applications, such as Document Summarization, essay scoring, organizing documents for query based information dissemination, email management and topic specific search engines.

The phenomenal rise of web based services such social networks, e-commerce and a variety of e-portals such as those for e-governance require well-organized management of documents. The first step for achieving this is an accurate, reliable and fast classification of relevant documents into a set of known categories. Different kinds of documents such as emails, blogs, posts and comments are rapidly populating the web. Emerging applications such as text based event detection, sentiment analysis and disaster management demand that Document Classification take into cognizance the meaning or semantics that are knitted into such small messages. These new paradigms pose the challenge of analyzing tightly worded documents with highly meaningful content.

This thesis focuses on the problems of Context Based Document Classification and Analysis. Our main aim is to investigate the urgent approaches and challenges that exist in the field of DC and enhance its power by developing a context based approaches that cork in symphony with traditional statistical approaches. This chapter introduces the basic notion comprising the area of DC and presents the motivation, objectives, organization and contribution of this thesis.

In section 1.1 we survey various aspects of the domain of DC such as the prevalent approaches, performance measures, general architecture and applications of DC. In section 1.2, we give our motivation to work on the thesis’ core problem “Context Based Document Classification and Analysis”. Section 1.3 enumerates the objectives of this
dissertation. In section 1.4, we provide the chapter plan of the thesis. In section 1.5, we enumerate the contributions of the thesis. In section 1.6, we summarize the chapter.

1.1) Document Classification

We first take a bird’s eye view of the basic concepts that cover the area of DC, the main approaches that are followed the performance metrics which are used to assess DC systems and the applications of DC.

1.1.1) Overview

The task of DC is accomplished by merging the fields of Information Retrieval (IR) and Machine Learning (ML) [1]. IR is the process of sieving out relevant information from a large collection of data to satisfy a query. Its tenets are applied in the area of DC to scoop out the relevant documents of one category from a corpus based on their content. Therefore, DC inherits itself as a specialization of IR.

ML is the process of designing algorithms to recognize complex patterns existing in given data and make intelligent decisions using these patterns for specific applications. Machine learning techniques for DC aim at extracting those features which are helpful in identifying the relevancy of a document to each category. The following two broad ML techniques are usually employed.

A) Supervised Learning: Supervised learning is a technique to guide the process of classification by providing prior information such as labeled documents and keywords to train the classifier [1]. The input corpus is parted into the two sets for these phases: the Training set that comprises labeled documents and the Testing set that comprises documents to be classified. In the Training Phase, the classifier learns by categorizing documents to the appropriate categories by using labeled information in the Training set. Having thus tuned its decision parameters, it enters the Testing Phase, when the classifier assigns categories to the documents of Testing set by utilizing its previously acquired knowledge. Some popular generic
methods of supervised ML classifiers are Naïve Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM) and Neural Networks (NN) [2].

B) Unsupervised Learning: Here the classifier does not have any labeled information of documents for learning [1]. K-Means clustering is the most popular technique for unsupervised technique. It classifies a given dataset by initializing a pre-decided number of clusters with seed data and then assigning the remaining data instances to one of them by using a suitable distance metric to calculate their similarity to each of the clusters.

1.1.2) Approaches for DC

Document Classification methods can be classified into broad approaches: Statistical and Context based.

A) Statistical approaches: Predominantly, Statistical Approaches have been applied for DC thus far. These approaches are based on frequency of word occurrences in a given document. Statistical techniques consider the words of a document as unordered or independent elements and simply compute the frequency of these feature items. The most popular being Term Frequency-Inverse Document Frequency (TF-IDF), Chi-square and Information Gain [3]. Statistical techniques believe that the independent words of a document have enough potential to represent a document and play significant role in classification [4]. Statistical approaches, therefore, is more concerned with data occurrence without explicit representation of the meaning of data. Several algorithms based on this approach have been investigated and have given good results in various web applications [5] [6] [7] [8] [9] [10] [11] [12] [13].

