A COMPREHENSIVE ANALYSIS OF USING WORDNET, PART-OF-SPEECH TAGGING, AND WORD SENSE DISAMBIGUATION IN TEXT CATEGORIZATION

by

Kerem Çelik

B.S., Computer Engineering, Bahçeşehir University, 2009

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2012

A COMPREHENSIVE ANALYSIS OF USING WORDNET, PART-OF-SPEECH TAGGING, AND WORD SENSE DISAMBIGUATION IN TEXT CATEGORIZATION

APPROVED BY:

Assoc. Prof. Tunga Güngör .....
(Thesis Supervisor)

Prof. Emin Anarım .....

Assist. Prof. Arzucan Özgür .....

DATE OF APPROVAL: 08.06.2012

ACKNOWLEDGEMENTS

I would like to express my special gratitude to my thesis advisor, Assoc. Prof. Tunga Güngör, for his mentorship, guidance and support throughout the whole process of this thesis.

I would like to thank to Prof. Emin Anarım and Assist. Prof. Arzucan Özgür for their participation to my thesis jury among their heavy program and also for their valuable comments.

I also would like to thank to TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu) for their financial support (BIDEB-2210 Fellowship) during my master study.

The last but not the least, I would like to thank family and my wife Melike Demir Çelik for her unconditional love and unwavering support.

ABSTRACT

A COMPREHENSIVE ANALYSIS OF USING WORDNET, PART-OF-SPEECH TAGGING, AND WORD SENSE DISAMBIGUATION IN TEXT CATEGORIZATION

By the huge increase of data volume in the digital environment and the machine learning techniques, studies on automatic categorization of text documents is increased. Text categorization is simply assigning predefined label to unseen documents by using some learning models. Traditional text categorization is based on statistical analysis of documents to represent the document with some vectors. And then, one of the machine learning techniques is used for categorization of documents.In addition to the traditional text categorization techniques, in this thesis, we group words by their part of speech tag and investigate the effect of each part of speech individually and jointly in the classification accuracy. Furthermore, we incorporate semantic features such as synonyms, hypernyms, hyponyms, meronyms and topics into the documents by using WordNet. Thus we add meaning of terms. One of the problems faced in this study is that not all the semantic features really related to the document, in other words synsets generate ambiguity. To solve the problem we introduce a new method to eliminate the ambiguity. In this thesis the main objective is to investigate the contribution of semantic features. By incorporating semantic features we add meaning to the documents and thus the classification accuracy increased.

ÖZET

METİN SINIFLANDIRMADA WORDNET, KELİME TÜRLERİ VE KELİME ANLAMI BELİRGİNLEŞTİRME KULLANIMININ KAPSAMLI ANALİZİ

Dijital ortamdaki metinler ve yapay öğrenme tekniklerindeki büyük artış, metinleri otomatik sınıflandırma çalışmalarının artmasına neden oldu. Metin sınıflandırma, temel olarak, öğrenme modellerini kullanarak, daha önceden görülmemiş dökümanları önceden belirlenmiş sınıflara atamaktır. Geleneksel metin sınıflandırma, herbir dökümanı, istatistiksel olarak inceleyerek belirli bir dizi haline getirmeyi hedefler ve ardından, metinleri sınıflandırmak için yapay öğrenme tekniklerini kullanır. Bu tez kapsamında, geleneksel metin sınıflandırma yöntemlerine ek olarak, metinlerde bulunan kelimeleri türlerine gore gruplandırıyoruz ve her bir türün sınıflandırma başarısındaki katkısını hem ayrı ayrı hem beraberce değerlendiriyoruz. Bunların yanı sıra, metinlere WordNet kullanarak, anlamsal özniteliklerden(semantic features) olan; eş anlamı(synonym), genel anlamı(hypernym), özel anlamı(hyponym), parça anlamı(meronyms) ve konuyu(topic) ekliyoruz. Bu sayede metinlere anlam(semantic) eklemiş oluyoruz. Bu aşamada yaşanılacak sorunlardan bir tanesi, bu anlamlar için anlam belirsizliği(ambiguity) oluşmasıydı. Bu problemi geliştirdiğimiz bir yöntem ile ortadan kaldırmaya çalıştık. Bu tezdeki temel amacımız, anlamsal özniteliklerin metin sınıflandırmaya olan katkılarını araştırmak ve bu sayede sınıflandırmadaki doğruluk başarısını arttırmaktır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

d	Document
d_i	Document i
dS_i	Set of synset for document i
N	Number of Document
C	Number of Category
w_i	Word i
t_i	Term i
s_i	Synset i
w_{ij}	Weight of term i in document j
$CHI(t_k, c_i)$	Chi-square statistic of term t_k in category c_i
$P(t_k, c_i)$	number of documents belonging to class c_i in which term t_k occurs
$P(\bar{t_k}, \bar{c_i})$	number of documents not belonging to class c_i in which term t_k does not occur
$P(\bar{t_k}, c_i)$	number of documents belonging to class c_i in which term t_k does not occur
$P(t_k, \bar{c_i})$	number of documents not belonging to class c_i in which term t_k occurs
$P(t_k)$	number of documents in which term t_k occurs
$P(\bar{t_k})$	number of documents in which term t_k does not occur
$P(c_i)$	number of documents belonging to class c_i
$P(\bar{c_i})$	number of documents not belonging to class c_i
$Score(s_i)$	Score of synset i
$Similarity(s_i, s_j)$	Similarity between synset i and synset j
$Hypernyms(s_i)$	Hypernym tems list of synset i
$Topics(s_i)$	Topic tems list of synset i
$CommonCount(.,.)$	Common Count of given two list i
ρ	Recall
π	Recall

LIST OF ACRONYMS/ABBREVIATIONS

Acc2	Accuracy2
Adj	Adjactive
Adv	Adverb
a.k.a	As Known As
AT&T	American Telephone and Telegraph
BOW	Bag of Word
CHI	Chi-square statistic
DF	Document Frequency
FN	False Negative
FP	False Positive
FS	Feature Selection
FW	Feature Weighting
HTML	Hyper Text Markup Language
Hype	Hypernym
Hypo	Hyponym
Idf	Inverse Document Frequency
IG	Information Gain
ISA	Is A Relation
k-NN	k Nearest Neighbour
L	Table/Figure on the Left
Macro-F	Macro F Measure
Mero	Meronym
Micro-F	Micro F Measure
N	Noun
NB	Naïve Bayes
POS	Part of Speech
POST	Part of Speech Tagging
R	Table/Figure on the Right

SMART	System for the Mechanical Analysis and Retrieval of Text
SVM	Support Vector Machine
Syn	Synonym
TC	Text Categorization
Tf	Term Frequency
TN	True Negative
Top	Topic
TP	True Positive
V	Verb

1. INTRODUCTION

Since the early 1990s, the accessibility and abundance of digital documents make text categorization an important and necessary research field. Today, text categorization is being applied in many contexts in order to organize and manipulate the documents. Arrival of the machine learning methods in text categorization is one of the essential factors that improve the effectiveness of text categorization in information retrieval systems.

Text categorization (a.k.a. text classification) is the task of assigning predefined categories to text documents. It can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively.

In order to classify documents, generally, each document is represented as a vector of terms. *Bag of words* is one of the simplest and best methods that can be used to represent the document as set of words in the document. When number of words in the document is considered, the high dimensionality is an important problem. To overcome the problem of high dimensionality, dimensionality reduction techniques can be used. Thus often leads to better performance and accuracy.

Some of the language dependent dimensionality reduction methods like stemming and stop word removal can be applied to reduce the dimensionality in a reasonable amount. Moreover, feature selection is another process that reduces the dimensionality by scoring the terms considering the importance of the term in the corpus, and selecting the terms with highest scores. Like language dependent reduction methods, feature

selection not only reduces the directionality but also improves the performance and the accuracy.

In text categorization there are two main policies to apply feature selection: local policy and global policy. The local policy, where a different set of features is selected from each class independent from other classes, gives equal weight to each class. Thus, it tends to optimize the classification performance on frequent and infrequent classes by selecting the most important features for each class. On the other hand, the global policy, where a single set of features is selected from all classes, provides a global view of the entire dataset by extracting a single global score from the local scores. Thus, the global policy tends to penalize the infrequent classes in highly skew datasets by selecting the most important features for the entire dataset. In this study we propose a new policy called document policy. By using this policy, for every document feature selection is done individually independent from other document or classes. Thus a document having fewer terms is not dominated by documents having many terms.

Considering all the traditional text categorization techniques, the meaning of text is missing in the picture. By using semantic features in categorization better results can be obtained. *Part of speech tags* of terms and relations in *WordNet* such as synonyms, hypernyms, hyponyms, meronyms and topics of terms are types of semantic features that can be used. Knowing part of speech tag of a term is important information since not all word forms tells same about the document content. For example nouns are usually more descriptive than adverbs. Incorporating *WordNet* features with document's features can be important contribution especially when there are fewer terms in the document. Thus the documents having less terms will have more terms and be represented better. Incorporating those features and information into document makes it richer in terms of content. Even though the running time performance may degrade a little bit but the better categorization accuracy will be taken.

1.1. Related Works

The arrival of the machine learning methods in the text categorization field is one of the most important factors that accelerate the improvement in this field by strong theoretical motivations. A growing number of machine learning methods have been used for text categorization such as probabilistic classifiers, decision trees, nearest neighbor classifiers and neural networks. [1]. In 1995 a new machine learning method Support Vector Machines (SVMs) were introduced by Vapnik [2]. In later years, many studies have explored the use of SVMs for text categorization with promising results [3–6]. One of the most basic studies that introduce SVMs for text categorization is presented by Joachims in 1998. In the study the performance of SVM using non-linear model is compared with four popular machine learning algorithms (Naïve Bayes (NB) classifier, Rocchio method, k-nearest neighbour (k-NN) classifier and C4.5 decision tree) on Reuters and Ohsumed datasets. The analysis concludes that SVM is very well suited for test categorization and significantly outperforms other methods.

Reducing dimensionality is another critical issue in text categorization. Feature selection is one of the effective methods that improves the efficiency and accuracy of the classifiers by selecting only more discriminative terms in a dataset as features. In the literature, various feature selection methods have been presented and analysed [3].

The use of semantic features in text categorization is usually done with WordNet in the literature. Chua and Kulathuramaiyer, 2004 studied on selecting features using WordNet and stated that using WordNet for feature selection is promising [7]. Zhang, 2004, uses semantic features for selecting features and asserted that using WordNet is a good resource for text classification. In the study, they tried to extract semantically related features. And this method can improve the accuracy of categorization on dataset that have semantically distinct classes [8]. Bloehdorn, 2004, introduced a new method called *AdaBoost*. AdaBoost, was proposed to perform the final classifications based on the classical word vector representations and the conceptual features. In the study, two features lists were integrated to get better results [9]. Li and Zhao, 2009 use semantic features in their studies. They use WordNet to find semantic features of

category names. Later category name's semantic features is used for labelling of the documents [10].

1.2. Motivation

The traditional text categorization techniques; feature selection, term weighting, classifiers are saturated field of study. It is not easy to contribute more. But the semantic in text categorization can be studied further. There are some studies in the literature that investigate the usage of semantic features in text categorization by using WordNet. Most of the studies that use WordNet state that the accuracy of classification increased reasonably.

Zhang *et al.*, 2004, propose a WordNet based approach for feature selection. Instead of evaluating the terms as isolated feature, they use the relationship of terms. For this purpose WordNet is used to find the most relevant synset. In this study the most relevant synsets are found by simple thresholds. For example; if a term has more than five meaning, remove the term. Moreover this study only uses nouns, other part of speech tagged terms are removed [8].

Mansury *et al.*, 2006, evaluate the use of WordNet in text categorization and consider not only synonyms, hypernyms and hyponyms but also meronyms and holonyms. In this study they create a model with WordNet features and see that synonyms and hypernyms increase the accuracy and hyponyms decreases [11].

Furthermore, use of semantic features introduces ambiguity. One word may have tens of meanings and only some of the meanings fits in the context. When you ask WordNet for its synonyms it will give you all the meaning it has. To solve this problem we need to consider the term with its environment document.

In this study we will investigate the contribution of semantic features in a vast perspective. We will use part of speech tags to tag every word and only consider nouns, verbs, adjectives, adverbs. And then, we will find semantic features by using WordNet

and apply our disambiguation method to overcome the ambiguity.

1.3. Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 describes our document presentation and preprocessing steps. Chapter 3 gives general information about WordNet. Chapter 4 gives details of used feature selection metrics and their formulations. Chapter 5 explains the contribution of semantic features in text categorization. In Chapter 6 the system architecture is explained by flow charts. In Chapter 7 we explain the created environment for the analysis. Chapter 8 discusses the taken results. At the end, Chapter 9 concludes the thesis and gives the future search.

2. DOCUMENT REPRESENTATION AND PREPROCESSING

Document representation is the process of transforming the unstructured text into a structured data as a vector in order to classify the text documents by applying machine learning techniques. The most widely used method for document representation is the *vector space model* introduced by Salton and associates in 1975 [6]. In vector space model, each document is represented as a vector d and each dimension in the vector corresponds to a distinct term in the term space of the document collection [12].

In this study, we use *bag of words* in vector space model which define each term as a distinct single word. The bag of words representation is a very simple and preferred method that makes the representation and learning highly efficient and easy by ignoring the order and meaning of distinct words [13]. Although it is a simple method, high dimensionality becomes an important issue when terms are defined as single words in the feature space. In order to reduce the high dimensionality, we apply some preprocessing methods which are described by the following sections:

Reducing dimensionality is another critical issue in text categorization. Feature selection is one of the effective methods that improves the efficiency and accuracy of the classifiers by selecting only more discriminative terms in a dataset as features. In the literature, various feature selection methods have been presented and analysed.

2.1. Parsing The Document

In the first step, all the HTML mark-up tags and non-alphabetic characters such as numerals, special characters and date are removed from the documents in the dataset. Then case-folding is applied to convert all characters into same case *lower case* in order to avoid the duplication of the same words.

2.2. Removing Stopwords

Overly common words, such as pronouns, prepositions and conjunctions in English, like *it*, *in* and *and*, occur so frequently that they cannot give any useful information about the content and be discriminatory for a specific class. These words are called *stopwords*. We use the stopword list that was built by Salton and Buckley for the SMART system at Cornell University to eliminate common words. The list consists of 571 words is given in Appendix A

2.3. Stemming

Removing stopwords causes an efficient reduction in the dimensionality of the feature space but we also need stemming word to reduce the dimensionality of the feature space to a reasonable number. Stemming is a preprocessing for finding the root morphemes of the words. In order to stem the words, we use Porter's Stemmer which is the most widely used algorithm for word stemming in English. Porter's Stemming Algorithm is a process for removing the common morphological and inflexional affixes from words [14, 15]. In other words, it is based on only morphological issues that are completely independent from the syntactic and semantic structure of the sentence. For example, the words *computer*, *computers*, *computing* and *computes* are stemmed the same root *comput*. After stemming, terms that we left with a single character are also removed since they cannot give any information about the content of a document.

2.4. Term Weighting

As already mentioned at the beginning of this section we represent each document as a vector d

$$d = (w_1, w_2, ..., w_n)$$

where w_i is the weight of term i in document d. There are several ways to compute these term weights [16]. There are three main assumptions that are valid for all computations

[17].

- Rare terms are no less important than frequent terms,
- Multiple appearances of a term in a document are more important than single appearances,
- Long documents are no more important than short documents.

The term frequency-inverse document frequency (*Tf-Idf*) weighting is one of the widely used weighting methods that take into account these properties. *Df* formula meets the first assumption, *Tf* formula meets the second assumption and length-normalization meets the third assumption which given above. Thus we apply *Tf-Idf* weighting method in this study whose formula is given below:

$$w_{ij} = tf_{ij} \log(\frac{N}{df_i})$$

where w_{ij} is the weight of a term i in document j, tf_{ij} denotes the frequency of the term i in document j, df_{ij} denotes the number of documents in which a term i occurs in the whole document collection and N is the total number of documents.

The $tf - idf$ weighting considers that if a term are often occurs in a document, it is more discriminative whereas if it appears in most of the documents, then it is less discriminative for the content.

3. WORDNET

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet's structure makes it a useful tool for computational linguistics and natural language processing [18].

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus do not follow any explicit pattern other than meaning similarity [18, 19].

3.1. Structure

The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms–words that denote the same concept and are interchangeable in many contexts–are grouped into unordered sets (synsets). Each of WordNet's 117 000 synsets is linked to other synsets by means of a small number of *conceptual relations*. Additionally, a synset contains a brief definition (*gloss*) and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique [18, 19].

3.2. Relations

The most frequently encoded relation among synsets is the super-subordinate relation (also called hyperonymy, hyponymy or *is a* relation). It links more general synsets like furniture, piece of furniture to increasingly specific ones like bed and bunkbed. Thus, WordNet states that the category furniture includes bed, which in turn includes bunkbed; conversely, concepts like bed and bunkbed make up the category furniture. All noun hierarchies ultimately go up the root node entity. Hyponymy relation is transitive: if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture. WordNet distinguishes among Types (common nouns) and Instances (specific persons, countries and geographic entities). Thus, armchair is a type of chair; Barack Obama is an instance of a president. Instances are always leaf (terminal) nodes in their hierarchies [18, 19].

Meronymy, the part-whole relation holds between synsets like chair and back, backrest, seat and leg. Parts are inherited from their super ordinates: if a chair has legs, then an armchair has legs as well. Parts are not inherited *upward* as they may be characteristic only of specific kinds of things rather than the class as a whole: chairs and kinds of chairs have legs, but not all kinds of furniture have legs [18, 19].

4. FEATURE SELECTION

Text categorization is a supervised learning task that assigns the predefined category labels to new documents based on the likelihood derived from a set of labeled training documents. In order to classify documents, each document should be transformed into a model that preserves as much of the original information as possible. The bag of words representation is one of the simple and preferred models that represents a document as a set of distinct words by ignoring the order and meaning of words. When the number of words in documents is considered, high dimensionality may become an inevitable problem. Since the data in text categorization are high-dimensional, naturally dimensionality reduction becomes a necessity for efficiency and accuracy.

Feature selection is one of the well-known processes that reduces the dimensionality by ranking all features according to their importance estimated by a metric and then selecting ones with the highest values. Feature selection not only reduces time and storage requirements but also improves the efficiency and accuracy of the classifiers. Feature selection makes applying classifiers on data more efficient by reducing the size of the effective features. In addition, feature selection often improves accuracy of the classification by eliminating noise features that are non-informative and misleading for classification and lead to incorrect generalization (over fitting) from the training set.

4.1. Global, Local and Document Policy

In text categorization there are two main policies to apply feature selection: local policy and global policy. In the first policy, a different set of features is selected from each category. In the second policy, a single set of features is selected from all categories.

The local policy, where a different set of features is selected from each class independent from other classes, gives equal weight to each class. Thus, the local policy tends to optimize the classification performance on frequent and infrequent classes by

selecting the most important features for each class. On the other hand, the global policy, where a single set of features is selected from all classes, provides a global view of the entire dataset by extracting a single global score from the local scores. Thus, the global policy tends to penalize the infrequent classes in highly skewed datasets by selecting the most important features for the entire dataset [20, 21].

