



Faculty of Computer Science and Information Technology

**USING WORDNET TO ENHANCE FEATURE SELECTION
IN AUTOMATED TEXT CATEGORIZATION**

Stephanie Chua Hui Li

**Master of Science
(Information Technology)
2004**

**USING WORDNET TO ENHANCE FEATURE SELECTION IN AUTOMATED
TEXT CATEGORIZATION**

P.KHIDMAT MAKLUMAT AKADEMIK

UNIMAS



1000246303

STEPHANIE CHUA HUI LI

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2004

Acknowledgment

First and foremost, I would like to extend my utmost gratitude to my supervisor, Associate Professor Narayanan Kulathuramaiyer for all his invaluable guidance and encouragement throughout this research. I would also like to thank Mr Bong Chih How for his resources, advice and ideas pertaining to this research. My acknowledgment also goes to the Pelan Tindakan Pembangunan Teknologi Perindustrian (PTPTP) for the financial support provided. I also have all my friends to thank for their kindness and understanding as well as the discussions and support given, especially during the difficult times. Finally, I am deeply grateful to my beloved family for their unconditional love, patience, support and encouragement.

Abstract

(In the field of automated text categorization, the large dimensionality of the feature space is a major problem as it involves extensive computations. Feature selection is one of the approaches to reduce the dimensionality of the feature space. This research explores the use of WordNet (Miller et al., 1990), a lexical database, for performing feature selection for an automated text categorization system. The WordNet-based approach employs lexical and semantics information for feature selection. WordNet allows the selection of terms that are lexically and semantically representative of a category of documents, as opposed to statistical approaches traditionally used for feature selection. ✓

We proposed three WordNet based approaches for feature selection. The first one is to use the WordNet nouns approach that selects all nouns in WordNet that occur in each category as features. The second approach is based on lexical semantics that selects synonymous terms that co-occur in a category while the third approach is a combination of the lexical semantics approach with statistical feature selection methods.

The lexical semantics approach performed better than the WordNet nouns approach with more than 40% of reduction in feature space in the experiments using the Reuters-21578 dataset. The lexical semantics approach also outperformed popular statistical feature selection methods, namely, Chi-Square (Chi2) and Information Gain (IG). The combined approach has improved the performance of the statistical methods. WordNet has successfully been used to enhance feature selection,

highlighting the possibility of determining semantic features automatically. The limitations of the lexical semantics approach are also highlighted, proposing an improved framework and an extension to overcome them.

Abstrak

Tajuk tesis ini ialah “Menggunakan WordNet untuk mempertingkatkan kaedah pemilihan teks dalam bidang klasifikasi teks secara automatik”. Dalam bidang klasifikasi teks secara automatik, dimensi teks yang besar merupakan suatu masalah yang besar kerana ia memerlukan kuasa komputasi yang tinggi. Pemilihan teks merupakan salah satu cara untuk mengurangkan saiz dimensi teks. Penyelidikan ini menggunakan WordNet (Miller et al., 1990), sebuah pangkalan data leksikal, untuk menjalankan proses pemilihan teks dalam sistem klasifikasi teks secara automatik ini. Kaedah pemilihan teks melalui WordNet ini menggunakan maklumat leksikal dan semantik untuk proses pemilihan teks dan ini membolehkan pemilihan perkataan yang selaras dengan topik kategori. Ini berbeza dengan kaedah pemilihan teks statistik yang selama ini digunakan untuk mengurangkan saiz dimensi teks.

Kami mencadangkan tiga kaedah pemilihan teks yang menggunakan WordNet. Kaedah pertama adalah kaedah kata nama WordNet yang memilih semua kata nama terkandung dalam WordNet yang wujud dalam setiap kategori untuk dijadikan perkataan wakil kategori tersebut. Kaedah kedua adalah berdasarkan maklumat leksikal semantik yang memilih sinonim yang wujud bersama dalam suatu kategori manakala kaedah ketiga ialah kombinasi kaedah leksikal semantik dengan kaedah pemilihan teks statistik.