However, statistical approaches do not take into account the semantic relationship between words on the context created by their relative positions. They completely ignore the fact that words at different positions may have different contributions to
the theme of a document or that a group of words can signify something meaningful.

**B) Context based approaches:** An alternative approach is *Context Based Document Classification (CBDC)* that takes into account how a word $w_1$ influences the occurrence of another word $w_2$ in a document. Thus, the presence or absence of $w_1$ affects a classification based on $w_2$. These methods exploit the relationship among the words of a document in order to evaluate their semantic relevance to a given category. Some recent papers have reported various techniques and algorithms for finding relevancy among words [4] [14] [15] [16] [17] [18]. Yet there is ample scope to investigate further into this arena.

Context feature extraction offers a lot of scope to experiment on the relevancy among words in a document in various ways and exploit them to achieve improved results. Some of them are enlisted below:

- **Semantic features:** The Lexical Semantics of a term are defined as a set of words that are hierarchically related to it within an ontological network of words that are bound together by lexical cohesiveness. They include *Synonyms, Hypernyms, Hyponyms, Meronyms etc* [19]. These features can be extracted by using knowledge based ontologies such as WordNet, Hownet and DBpedia. Appendix A describes different kinds of lexical cohesion that are incorporated in WordNet.

Many semantic features which cannot be obtained by traversing through an ontology’s hierarchical relations can be obtained by using Wikipedia. Wikipedia is used to extract the relevant terms that are present as hyperlinks in related *Article Pages* and in their systematically organized *Category Pages*. As an example ‘Windows’ is not lexically related to the term ‘Computer’. But when a document contains information about operating systems, ‘Windows’ may be an important feature which can be conveniently extracted from the relevant
Wikipedia Article/Category pages. Appendix B exemplifies the process of extracting Wikipedia-referring terms from its Article/Category pages.

- **Syntactic features** There are various ways to track the contextual significance of groups of words. One way is to count the number of times a term co-occurs with another term (Bigram) or with a set of terms (N-gram). A **Lexical unit** such as a phrase or group of related words that form an association represents another type of context. For example “Traffic light”, “take care of” and “by the way” represent meaningful word-associations. **Grammatically related** groups of words such as noun phrases and verb phrases signify Parts of Speech (POS). These can be extracted with the help of a POS-tagger [20].

- **Structural features**: A document’s structural characteristics add another dimension to its feature space. They include parameters such as the total number of words, total number of sentences, average length of sentences and relative positions of words in a document.

C) **Collaborative approaches**: Some research efforts have been made to combine statistical and contextual features by extending the frequency based method to augment a contextual score such as by counting joint occurrences of groups of words within a document [21] [22] [23] [24] [25].

### 1.1.3) Performance Measures

A performance measure provides a basis for quantitative analysis of Document Classification. It is a scale to measure the quality against a desired goal. To define different performance indices, consider a binary classifier, which has two categories for Document Classification: 1) Positive and 2) Negative. The following parameters are important performance indices:

- **True Positive (TP)**: It is the total number of documents that belong to the actual ‘Positive’ category and have been predicted to belong to the ‘Positive category’.
- **False Positive (FP):** It is the total number of documents that belong to the actual ‘Negative’ category and have been predicted to belong to the ‘Positive’ category.

- **False Negative (FN):** It is the total number of documents that belong to the actual ‘Positive’ class and have been predicted to belong to the ‘Negative’ category.

- **True Negative (TN):** It is the total number of documents that belong to the actual ‘Negative’ category and have been predicted to belong to the ‘Negative’ category. Table 1.1 presents the attributes to calculate the performance indices in DC.