We propose a new policy called document policy. In this policy, as opposed to the given policies, selection is done for every document individually. By this policy a document having few terms will not be dominated by documents having many terms.

There are several ways to obtain global score from the local scores: maximization, averaging, weighted averaging and weighted maximum are the most popular globalization techniques. Maximization, averaging and weighted averaging were presented by Yang and Pedersen in 1997 and weighted maximum was proposed by Calvo and Ceccatto in 2000 [22, 23]. We selected maximization as a globalization technique, since it consistently outperformed other globalization techniques in the study of Debole and Sebastiani. In their paper, the success of the maximization was explained that it prefers to select terms that are good separator even on a single category rather than terms that are only fair separators on many categories. The formulation of computing the maximization is given below.

$$f_{max}(t_k) = \max_{1 \leq i \leq |C|} f(t_k, c_i)$$

Where C is the set of categoris, t_k term k and c_i is category i. $f_{max}(t_k)$ calculates the maximum category score of term t_k.

In this study, first of all the five widely used feature selection metrics: term frequency-inverse document frequency (tf-idf), chi-square statistics (CHI), information gain (IG), Accuracy2 (Acc2) and document frequency thresholding (DF) are analyzed. But since we only want to measure the contribution of semantic features, CHI is selected as it outperforms in most of the scenarios.

4.2. Chi-square Statistics (CHI)

In experimental sciences, chi-square statistics is frequently used to measure how the observation results differ from the expected results. In other words, it measures the independence of two random variables.

$$CHI = \sum_{ij} \frac{(Observed_{ij} - Expected_{ij})^2}{Expected_{ij}}$$

Chi-square statistics is also widely used in text categorization [3, 20, 24]. In text categorization, the two random variables are occurrence of term t_k and occurrence of class c_i and chi-square statistics measures the independence between t_k and c_i. The formula for chi-square score is:

$$CHI(t_k, c_i) = N \times \frac{[P(t_k, c_i)P(\bar{t_k}, \bar{c_i}) - P(\bar{t_k}, c_i)P(t_k, \bar{c_i})]^2}{P(t_k)P(\bar{t_k})P(c_i), P(\bar{c_i})}$$

where $P(t_k)$ is the percentage of documents in which term t_k occurs, $P(\bar{t_k})$ is the percentage of documents in which term t_k does not occur, $P(c_i)$ is the percentage of documents belonging to class c_i $P(\bar{c_i})$ is the percentage of documents not belonging to class c_i , $P(t_k, c_i)$ is the percentage of documents belonging to class c_i in which term t_k occurs, $P(\bar{t_k}, \bar{c_i})$ is the percentage of documents not belonging to class c_i in which term t_k does not occur, $P(\bar{t_k}, c_i)$ is the percentage of documents belonging to class c_i in which term t_k does not occur and $P(t_k, \bar{c_i})$ is the percentage of documents not belonging to class c_i in which term t_k occurs. If chi-square score of a term t_k is low value, this means t_k is independent from the class ci and if chi-square score of a term t_k is high value, this means t_k is dependent of the class c_i . Thus the chi-square feature selection method selects the terms with the highest chi-square score which are more informative for classification.

5. CONTRIBUTION OF SEMANTIC FEATURES

This section discusses the contribution of semantic features in text categorization. Various researches have been done to improve the performance of text categorization. In the field of text categorization the majority of studies focused on feature selection metrics, classifiers and actually the studies in the field is quite saturated. We will focus on contribution of semantic features rather than feature selection and machine learning techniques. By incorporating the semantic features and information into text categorization we can improve the accuracy of the classification.

Traditional text categorization techniques are not aware of the language; terms are evaluated as meaningless symbols. Incorporating semantic features into text categorization will add meaning into categorization process. It will be good to know part of speech tag of a word as not all the word forms contribute the same. It can be said that nouns have more meaning than adverbs. In addition, synonyms, hypernyms, hyponyms, meronyms and topic information about every single word can add more in the performance of categorization.

There are very powerful tools and techniques to be used in the name of semantic. WordNet is one of the most powerful tools that has a stable English database (There are also other language databases available that implement the same model such as Spanish, Chinese, etc.). In addition there are part of speech tagging libraries available to be used easily.

5.1. Part of Speech Tag

Every term in the document has a part of speech tag such as noun, verb, adjective and adverb. As human, we can see that not all the word forms contribute to the meaning of a document in the same amount. For example it is expected that adverbs are kind of transition words and do not tell much about the content in the document, whereas nouns tells much more. Thus we analyze the contribution of each word form

in text categorization.

Part of speech tagger is a lexicon based library developed by Mark Watson [25]. The library accepts at least one sentence and returns the tokenized terms with part of speech information.

5.2. WordNet Relations

We will use WordNet to find the synonym, hypernym, hyponym, meronym and topic information of a given word. WordNet stores terms in synsets, and every synset has relations to other synsets. The relations can be hypernym, hyponym, meronym and topic. WordNet has many more relations but in this study we will use the specified ones. In Figure 5.1 the semantic features that are used in this study can be seen. We add those semantic features into the term list just as other terms found in the document.

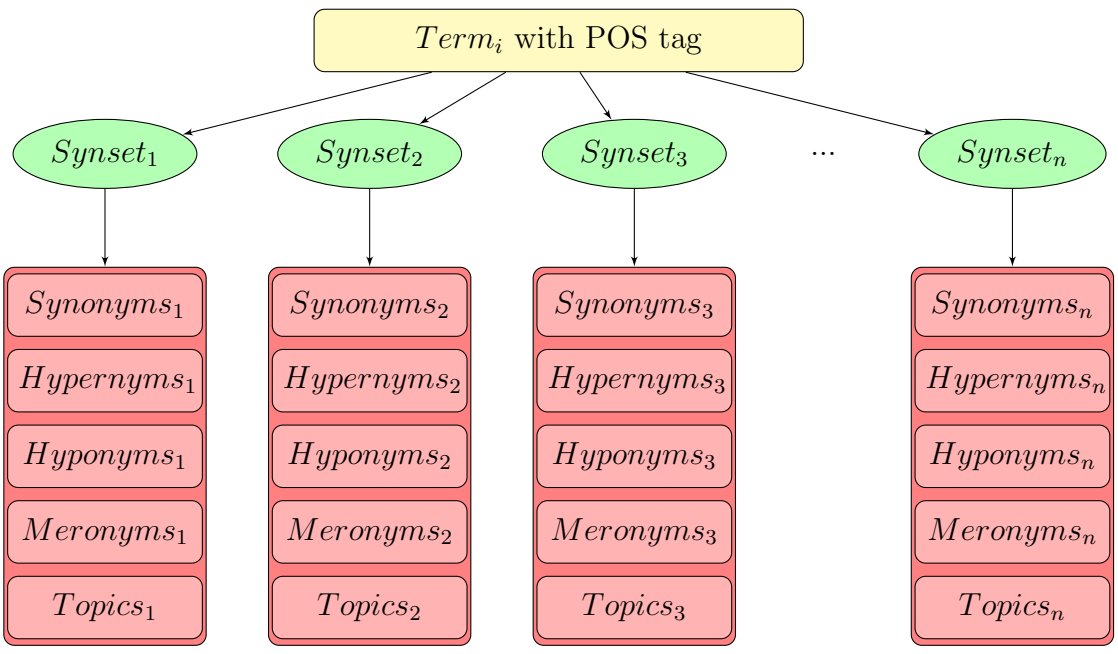

Figure 5.1. Term's Semantic Features.

One of the problems can be faced when we add those semantic features into the term list is that not all synsets are really related to the context of the document. Thus disambiguation is applied. There is detailed information about disambiguation in the

next section in this chapter.

5.3. Disambiguation

Disambiguation in Wikipedia is the process of resolving the conflicts that arise when a single term is ambiguous [26]. In this study there are many synsets for every word, we cannot really say that all the synsets are related to the context of the document. WordNet tries to do the best it can to disambiguate the irrelevant synsets. But it still does not know the context the term is in. We have to find a mechanism to eliminate the non-related synsets for the document.

In Figure 5.1 a single term's synsets and their semantic features can be seen. When we think of a document; there are other terms and their synsets. We will consider all the terms and their synsets to do disambiguation. In consequence a term can be represented as

$$t_i = \{s_1, s_2, ..., s_m\}$$

where t_i is term synset set and s_i is any synset that contains synonyms, hypernyms, hyponyms, meronyms and topics. And a document can be represented as

$$d_i = \{t_1, t_2, ..., t_n\}$$

And as a union of synset of each terms, document synset can be represented as

$$dS_i = \{s_1, s_2, ..., s_k\}$$

We come up with a scoring metric to calculate a score for every synsets in dS_i and then, apply a threshold to select the synsets that receives the best scores. Score calculation is described in next section of this chapter.

5.3.1. Disambiguation Score Calculation

In this section score of each synset in the document will be calculated. Every synset has synonyms, hypernyms, hyponyms, meronyms and topics that are found by using WordNet. We will use those features to identify the synset. Hypernyms and topics of synsets can be used for identification. Hypernyms tells us root or more general concept of the words. Where as topics tells us topic information of words. We will analyze both of them for disambiguation process. Once we identify the synset we will calculate score of every synset by calculating the similarity of the synset with all other synsets. Thus, we will use the total similarity as score of the synset. After scoring phase, we will simply apply a threshold to select the synsets of documents.

$$Score(s_i) = \sum_{j=0, i \neq j}^{k} Similarity(s_i, s_j)$$

where $Score(s_i)$ denotes the scores of s_i in dS_i and $Similarity(s_i, s_j)$ is defined as

$$Similarity(s_i, s_j) = CommonCount(Hypernyms(s_i), Hypernyms(s_j))$$

$$Similarity(s_i, s_j) = CommonCount(Topics(s_i), Topics(s_j))$$

where $Hypernyms(s_i)$ and $Topics(s_i)$ denotes the hypernym term list and topic term list respectively for synset s_i and $CommonCount(Hypernyms(s_i), Hypernyms(s_j))$ denotes the number of common terms in given two list. The matching rules is a little bit loose; If one term contains in another term's content we can say it is match.

6. SYSTEM ARCHITECTURE

In this section we present our text categorization system that consist of the different blocks. We characterize the functionality of each block and describe the iterations between the blocks. Figure 6.1 describes the overall flow of the system. And in the following sections we will describe each individual block in detail.

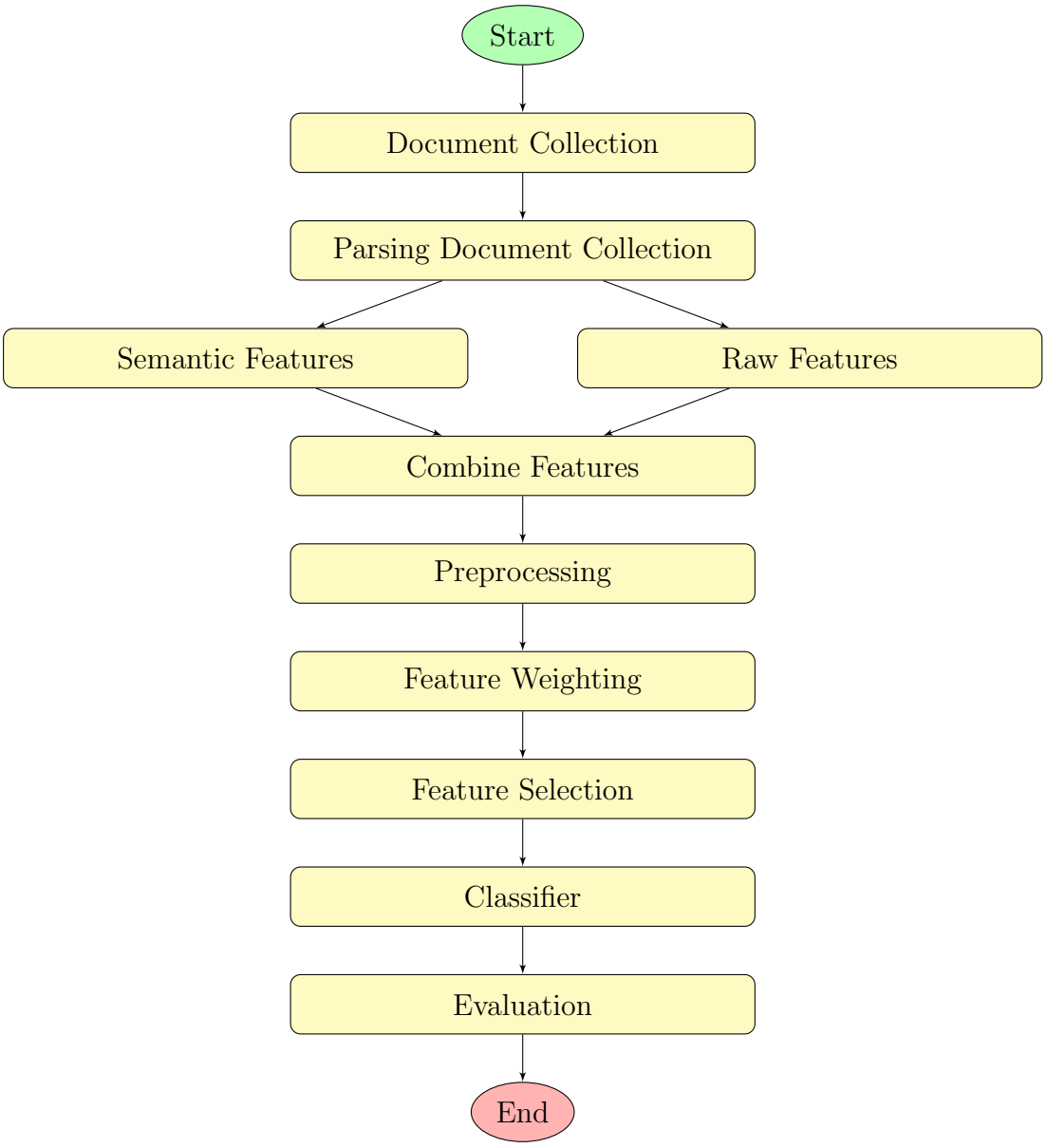

Figure 6.1. High Level System Architecture.

6.1. Document Collection

The data can be in any format; some of the datasets stored in a single file managed by HTML tags, while some have a distinct file for every document. In the system, the data parsing is implemented differently for every data collection. At the end of this process, each document represented as a single string.

6.2. Semantic Features

Finding semantic features is done for every document individually. This process starts with Part of Speech Tagging and then for nouns, verbs, adjectives and adverbs, by using WordNet synonyms, hypernyms, hyponyms, meronyms and topics are found. The flow can be seen in Figure 6.2.

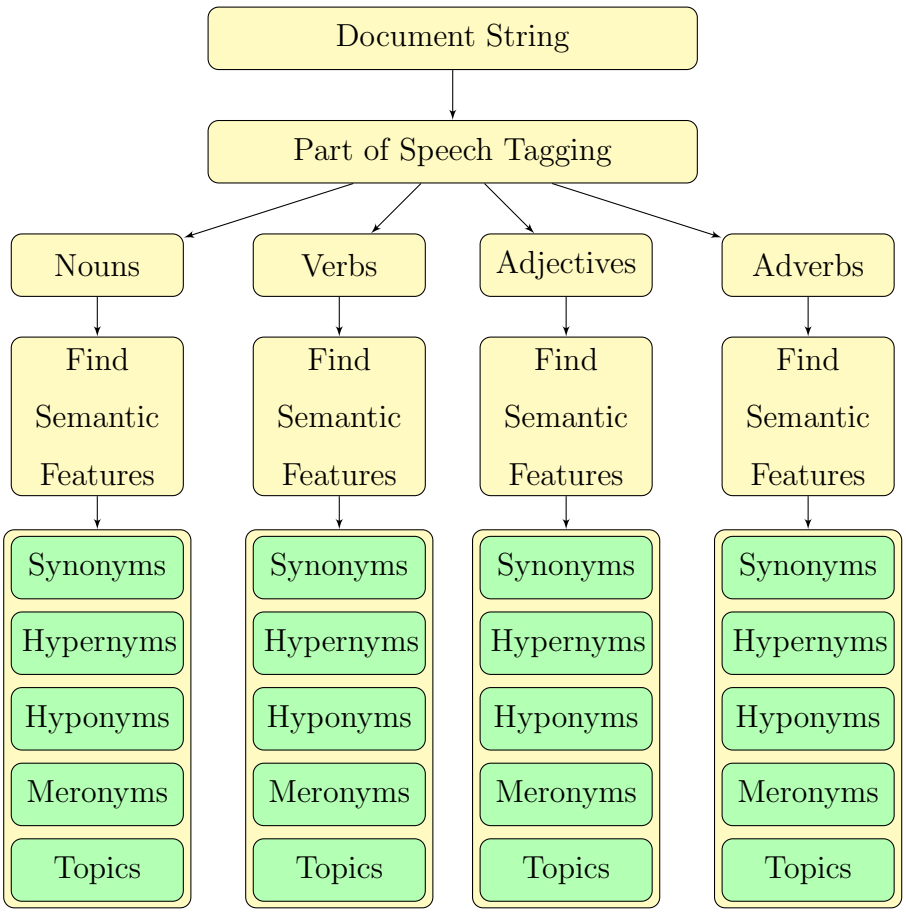

Figure 6.2. Semantic Features System Architecture.

In Figure 6.2, it seems that all the semantic features are managed in a single list. But they are managed in synsets. Each synset contains synonyms and relations (hypernyms, hyponyms, meronyms and topics) of this term. We will use these relations for eliminating the ambiguity. It is not shown here in the architecture, but at the end of the process in Figure 6.2 the disambiguation process takes place.

6.3. Raw Features

In this section the raw features will be extracted from the document string. This process simply parse the document string to remove non-alphabetic characters such as numerals, special characters and date. Then case-folding is applied to convert all characters into same case, in this study to lower.

6.4. Combine Features

In this section features found in raw features and semantic features are combined, by a given configuration. This configuration may state something like that, only get raw features, noun and noun's synonyms. At the end of this section a single list of terms is obtained to be used in the next section.

6.5. Preprocessing

In this section, the term list prepared in the previous section will be processed. In this process the stopwords will be removed and stemming will be applied. The flow can be seen in the Figure 6.3.

6.6. Feature Weighting

In this section the weight of features are calculated by *Tf-Idf* feature weighting method. The details can be found in Chapter 2.

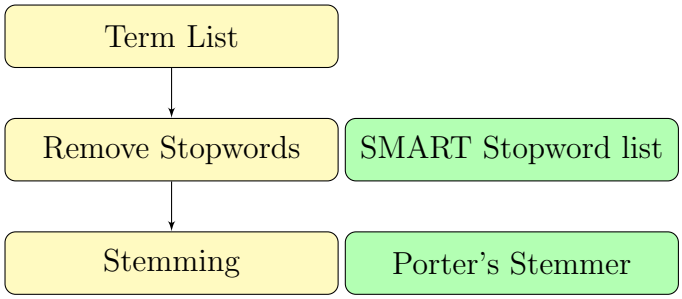

Figure 6.3. Preprocessing System Architecture.