Kaedah leksikal semantik adalah lebih efektif berbanding dengan kaedah kata nama WordNet dengan lebih daripada 40% pengurangan saiz dimensi perkataan dalam eksperimen yang menggunakan set data Reuters-21578. Kaedah leksikal semantik juga

berjaya memperoleh keputusan yang lebih baik daripada kaedah statistik popular, iaitu, Chi-square (Chi2) dan Information Gain (IG). Kaedah kombinasi telah berjaya mempertingkatkan keputusan yang dicapai oleh kaedah statistik. WordNet telah berjaya digunakan untuk mempertingkatkan kualiti proses pemilihan teks dengan kaedah leksikal semantik ini lebih efektif berbanding dengan kaedah statistik yang sedia ada, iaitu, Chi2 dan IG. Dengan itu, kaedah ini dilihat sebagai satu kaedah pemilihan teks yang berpotensi. Limitasi kaedah leksikal semantik ini telah dikenalpasti dan rangka kerja yang lebih baik serta lanjutan untuk mengatasi limitasi kaedah tersebut telah dicadangkan.

Table of Contents

	Page
Acknowledgment	i
Abstract	ii
Abstrak	iv
Table of Contents	vi
List of Figures	xii
List of Tables	xiv
Chapter 1 Introduction	
1.1 Introduction to Text Categorization	1
1.2 Machine Learning Approach to Text Categorization	2
1.3 Framework for Automated Text Categorization	3
1.4 Dimensionality Problem	5
1.5 Feature Selection vs Feature Extraction	6
1.5.1 Statistical Feature Selection	7
1.5.2 Natural Language Processing (NLP) Approach to Feature Selection	8
1.6 Problem Statement	9
1.7 Research Objectives	10
1.8 Scope	11
1.9 Chapters Overview	12

Chapter 2 Literature Review

2.1	Introduction	14
2.2	Text Representation for Text Categorization	14
2.3	Automatic Approach to Text Categorization	15
	2.3.1 Knowledge Engineering	15
	2.3.2 Machine Learning	16
2.4	Machine Learning Approach to Text Categorization	17
	2.4.1 Naïve Bayes (NB) Classifier	19
	2.4.2 Single-label vs Multi-label Categorization	22
2.5	Dimensionality Reduction by Feature Extraction	22
	2.5.1 Term Clustering	23
	2.5.2 Latent Semantic Indexing (LSI)	24
2.6	Dimensionality Reduction by Feature Selection	25
	2.6.1 Information Gain (IG)	26
	2.6.2 Chi-Square (Chi2)	27
	2.6.3 Discussion on Information Gain and Chi-Square	28
2.7	Word Sense Disambiguation	29
2.8	Introduction to WordNet	31
2.9	Incorporating WordNet in the Task of Text Categorization	32
	2.9.1 WordNet in Indexing	33
	2.9.2 WordNet in Text Representation	34
	2.9.3 WordNet in Semantic Distance Measurement	36
	2.9.4 WordNet in Feature Selection	38
2.10	Comparison and Discussion	39
2.11	Evaluation Measures	41

2.11.1	Performance Measures	41
2.11.2	Qualitative Analysis	46
2.12	Summary	46
 Chapter 3 Design		
3.1	Introduction	48
3.2	Text Categorization Framework	49
3.2.1	Document Preprocessing	50
3.2.2	Feature Selection	52
3.2.2.1	WordNet Nouns Approach	52
3.2.2.2	WordNet Lexical Semantics Approach	53
3.2.2.3	Categorical Sense Determination	55
3.2.2.4	Combined Approach	56
3.2.3	Machine Learning for Automated Text Categorization	57
3.3	Summary	60
 Chapter 4 Implementation		
4.1	Introduction	61
4.2	Implementation	61
4.2.1	Document Preprocessing	64
4.2.2	Feature Selection	65
4.2.2.1	WordNet Nouns Approach	65
4.2.2.2	WordNet Lexical Semantics Approach	66
4.2.2.3	Combined Approach	67
4.2.3	Machine Learning	68

4.2.3.1	Machine Learning Tool: Waikato Environment for Knowledge Analysis (WEKA)	71
4.3	Summary	72
Chapter 5 Results and Analysis		
5.1	Introduction	73
5.2	Dataset	73
5.3	Experiments	74
5.4	Results and Analysis for Reuters-21578 Dataset	76
5.4.1	Dimensionality Reduction	76
5.4.2	Experiment 1: WN vs WLS	77
5.4.2.1	F-measure Analysis	77
5.4.2.2	Dimensionality Reduction	79
5.4.2.3	Qualitative Analysis	80
5.4.3	Experiment 2: Comparison between WN, WLS, Chi2 and IG	81
5.4.3.1	F-measure Analysis	81
5.4.3.2	Qualitative Analysis	86
5.4.3.3	Discussion	89
5.4.4	Experiment 3: Comparison between Chi2, IG and Combined Approaches	92
5.4.4.1	F-measure Analysis	92
5.4.4.2	Discussion	94
5.5	Summary	95