We can understand these parameters more clearly by the following table:

Let $N$ be the sum total of $TP$, $FP$, $FN$ and $TN$. The following scores are defined:

<table>
<thead>
<tr>
<th>Predicted Categories</th>
<th>Actual Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual Positive Class</td>
</tr>
<tr>
<td>Predicted Positive Class</td>
<td>True Positive ($TP$)</td>
</tr>
<tr>
<td>Predicted Negative Class</td>
<td>False Negative ($FN$)</td>
</tr>
</tbody>
</table>

**Table 1.1 Attributes for Calculating Performance Measures**

1. **Precision (PR):** It is the fraction of the number of correctly predicted documents for ‘Positive’ class which is divided by total number of predicted documents for ‘Positive’ class.

   \[ PR = \frac{TP}{TP + FP} \]  

2. **Recall (RE):** It is the fraction of the number of correctly predicted documents for ‘Positive’ class which is divided by total number of documents of ‘Positive’ class.
\[ RE = \frac{TP}{TP + FN} \]  

1.2

3. **F1-measure**: The harmonic mean of precision and recall is called as F score. Here both are evenly weighted in the computation, thus it is also called as F1 measure.

This is defined as

\[ F1 = 2 \left( \frac{1}{RE} + \frac{1}{PR} \right)^{-1} \]  

1.3

4. **Accuracy (AC)**: It is the average of the number of correctly classified documents.

\[ AC = \frac{(TP + TN)}{N} \]  

1.4

5. **Error (ER)**: It is the average of the number of incorrectly classified documents.

\[ ER = \frac{(FP + FN)}{N} \]  

1.5

These performance indices can be measured in two ways:

**A) Micro-averaged results**: Micro-averaged results first calculate the category-wise true positive, false positive and false negative values to calculate the Precision, Recall and F1-measure for each individual category. These category-wise metrics are then averaged over all categories to get the corresponding overall values.

**B) Macro-averaged results**: Macro-averaged results consider the true positive, false positive and false negative values for all documents belonging to any of the categories in order to compute the overall Precision, Recall and F1-measure.
1.1.4) General flow of DC

DC is usually a two-step process. First, the text contained in a document is analyzed and a compact set of features that characterize the document are generated. Next, based on their extracted features, the documents are classified to their respective categories by employing a suitable machine learning technique.

DC undergoes a sequence of tasks to achieve its aim. The tasks involved in the DC process for both supervised and unsupervised techniques are depicted in Figure 1.1.

![Diagram showing the general flow of DC, with supervised and unsupervised flow processes shown separately.](image-url)
If we follow a supervised approach, the classifier must be trained by a set of training documents. Thus the corpus is split into training and testing sets. This step is not required in unsupervised classification. The individual steps are briefly described below:

- **Pre-process Documents**: The raw documents must be pre-processed to convert them into an array of non-trivial tokens. Pre-processing comprises of the following steps:

  - **a) Stop word removal**: Trivial words, such as ‘a’, ‘am’, ‘above’, ‘across’, ‘alongside’ ‘but’ etc, which do not play any important role in classifying a document are removed.

  - **b) Stemming**: Inflected words are converted to their root forms. These are called tokens. The objective of this step is to avoid treatment of the same words in different forms. For example, both ‘celebration’ and ‘celebrating’ are converted to the same base form ‘celebrate’, thereby dealing with a single word. Stemming reduces the size of the feature set.

- **Feature Selection and Extraction**: This step is required to reduce the large dimensionality of the feature space that is typical of text documents. It helps to eliminate unimportant tokens from documents [26]. Dimension reduction is accomplished by two methods:

  i) **Feature selection**: In this method, we select a subset from the original features. Since the number of selected features are less than the original features, it reduce great dimension of feature space which helps to restrain cost and time complexity.

  ii) **Feature Extraction**: In feature extraction, we combine promising new features to the original ones. As a resultant, we receive a higher dimension projected feature space. Now we extract a set of features from this feature space.
➢ **Train the Classifier:** The labeled features sets are applied to a classifier which accordingly adjusts its internal decision parameters so as to be able to discern the category of any new documents.

➢ **Label Test Documents:** The test documents are supplied to a trained classifier. These documents are labeled with their respective categories according to the classifier’s acquired discerning ability.