6.7. Feature Selection

In this section features are scored by CHI-Square feature selection metric. It is a category terms' based scoring metric. After scoring the selection is done by given threshold and policy. The policy can be global, local and document policy. The flow can be seen in the Figure 6.4

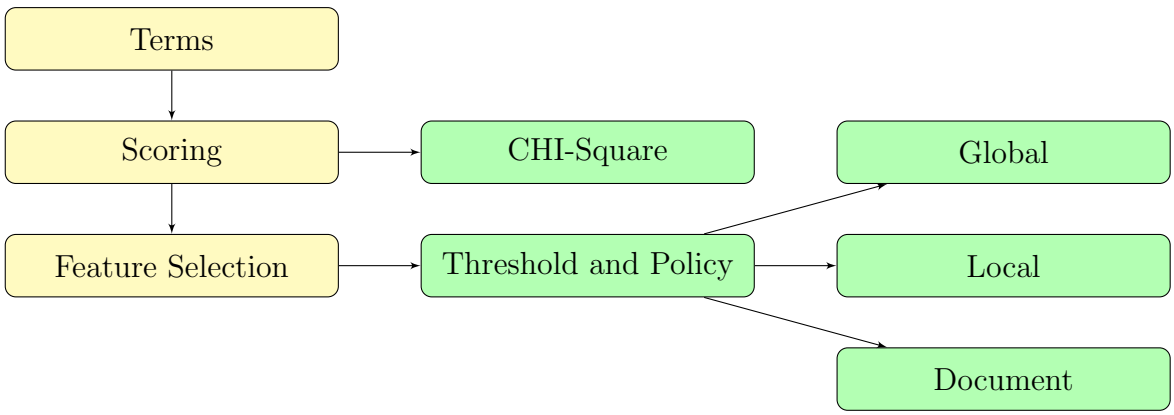

Figure 6.4. Feature Selection System Architecture.

6.8. Classfier

In this section the classification procedure is explained. *SVM-Light* is used as classifer. In Figure 6.5 the flow can be seen.

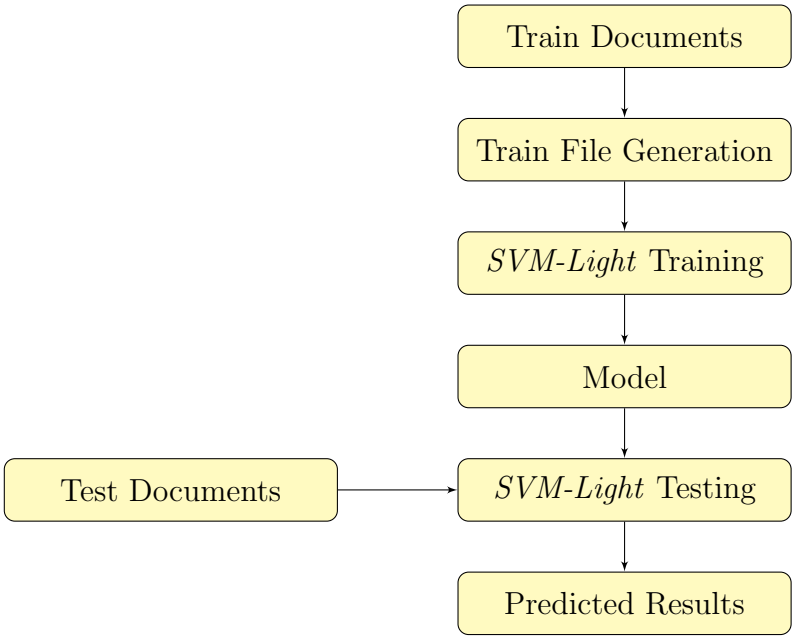

Figure 6.5. Classifier System Architecture.

6.9. Evaluation

In this section the success measurement is done. The flow can be seen in the Figure 6.6.

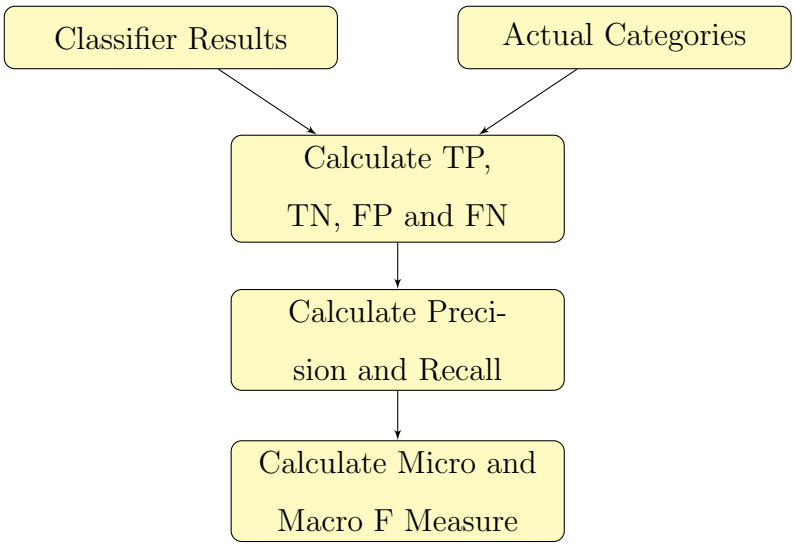

Figure 6.6. Evaluation System Architecture.

7. EXPERIMENTAL SETUP

7.1. Classifier

In this study, we use SVM as a classifier which outperformed other classification methods in text categorization consistently in previous studies [3–6, 27–30].

Support Vector Machine was introduced as a statistical learning theory by Vapnik in 1995 at AT&T Bell Laboratories [31]. It is based on the Structural Risk Minimization principle from computational learning theory [2]. The basic idea of this principle is to find a hypothesis for guarantee the minimum true error. In here, true error means the probability that a hypothesis will make an error on an unseen and randomly selected test example [6]. SVM is designed for solving two-class problems and the idea behind SVM is to find a hyperplane in n-dimensional space that separates the positive training examples from the negative examples with the largest possible margin in order to determine the best separation between the two classes. In text categorization, learned hyperplanes separates the documents in input space that belong to different topics.

One of the reasons the success of the SVM in text categorization is its capability in very high dimensional feature vectors, given that these vectors are sparse [28]. Because the learning process of SVM is independent from the dimensionality of the feature space while it measures the complexity of hypothesis according to the margin which means it separates the data instead of the number of feature [6]. Another feature that distinguishes SVM from other classifiers is the generalization ability. The decision function is determined by assuming that the training data which belongs to different classes does not overlap with each other. So the distance from the training data is maximized and in this way SVM prevents overfitting to training data [31]. In addition to, SVM provide a fast and effective classification that can easily incorporate new documents [27]. Thus, we can say that SVM ideally suitable for text categorization.

In the study, we use SVM_{light} with the default parameter settings that a linear

kernel has been used. The SVM_{light} system is a very efficient implementation of SVMs that was developed by Joachims, 1999, at the University of Dortmund and has been commonly used in previous studies

7.2. Datasets

We perform our experiments on five standard datasets, widely used in text categorization research. The properties of these datasets are summarized in 7.1. We divide these five datasets into 3 categories according to their skewness. The skewness is calculated by dividing the standard deviation of the class distribution by the mean of the distribution. The first two datasets: 20Newsgroup and Classic3 are homogeneous datasets. One difference between them is the class relatedness. In Classic3 all the classes are nearly equally well represented in the training set and each class is disjoint from each other clearly, whereas in 20Newsgroup the classes are closely related to each other. 7Sectors is categorized as skew datasets in our study because it is neither homogeneous as Classic3 nor highly skew as the WebKB and Reuters. Finally, WebKB and Reuters-21578 are categorized as a highly skewed dataset with varying class distributions. These datasets are particularly hard to categorize since the rare classes are dominated by the common classes. In addition, there is a strong semantic overlap between the topics since both WebKB and Reuters consist of general topics that are very close to each other and share many common terms. In order to divide the Reuters-21578 dataset into training and test sets, we use ModApte splitting method that has been mostly used in the literature. In Chapter 8, we discuss the property of each dataset in more detail.

7.3. Performance Measures

To evaluate the performance of the contribution of text categorization, we use the commonly used F-measure metric which is equal to the harmonic mean of recall ρ

Table 7.1. Properties of Datasets.

Datasets	# of documents	# of train documents	# of test documents	# of terms	# of classes	min class size	max class size	Skewness (sd/mean)
20Newsgroup	18846	11314	7532	90812	20	628	999	homogeneous (0.11)
Classic3	3891	2699	1192	10930	3	1033	1460	homogeneous (0.14)
7Sectors	3308	2181	1127	56314	7	290	949	skew (0.45)
WebKB	5396	4740	656	102285	4	182	3160	highly skew (0.81)
Reuters-21578	12902	9603	3299	20308	90	2	2964	highly skew (3.32)

and precision π [32]. They are defined as:

$$\pi = \frac{TP}{TP + FP}, \rho = \frac{TP}{TP + FN}$$

$$F = \frac{2\pi\rho}{\pi + \rho}$$

The idea behind the F-measure can be explained in Figure 7.1. The right circle represents the all defective set and the left represents the set that were classified as defective by a classifier. The intersection between these sets represents the true positive (TP) while the remaining parts represent false negative (FN) and false positive (FP). Accuracy of the classifier is defined by measuring the extent of the intersection between the two sets [33].

Since the absolute size is not meaningful, this value should be normalized by the proportional area. The F-measure is defined as:

$$F = \frac{2(TP)}{FP + FN + 2(TP)}$$

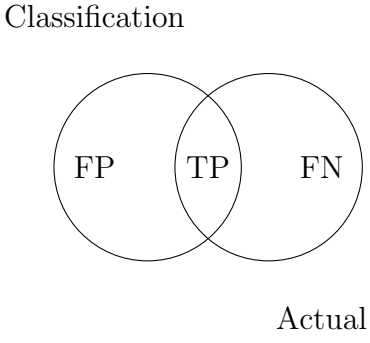

Figure 7.1. F-Measure Demonstration.

The F-measure values are in the interval [0-1]. When the two sets are identical, F-measure obtains the highest value and it obtains the lowest value when the two sets are mutually exclusive. Thus, larger F-measure values correspond to higher classification quality. F-measure can be computed by two different alternatives, micro-averaged F-measure and macro-averaged F-measure. In this way, the overall F-measure score of the entire classification problem can be computed by using these different types of averaging methods [32].

Micro-averaged F-measure gives equal weight to each document and therefore it tends to be dominated by the classifier's performance on common categories while reflects the overall accuracy better. Precision and recall are obtained by summing over all individual decision:

$$\pi = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}, \rho = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i}$$

where C indicates the number of categories.

$$\textit{Micro-average F-Measure} = \frac{2\pi\rho}{\pi + \rho}$$

On the other hand Macro-averaged F-measure gives equal weight to each category regardless of its frequency and thus it is influenced more by the classifier's performance on rare categories. Precision and recall are first computed locally for each category and

then F-measure is computed globally by averaging over the decisions of all categories:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}, \rho_i = \frac{TP_i}{TP_i + FN_i}$$

$$F_i = \frac{2\pi_i \rho_i}{\pi_i + \rho_i}$$

$$\textit{Macro-averaged F-Measure} = \frac{\sum_{i=1}^{M} F_i}{M}$$

In text classification, TP_i is the number of documents that are assigned correctly to class i. FP_i is the number of documents that are assigned incorrectly to class i by the classifier but which actually do not belong to class i and FN_i is the number of documents not assigned to class i by the classifier but which actually belong to class i.

8. RESULTS AND DISCUSSION

In this section the detailed results for all the dataset used will be given and discussed. Firstly, for every dataset, the details about the dataset will be given. And then, we will demonstrate the contribution of semantic in text categorization in four distinct sub-sections:

- Analysis of Existing Metrics.
- Contribution of POS.
- Contribution of WordNet features.
- Contribution of disambiguation.

In analysis of existing metrics section, we will compare the use of different policies with different thresholds. We manage the thresholds in two ways. In static; we select the given number of features, while in percentage; we select the given percentage of the terms. In addition we have three policies: global, local and document policies. The aim in this section is to find the best policy with best threshold.

In contribution of POS section, we will measure the contribution of being aware of part of speech tags of terms. In this study we only focus on four POS: noun, verb, adjective and adverb. In the results we will discuss the contribution of each of them both individually and jointly. For example in one configuration, we can say use nouns and verbs, while in another configuration we may say use nouns and adjectives. In this section, the aim is to find the best combination.

In contribution of WordNet features section, we will measure the contribution of using WordNet features. We use synonyms, hypernyms, hyponyms, meronyms and topics features. In the results for every dataset, we will use global policy with 1000, 1500 and 2000 feature selected configuration. The aim is to compare the use of WordNet feature with no semantic features option.

In contribution of disambiguation section, we will apply our disambiguation method to eliminate the ambiguity. Disambiguation process works with a threshold value that represent the elimination of synset amount. The more the value low the more synset will be eliminated. The aim in this section is to measure success of classifier with the removal of ambiguity.

8.1. 20 Newsgroup Dataset

8.1.1. Property of the Dataset

As explained in the previous chapter, this dataset classified as homogeneous dataset. The dataset consists of 20000 messages taken from 20 newsgroups. 1000 articles were taken from each of the following 20 newsgroups:

- alt.atheism
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast

Table 8.1. Static FS(CHI), Micro-F Measure for 20Newsgroup Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.8757	0.9022	0.9110	0.9116	0.9120	0.9122	0.9122	0.9122	0.9120	0.9120
Local	0.7837	0.8110	0.8254	0.8442	0.8531	0.8708	0.8802	0.8843	0.8870	0.9120
Global	0.6612	0.6853	0.7175	0.7464	0.7820	0.8286	0.8529	0.8663	0.8753	0.9120

Table 8.2. Static FS(CHI), Macro-F Measure for 20Newsgroup Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.8631	0.8908	0.9000	0.9015	0.9024	0.9030	0.9028	0.9029	0.9026	0.9026
Local	0.7628	0.7943	0.8108	0.8302	0.8386	0.8575	0.8676	0.8704	0.8733	0.9026
Global	0.6457	0.6499	0.6795	0.7098	0.7413	0.8033	0.8366	0.8513	0.8617	0.9026

- talk.politics.misc
- talk.religion.misc

Approximately 4% of the articles are cross-posted. The articles are typical postings and thus have headers including subject lines, signature files, and quoted portions of other articles. It can be observed that the marginal distributions of the articles among different newsgroups are not identical. There exists distribution shift from one newsgroup to any other newsgroups. However, we observe that some newsgroups are related. For example, the newsgroups rec.autos and rec.motorcycles are related to car. The newsgroups comp.sys.mac.hardware and comp.sys.ibm.pc.hardware are related to hardware, etc.

8.1.2. Analysis of Existing Metrics

In Table 8.1 and 8.2, it can be seen that document policy gives better results. For selection amount 500 both Micro and Macro F measures give better results than no selection option and others.

In Table 8.3 and 8.4, global policy gives better results than other policies and for selection amount greater than 50%, it gives better results than no selection option.

Table 8.3. Percentage FS(CHI), Micro-F Measure for 20Newsgroup Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.8430	0.8625	0.8782	0.8881	0.8949	0.8978	0.9022	0.9064	0.9084	0.9110	0.9120
Local	0.8720	0.8793	0.8863	0.8908	0.8924	0.8976	0.9004	0.9048	0.9071	0.9105	0.9120
Global	0.8944	0.9017	0.9067	0.9122	0.9120	0.9125	0.9116	0.9123	0.9125	0.9125	0.9120

Table 8.4. Percentage FS(CHI), Macro-F Measure for 20Newsgroup Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.8319	0.8515	0.8663	0.8767	0.8845	0.8870	0.8917	0.8963	0.8986	0.9012	0.9026
Local	0.8589	0.8659	0.8720	0.8788	0.8807	0.8872	0.8897	0.8943	0.8970	0.9009	0.9026
Global	0.8835	0.8912	0.8967	0.9025	0.9024	0.9030	0.9021	0.9030	0.9033	0.9033	0.9026

Table 8.5. Contribution of POS for 20Newsgroup - Micro-F Measure.

Contribution	1000	1500	2000
Raw	0.8529	0.8663	0.8753
Noun	0.8391	0.8523	0.8620
Verb	0.5414	0.5604	0.5668
Adjactive	0.4700	0.4736	0.4845
Adverb	0.1994	0.2105	0.2183
N+V	0.8413	0.8540	0.8644
N+Adj	0.8387	0.8510	0.8610
N+V+Adj	0.8422	0.8545	0.8648
N+V+Adj+Adv	0.8420	0.8547	0.8650

Contribution	1000	1500	2000
Raw	0.8529	0.8663	0.8753
Raw+Noun	0.8596	0.8721	0.8813
Raw+Verb	0.8523	0.8646	0.8755
Raw+Adjactive	0.8532	0.8649	0.8751
Raw+Adverb	0.8527	0.8653	0.8751
Raw+N+V	0.8582	0.8715	0.8799
Raw+N+Adj	0.8590	0.8724	0.8796
Raw+N+V+Adj	0.8573	0.8730	0.8795
Raw+N+V+Adj+Adv	0.8583	0.8731	0.8798

8.1.3. Contribution of POS

In Table 8.5, the contribution of any of the word forms does not results in better results. But with raw features the results are better. Same thing in Table 8.6 : adding word forms only does not gives good results, whereas using word forms with raw terms increases the Macro-F measure. When we evaluate them together, using only given POS, the measures are not increased, but with raw features, the use of noun, adjactive and verbs increases the measures.

8.1.4. Contribution of WordNet Features

In Table 8.7, incorporating WordNet features usually results in better results. When we compare the configurations with the reference configuration, *No Semantic Features*, using WordNet features for the left side table, always increases the results

Table 8.6. Contribution of POS for 20Newsgroup - Macro-F Measure.

Contribution	1000	1500	2000
Raw	0.8366	0.8513	0.8617
Noun	0.8207	0.8355	0.8458
Verb	0.5118	0.5339	0.5413
Adjactive	0.4517	0.4543	0.4645
Adverb	0.1893	0.2016	0.2085
N+V	0.8223	0.8367	0.8480
N+Adj	0.8216	0.8354	0.8463
N+V+Adj	0.8243	0.8389	0.8498
N+V+Adj+Adv	0.8238	0.8389	0.8501

Contribution	1000	1500	2000
Raw	0.8366	0.8513	0.8617
Raw+Noun	0.8454	0.8593	0.8693
Raw+Verb	0.8360	0.8491	0.8619
Raw+Adjactive	0.8384	0.8505	0.8623
Raw+Adverb	0.8364	0.8499	0.8616
Raw+N+V	0.8435	0.8588	0.8677
Raw+N+Adj	0.8455	0.8600	0.8681
Raw+N+V+Adj	0.8436	0.8604	0.8680
Raw+N+V+Adj+Adv	0.8445	0.8609	0.8683

Table 8.7. WordNet Features for 20Newsgroup - Noun(L) and Raw + Noun(R).

Contribution	**Micro-F**	**Macro-F**	**Contribution**	**Micro-F**	**Macro-F**
Synonyms	0.8787	0.8439	Synonyms	0.8903	0.8805
Hypernyms	0.8850	0.8529	Hypernyms	0.9090	0.8891
Hyponyms	0.8912	0.8526	Hyponyms	0.9028	0.8981
Meronyms	0.8812	0.8560	Meronyms	0.8723	0.8846
Topics	0.8736	0.8464	Topics	0.8966	0.8732
Syn+Hype	0.8710	0.8471	Syn+Hype	0.8852	0.8855
Hype+Top	0.8862	0.8458	Hype+Top	0.9016	0.8872
Syn+Hype+Top	0.8812	0.8472	Syn+Hype+Top	0.8736	0.8836
Syn+Hype+Hypo+Mero+Top	0.8710	0.8623	Syn+Hype+Hypo+Mero+Top	0.8978	0.8994
No Semantic Features	0.8620	0.8458	No Semantic Features	0.8813	0.8693

Table 8.8. WordNet Features for 20Newsgroup - Noun + Verb(L) and Raw + Noun + Verb(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.9242	0.8708	Synonyms	0.8607	0.8637
Hypernyms	0.8745	0.8382	Hypernyms	0.8962	0.8871
Hyponyms	0.8992	0.8531	Hyponyms	0.8824	0.8872
Meronyms	0.8833	0.8553	Meronyms	0.8710	0.8867
Topics	0.8833	0.8569	Topics	0.8875	0.8715
Syn+Hype	0.8720	0.8327	Syn+Hype	0.8736	0.8775
Hype+Top	0.8857	0.8575	Hype+Top	0.8937	0.8827
Syn+Hype+Top	0.9088	0.8480	Syn+Hype+Top	0.8723	0.8784
Syn+Hype+Hypo+Mero+Top	0.8770	0.8517	Syn+Hype+Hypo+Mero+Top	0.9023	0.8959
No Semantic Features	0.8644	0.8480	No Semantic Features	0.8799	0.8677

except adding synonyms for Macro-F measure. In Table 8.8, it can be seen that adding semantic features usually increases the success. Micro-F measures always increased, while Macro-F measure decreases only with hypernyms and syn+hypo configuration. On the other hand, in the results shown on the right side, we can conclude that hypernyms, hyponyms and topics increases the success. The result shown in Table 8.9 shows that using semantic features always increases Micro-F measure, and usually increases Macro-F measure. Same thing is true for other Tables 8.10 and 8.11.

Table 8.9. WordNet Features for 20Newsgroup - Noun + Adj(L) and Raw + Noun + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8897	0.8459	Synonyms	0.8848	0.8606
Hypernyms	0.9010	0.8559	Hypernyms	0.9066	0.8751
Hyponyms	0.9097	0.8554	Hyponyms	0.9028	0.8821
Meronyms	0.8833	0.8600	Meronyms	0.8887	0.8731
Topics	0.8858	0.8424	Topics	0.8874	0.8799
Syn+Hype	0.8807	0.8399	Syn+Hype	0.9129	0.8652
Hype+Top	0.9035	0.8503	Hype+Top	0.8809	0.8772
Syn+Hype+Top	0.8820	0.8389	Syn+Hype+Top	0.8887	0.8802
Syn+Hype+Hypo+Mero+Top	0.8820	0.8605	Syn+Hype+Hypo+Mero+Top	0.8874	0.8661
No Semantic Features	0.8610	0.8463	No Semantic Features	0.8796	0.8681

Table 8.10. WordNet Features for 20Newsgroup - Noun + Verb + Adj(L) and Raw + Noun + Verb + Adj(R).

Contribution	Micro-F	Macro-F
Synonyms	0.8830	0.8467
Hypernyms	0.9032	0.8529
Hyponyms	0.8944	0.8606
Meronyms	0.8817	0.8590
Topics	0.8753	0.8420
Syn+Hype	0.9044	0.8683
Hype+Top	0.9167	0.8552
Syn+Hype+Top	0.9020	0.8677
Syn+Hype+Hypo+Mero+Top	0.9302	0.8581
No Semantic Features	0.8648	0.8498

Contribution	Micro-F	Macro-F
Synonyms	0.8860	0.8686
Hypernyms	0.9063	0.8840
Hyponyms	0.8886	0.8794
Meronyms	0.8834	0.8842
Topics	0.8950	0.8719
Syn+Hype	0.8924	0.8779
Hype+Top	0.8975	0.8780
Syn+Hype+Top	0.8834	0.8756
Syn+Hype+Hypo+Mero+Top	0.9088	0.8942
No Semantic Features	0.8795	0.8680

Table 8.11. WordNet Features for 20Newsgroup - Noun + Verb + Adj + Adv(L) and Raw + Noun + Verb + Adj + Adv(R).

Contribution	Micro-F	Macro-F
Synonyms	0.8727	0.8368
Hypernyms	0.8903	0.8462
Hyponyms	0.8940	0.8587
Meronyms	0.9050	0.8488
Topics	0.8903	0.8577
Syn+Hype	0.8841	0.8545
Hype+Top	0.8752	0.8574
Syn+Hype+Top	0.9266	0.8704
Syn+Hype+Hypo+Mero+Top	0.8790	0.8395
No Semantic Features	0.8650	0.8501

Contribution	Micro-F	Macro-F
Synonyms	0.8785	0.8645
Hypernyms	0.9003	0.8725
Hyponyms	0.8901	0.8773
Meronyms	0.8772	0.8764
Topics	0.8901	0.8605
Syn+Hype	0.8863	0.8665
Hype+Top	0.8952	0.8750
Syn+Hype+Top	0.8798	0.8648
Syn+Hype+Hypo+Mero+Top	0.9053	0.8895
No Semantic Features	0.8798	0.8683

8.1.5. Contribution of Disambiguation

In Tables 8.13 , 8.15 , 8.17 , 8.19, if we do not consider a few of the results, Macro-F measures always decreased when disambiguation is applied.

In Tables 8.12 , 8.14 , 8.16 , 8.18 we can say that disambiguation with threshold 70%, increases the Micro-F measure. When we compare the use of topics and hypernyms we can say that using topics in disambiguation gives slightly better results.

8.2. Classic3 Dataset

8.2.1. Property of the Dataset

The Classic3 dataset is a well known collection of documents composed of 3891 abstracts from 3 disjoint research fields as shown in Table 8.20 1398 CRANFIELD

Table 8.12. Hypernyms Disambiguation for 20Newsgroup - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8830	0.8866	0.8883	0.8873
Hypernyms	0.9032	0.8981	0.8998	0.8998
Hyponyms	0.8944	0.8924	0.8979	0.9038
Meronyms	0.8817	0.8892	0.8879	0.8874
Topics	0.8753	0.8747	0.8762	0.8771
Syn+Hype	0.9044	0.9032	0.9015	0.9027
Hype+Top	0.9167	0.9107	0.9144	0.9136
Syn+Hype+Top	0.9020	0.9005	0.8980	0.9001
Syn+Hype+Hypo+Mero+Top	0.9302	0.9342	0.9398	0.9414

Table 8.13. Hypernyms Disambiguation for 20Newsgroup - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8467	0.7859	0.7965	0.8125
Hypernyms	0.8529	0.7774	0.7846	0.7850
Hyponyms	0.8606	0.8363	0.8484	0.8451
Meronyms	0.8590	0.8405	0.8678	0.8209
Topics	0.8420	0.8189	0.8201	0.8306
Syn+Hype	0.8683	0.7918	0.7875	0.7664
Hype+Top	0.8552	0.7978	0.8028	0.8186
Syn+Hype+Top	0.8677	0.8163	0.8074	0.8003
Syn+Hype+Hypo+Mero+Top	0.8581	0.8101	0.8376	0.8222

Table 8.14. Hypernyms Disambiguation for 20Newsgroup - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8860	0.8911	0.8933	0.8907
Hypernyms	0.9063	0.9004	0.9055	0.9054
Hyponyms	0.8886	0.8936	0.8938	0.9041
Meronyms	0.8834	0.8869	0.8873	0.8866
Topics	0.8950	0.8944	0.8948	0.8967
Syn+Hype	0.8924	0.8965	0.8964	0.8949
Hype+Top	0.8975	0.8910	0.8959	0.8946
Syn+Hype+Top	0.8834	0.8843	0.8869	0.8849
Syn+Hype+Hypo+Mero+Top	0.9088	0.9219	0.9233	0.9305

Table 8.15. Hypernyms Disambiguation for 20Newsgroup - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8686	0.8067	0.8189	0.8115
Hypernyms	0.8840	0.8212	0.8384	0.8395
Hyponyms	0.8794	0.8541	0.8120	0.8486
Meronyms	0.8842	0.8849	0.9039	0.8728
Topics	0.8719	0.8885	0.8783	0.8932
Syn+Hype	0.8779	0.8769	0.8578	0.8510
Hype+Top	0.8780	0.8077	0.8173	0.8252
Syn+Hype+Top	0.8756	0.8605	0.8551	0.8540
Syn+Hype+Hypo+Mero+Top	0.8942	0.8594	0.8428	0.8798

Table 8.16. Topics Disambiguation for 20Newsgroup - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8830	0.8851	0.8849	0.8879
Hypernyms	0.9032	0.9000	0.9043	0.9030
Hyponyms	0.8944	0.9080	0.9078	0.9065
Meronyms	0.8817	0.8898	0.8896	0.8890
Topics	0.8753	0.8759	0.8753	0.8771
Syn+Hype	0.9044	0.9021	0.9034	0.9029
Hype+Top	0.9167	0.9130	0.9140	0.9165
Syn+Hype+Top	0.9020	0.8980	0.8991	0.8989
Syn+Hype+Hypo+Mero+Top	0.9302	0.9496	0.9506	0.9520

Table 8.17. Topics Disambiguation for 20Newsgroup - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8467	0.7721	0.7686	0.7686
Hypernyms	0.8529	0.7768	0.7854	0.7765
Hyponyms	0.8606	0.8490	0.8391	0.8649
Meronyms	0.8590	0.8394	0.8404	0.8301
Topics	0.8420	0.8267	0.8214	0.8326
Syn+Hype	0.8683	0.7726	0.7679	0.7516
Hype+Top	0.8552	0.8005	0.7948	0.7895
Syn+Hype+Top	0.8677	0.7960	0.7821	0.7705
Syn+Hype+Hypo+Mero+Top	0.8581	0.8545	0.8496	0.8627

Table 8.18. Topics Disambiguation for 20Newsgroup - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8860	0.8922	0.8924	0.8913
Hypernyms	0.9063	0.9067	0.9072	0.9089
Hyponyms	0.8886	0.9051	0.9062	0.9062
Meronyms	0.8834	0.8888	0.8899	0.8886
Topics	0.8950	0.8954	0.8954	0.8955
Syn+Hype	0.8924	0.8962	0.8984	0.9016
Hype+Top	0.8975	0.8970	0.8973	0.8979
Syn+Hype+Top	0.8834	0.8854	0.8876	0.8887
Syn+Hype+Hypo+Mero+Top	0.9088	0.9346	0.9354	0.9377

Table 8.19. Topics Disambiguation for 20Newsgroup - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8686	0.8200	0.8177	0.8164
Hypernyms	0.8840	0.8646	0.8569	0.8718
Hyponyms	0.8794	0.8869	0.8953	0.8909
Meronyms	0.8842	0.8924	0.8970	0.8950
Topics	0.8719	0.8856	0.8941	0.8929
Syn+Hype	0.8779	0.8729	0.8768	0.8786
Hype+Top	0.8780	0.8396	0.8383	0.8446
Syn+Hype+Top	0.8756	0.8714	0.8664	0.8642
Syn+Hype+Hypo+Mero+Top	0.8942	0.9132	0.9274	0.9256

Table 8.20. Properties of Classic3 Dataset.

Category	**test documents**	**train documents**	**total documents**
Cranfield	427	971	1398
Medline	304	729	1033
Cisi	461	999	1460

Table 8.21. Static FS(CHI), Micro-F Measure for Classic3 Dataset.

Policy	**10**	**30**	**50**	**100**	**200**	**500**	**1000**	**1500**	**2000**	**NoSelection**
Document	0.9955	0.9994	0.9987	0.9987	0.9987	0.9987	0.9987	0.9987	0.9987	0.9987
Local	0.9777	0.9777	0.9804	0.9896	0.9916	0.9942	0.9974	0.9987	0.9981	0.9987
Global	0.9804	0.9750	0.9723	0.9797	0.9870	0.9961	0.9974	0.9981	0.9987	0.9987

documents from aeronautical system papers, 1033 MEDLINE documents from medical journals, and 1460 CISI documents from information retrieval papers.

Classic3 has been used by many researchers [12, 24] in text mining and it is chosen as a *homogenous dataset* in our study, where all the classes are nearly equally well represented in the training set. First two thirds of each class is selected for the training set and the remaining one third is used for testing.

The most significant feature of the Classic3 dataset is that each class is disjoint from each other clearly, which means about 50 percent of the terms occur in only one class and the documents that share many common terms belong to the same class in the dataset. Since the classes are disjoint from each other, the Classic3 dataset is relatively easy to classify among other datasets in our study.

8.2.2. Analysis of Existing Metrics

Since this dataset is homogeneous and the results are quite high, we may cannot compare the results successfully. In Table 8.21-24, we can say that using document policy in static selection gives better results, while using global policy in the percentage selection is better.

Table 8.22. Static FS(CHI), Macro-F Measure for Classic3 Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.9952	0.9992	0.9986	0.9986	0.9986	0.9986	0.9986	0.9986	0.9986	0.9986
Local	0.9773	0.9760	0.9790	0.9890	0.9905	0.9936	0.9970	0.9984	0.9978	0.9986
Global	0.9795	0.9748	0.9711	0.9770	0.9859	0.9961	0.9977	0.9983	0.9989	0.9986

Table 8.23. Percentage FS(CHI), Micro-F Measure for Classic3 Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.9817	0.9890	0.9968	0.9968	0.9961	0.9974	0.9974	0.9981	0.9987	0.9987	0.9987
Local	0.9948	0.9942	0.9968	0.9981	0.9981	0.9981	0.9974	0.9961	0.9968	0.9940	0.9987
Global	0.9987	0.9987	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9981	0.9987	0.9987

8.2.3. Contribution of POS

Same problem here, the comparison is not easy for this dataset, using word forms for this dataset does not give better results. But we can say the results are not bad. The results can be seen in Table 8.25 and 8.26.

8.2.4. Contribution of WordNet Features

As can be seen in the results in Table 8.27-31, the comparision is really hard. Since the results are not very different than the no semantic feature option, we can not say something helpfull here.

8.2.5. Contribution of Disambiguation

Same problem here occur, the results are between $[0.99, 1]$, and there is no pattern here to say something about the behaviour. The results can be seen in Table 8.32-39.

Table 8.24. Percentage FS(CHI), Macro-F Measure for Classic3 Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.9812	0.9887	0.9964	0.9966	0.9958	0.9972	0.9972	0.9980	0.9986	0.9986	0.9986
Local	0.9947	0.9935	0.9964	0.9980	0.9980	0.9980	0.9970	0.9958	0.9964	0.9935	0.9986
Global	0.9988	0.9989	0.9980	0.9980	0.9980	0.9980	0.9980	0.9980	0.9980	0.9986	0.9986

Table 8.25. Contribution of POS for Classic3 - Micro-F Measure.

Contribution	1000	1500	2000
Raw	0.9974	0.9981	0.9987
Noun	0.9961	0.9968	0.9974
Verb	0.9380	0.9444	0.9401
Adjactive	0.9750	0.9743	0.9750
Adverb	0.7385	0.7385	0.7385
N+V	0.9974	0.9974	0.9974
N+Adj	0.9955	0.9968	0.9968
N+V+Adj	0.9961	0.9981	0.9974
N+V+Adj+Adv	0.9974	0.9981	0.9974

Contribution	1000	1500	2000
Raw	0.9974	0.9981	0.9987
Raw+Noun	0.9974	0.9987	0.9981
Raw+Verb	0.9981	0.9987	0.9981
Raw+Adjactive	0.9987	0.9974	0.9974
Raw+Adverb	0.9974	0.9987	0.9974
Raw+N+V	0.9981	0.9994	0.9987
Raw+N+Adj	0.9968	0.9974	0.9974
Raw+N+V+Adj	0.9974	0.9987	0.9981
Raw+N+V+Adj+Adv	0.9968	0.9987	0.9974

Table 8.26. Contribution of POS for Classic3 - Macro-F Measure.

Contribution	1000	1500	2000
Raw	0.9977	0.9983	0.9989
Noun	0.9961	0.9966	0.9975
Verb	0.9339	0.9410	0.9366
Adjactive	0.9738	0.9732	0.9741
Adverb	0.7203	0.7203	0.7203
N+V	0.9972	0.9970	0.9970
N+Adj	0.9957	0.9966	0.9969
N+V+Adj	0.9961	0.9980	0.9972
N+V+Adj+Adv	0.9972	0.9980	0.9972

Contribution	1000	1500	2000
Raw	0.9977	0.9983	0.9989
Raw+Noun	0.9977	0.9988	0.9983
Raw+Verb	0.9983	0.9989	0.9980
Raw+Adjactive	0.9989	0.9972	0.9972
Raw+Adverb	0.9977	0.9989	0.9975
Raw+N+V	0.9983	0.9994	0.9988
Raw+N+Adj	0.9971	0.9977	0.9977
Raw+N+V+Adj	0.9977	0.9989	0.9983
Raw+N+V+Adj+Adv	0.9971	0.9989	0.9977

Table 8.27. WordNet Features for Classic3 - Noun(L) and Raw + Noun(R).

Contribution	Micro-F	Macro-F
Synonyms	0.9903	0.9893
Hypernyms	0.9935	0.9935
Hyponyms	0.9817	0.9809
Meronyms	0.9922	0.9911
Topics	0.9961	0.9961
Syn+Hype	0.9903	0.9896
Hype+Top	0.9948	0.9947
Syn+Hype+Top	0.9922	0.9918
Syn+Hype+Hypo+Mero+Top	0.9850	0.9839
No Semantic Features	0.9974	0.9975

Contribution	Micro-F	Macro-F
Synonyms	0.9942	0.9939
Hypernyms	0.9968	0.9966
Hyponyms	0.9903	0.9894
Meronyms	0.9981	0.9983
Topics	0.9981	0.9982
Syn+Hype	0.9935	0.9935
Hype+Top	0.9981	0.9983
Syn+Hype+Top	0.9942	0.9941
Syn+Hype+Hypo+Mero+Top	0.9909	0.9902
No Semantic Features	0.9981	0.9983

Table 8.28. WordNet Features for Classic3 - Noun + Verb(L) and Raw + Noun + Verb(R).