➢ **Evaluate the Classifier:** Now, the predicted categories of test documents are matched with their originally labeled categories. Based on the number of correct and incorrect matches, the performance of the classifier is evaluated by calculating the various performance indices explained in the previous subsection.

### 1.1.5) Applications of DC

Document Classification is being applied to an ever-growing set of applications. Some representative applications are given below.

➢ **E-mail Management:** Email affects every user of the Internet. However, emails also flood the Inbox quickly leading to a morass of unorganized information. Even though many email providers allow the creation of folders and sub-folders where emails can be routed based on the sender’s address, date, and subject. There is an urgent need for automatically segregating emails based on their relevance to the user [27].

As a basic need, spam filtering classifies messages into two categories, *viz.* Spam and *Non-Spam*. Besides being undesired, spam email consumes a lot of network bandwidth. Over time, spammers resort to deceptive and deluging methods to get around anti-spam software thereby leading to a gradual degeneration of the filter’s efficacy. To counter this, innovative DC approaches with good generalization, continuous adaptive learning and context sensitivity need to be applied. Extending this concept to the general case of filtering emails into several categories based on
their relevance to the user, we can investigate DC approaches for personalized management of all emails.

➢ **Text Summarization:** Summarization is the process of filtering the most useful information from various information sources to generate a compact version for a particular user or task [28].

There are two kinds of summarization systems:

(a) **Structural:** Structural summary uses the logical structure of a document to compress it with the help of headings, subheadings, titles, subtitles etc.

(b) **Semantic:** Semantic summary deals with the semantic relation of words in a document to extract a meaningful gist. It has been observed that semantic based document summarization is fast evolving as an interesting and popular area of research. Lexical chains are a popular technique for text summarization [29] [30] [31].

➢ **Search Engines:** Search Engines are invaluable tools for searching relevant web pages according to the user’s desired topic [32]. This is the most popular tool for getting tomes of information from World Wide Web (www) in a systematic manner with page ranking. Present search engines largely use statistical approaches. A search engine extracts keywords from a sentence entered by the user and lists all related web pages matching those keywords. Now the scenario is changing to the use of search engines based on semantic models. These search engines meaningfully relate the keywords of a user’s query with tokens of web pages to determine their contextual relevance to the query.

➢ **Sentiment Analysis:** Sentiment Analysis, also known as opinion mining, has of recent gathered serious attention from researches [33]. It is used to rank movies, analyze a new consumable product and observe the sensitivity of current affairs by analyzing the emotions of people as gathered from text in documents that are posted by netizens on various social web sites and in the blogosphere.
Web News Classification: WWW is full of news on various issues gathered from different sources. To extract the news on a single episode, DC techniques are utilized [34].

1.2) Motivation

Over the years, numerous statistical approaches for DC have been successfully devised. As stated before, they simply count the occurrences of words in a document. Statistical techniques have been widely applied to web applications. They have been harnessed to capacity. Their applicability in further improving the quality of DC seems to have reached a saturation point. It can no longer handle the increasing complexities posed by polysemous words, multi-word expressions and closely related categories that require semantic analysis.

Recently, researchers have illustrated the potential for applying context-oriented approaches to improve upon the quality of Document Classification [24] [16] [18].

The World Wide Web is fast getting populated by websites that house a wide variety of documents. They are unique in themselves; quite different from traditional documents. For example Twitter hosts short length massages with a high degree of meaningful context encapsulated within 140 characters. Blogs posted in the Blogosphere, posts comments and replies posted on social networks like Facebook and LinkedIn and emails are other types of interactively generated documents where words seldom repeat. Statistical approaches would not suffice for representing them; we need to analyze the contextual content for classifying such documents that can intelligently derive the contextual meaning implied by related words.

Moreover, context can be interpreted in a variety of intuitively appealing ways. The contextual features that can be tapped were enumerated in subsection 1.1.2. They include lexical cohesion, Part of Speech, phrased expressions and related references found in Wikipedia Articles and Category pages.
We derive the prime motivation for investigating the central theme of the thesis viz *Context Based Document Classification and Analysis* from the above considerations. Evidently, both statistical as well as contextual approaches for DC possess their own strengths and weaknesses *vis-à-vis* different corpora. The challenge therefore, is to tap the power of both these approaches to classify documents in the best possible way according to own expressive characteristics.