Contribution	Micro-F	Macro-F
Synonyms	0.9916	0.9906
Hypernyms	0.9948	0.9944
Hyponyms	0.9817	0.9796
Meronyms	0.9942	0.9931
Topics	0.9974	0.9972
Syn+Hype	0.9909	0.9902
Hype+Top	0.9955	0.9952
Syn+Hype+Top	0.9916	0.9910
Syn+Hype+Hypo+Mero+Top	0.9877	0.9862
No Semantic Features	0.9974	0.9970

Contribution	Micro-F	Macro-F
Synonyms	0.9942	0.9941
Hypernyms	0.9974	0.9975
Hyponyms	0.9909	0.9897
Meronyms	0.9981	0.9983
Topics	0.9994	0.9994
Syn+Hype	0.9955	0.9955
Hype+Top	0.9981	0.9983
Syn+Hype+Top	0.9935	0.9937
Syn+Hype+Hypo+Mero+Top	0.9909	0.9904
No Semantic Features	0.9987	0.9988

Table 8.29. WordNet Features for Classic3 - Noun + Adj(L) and Raw + Noun + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.9903	0.9896	Synonyms	0.9935	0.9937
Hypernyms	0.9961	0.9958	Hypernyms	0.9961	0.9958
Hyponyms	0.9863	0.9853	Hyponyms	0.9929	0.9922
Meronyms	0.9942	0.9936	Meronyms	0.9981	0.9983
Topics	0.9961	0.9960	Topics	0.9981	0.9982
Syn+Hype	0.9909	0.9907	Syn+Hype	0.9916	0.9913
Hype+Top	0.9948	0.9947	Hype+Top	0.9974	0.9977
Syn+Hype+Top	0.9916	0.9913	Syn+Hype+Top	0.9909	0.9911
Syn+Hype+Hypo+Mero+Top	0.9857	0.9845	Syn+Hype+Hypo+Mero+Top	0.9857	0.9844
No Semantic Features	0.9968	0.9969	No Semantic Features	0.9974	0.9977

Table 8.30. WordNet Features for Classic3 - Noun + Verb + Adj(L) and Raw + Noun + Verb + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.9948	0.9946	Synonyms	0.9948	0.9944
Hypernyms	0.9961	0.9958	Hypernyms	0.9968	0.9969
Hyponyms	0.9844	0.9828	Hyponyms	0.9909	0.9897
Meronyms	0.9961	0.9956	Meronyms	0.9987	0.9989
Topics	0.9981	0.9980	Topics	0.9994	0.9994
Syn+Hype	0.9929	0.9924	Syn+Hype	0.9942	0.9941
Hype+Top	0.9961	0.9960	Hype+Top	0.9981	0.9983
Syn+Hype+Top	0.9922	0.9918	Syn+Hype+Top	0.9922	0.9923
Syn+Hype+Hypo+Mero+Top	0.9890	0.9882	Syn+Hype+Hypo+Mero+Top	0.9903	0.9896
No Semantic Features	0.9974	0.9972	No Semantic Features	0.9981	0.9983

Table 8.31. WordNet Features for Classic3 - Noun + Verb + Adj + Adv(L) and Raw + Noun + Verb + Adj + Adv(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.9929	0.9929	Synonyms	0.9935	0.9935
Hypernyms	0.9955	0.9952	Hypernyms	0.9981	0.9983
Hyponyms	0.9850	0.9834	Hyponyms	0.9916	0.9906
Meronyms	0.9961	0.9956	Meronyms	0.9987	0.9989
Topics	0.9974	0.9972	Topics	0.9981	0.9983
Syn+Hype	0.9929	0.9924	Syn+Hype	0.9942	0.9941
Hype+Top	0.9961	0.9960	Hype+Top	0.9981	0.9983
Syn+Hype+Top	0.9929	0.9924	Syn+Hype+Top	0.9942	0.9941
Syn+Hype+Hypo+Mero+Top	0.9877	0.9867	Syn+Hype+Hypo+Mero+Top	0.9909	0.9904
No Semantic Features	0.9974	0.9972	No Semantic Features	0.9974	0.9977

Table 8.32. Hypernyms Disambiguation for Classic3 - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9948	0.9961	0.9961	0.9974
Hypernyms	0.9961	0.9955	0.9942	0.9955
Hyponyms	0.9844	0.9903	0.9929	0.9929
Meronyms	0.9961	0.9974	0.9981	0.9981
Topics	0.9981	0.9974	0.9981	0.9987
Syn+Hype	0.9929	0.9935	0.9935	0.9955
Hype+Top	0.9961	0.9948	0.9942	0.9955
Syn+Hype+Top	0.9922	0.9935	0.9942	0.9955
Syn+Hype+Hypo+Mero+Top	0.9890	0.9916	0.9903	0.9935

Table 8.33. Hypernyms Disambiguation for Classic3 - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9946	0.9958	0.9958	0.9972
Hypernyms	0.9958	0.9950	0.9936	0.9950
Hyponyms	0.9828	0.9891	0.9927	0.9929
Meronyms	0.9956	0.9970	0.9978	0.9980
Topics	0.9980	0.9972	0.9980	0.9986
Syn+Hype	0.9924	0.9930	0.9933	0.9952
Hype+Top	0.9960	0.9942	0.9936	0.9950
Syn+Hype+Top	0.9918	0.9928	0.9938	0.9954
Syn+Hype+Hypo+Mero+Top	0.9882	0.9911	0.9896	0.9935

Table 8.34. Hypernyms Disambiguation for Classic3 - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9948	0.9974	0.9981	0.9987
Hypernyms	0.9968	0.9961	0.9968	0.9981
Hyponyms	0.9909	0.9935	0.9955	0.9974
Meronyms	0.9987	0.9987	0.9994	0.9987
Topics	0.9994	0.9981	0.9981	0.9987
Syn+Hype	0.9942	0.9968	0.9948	0.9974
Hype+Top	0.9981	0.9955	0.9974	0.9987
Syn+Hype+Top	0.9922	0.9955	0.9955	0.9974
Syn+Hype+Hypo+Mero+Top	0.9903	0.9935	0.9935	0.9948

Table 8.35. Hypernyms Disambiguation for Classic3 - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9944	0.9972	0.9983	0.9988
Hypernyms	0.9969	0.9963	0.9969	0.9982
Hyponyms	0.9897	0.9933	0.9953	0.9975
Meronyms	0.9989	0.9989	0.9994	0.9989
Topics	0.9994	0.9983	0.9983	0.9988
Syn+Hype	0.9941	0.9966	0.9949	0.9974
Hype+Top	0.9983	0.9955	0.9977	0.9988
Syn+Hype+Top	0.9923	0.9952	0.9957	0.9974
Syn+Hype+Hypo+Mero+Top	0.9896	0.9930	0.9933	0.9947

Table 8.36. Topics Disambiguation for Classic3 - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9948	0.9981	0.9974	0.9987
Hypernyms	0.9961	0.9974	0.9974	0.9974
Hyponyms	0.9844	0.9961	0.9968	0.9961
Meronyms	0.9961	0.9981	0.9981	0.9981
Topics	0.9981	0.9987	0.9981	0.9981
Syn+Hype	0.9929	0.9961	0.9968	0.9974
Hype+Top	0.9961	0.9974	0.9974	0.9974
Syn+Hype+Top	0.9922	0.9968	0.9955	0.9968
Syn+Hype+Hypo+Mero+Top	0.9890	0.9955	0.9948	0.9948

Table 8.37. Topics Disambiguation for Classic3 - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9946	0.9980	0.9970	0.9986
Hypernyms	0.9958	0.9972	0.9972	0.9972
Hyponyms	0.9828	0.9958	0.9964	0.9958
Meronyms	0.9956	0.9980	0.9980	0.9980
Topics	0.9980	0.9986	0.9980	0.9980
Syn+Hype	0.9924	0.9963	0.9964	0.9972
Hype+Top	0.9960	0.9975	0.9972	0.9970
Syn+Hype+Top	0.9918	0.9971	0.9952	0.9964
Syn+Hype+Hypo+Mero+Top	0.9882	0.9957	0.9947	0.9944

Table 8.38. Topics Disambiguation for Classic3 - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9948	0.9987	0.9987	0.9981
Hypernyms	0.9968	0.9987	0.9981	0.9987
Hyponyms	0.9909	0.9968	0.9981	0.9981
Meronyms	0.9987	0.9987	0.9987	0.9987
Topics	0.9994	0.9987	0.9987	0.9981
Syn+Hype	0.9942	0.9987	0.9987	0.9981
Hype+Top	0.9981	0.9987	0.9987	0.9994
Syn+Hype+Top	0.9922	0.9987	0.9987	0.9987
Syn+Hype+Hypo+Mero+Top	0.9903	0.9981	0.9974	0.9987

Table 8.39. Topics Disambiguation for Classic3 - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.9944	0.9989	0.9989	0.9983
Hypernyms	0.9969	0.9988	0.9983	0.9988
Hyponyms	0.9897	0.9969	0.9983	0.9983
Meronyms	0.9989	0.9989	0.9989	0.9989
Topics	0.9994	0.9988	0.9988	0.9983
Syn+Hype	0.9941	0.9989	0.9989	0.9983
Hype+Top	0.9983	0.9988	0.9988	0.9994
Syn+Hype+Top	0.9923	0.9988	0.9988	0.9988
Syn+Hype+Hypo+Mero+Top	0.9896	0.9983	0.9977	0.9988

8.3. 7 Sectors Dataset

8.3.1. Property of the Dataset

The 7 Sectors dataset contains web pages gathered from 7 different sectors: basic materials, energy, financial, healthcare, technology, transportation and utilities. And every category has several different sub categories. Here are the sub-categories of categories;

- basic materials(Train Documents:650, Test Documents:299, Total: 949)
  - i chemical manufacturing industry.
  - ii chemicals plastics and rubber industry.
  - iii containers and packaging industry.
  - iv fabricated plastic and rubber industry.
  - v forestry and wood products industry.
  - vi gold and silver industry.
  - vii iron and steel industry.
  - viii metal and mining industry.
  - ix misc fabricated products industry.
  - x non metallic mining industry.
  - xi paper and paper products industry.
- energy(Train Documents:253, Test Documents:102, Total: 355)
  - i coal industry.
  - ii oil and gas integrated industry.
  - iii oil and gas operations industry.
  - iv oil well services and equipment industry.
- financial(Train Documents:100, Test Documents:190, Total: 290)
  - i banking sector.
  - ii consumer financial services industry.
  - iii insurance sector.
  - iv investment services industry.
  - v misc financial services industry.

- healthcare(Train Documents:299, Test Documents:100, Total: 399)
  - i biotechnology and drugs industry.
  - ii healthcare facilities industry.
  - iii major drugs industry.
  - iv medical equipment and supplies industry.
- technology(Train Documents:300, Test Documents:200, Total: 500)
  - i communications equipment industry.
  - ii computer sector.
  - iii electronic instruments and controls industry.
  - iv office equipment industry.
  - v scientific and technical instruments industry.
  - vi semiconductors industry.
- transportation(Train Documents:379, Test Documents:136, Total: 515)
  - i air courier industry.
  - ii airline industry.
  - iii misc transportation industry.
  - iv railroad industry.
  - v trucking industry.
  - vi water transportation industry.
- utilities(Train Documents:200, Test Documents:100, Total: 300)
  - i electric utilities industry.
  - ii natural gas industry.
  - iii water utilities industry.

This dataset categorised as skew dataset and each category has a distinct profile in terms of the terms and topics.

8.3.2. Analysis of Existing Metrics

The results of existing metrics can be seen in Table 8.40-43. The results in this section can be used for comparison purposes as in the following sections for this data set, the semantic contribution will be discusses.

Table 8.40. Static FS(CHI), Micro-F Measure for 7Sectors Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.5625	0.5505	0.5689	0.5797	0.5770	0.5877	0.5859	0.5859	0.5833	0.5859
Local	0.5761	0.6298	0.6281	0.6256	0.6163	0.6069	0.5680	0.5680	0.5680	0.5859
Global	0.4171	0.4883	0.4933	0.5698	0.5991	0.5877	0.5680	0.5770	0.5886	0.5859

Table 8.41. Static FS(CHI), Macro-F Measure for 7Sectors Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.4734	0.4526	0.4789	0.5020	0.5042	0.5189	0.5175	0.5180	0.5143	0.5156
Local	0.5489	0.5734	0.5691	0.5663	0.5646	0.5525	0.4864	0.4790	0.4853	0.5156
Global	0.3507	0.4714	0.4797	0.5489	0.5762	0.5309	0.4881	0.5047	0.5312	0.5156

8.3.3. Contribution of POS

The results of using POS information in text categorization can be seen in Table 8.44 and 8.45. As in the first table the using POS information gives better results than raw features except verb, adjective and adverb only usages for 1000 term selection case, and the contribution decreases when selected number of term decreased.

8.3.4. Contribution of WordNet Features

The results of incorporating WordNet features in text categorization results can be seen in this section. In Table 8.46 the results show that using hypernyms and topics increases the Micro-F measure, while using synonyms, hypernyms and topics increases Macro-F measure. The results in Table 8.47 it can be seen that using synonyms and hypernyms increases both Micro-F and Macro-F measure no matter the raw terms used or not. In Table 8.48 the using hypernyms and topics increases Micro-F and Macro-F measure when there is no raw terms, while there is no configuration that increases Micro and Macro F measure for configuration contains raw terms. In Table 8.49, we can see that using synonyms and hypernyms make the results better for configurations do not contain raw terms, while incorporating all the semantic features increases Micro and Macro-F measure for configuration contains raw terms. In Table 8.50, meronyms

Table 8.42. Percentage FS(CHI), Micro-F Measure for 7Sectors Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.6163	0.5939	0.5707	0.5634	0.5634	0.5551	0.5579	0.5698	0.5779	0.5725	0.5859
Local	0.5671	0.5542	0.5073	0.5161	0.5112	0.5132	0.5345	0.5662	0.5542	0.5652	0.5859
Global	0.5815	0.5833	0.5761	0.5698	0.5689	0.5716	0.5616	0.5788	0.5859	0.5859	0.5859

Table 8.43. Percentage FS(CHI), Macro-F Measure for 7Sectors Dataset.

Policy	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**NoSelection**
Document	0.5715	0.5394	0.4991	0.4808	0.4779	0.4639	0.4748	0.5009	0.5123	0.5070	0.5156
Local	0.4842	0.4370	0.3885	0.3954	0.3872	0.3953	0.4265	0.4900	0.4729	0.4950	0.5156
Global	0.5176	0.5158	0.5004	0.4912	0.4921	0.4953	0.4830	0.5072	0.5156	0.5156	0.5156

Table 8.44. Contribution of POS for 7Sectors - Micro-F Measure.

Contribution	**1000**	**1500**	**2000**
Raw	0.5680	0.5770	0.5886
Noun	0.5707	0.5662	0.5734
Verb	0.4434	0.4509	0.4498
Adjactive	0.4770	0.4811	0.4913
Adverb	0.3969	0.3969	0.3969
N+V	0.5707	0.5643	0.5752
N+Adj	0.5913	0.5607	0.5625
N+V+Adj	0.5833	0.5707	0.5561
N+V+Adj+Adv	0.5833	0.5743	0.5643

Contribution	**1000**	**1500**	**2000**
Raw	0.5680	0.5770	0.5886
Raw+Noun	0.5707	0.5616	0.5634
Raw+Verb	0.5634	0.5725	0.5725
Raw+Adjactive	0.5698	0.5716	0.5734
Raw+Adverb	0.5652	0.5770	0.5851
Raw+N+V	0.5634	0.5551	0.5486
Raw+N+Adj	0.5698	0.5533	0.5597
Raw+N+V+Adj	0.5634	0.5523	0.5579
Raw+N+V+Adj+Adv	0.5625	0.5505	0.5570

Table 8.45. Contribution of POS for 7Sectors - Macro-F Measure.

Contribution	1000	1500	2000
Raw	0.4881	0.5047	0.5312
Noun	0.5166	0.5107	0.5142
Verb	0.3357	0.3380	0.3527
Adjactive	0.3792	0.3822	0.3808
Adverb	0.2245	0.2245	0.2245
N+V	0.5077	0.5013	0.5099
N+Adj	0.5371	0.5096	0.5076
N+V+Adj	0.5269	0.5071	0.4871
N+V+Adj+Adv	0.5261	0.5125	0.5006

Contribution	1000	1500	2000
Raw	0.4881	0.5047	0.5312
Raw+Noun	0.4841	0.4947	0.5037
Raw+Verb	0.4753	0.4994	0.5138
Raw+Adjactive	0.4899	0.4966	0.5113
Raw+Adverb	0.4787	0.5010	0.5289
Raw+N+V	0.4745	0.4874	0.4853
Raw+N+Adj	0.4830	0.4776	0.5012
Raw+N+V+Adj	0.4756	0.4788	0.4948
Raw+N+V+Adj+Adv	0.4731	0.4809	0.4929

Table 8.46. WordNet Features for 7Sectors - Noun(L) and Raw + Noun(R).

Contribution	**Micro-F**	**Macro-F**
Synonyms	0.5268	0.4782
Hypernyms	0.5716	0.5251
Hyponyms	0.5083	0.4500
Meronyms	0.5725	0.5198
Topics	0.5514	0.4895
Syn+Hype	0.5671	0.5277
Hype+Top	0.5797	0.5347
Syn+Hype+Top	0.5643	0.5247
Syn+Hype+Hypo+Mero+Top	0.5229	0.4716
No Semantic Features	0.5734	0.5142

Contribution	**Micro-F**	**Macro-F**
Synonyms	0.5220	0.4335
Hypernyms	0.5316	0.4457
Hyponyms	0.5181	0.4688
Meronyms	0.5743	0.4957
Topics	0.5523	0.4879
Syn+Hype	0.5505	0.4758
Hype+Top	0.5354	0.4496
Syn+Hype+Top	0.5725	0.5041
Syn+Hype+Hypo+Mero+Top	0.5392	0.4733
No Semantic Features	0.5634	0.5037

Table 8.47. WordNet Features for 7Sectors - Noun + Verb(L) and Raw + Noun + Verb(R).

Contribution	**Micro-F**	**Macro-F**
Synonyms	0.5258	0.4804
Hypernyms	0.5761	0.5287
Hyponyms	0.4963	0.4436
Meronyms	0.5680	0.5111
Topics	0.5579	0.4851
Syn+Hype	0.5806	0.5390
Hype+Top	0.5779	0.5271
Syn+Hype+Top	0.5779	0.5353
Syn+Hype+Hypo+Mero+Top	0.5316	0.4716
No Semantic Features	0.5752	0.5099

Contribution	**Micro-F**	**Macro-F**
Synonyms	0.5181	0.4276
Hypernyms	0.5542	0.4693
Hyponyms	0.5033	0.4576
Meronyms	0.5643	0.4736
Topics	0.5430	0.4780
Syn+Hype	0.5707	0.5015
Hype+Top	0.5505	0.4676
Syn+Hype+Top	0.5806	0.5173
Syn+Hype+Hypo+Mero+Top	0.5505	0.4846
No Semantic Features	0.5486	0.4853

Table 8.48. WordNet Features for 7Sectors - Noun + Adj(L) and Raw + Noun + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.5449	0.4845	Synonyms	0.5249	0.4407
Hypernyms	0.5734	0.5201	Hypernyms	0.5220	0.4283
Hyponyms	0.5171	0.4655	Hyponyms	0.5210	0.4719
Meronyms	0.5833	0.5251	Meronyms	0.5689	0.4898
Topics	0.5467	0.4864	Topics	0.5430	0.4689
Syn+Hype	0.5200	0.4255	Syn+Hype	0.5316	0.4375
Hype+Top	0.5716	0.5197	Hype+Top	0.5200	0.4199
Syn+Hype+Top	0.5287	0.4331	Syn+Hype+Top	0.5316	0.4323
Syn+Hype+Hypo+Mero+Top	0.5364	0.4740	Syn+Hype+Hypo+Mero+Top	0.5495	0.4860
No Semantic Features	0.5625	0.5076	No Semantic Features	0.5597	0.5012

Table 8.49. WordNet Features for 7Sectors - Noun + Verb + Adj(L) and Raw + Noun + Verb + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.5316	0.4763	Synonyms	0.5161	0.4267
Hypernyms	0.5607	0.5018	Hypernyms	0.5402	0.4481
Hyponyms	0.5043	0.4499	Hyponyms	0.5003	0.4547
Meronyms	0.5939	0.5363	Meronyms	0.5616	0.4730
Topics	0.5392	0.4655	Topics	0.5383	0.4611
Syn+Hype	0.5707	0.5192	Syn+Hype	0.5467	0.4572
Hype+Top	0.5671	0.5132	Hype+Top	0.5268	0.4303
Syn+Hype+Top	0.5680	0.5143	Syn+Hype+Top	0.5335	0.4402
Syn+Hype+Hypo+Mero+Top	0.5523	0.4958	Syn+Hype+Hypo+Mero+Top	0.5652	0.5036
No Semantic Features	0.5561	0.4871	No Semantic Features	0.5579	0.4948

increase Micro-F measure and using synonyms and hypernyms increases Micro-F measure for no raw terms option. However using all semantic features for configuration contains raw features increases both Micro and Macro F measure.

Table 8.50. WordNet Features for 7Sectors - Noun + Verb + Adj + Adv(L) and Raw + Noun + Verb + Adj + Adv(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.5297	0.4738	Synonyms	0.5152	0.4250
Hypernyms	0.5689	0.5123	Hypernyms	0.5420	0.4492
Hyponyms	0.5073	0.4546	Hyponyms	0.5003	0.4535
Meronyms	0.5859	0.5275	Meronyms	0.5689	0.4867
Topics	0.5392	0.4644	Topics	0.5345	0.4558
Syn+Hype	0.5680	0.5199	Syn+Hype	0.5477	0.4630
Hype+Top	0.5680	0.5137	Hype+Top	0.5287	0.4349
Syn+Hype+Top	0.5662	0.5113	Syn+Hype+Top	0.5326	0.4351
Syn+Hype+Hypo+Mero+Top	0.5523	0.4966	Syn+Hype+Hypo+Mero+Top	0.5662	0.5026
No Semantic Features	0.5643	0.5006	No Semantic Features	0.5570	0.4929

Table 8.51. Hypernyms Disambiguation for 7Sectors - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.5316	0.5533	0.5689	0.5561
Hypernyms	0.5607	0.5797	0.5588	0.5514
Hyponyms	0.5043	0.5335	0.5458	0.5707
Meronyms	0.5939	0.5671	0.5486	0.5662
Topics	0.5392	0.5643	0.5616	0.5579
Syn+Hype	0.5707	0.5616	0.5597	0.5579
Hype+Top	0.5671	0.5770	0.5652	0.5570
Syn+Hype+Top	0.5680	0.5662	0.5652	0.5588
Syn+Hype+Hypo+Mero+Top	0.5523	0.5411	0.5449	0.5486

Table 8.52. Hypernyms Disambiguation for 7Sectors - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4763	0.4878	0.5085	0.4891
Hypernyms	0.5018	0.5129	0.4910	0.4885
Hyponyms	0.4499	0.4721	0.4547	0.5129
Meronyms	0.5363	0.5114	0.4756	0.5044
Topics	0.4655	0.5025	0.4961	0.4934
Syn+Hype	0.5192	0.5125	0.4972	0.4932
Hype+Top	0.5132	0.5130	0.5021	0.4990
Syn+Hype+Top	0.5143	0.5173	0.5019	0.4955
Syn+Hype+Hypo+Mero+Top	0.4958	0.4706	0.4872	0.4809

8.3.5. Contribution of Disambiguation

In this section the contribution of disambiguation will be discussed for 7 sectors dataset. In Tables 8.51-54, it can be clearly seen, almost for all the contributions, applying disambiguation, by using hypernyms, increases Micro and Macro F measure.

In Tables 8.55-58, results of disambiguation by using topic information can be seen. Applying disambiguation with 70% threshold gives better results than no selection cases both for Micro and Macro F measures.

Table 8.53. Hypernyms Disambiguation for 7Sectors - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.5161	0.5449	0.5449	0.5258
Hypernyms	0.5402	0.5373	0.5449	0.5392
Hyponyms	0.5003	0.5542	0.5458	0.5420
Meronyms	0.5616	0.5514	0.5326	0.5316
Topics	0.5383	0.5449	0.5570	0.5458
Syn+Hype	0.5467	0.5354	0.5551	0.5597
Hype+Top	0.5268	0.5402	0.5439	0.5542
Syn+Hype+Top	0.5335	0.5326	0.5542	0.5268
Syn+Hype+Hypo+Mero+Top	0.5652	0.5383	0.5326	0.5306

Table 8.54. Hypernyms Disambiguation for 7Sectors - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4267	0.4707	0.4840	0.4440
Hypernyms	0.4481	0.4624	0.4729	0.4620
Hyponyms	0.4547	0.4613	0.4459	0.4789
Meronyms	0.4730	0.4864	0.4545	0.4523
Topics	0.4611	0.4767	0.4946	0.4440
Syn+Hype	0.4572	0.4353	0.4813	0.4969
Hype+Top	0.4303	0.4627	0.4741	0.4927
Syn+Hype+Top	0.4402	0.4267	0.4816	0.4497
Syn+Hype+Hypo+Mero+Top	0.5036	0.4582	0.4411	0.4307

Table 8.55. Topics Disambiguation for 7Sectors - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.5316	0.5579	0.5523	0.5505
Hypernyms	0.5607	0.5634	0.5616	0.5734
Hyponyms	0.5043	0.5725	0.5779	0.5725
Meronyms	0.5939	0.5607	0.5616	0.5634
Topics	0.5392	0.5345	0.5495	0.5523
Syn+Hype	0.5707	0.5597	0.5523	0.5542
Hype+Top	0.5671	0.5430	0.5467	0.5652
Syn+Hype+Top	0.5680	0.5268	0.5345	0.5467
Syn+Hype+Hypo+Mero+Top	0.5523	0.5514	0.5439	0.5671

Table 8.56. Topics Disambiguation for 7Sectors - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4763	0.4927	0.4855	0.4798
Hypernyms	0.5018	0.4965	0.4972	0.5103
Hyponyms	0.4499	0.5094	0.5244	0.5168
Meronyms	0.5363	0.4982	0.4986	0.4991
Topics	0.4655	0.4495	0.4799	0.4866
Syn+Hype	0.5192	0.4981	0.4882	0.4813
Hype+Top	0.5132	0.4673	0.4797	0.5013
Syn+Hype+Top	0.5143	0.4502	0.4635	0.4734
Syn+Hype+Hypo+Mero+Top	0.4958	0.4736	0.4774	0.5146

Table 8.57. Topics Disambiguation for 7Sectors - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.5161	0.5420	0.5533	0.5561
Hypernyms	0.5402	0.5616	0.5551	0.5616
Hyponyms	0.5003	0.5533	0.5458	0.5420
Meronyms	0.5616	0.5392	0.5411	0.5467
Topics	0.5383	0.5449	0.5542	0.5523
Syn+Hype	0.5467	0.5449	0.5561	0.5662
Hype+Top	0.5268	0.5477	0.5561	0.5597
Syn+Hype+Top	0.5335	0.5523	0.5477	0.5579
Syn+Hype+Hypo+Mero+Top	0.5652	0.5210	0.5364	0.5486

Table 8.58. Topics Disambiguation for 7Sectors - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4267	0.4708	0.4847	0.4894
Hypernyms	0.4481	0.5027	0.4972	0.5000
Hyponyms	0.4547	0.4835	0.4762	0.4734
Meronyms	0.4730	0.4688	0.4720	0.4828
Topics	0.4611	0.4718	0.4920	0.4931
Syn+Hype	0.4572	0.4754	0.4959	0.5069
Hype+Top	0.4303	0.4791	0.4955	0.4983
Syn+Hype+Top	0.4402	0.4826	0.4849	0.4986
Syn+Hype+Hypo+Mero+Top	0.5036	0.4141	0.4354	0.4731

Table 8.59. Properties of WebKB Dataset.

Category	test documents	train documents	total documents
course	38	892	930
department	32	149	182
faculty	46	1078	1124
student	571	2589	3160

8.4. WebKB Dataset

8.4.1. Property of the Dataset

The WebKB dataset contains web pages gathered from university computer science departments. The pages are divided into seven categories: student, faculty, staff, course, project, department and other. In this study, we use only four most populous entity-representing categories: department, faculty, course and student, all together containing 5396 pages. This dataset has been used in the studies [34–36]. The details about dataset can be seen in Table 8.59. As seen in the dataset details, the documents are closely related. The number of shared terms in the categories is very high. Thus the classification is very hard for this dataset.

8.4.2. Analysis of Existing Metrics

In this section we will discuss the results of using existing metrics for WebKB dataset. The results can be seen in Tables 8.60-63. For this dataset, when number of

Table 8.60. Static FS(CHI), Micro-F Measure for WebKB Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.8757	0.8551	0.8491	0.8338	0.8094	0.7996	0.7996	0.8007	0.8007	0.7963
Local	0.8201	0.6072	0.6599	0.6432	0.6801	0.7299	0.7481	0.7738	0.7646	0.7963
Global	0.9112	0.8699	0.8451	0.7623	0.7481	0.7360	0.7493	0.7623	0.7634	0.7963

Table 8.61. Static FS(CHI), Macro-F Measure for WebKB Dataset.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.5453	0.8607	0.8819	0.9113	0.9066	0.9077	0.9077	0.9080	0.9080	0.9067
Local	0.8568	0.8461	0.8602	0.8435	0.8534	0.8410	0.8807	0.8843	0.8799	0.9067
Global	0.5971	0.6568	0.6675	0.9027	0.9032	0.8923	0.8925	0.8964	0.8968	0.9067

selected item decreased local policy gives better results, in contrast, when increased; global policy better results than others.

8.4.3. Contribution of POS

In this section, we will discuss the contribution of POS for WebKB dataset. In the results given in Tables 8.64 and 8.65 it can be seen that Micro-F measures increased by using POS, but we cannot say the same for Macro-F measure; it almost behaves the same with base line.

8.4.4. Contribution of WordNet Features

In this section we will discuss the contribution of WordNet semantic features for WebKB dataset.

Results In Table 8.66 shows that using all the semantic features for nouns, together or individually, increases both Micro and Macro F measure. For the configuration contains raw feature we can say the same result with a few exceptions. Results In Table 8.67 shows that using all the semantic features for nouns and verbs, together or individually, increases Micro-F measure, but using hypernyms and synonyms + hy-

Table 8.62. Percentage FS(CHI), Micro-F Measure for WebKB Dataset.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.7863	0.7750	0.7829	0.7941	0.8051	0.8073	0.8201	0.8116	0.7886	0.7874	0.7963
Local	0.8389	0.8491	0.8420	0.8389	0.8410	0.8389	0.8338	0.7919	0.7807	0.7773	0.7963
Global	0.7952	0.8105	0.8084	0.8094	0.8084	0.8018	0.7963	0.7952	0.7941	0.7974	0.7963

Table 8.63. Percentage FS(CHI), Macro-F Measure for WebKB Dataset.

Policy	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**NoSelection**
Document	0.8253	0.8575	0.8730	0.8849	0.8866	0.8882	0.8893	0.8868	0.8870	0.8778	0.9067
Local	0.8828	0.8732	0.8712	0.8703	0.8709	0.8703	0.8688	0.8765	0.8732	0.8819	0.9067
Global	0.9064	0.9148	0.9063	0.9066	0.9103	0.9083	0.9067	0.9064	0.9060	0.9070	0.9067

Table 8.64. Contribution of POS for WebKB - Micro-F Measure.

Contribution	**1000**	**1500**	**2**000
Raw	0.7493	0.7623	0.7634
Noun	0.7681	0.7669	0.7681
Verb	0.7225	0.7287	0.7336
Adjactive	0.6761	0.6867	0.6893
Adverb	0.5646	0.5630	0.5785
N+V	0.7669	0.7715	0.7750
N+Adj	0.7588	0.7646	0.7599
N+V+Adj	0.7658	0.7669	0.7646
N+V+Adj+Adv	0.7715	0.7738	0.7715

Contribution	**1000**	**1500**	**2**000
Raw	0.7493	0.7623	0.7634
Raw+Noun	0.7564	0.7761	0.7795
Raw+Verb	0.7552	0.7552	0.7564
Raw+Adjactive	0.7505	0.7529	0.7576
Raw+Adverb	0.7552	0.7704	0.7646
Raw+N+V	0.7634	0.7773	0.7841
Raw+N+Adj	0.7529	0.7681	0.7715
Raw+N+V+Adj	0.7623	0.7692	0.7738
Raw+N+V+Adj+Adv	0.7588	0.7681	0.7750

Table 8.65. Contribution of POS for WebKB - Macro-F Measure.

Contribution	**1000**	**1500**	**2**000
Raw	0.8925	0.8964	0.8968
Noun	0.8918	0.8914	0.8877
Verb	0.5269	0.5286	0.5419
Adjactive	0.5498	0.5529	0.5599
Adverb	0.3633	0.3628	0.3672
N+V	0.8898	0.8912	0.8922
N+Adj	0.8904	0.8922	0.8908
N+V+Adj	0.8886	0.8889	0.8882
N+V+Adj+Adv	0.8943	0.8949	0.8943

Contribution	**1000**	**1500**	**2**000
Raw	0.8925	0.8964	0.8968
Raw+Noun	0.8824	0.8840	0.8850
Raw+Verb	0.8879	0.8879	0.8883
Raw+Adjactive	0.8967	0.9011	0.9025
Raw+Adverb	0.8943	0.8989	0.8972
Raw+N+V	0.8846	0.8843	0.8863
Raw+N+Adj	0.8856	0.8943	0.8953
Raw+N+V+Adj	0.8884	0.8881	0.8894
Raw+N+V+Adj+Adv	0.8874	0.8943	0.8963

Table 8.66. WordNet Features for WebKB - Noun(L) and Raw + Noun(R).

Contribution	**Micro-F**	**Macro-F**	**Contribution**	**Micro-F**	**Macro-F**
Synonyms	0.7829	0.8857	Synonyms	0.7874	0.8964
Hypernyms	0.7886	0.8952	Hypernyms	0.8040	0.9052
Hyponyms	0.7941	0.8948	Hyponyms	0.7985	0.9144
Meronyms	0.7852	0.8983	Meronyms	0.7715	0.9006
Topics	0.7784	0.8883	Topics	0.7930	0.8890
Syn+Hype	0.7761	0.8891	Syn+Hype	0.7829	0.9015
Hype+Top	0.7897	0.8877	Hype+Top	0.7974	0.9032
Syn+Hype+Top	0.7852	0.8891	Syn+Hype+Top	0.7727	0.8995
Syn+Hype+Hypo+Mero+Top	0.7761	0.9050	Syn+Hype+Hypo+Mero+Top	0.7941	0.9157
No Semantic Features	0.7681	0.8877	No Semantic Features	0.7795	0.8850

Table 8.67. WordNet Features for WebKB - Noun + Verb(L) and Raw + Noun + Verb(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8286	0.9162	Synonyms	0.7669	0.8822
Hypernyms	0.7841	0.8819	Hypernyms	0.7985	0.9062
Hyponyms	0.8062	0.8976	Hyponyms	0.7863	0.9062
Meronyms	0.7919	0.8999	Meronyms	0.7761	0.9057
Topics	0.7919	0.9016	Topics	0.7908	0.8902
Syn+Hype	0.7818	0.8761	Syn+Hype	0.7784	0.8963
Hype+Top	0.7941	0.9022	Hype+Top	0.7963	0.9017
Syn+Hype+Top	0.8148	0.8923	Syn+Hype+Top	0.7773	0.8972
Syn+Hype+Hypo+Mero+Top	0.7863	0.8961	Syn+Hype+Hypo+Mero+Top	0.8040	0.9151
No Semantic Features	0.7750	0.8922	No Semantic Features	0.7841	0.8863

Table 8.68. WordNet Features for WebKB - Noun + Adj(L) and Raw + Noun + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.7852	0.8903	Synonyms	0.7761	0.8876
Hypernyms	0.7952	0.9009	Hypernyms	0.7952	0.9026
Hyponyms	0.8029	0.9003	Hyponyms	0.7919	0.9097
Meronyms	0.7795	0.9052	Meronyms	0.7795	0.9005
Topics	0.7818	0.8867	Topics	0.7784	0.9074
Syn+Hype	0.7773	0.8840	Syn+Hype	0.8007	0.8923
Hype+Top	0.7974	0.8950	Hype+Top	0.7727	0.9047
Syn+Hype+Top	0.7784	0.8830	Syn+Hype+Top	0.7795	0.9078
Syn+Hype+Hypo+Mero+Top	0.7784	0.9057	Syn+Hype+Hypo+Mero+Top	0.7784	0.8933
No Semantic Features	0.7599	0.8908	No Semantic Features	0.7715	0.8953

pernyms decreases Macro-F measure. For the configuration contains raw feature we can say the same result with a few exceptions. We can infer same results for results in Tables 8.68-70.

Table 8.69. WordNet Features for WebKB - Noun + Verb + Adj(L) and Raw + Noun + Verb + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.7807	0.8850	Synonyms	0.7795	0.8901
Hypernyms	0.7985	0.8915	Hypernyms	0.7974	0.9058
Hyponyms	0.7908	0.8995	Hyponyms	0.7818	0.9012
Meronyms	0.7795	0.8979	Meronyms	0.7773	0.9060
Topics	0.7738	0.8801	Topics	0.7874	0.8935
Syn+Hype	0.7996	0.9076	Syn+Hype	0.7852	0.8996
Hype+Top	0.8105	0.8938	Hype+Top	0.7897	0.8997
Syn+Hype+Top	0.7974	0.9069	Syn+Hype+Top	0.7773	0.8972
Syn+Hype+Hypo+Mero+Top	0.8224	0.3847	Syn+Hype+Hypo+Mero+Top	0.7996	0.9164
No Semantic Features	0.7646	0.8882	No Semantic Features	0.7738	0.8894

Table 8.70. WordNet Features for WebKB - Noun + Verb + Adj + Adv(L) and Raw + Noun + Verb + Adj + Adv(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.7784	0.8803	Synonyms	0.7738	0.8923
Hypernyms	0.7941	0.8902	Hypernyms	0.7930	0.9007
Hyponyms	0.7974	0.9032	Hyponyms	0.7841	0.9055
Meronyms	0.8073	0.8929	Meronyms	0.7727	0.9047
Topics	0.7941	0.9022	Topics	0.7841	0.8883
Syn+Hype	0.7886	0.8989	Syn+Hype	0.7807	0.8944
Hype+Top	0.7807	0.9019	Hype+Top	0.7886	0.9032
Syn+Hype+Top	0.8265	0.9156	Syn+Hype+Top	0.7750	0.8927
Syn+Hype+Hypo+Mero+Top	0.7841	0.8831	Syn+Hype+Hypo+Mero+Top	0.7974	0.9182
No Semantic Features	0.7715	0.8943	No Semantic Features	0.7750	0.8963

Table 8.71. Hypernyms Disambiguation for WebKB - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.7807	0.7963	0.7996	0.7829
Hypernyms	0.7985	0.8105	0.8062	0.7952
Hyponyms	0.7908	0.8018	0.7919	0.7841
Meronyms	0.7795	0.7852	0.7773	0.7818
Topics	0.7738	0.7795	0.7681	0.7738
Syn+Hype	0.7996	0.8084	0.8127	0.8029
Hype+Top	0.8105	0.8180	0.8127	0.7886
Syn+Hype+Top	0.7974	0.8062	0.8094	0.7941
Syn+Hype+Hypo+Mero+Top	0.8224	0.8191	0.8073	0.8084

8.4.5. Contribution of Disambiguation

In this section we will discuss the contribution of applying disambiguation.