**1.3) Objectives of the Thesis**

We set following objectives to be achieved through the dissertation:

- To investigate the potential, the bottlenecks and the challenges in the arena of DC
- To incorporate innovative context oriented features and approaches that can be effectively utilized for DC, and develop fast and high performance classifiers.
- To optimize the contribution of different contextual features so as to achieve the best classification accuracy for a given corpus.
- To combine Contextual and Statistical approaches in a tailored manner for given corpora.

**1.4) Thesis Layout and Chapters’ theme**

The main body of this thesis is classified in Figure 1.2. In this section, we outline the central theme of each chapter that follows this introductory chapter and give the complete thesis layout.

- **Chapter 2- Survey of Literature:** To achieve our first objective, we investigated and comparatively evaluated various approaches reported by the authors in the field of DC. In Chapter 2, we present an exhaustive review of literature, which guided us to move forward on the path of the theme of this thesis.
Figure 1.1 Thesis Layout
Chapter 3- Keyword Enrichment, Overlapped Semantics and Belongingness:
In Chapter 3, we present a knowledge based DC scheme. The first concept in this scheme is Semantic Keyword Enrichment. A list of keywords for each category is automatically extracted by using Lexical Semantics from WordNet and hyperlinks from Wikipedia Article Pages. Later the keywords list in each category is further enhanced by acquiring fresh keywords from classified documents. A **Keyword Strength (KS)** matrix computes the extent of belongingness of the keywords to various categories.

The dimensionality of pre-processed documents is reduced by applying the concept of **Overlapped Semantics (OS)** [35]. By OS, we extract highly cohesive words of the document, which are important for classification.

The degree of similarity between the features of the document and keyword lists of each category is quantified by a metric called **Belongingness**. Notably, it is not restricted to a binary value but can vary anywhere between 0 and 1. A document possesses some belongingness to each category. Classification is then achieved by attributing the document to the category with the maximum **Belongingness**.

Chapter 4- Dual Lexical Chaining: In Chapter 4, we propose a novel approach to categorize text documents using a **Dual Lexical Chaining** (DLC) technique. This is amalgamated with the concept of Keyword Enrichment introduced in the previous chapter. Highly cohesive terms of a document are woven together into two separate Lexical Chains; one for their noun senses and another for their verb senses. This segregation enables a better expression of word cohesiveness by treating concept terms and verbs distinctively.

We further propose a new metric to calculate the strength of a lexical chain. It includes a statistical part given by **Term Frequency_Inverse Document Frequency_Relative Category Frequency (TF_IDF_RCF)** which is an improvement upon the conventional **TF_IDF** measure [3]. The chain’s contextual strength is determined by its degree of lexical matching with category-keywords.
as well as by the relative positions of its constituent terms. Finally, the *Belongingness* of a document to a category is determined by all its Lexical Chains weighted by their respective lengths.

- **Chapter 5- Lexical Semantics with M-SVM:** Having identified a set of features that are well suited for context oriented DC, we next focus upon the power of Lexical Semantics solely on their own merit and utilized a generalized ML technique to improve performance of the classifier. In Chapter 5, we chose a Multiclass Support Vector Machine \([36]\) as a classifier. In general SVM utilizes useful localized information from the data set called support vectors to construct a separating hyperplane. Intuitively, this appears more suitable for context based classification as compared to classifiers that learn from global statistical characteristics of the data set.

In this scheme, we prepare a very short feature vector by recording the intersection count between the Lexical Semantics of tokens in a document and those of category-keywords.

Our experimental results revealed that using only Lexical Semantics as features, the classification accuracy saturated at 70% only. But it does signify the strength of Lexical Semantics without the help of any other features. The results led us to integrate other complimentary features along with Lexical Semantics and utilize an M-SVM to boost classification accuracy further.