Results in Tables 8.71 and 8.72 shows that applying disambiguation increases both Micro and Macro F measure.

In Tables 8.73 and 8.74 we can observe the applying disambiguation for noun, verb, adjactive and raw terms. We can conclude that disambiguation increases Micro-F measure, bur Macro-F measure decreased.

Table 8.72. Hypernyms Disambiguation for WebKB - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8850	0.9083	0.9049	0.8884
Hypernyms	0.8915	0.9118	0.9140	0.9036
Hyponyms	0.8995	0.9045	0.9035	0.8914
Meronyms	0.8979	0.8996	0.8927	0.8947
Topics	0.8801	0.8966	0.8893	0.8910
Syn+Hype	0.9076	0.9138	0.9115	0.9031
Hype+Top	0.8938	0.9112	0.9124	0.9006
Syn+Hype+Top	0.9069	0.9041	0.9078	0.9005
Syn+Hype+Hypo+Mero+Top	0.8947	0.9096	0.9044	0.8960

Table 8.73. Hypernyms Disambiguation for WebKB - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.7795	0.7852	0.7952	0.7818
Hypernyms	0.7974	0.8018	0.7930	0.7784
Hyponyms	0.7818	0.7874	0.7874	0.7818
Meronyms	0.7773	0.7807	0.7897	0.7715
Topics	0.7874	0.7738	0.7564	0.7704
Syn+Hype	0.7852	0.8040	0.8040	0.7897
Hype+Top	0.7897	0.8127	0.8018	0.7795
Syn+Hype+Top	0.7773	0.7974	0.8116	0.7829
Syn+Hype+Hypo+Mero+Top	0.7996	0.8084	0.7874	0.7874

Table 8.74. Hypernyms Disambiguation for WebKB - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8901	0.8957	0.9039	0.8918
Hypernyms	0.9058	0.9109	0.8941	0.8908
Hyponyms	0.9012	0.8990	0.9110	0.9037
Meronyms	0.9060	0.8889	0.9098	0.8928
Topics	0.8935	0.8894	0.8842	0.8884
Syn+Hype	0.8996	0.8890	0.8974	0.8890
Hype+Top	0.8997	0.9166	0.8993	0.8844
Syn+Hype+Top	0.8972	0.8913	0.8981	0.8828
Syn+Hype+Hypo+Mero+Top	0.9164	0.9141	0.9136	0.8883

In Table 8.75-78, results of applying disambiguation by using topic information can be seen. We can observe that using topic information does not increase the Micro and Macro F measures with an exception: Micro-F measure increased a little bit for configuration that includes raw terms.

Table 8.75. Topics Disambiguation for WebKB - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.7807	0.7773	0.7738	0.7692
Hypernyms	0.7985	0.7750	0.7930	0.7588
Hyponyms	0.7908	0.7874	0.7829	0.7761
Meronyms	0.7795	0.7658	0.7669	0.7727
Topics	0.7738	0.7773	0.7738	0.7807
Syn+Hype	0.7996	0.7773	0.7829	0.7646
Hype+Top	0.8105	0.7773	0.7941	0.7750
Syn+Hype+Top	0.7974	0.7807	0.7908	0.7841
Syn+Hype+Hypo+Mero+Top	0.8224	0.7841	0.7930	0.7738

Table 8.76. Topics Disambiguation for WebKB - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8850	0.8879	0.8910	0.8813
Hypernyms	0.8915	0.8913	0.8926	0.8865
Hyponyms	0.8995	0.8910	0.8911	0.8995
Meronyms	0.8979	0.8925	0.8929	0.8906
Topics	0.8801	0.8837	0.8801	0.8821
Syn+Hype	0.9076	0.8894	0.8870	0.8896
Hype+Top	0.8938	0.8811	0.8861	0.8887
Syn+Hype+Top	0.9069	0.8904	0.8877	0.8888
Syn+Hype+Hypo+Mero+Top	0.8947	0.8873	0.8900	0.8884

Table 8.77. Topics Disambiguation for WebKB - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.7795	0.7750	0.7738	0.7692
Hypernyms	0.7974	0.7963	0.8062	0.7795
Hyponyms	0.7818	0.7727	0.7773	0.7704
Meronyms	0.7773	0.7727	0.7692	0.7681
Topics	0.7874	0.7738	0.7750	0.7818
Syn+Hype	0.7852	0.7829	0.7863	0.7795
Hype+Top	0.7897	0.7941	0.7996	0.7886
Syn+Hype+Top	0.7773	0.7952	0.7841	0.7818
Syn+Hype+Hypo+Mero+Top	0.7996	0.7919	0.8018	0.7829

Table 8.78. Topics Disambiguation for WebKB - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8901	0.8978	0.8960	0.8946
Hypernyms	0.9058	0.9002	0.9071	0.8952
Hyponyms	0.9012	0.8849	0.8887	0.8884
Meronyms	0.9060	0.8891	0.8905	0.8902
Topics	0.8935	0.8894	0.8872	0.8918
Syn+Hype	0.8996	0.8921	0.8997	0.8977
Hype+Top	0.8997	0.8955	0.8971	0.8938
Syn+Hype+Top	0.8972	0.8958	0.8925	0.8943
Syn+Hype+Hypo+Mero+Top	0.9164	0.8880	0.8935	0.8879

8.5. Reuters-21578 Dataset

8.5.1. Property of the Dataset

The final dataset used in this study is Reuters-21578 which is one of the most popular data collections. The Reuters-21578 dataset compiled by David Lewis contains news-wire stories in 1987. These documents were manually categorized by the personnel from Reuters Ltd. and Carnegie Group Inc. in 1987. The collection was made available for scientific research in 1990. Many different versions have been used in past studies and it is considered as the standard benchmark for automatic document organization systems.

In order to divide the Reuters-21578 dataset into training and test sets, we use ModApte splitting method that has been mostly used in the literature [20, 24, 31]. Originally, the Reuters-21578 dataset consists of 21,578 documents that divided into 135 different categories. But with ModApte split the training set consists of 9,603 documents, the test set consists of 3,299 documents and 8,676 documents were unused. The splitting criteria are:

- The training set consists of any document in the dataset that has at least one category assigned and is dated earlier than April 7th, 1987;
- The test set consists of any document in the dataset that has at least on category assigned and is dated April 7th, 1987 or later; and
- The unused set consists of any documents that has no categories assigned to them.

After removing the categories that do not exist both in the training set and in the test set, remaining with 90 classes out of 135. Table 8.79 details the categories of the dataset with ModApte split.

Table 8.79. Properties of Reuters-21578 Dataset.

cocoa	18	55	73	crude	189	389	578	gas	17	37	54
grain	149	433	582	nat-gas	30	75	105	jobs	21	46	67
wheat	71	212	283	cpi	28	69	97	lei	3	12	15
corn	56	181	237	gnp	35	101	136	yen	14	45	59
barley	14	37	51	money-fx	179	538	717	zinc	13	21	34
oat	6	8	14	interest	131	347	478	orange	11	16	27
sorghum	10	24	34	bop	30	75	105	pet-chem	12	20	32
veg-oil	37	87	124	rice	24	35	59	fuel	10	13	23
lin-oil	1	1	2	rubber	12	37	49	wpi	10	19	29
soy-oil	11	14	25	copra-cake	1	2	3	potato	3	3	6
sun-oil	2	5	7	palm-oil	10	30	40	lead	14	15	29
soybean	33	78	111	palmkernel	1	2	3	groundnut	4	5	9
oilseed	47	124	171	tea	4	9	13	income	7	9	16
sunseed	5	11	16	alum	23	35	58	palladium	1	2	3
earn	1087	2877	3964	gold	30	94	124	nickel	1	8	9
acq	719	1650	2369	platinum	7	5	12	lumber	6	10	16
copper	18	47	65	strategic-metal	11	16	27	jet	1	4	5
housing	4	16	20	tin	12	18	30	instal-debt	1	5	6
money-supply	34	140	174	rapeseed	9	18	27	dfl	1	2	3
coffee	28	111	139	groundnut-oil	1	1	2	dmk	4	10	14
ship	89	197	286	rape-oil	3	5	8	coconut-oil	3	4	7
sugar	36	126	162	dlr	44	131	175	cpu	1	3	4
trade	117	369	486	l-cattle	2	6	8	cotton-oil	2	1	3
reserves	18	55	73	retail	2	23	25	naphtha	4	2	6
meal-feed	19	30	49	ipi	12	41	53	nzdlr	2	2	4
soy-meal	13	13	26	silver	8	21	29	rand	1	2	3
rye	1	1	2	iron-steel	14	40	54	coconut	2	4	6
cotton	20	39	59	hog	6	16	22	castor-oil	1	1	2
carcass	18	50	68	propane	3	3	6	nkr	2	1	3
livestock	24	75	99	heat	5	14	19	sun-meal	1	1	2

Table 8.80. Static FS(CHI), Micro-F Measure for Reuters-21578.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.8252	0.8381	0.8447	0.8460	0.8488	0.8474	0.8474	0.8474	0.8474	0.8474
Local	0.8074	0.8142	0.8062	0.8187	0.8222	0.8308	0.8356	0.8406	0.8413	0.8474
Global	0.6993	0.7250	0.7383	0.7421	0.7891	0.8244	0.8383	0.8445	0.8467	0.8474

Table 8.81. Static FS(CHI), Macro-F Measure for Reuters-21578.

Policy	10	30	50	100	200	500	1000	1500	2000	NoSelection
Document	0.3149	0.3729	0.3912	0.3975	0.4044	0.4037	0.4037	0.4037	0.4037	0.4037
Local	0.3207	0.3402	0.3026	0.3270	0.3510	0.3793	0.4090	0.4071	0.4030	0.4037
Global	0.2117	0.2176	0.2402	0.2463	0.3370	0.3602	0.3810	0.3835	0.3837	0.4037

These 90 categories are very close to each other thus some of the documents are assigned to more than one category. The maximum number of categories assigned to a document is 14 and the average number of categories per document is 1.24. The 10 top categories which are shown in bold in Table 8.79 constitute about 75 percent of the dataset and the remaining 80 categories constitute only about 25 percent of all documents. In addition, the 2 top categories *earn* and *acq* constitute about 48 percent of the dataset.

8.5.2. Analysis of Existing Metrics

In Table 8.80 and 8.81 it can be seen that document policy dominates others and for threshold 200 both for Micro and Macro F measure, the results are better than no selection and other options. In Table 8.82 and Table 8.83, it can seen that global policy usually gives better results than local and document policies.

As a result, when the number of terms selected increased the global policy gives better results. In contrast when number of selected terms is decreased document policy gives better results. But it should not be forgotten that in document policy, the selected amount of terms is done for every document individually. Thus the number of total selected term is much more.

Table 8.82. Percentage FS(CHI), Micro-F Measure for Reuters-21578.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.8222	0.8281	0.8299	0.8307	0.8292	0.8307	0.8337	0.8366	0.8392	0.8419	0.8474
Local	0.8104	0.8138	0.8209	0.8228	0.8215	0.8314	0.8363	0.8401	0.8438	0.8451	0.8474
Global	0.8392	0.8456	0.8470	0.8472	0.8479	0.8479	0.8469	0.8477	0.8481	0.8474	0.8474

Table 8.83. Percentage FS(CHI), Macro-F Measure for Reuters-21578.

Policy	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	NoSelection
Document	0.3555	0.3414	0.3474	0.3395	0.3471	0.3544	0.3679	0.3758	0.3792	0.3825	0.4037
Local	0.2577	0.2816	0.2963	0.3206	0.3339	0.3526	0.3740	0.3853	0.3996	0.4039	0.4037
Global	0.3891	0.3870	0.4041	0.4043	0.3982	0.4019	0.4006	0.4039	0.4027	0.4037	0.4037

Table 8.84. Contribution of POS for Reuters-21578 - Micro-F Measure.

Contribution	1000	1500	2000
Raw	0.8383	0.8445	0.8467
Noun	0.8359	0.8395	0.8415
Verb	0.6457	0.6535	0.6573
Adjactive	0.6648	0.6600	0.6593
Adverb	0.2029	0.2029	0.2029
N+V	0.8374	0.8419	0.8419
N+Adj	0.8343	0.8377	0.8431
N+V+Adj	0.8359	0.8429	0.8460
N+V+Adj+Adv	0.8356	0.8442	0.8465

Contribution	1000	1500	2000
Raw	0.8383	0.8445	0.8467
Raw+Noun	0.8445	0.8495	0.8513
Raw+Verb	0.8345	0.8415	0.8428
Raw+Adjactive	0.8363	0.8429	0.8461
Raw+Adverb	0.8377	0.8444	0.8469
Raw+N+V	0.8436	0.8470	0.8502
Raw+N+Adj	0.8451	0.8469	0.8532
Raw+N+V+Adj	0.8442	0.8477	0.8511
Raw+N+V+Adj+Adv	0.8444	0.8474	0.8507

8.5.3. Contribution of POS

In this section we will discuss the results of using POS of terms.

In Table 8.84 and 8.85 we can say that using POS information only option does not have better results than not to use option. But when we add raw terms into the categorization, on the right table, contribution that contains nouns gives better results.

8.5.4. Contribution of WordNet Features

In this section we will discuss the contribution of WordNet features for Reuters-21578 dataset.

In Table 8.86, the contribution of WordNet features for using only nouns is given.

Table 8.85. Contribution of POS for Reuters-21578 - Macro-F Measure.

Contribution	1000	1500	2000
Raw	0.3810	0.3835	0.3837
Noun	0.3856	0.3943	0.3781
Verb	0.1478	0.1458	0.1395
Adjactive	0.1670	0.1600	0.1603
Adverb	0.0477	0.0477	0.0477
N+V	0.3745	0.3865	0.3778
N+Adj	0.3732	0.3821	0.3894
N+V+Adj	0.3789	0.3786	0.3841
N+V+Adj+Adv	0.3726	0.3804	0.3838

Contribution	1000	1500	2000
Raw	0.3810	0.3835	0.3837
Raw+Noun	0.4166	0.4197	0.4161
Raw+Verb	0.3789	0.3887	0.3804
Raw+Adjactive	0.3878	0.3893	0.3904
Raw+Adverb	0.3767	0.3831	0.3922
Raw+N+V	0.4119	0.4173	0.4218
Raw+N+Adj	0.4162	0.4212	0.4243
Raw+N+V+Adj	0.4157	0.4233	0.4268
Raw+N+V+Adj+Adv	0.4099	0.4207	0.4227

Table 8.86. WordNet Features for Reuters-21578 - Noun(L) and Raw + Noun(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8330	0.4004	Synonyms	0.8445	0.4257
Hypernyms	0.8366	0.3963	Hypernyms	0.8483	0.4289
Hyponyms	0.8190	0.3617	Hyponyms	0.8305	0.3967
Meronyms	0.8316	0.3809	Meronyms	0.8438	0.4229
Topics	0.8381	0.3850	Topics	0.8507	0.4176
Syn+Hype	0.8372	0.4205	Syn+Hype	0.8467	0.4368
Hype+Top	0.8361	0.3855	Hype+Top	0.8495	0.4267
Syn+Hype+Top	0.8361	0.4110	Syn+Hype+Top	0.8460	0.4346
Syn+Hype+Hypo+Mero+Top	0.8121	0.3891	Syn+Hype+Hypo+Mero+Top	0.8153	0.3852
No Semantic Features	0.8415	0.3781	No Semantic Features	0.8513	0.4161

Table 8.87. WordNet Features for Reuters-21578 - Noun + Verb(L) and Raw + Noun + Verb(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8399	0.4181	Synonyms	0.8486	0.4456
Hypernyms	0.8440	0.4095	Hypernyms	0.8514	0.4374
Hyponyms	0.8281	0.3688	Hyponyms	0.8319	0.4135
Meronyms	0.8365	0.3874	Meronyms	0.8451	0.4224
Topics	0.8431	0.3899	Topics	0.8495	0.4168
Syn+Hype	0.8399	0.4273	Syn+Hype	0.8458	0.4401
Hype+Top	0.8442	0.4091	Hype+Top	0.8500	0.4415
Syn+Hype+Top	0.8393	0.4235	Syn+Hype+Top	0.8461	0.4445
Syn+Hype+Hypo+Mero+Top	0.8196	0.3751	Syn+Hype+Hypo+Mero+Top	0.8252	0.3859
No Semantic Features	0.8419	0.3778	No Semantic Features	0.8502	0.4218

We can say that semantic features does not increase Micro-F measures, but when it comes to Macro-F we can clearly say that using semantic features increases the measure. In Table 8.87, the contribution of semantic features for nouns and verbs can be seen. We can say that using Hypernyms and Topics together, increase Micro-F measure for non-raw terms options. For raw term included option, we can still say that hypernyms increases the measure, but topics decreases. But when we look at Macro-F measure using semantic features always increased. In Table 8.88, the results of using semantic features for noun and adjectives can be seen. The results shows that there is no configuration that increases Micro-F measure. But when we look at Macro-F measure, we can say that almost all the configurations increased the measure. Same inferences can be done for other configurations in Tables 8.89 and 8.90.

8.5.5. Contribution of Disambiguation

In this section we will show the results of applying disambiguation for Reuters-21578 dataset. In Table 8.91, we can conclude that, applying 70% disambiguation

Table 8.88. WordNet Features for Reuters-21578 - Noun + Adj(L) and Raw + Noun + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8328	0.4162	Synonyms	0.8429	0.4462
Hypernyms	0.8408	0.4013	Hypernyms	0.8479	0.4357
Hyponyms	0.8202	0.3644	Hyponyms	0.8327	0.4011
Meronyms	0.8361	0.3927	Meronyms	0.8444	0.4209
Topics	0.8411	0.3910	Topics	0.8516	0.4235
Syn+Hype	0.8397	0.4209	Syn+Hype	0.8438	0.4387
Hype+Top	0.8390	0.3928	Hype+Top	0.8492	0.4319
Syn+Hype+Top	0.8393	0.4188	Syn+Hype+Top	0.8435	0.4320
Syn+Hype+Hypo+Mero+Top	0.8123	0.3865	Syn+Hype+Hypo+Mero+Top	0.8153	0.3964
No Semantic Features	0.8431	0.3894	No Semantic Features	0.8532	0.4243

Table 8.89. WordNet Features for Reuters-21578 - Noun + Verb + Adj(L) and Raw + Noun + Verb + Adj(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8401	0.4215	Synonyms	0.8429	0.4499
Hypernyms	0.8460	0.4216	Hypernyms	0.8516	0.4350
Hyponyms	0.8301	0.3849	Hyponyms	0.8332	0.4143
Meronyms	0.8381	0.3894	Meronyms	0.8454	0.4238
Topics	0.8445	0.3888	Topics	0.8499	0.4201
Syn+Hype	0.8426	0.4309	Syn+Hype	0.8436	0.4272
Hype+Top	0.8465	0.4118	Hype+Top	0.8504	0.4348
Syn+Hype+Top	0.8438	0.4213	Syn+Hype+Top	0.8444	0.4321
Syn+Hype+Hypo+Mero+Top	0.8224	0.3847	Syn+Hype+Hypo+Mero+Top	0.8263	0.3937
No Semantic Features	0.8460	0.3841	No Semantic Features	0.8511	0.4268

Table 8.90. WordNet Features for Reuters-21578 - Noun + Verb + Adj + Adv(L) and Raw + Noun + Verb + Adj + Adv(R).

Contribution	Micro-F	Macro-F	Contribution	Micro-F	Macro-F
Synonyms	0.8401	0.4165	Synonyms	0.8436	0.4502
Hypernyms	0.8453	0.4106	Hypernyms	0.8504	0.4382
Hyponyms	0.8297	0.3757	Hyponyms	0.8332	0.4173
Meronyms	0.8395	0.3935	Meronyms	0.8449	0.4197
Topics	0.8444	0.3846	Topics	0.8504	0.4174
Syn+Hype	0.8420	0.4314	Syn+Hype	0.8440	0.4287
Hype+Top	0.8465	0.4127	Hype+Top	0.8516	0.4462
Syn+Hype+Top	0.8433	0.4230	Syn+Hype+Top	0.8458	0.4314
Syn+Hype+Hypo+Mero+Top	0.8228	0.3792	Syn+Hype+Hypo+Mero+Top	0.8233	0.4123
No Semantic Features	0.8465	0.3838	No Semantic Features	0.8507	0.4227

Table 8.91. Hypernyms Disambiguation for Reuters-21578 - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8401	0.8435	0.8451	0.8442
Hypernyms	0.8460	0.8411	0.8428	0.8428
Hyponyms	0.8301	0.8283	0.8334	0.8388
Meronyms	0.8381	0.8453	0.8440	0.8435
Topics	0.8445	0.8440	0.8454	0.8463
Syn+Hype	0.8426	0.8415	0.8399	0.8410
Hype+Top	0.8465	0.8410	0.8444	0.8436
Syn+Hype+Top	0.8438	0.8424	0.8401	0.8420
Syn+Hype+Hypo+Mero+Top	0.8224	0.8259	0.8308	0.8323

Table 8.92. Hypernyms Disambiguation for Reuters-21578 - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4215	0.3912	0.3965	0.4044
Hypernyms	0.4216	0.3843	0.3878	0.3880
Hyponyms	0.3849	0.3740	0.3794	0.3780
Meronyms	0.3894	0.3809	0.3933	0.3721
Topics	0.3888	0.3781	0.3786	0.3835
Syn+Hype	0.4309	0.3930	0.3908	0.3804
Hype+Top	0.4118	0.3842	0.3866	0.3942
Syn+Hype+Top	0.4213	0.3964	0.3921	0.3886
Syn+Hype+Hypo+Mero+Top	0.3847	0.3632	0.3755	0.3686

increased Micro-F measure for synonyms, meronyms and all the combinations. And we can say that when the disambiguation ratio decreased the measure increased slightly. In contrast, we cannot say same thing for Macro-F measure, as can be seen in Table 8.92, applying disambiguation decreased the measure. We can say the same for raw data option as the results can be seen in Table 8.93 and 8.94.

Using topic information for disambiguation results shows that Micr0-F measure increased, but Macro-F measures decreased as in the hypernym usage. The results can

Table 8.93. Hypernyms Disambiguation for Reuters-21578 - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8436	0.8484	0.8506	0.8481
Hypernyms	0.8504	0.8449	0.8497	0.8495
Hyponyms	0.8332	0.8379	0.8381	0.8477
Meronyms	0.8449	0.8483	0.8486	0.8479
Topics	0.8504	0.8499	0.8502	0.8520
Syn+Hype	0.844	0.8479	0.8477	0.8463
Hype+Top	0.8516	0.8454	0.8500	0.8488
Syn+Hype+Top	0.8458	0.8467	0.8492	0.8472
Syn+Hype+Hypo+Mero+Top	0.8233	0.8352	0.8365	0.8429

Table 8.94. Hypernyms Disambiguation for Reuters-21578 - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4502	0.4181	0.4244	0.4206
Hypernyms	0.4382	0.4071	0.4156	0.4162
Hyponyms	0.4173	0.4053	0.3853	0.4027
Meronyms	0.4197	0.4201	0.4291	0.4143
Topics	0.4174	0.4253	0.4205	0.4276
Syn+Hype	0.4287	0.4282	0.4189	0.4156
Hype+Top	0.4462	0.4105	0.4153	0.4194
Syn+Hype+Top	0.4314	0.4240	0.4213	0.4207
Syn+Hype+Hypo+Mero+Top	0.4123	0.3963	0.3886	0.4056

Table 8.95. Topics Disambiguation for Reuters-21578 - Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8401	0.8420	0.8419	0.8447
Hypernyms	0.8460	0.8429	0.8470	0.8458
Hyponyms	0.8301	0.8428	0.8426	0.8413
Meronyms	0.8381	0.8458	0.8456	0.8451
Topics	0.8445	0.8451	0.8445	0.8463
Syn+Hype	0.8426	0.8404	0.8417	0.8411
Hype+Top	0.8465	0.8431	0.8440	0.8463
Syn+Hype+Top	0.8438	0.8401	0.8411	0.8410
Syn+Hype+Hypo+Mero+Top	0.8224	0.8395	0.8404	0.8417

be seen in Tables 8.95-98.

8.6. Summary of the Results

In this section, all the results will be evaluated together and the inferences will be given.

In analysis of existing metric sections we gave the results of using three different policies with different threshold. We can conclude that using our proposed document

Table 8.96. Topics Disambiguation for Reuters-21578 - Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4215	0.3843	0.3826	0.3826
Hypernyms	0.4216	0.3840	0.3882	0.3839
Hyponyms	0.3849	0.3797	0.3753	0.3868
Meronyms	0.3894	0.3805	0.3809	0.3762
Topics	0.3888	0.3817	0.3792	0.3844
Syn+Hype	0.4309	0.3834	0.3811	0.3730
Hype+Top	0.4118	0.3855	0.3828	0.3802
Syn+Hype+Top	0.4213	0.3865	0.3798	0.3741
Syn+Hype+Hypo+Mero+Top	0.3847	0.3831	0.3809	0.3868

Table 8.97. Topics Disambiguation for Reuters-21578 - Raw + Noun + Verb + Adj - Micro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.8436	0.8495	0.8497	0.8486
Hypernyms	0.8504	0.8507	0.8513	0.8529
Hyponyms	0.8332	0.8486	0.8497	0.8497
Meronyms	0.8449	0.8500	0.8511	0.8499
Topics	0.8504	0.8507	0.8507	0.8509
Syn+Hype	0.844	0.8476	0.8497	0.8527
Hype+Top	0.8516	0.8511	0.8514	0.8520
Syn+Hype+Top	0.8458	0.8477	0.8499	0.8509
Syn+Hype+Hypo+Mero+Top	0.8233	0.8467	0.8474	0.8495

Table 8.98. Topics Disambiguation for Reuters-21578 - Raw + Noun + Verb + Adj - Macro-F.

Contribution	No Disambiguation	70%	50%	30%
Synonyms	0.4502	0.4250	0.4238	0.4231
Hypernyms	0.4382	0.4286	0.4248	0.4322
Hyponyms	0.4173	0.4208	0.4248	0.4228
Meronyms	0.4197	0.4236	0.4258	0.4248
Topics	0.4174	0.4240	0.4280	0.4275
Syn+Hype	0.4287	0.4262	0.4281	0.4290
Hype+Top	0.4462	0.4267	0.4260	0.4292
Syn+Hype+Top	0.4314	0.4293	0.4269	0.4258
Syn+Hype+Hypo+Mero+Top	0.4123	0.4211	0.4276	0.4268

policy cannot be an alternative to resolve the problem of high dimensionality. Using document policy may increase the Micro and Macro F measure, because it uses many terms as opposed to other policies. As the number of total selected terms are not equal, it will not be fair to compare them together.

In contribution of POS sections we gave the result of using POS information of terms. The results show that using POS of term increases both Micro and Macro F measures. And it can be concluded that using POS with raw terms gives betters results than not to use them. In addition, using noun, adjective and verb results in better results. It can be said that using adverb decreases the categorization performance.

In contribution of WordNet features sections we gave the results of using WordNet features. It can be said that, for all dataset, using WordNet features increases both Micro and Macro F measure. There are five different features in the results; synonym, hypernym, hyponym, meronym and topic. We can say that using all of them together gives better results than using individually or combination of them. But if we have

to give the order of importance, when all the results are evaluated, it is: hypernym, synonym, topic, hyponym and meronym.

In contribution of disambiguation sections we gave results of applying disambiguation compared with no disambiguation. The results show that applying disambiguation increases the both Micro and Macro F measure. When the disambiguation rate increases (The given rate gives the amount of best synsets) the better results are taken.

9. CONCLUSION

In this study, we present a comprehensive analysis of using semantic features in text categorization. Firstly, we analyze the existing metrics with global, local and document policy with different thresholds. We can conclude that global and local policies achieve better results than document policy.

Secondly, we analyze using POS tag of word with and without raw terms and evaluate the performance of classification. The results show that using POS without raw features rarely gives better results, but, with raw features achieve best results.

In adition, we analyze the use of WordNet features; synonyms, hypernyms, hyponyms, meronyms and topics. And results show that using synonyms, hypernyms, hyponyms and topics gives better results.

Finally, to eliminate the ambiguity, we propose a disambiguation method that gain better results, especially in Micro-F measure, when compared to no disambiguation option.

As future work, we will make use of other WordNet's features such as holonyms, troponym, entailment, etc. In addition, greedy based disambiguation mechanism that calculates the score of synsets by measuring the similarity, can be improved. Moreover, more datasets can be used to show success of the proposed methodologies.

APPENDIX A: STOP WORD LIST

a	able	about	above	according	accordingly	across	actually
after	afterwards	again	against	ain't	all	allow	allows
almost	alone	along	already	also	although	always	am
among	amongst	an	and	another	any	anybody	anyhow
anyone	anything	anyway	anyways	anywhere	apart	appear	appreciate
appropriate	are	aren't	around	as	a's	aside	ask
asking	associated	at	available	away	awfully	b	be
became	because	become	becomes	becoming	been	before	beforehand
behind	being	believe	below	beside	besides	best	better
between	beyond	both	brief	but	by	c	came
can	cannot	cant	can't	cause	causes	certain	certainly
changes	clearly	c'mon	co	com	come	comes	concerning
consequently	consider	considering	contain	containing	contains	corresponding	could
couldn't	course	c's	currently	d	definitely	described	despite
did	didn't	different	do	does	doesn't	doing	done
don't	down	downwards	during	e	each	edu	eg
eight	either	else	elsewhere	enough	entirely	especially	et
etc	even	ever	every	everybody	everyone	everything	everywhere
ex	exactly	example	except	f	far	few	fifth
first	five	followed	following	follows	for	former	formerly
forth	four	from	further	furthermore	g	get	gets
getting	given	gives	go	goes	going	gone	got
gotten	greetings	h	had	hadn't	happens	hardly	has
hasn't	have	haven't	having	he	hello	help	hence
her	here	hereafter	hereby	herein	here's	hereupon	hers
herself	he's	hi	him	himself	his	hither	hopefully
how	howbeit	however	i	i'd	ie	if	ignored
i'll	i'm	immediate	in	inasmuch	inc	indeed	indicate
indicated	indicates	inner	insofar	instead	into	inward	is
isn't	it	it'd	it'll	its	it's	itself	i've
j	just	k	keep	keeps	kept	know	known
knows	l	last	lately	later	latter	latterly	least
less	lest	let	let's	like	liked	likely	little
look	looking	looks	ltd	m	mainly	many	may
maybe	me	mean	meanwhile	merely	might	more	moreover
most	mostly	much	must	my	myself	n	name
namely	nd	near	nearly	necessary	need	needs	neither
never	nevertheless	new	next	nine	no	nobody	non
none	noone	nor	normally	not	nothing	novel	now
nowhere	o	obviously	of	off	often	oh	ok
okay	old	on	once	one	ones	only	onto
or	other	others	otherwise	ought	our	ours	ourselves
out	outside	over	overall	own	p	particular	particularly
per	perhaps	placed	please	plus	possible	presumably	probably
provides	q	que	quite	qv	r	rather	rd
re	really	reasonably	regarding	regardless	regards	relatively	respectively
reuter	right	s	said	same	saw	say	saying
says	second	secondly	see	seeing	seem	seemed	seeming
seems	seen	self	selves	sensible	sent	serious	seriously
seven	several	shall	she	should	shouldn't	since	six
so	some	somebody	somehow	someone	something	sometime	sometimes
somewhat	somewhere	soon	sorry	specified	specify	specifying	still
sub	such	sup	sure	t	take	taken	tell
tends	th	than	thank	thanks	thanx	that	thats
that's	the	their	theirs	them	themselves	then	thence
there	thereafter	thereby	therefore	therein	theres	there's	thereupon
these	they	they'd	they'll	they're	they've	think	third

this	thorough	thoroughly	those	though	three	through	throughout
thru	thus	to	together	too	took	toward	towards
tried	tries	truly	try	trying	t's	twice	two
u	un	under	unfortunately	unless	unlikely	until	unto
up	upon	us	use	used	useful	uses	using
usually	uucp	v	value	various	very	via	viz
vs	w	want	wants	was	wasn't	way	we
we'd	welcome	well	we'll	went	were	we're	weren't
we've	what	whatever	what's	when	whence	whenever	where
whereafter	whereas	whereby	wherein	where's	whereupon	wherever	whether
which	while	whither	who	whoever	whole	whom	who's
whose	why	will	willing	wish	with	within	without
wonder	won't	would	would	wouldn't	x	y	yes
yet	you	you'd	you'll	your	you're	yours	yourself
yourselves	you've	z	zero				

REFERENCES

1. Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47, 2002.

2. Cortes, C. and V. Vapnik, "Support-Vector Networks", *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

3. Forman, G., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", *Journal of Machine Learning Research*, Vol. 3, pp. 1289–1305, 2003.

4. Dumais, S., J. Platt, D. Heckerman and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization", *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, pp. 148–155, ACM, New York, NY, USA, 1998.

5. Yang, Y. and X. Liu, "A Re-examination of Text Categorization Methods", *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pp. 42–49, ACM, New York, NY, USA, 1999.

6. Salton, G., A. Wong and C. S. Yang, "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620, 1975.

7. Chua, S. and N. Kulathuramaiyer, "Semantic Feature Selection Using WordNet", *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pp. 166–172, IEEE Computer Society, Washington, DC, USA, 2004.

8. Zhang, K., J. Sun and B. Wang, "Intelligent information processing II", chap. A WordNet-based Approach to Feature Selection in Text Categorization, pp. 475–

484, Springer-Verlag, London, UK, UK, 2005.

9. Bloehdorn, S. and A. Hotho, "Boosting for Text Classification with Semantic Features", *Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis*, WebKDD'04, pp. 149–166, Springer-Verlag, Berlin, Heidelberg, 2006.

10. Li, J., Y. Zhao and B. Liu, "Fully Automatic Text Categorization by Exploiting WordNet", *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '09, pp. 1–12, Springer-Verlag, Berlin, Heidelberg, 2009.

11. Mansuy, T. N. and R. J. Hilderman, "Evaluating WordNet Features in Text Classification Models", *The Florida AI Research Society Conference*, pp. 568–573, 2006.

12. Özgür, A., L. Özgür and T. Güngör, "Text Categorization with Class-based and Corpus-based Keyword Selection", *Proceedings of the 20th international conference on Computer and Information Sciences*, ISCIS'05, pp. 606–615, Springer-Verlag, Berlin, Heidelberg, 2005.

13. Zhang, W., T. Yoshida and X. Tang, "Text Classification Based on Multi-word with Support Vector Machine", *Knowledge-Based Systems*, Vol. 21, No. 8, pp. 879–886, 2008.

14. Porter, M. F., "Readings in information retrieval", chap. An Algorithm for Suffix Stripping, pp. 313–316, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

15. Porter, M. F., "POS Tagging", `http://tartarus.org/ martin/PorterStemmer/`, accessed April 4, 2012.

16. Salton, G. and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval", *Information Processing And Management*, Vol. 24, No. 5, pp. 513–523,

1988.

17. Zobel, J. and A. Moffat, "Exploring the Similarity Space", *SIGIR Forum*, Vol. 32, No. 1, pp. 18–34, 1998.

18. Miller, G. A., "WordNet: A Lexical Database for English", *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

19. "Review of "WordNet: An Electronic Lexical Database" by Christiane Fellbaum", *Computational Linguistics*, Vol. 25, No. 2, pp. 292–296, 1999, reviewer-Lin, Dekang.

20. Debole, F. and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization", *Proceedings of the 2003 ACM symposium on Applied computing*, SAC '03, pp. 784–788, ACM, New York, NY, USA, 2003.

21. Wanas, N. M., D. A. Said, N. H. Hegazy and N. M. Darwish, "A Study of Local and Global Thresholding Techniques in Text Categorization", *Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61*, AusDM '06, pp. 91–101, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2006.

22. Yang, Y. and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pp. 412–420, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

23. Calvo, R. A. and H. A. Ceccatto, "Intelligent Document Classification", *Intelligent Data Analysis*, Vol. 4, No. 5, pp. 411–420, 2000.

24. Tasci, S., *An Evaluation of Existing and New Feature Selection Metrics in Text Categorization*, Ph.D. Thesis, Bogaziçi University, 2006.

25. Watson, M., "POS Tagging", http://www.markwatson.com/opensource/, accessed May 14, 2012.

26. "Disambiguation", http://en.wikipedia.org/wiki/Wikipedia:Disambiguation/, accessed March 20, 2012.

27. Yizhang, G., "Methods for Pattern Classification", *New Advances In Machine Learning*, pp. 49–74, InTech Press, 2010.

28. Leopold, E. and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?", *Machine Learning*, Vol. 46, No. 1-3, pp. 423–444, 2002.

29. Clifton, C., R. Cooley and J. Rennie, "TopCat: Data Mining for Topic Identification in a Text Corpus", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 8, pp. 949–964, 2004.

30. Kwok, J. T., "Automated Text Categorization Using Support Vector Machine", *In Proceedings of the International Conference on Neural Information Processing (ICONIP*, pp. 347–351, 1998.

31. Joachims, T., "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pp. 137–142, Springer-Verlag, London, UK, UK, 1998.

32. Manning, C. D., P. Raghavan and H. Schtze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

33. Rokach, L. and O. Maimon, *Data Mining with Decision Trees: Theroy and Applications*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008.

34. Jones, R., A. Mccallum, K. Nigam and E. Riloff, "Bootstrapping for Text Learning Tasks", *In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and*

Applications, pp. 52–63, 1999.

35. Nigam, K. and R. Ghani, "Analyzing the Effectiveness and Applicability of Co-training", *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM '00, pp. 86–93, ACM, New York, NY, USA, 2000.

36. Nigam, K., A. K. McCallum, S. Thrun and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning*, Vol. 39, No. 2-3, pp. 103–134, 2000.