- **Chapter 6- GA driven Exploration for Collaborative DC:** The problem of saturation of average accuracy reported in the previous work motivated us to seek an automated solution for a collaborative approach that brings an array of statistical and contextual features under the same ambit. In this chapter, we opted for an exploration process that would tune the process of classification for a given corpus by finding the best weighted combination of a variety of features to yield the highest possible accuracy. This process could adapt the features for specific corpora, their sizes and writing patterns. We used a Genetic Algorithm for
optimizing the feature weights [37]. We embedded a Multiclass Support Vector Machine into its fitness function for the purpose of evaluating the accuracy of DC.

For statistical feature, we used the popular TF-IDF metric. For context based features, we used sixteen semantic relations. Among 16 relations, 13 are Lexical Semantics retrieved from WordNet named as *Synonym, Hypernym, Hyponym, Instance_Hypernym, Instance_Hyponym, Member_meronym, Member_holonym, Part_meronym, Part_holonym, Regions, Substance_meronym, Substance_holonyms* and *topics*. The next three features are derived from Wikipedia’s first-level, second-level and N-Depth referring terms.

The experimental results revealed interesting observations. We found that a statistical analysis or context analysis by themselves are not good enough for classifying documents of any corpus nicely. A fine-tuned balance of various features is essential.

- **Chapter 7- Conclusions & Future Scope:** In this chapter, we conclude with our findings, examine the limitations and outline future scope and directions.
- **Appendices:** Appendix A describes the WordNet usage as an ontology based on Lexical Semantics. Appendix B describes the utility of Wikipedia article pages and category pages as knowledge sources. Appendix C overviews the basic working and characteristics of Support Vector Machines.

### 1.5) Contributions of the Thesis

The contributions of the thesis are summarized below.

- We used GA based optimization to **fine-tune the balance between Statistical and Contextual features** in DC. We further refined the context based features into various components. The GA searched for the **optimal weights of these contextual features** for yielding the best classification accuracy for a given corpus. As a result of this optimization, the classifier reported good accuracy on various corpora.
We used **segregated noun and verb sense Lexical Chains** to represent a document. *Concept-terms* (nouns) are known to have a greater bearing on a document’s meaning than action terms (verbs). This segregation allows the classifier to differentiate between and attribute different weights to noun chains and verb chains.

We proposed a new **Belongingness** metric which calculates degree of belongingness of a document to predefined categories. The document is ascribed to the category for which it obtains maximum value of belongingness.

In addition to the oft-used Lexical Semantics such as *Synonyms, Hypernyms* and *Hyponyms*, we used an array of alternative Lexical Semantics such as *Holonyms, Co-ordinate terms and Topics* for improving the strength of lexical cohesion.

We propose a new statistical metric called **Relative Category Frequency (RCF)** which when coupled with the TF-IDF measure gives us **TF_IDF_RCF**. For supervised learning, this metric is shown to generate much **better classification accuracy** than the conventional **TF_IDF**. In essence, it appends extra strength of those terms obtained from training documents that have a marked presence in a particular category.

We augmented the lexical keyword list derived from WordNet with referential **keywords from N-Depth Wikipedia** as a source in order to generate a **semantically strong keyword list**.

In order to achieve the objective we propose a metric **Keyword Strength (KS)**, which helps to compute the importance of keywords for categories.

We introduced the concepts of (i) enriching keywords of a category by utilizing the tokens of high **Belongingness** classified documents, (ii) category-wise pruning of keywords that are not semantically related to labeled training documents of that category, thereby improving **Keyword Strength**.
1.6) Summary

In this chapter we presented an overview of Document Classification (DC). We elaborated the approaches, applications, and performance metrics of DC. We discussed our motivation, objectives and approaches for undertaking the challenge “Context Based Document Classification and Analysis” as the theme of this dissertation.

In the next chapter, we will review and discuss the research works of various authors in the field of Document Classification.