Chapter 6 Summary and Future Works

6.1	Introduction	97
6.2	Research Contributions	97
6.3	Limitations	99
6.4	Future Works	100
	6.4.1 Concepts Identification Based on WordNet's Hypernyms	100
	6.4.2 Synonymization Based on WordNet's Synsets	104
	6.4.3 Improved Framework for WordNet based Automated Text Categorization	105
6.5	Conclusion	106

List of References	107
---------------------------	------------

Appendices

APPENDIX A – Comparison of WLS (with singular and plural nouns) vs WLS (with singular nouns only)	117
APPENDIX B – Stop Words List	118
APPENDIX C – Index File Format	124
APPENDIX D – Comparison of WLS (nouns only) vs WLS (all POS)	127
APPENDIX E – ARFF File Format	128
APPENDIX F – Correlation Test Results	132
APPENDIX G – Extract of Codes	134
APPENDIX H – Conference Paper Accepted for the 2 nd International Conference on Artificial Intelligence in Engineering and Technology (ICAIET 2004), Kota Kinabalu, Sabah, 3 rd – 5 th August, 2004	136

International Conference on Web Intelligence (WI 2004), Beijing,
China, 20th – 24th September, 2004

List of Figures

	Page
Figure 1.1: Framework for automated text categorization	5
Figure 2.1: A HCG showing the concept “car” (extracted from Richardson and Smeaton, 1995)	38
Figure 3.1: Text categorization framework	50
Figure 3.2: Document preprocessing	51
Figure 3.3: An extract of the Reuters-21578 index file	51
Figure 3.4: The WordNet nouns approach	53
Figure 3.5: The WordNet lexical semantics approach	54
Figure 3.6: The combined approach	57
Figure 3.7: The machine learning approach for text categorization	59
Figure 4.1: Context diagram for the text categorization process	63
Figure 4.2: Child diagram for process 2 – Create index file for dataset	65
Figure 4.3: Child diagram of process 6 – Cross-referencing	67
Figure 4.4: Examples of synsets	67
Figure 4.5: Child diagram of process 4 – Perform statistical feature selection	68
Figure 4.6: Child diagram for process 9 – Create ARFF files	70
Figure 4.7: An extract of an ARFF file	70
Figure 4.8: Screenshot of the output from the machine learning process	72
Figure 5.1: Macro-averaged f-measure comparison between WLS and WN for the Reuters-21578 dataset	78

Figure 5.2:	Micro-averaged f-measure comparison between LS and WN for the Reuters-21578 dataset	78
Figure 5.3:	Macro-averaged f-measure for the Reuters-21578 dataset	82
Figure 5.4:	Micro-averaged f-measure for the Reuters-21578 dataset	83
Figure 5.5:	Macro-averaged f-measure for the Reuters-21578 top ten categories	84
Figure 5.6:	Micro-averaged f-measure for the Reuters-21578 top ten categories	85
Figure 5.7:	Comparison of the performance of WLS, Chi2 and IG in different categories at term size 100 for the Reuters-21578 top ten categories	89
Figure 5.8:	Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 for the Reuters-21578 top ten categories	93
Figure 5.9:	Comparison of micro-averaged f-measure between IG and WLS-IG for the Reuters-21578 top ten categories	93
Figure 6.1:	List of 25 unique beginners for WordNet nouns (extracted from Miller et al., 1990)	101
Figure 6.2:	Diagrammatic representation of hyponymic relations among seven unique beginners denoting different kinds of tangible things (extracted from Miller et al., 1990)	102
Figure 6.3:	Proposed framework to improve automated text categorization using the WordNet-based approaches	106

List of Tables

	Page
Table 2.1: Mathematical formula for feature selection methods (extracted from Debole and Sebastiani, 2003)	26
Table 2.2: Criteria for performance of feature selection approaches (extracted from Yang and Pedersen, 1997)	29
Table 2.3: The semantic relationship in WordNet	32
Table 2.4: Comparison between works using WordNet	40
Table 2.5: Contingency table for binary categorization (extracted from Witten and Frank, 2000)	42
Table 2.6: Computation for the performance metrics (extracted from Sebastiani, 1999)	43
Table 3.1: List of terms and their synsets in all senses	55
Table 5.1: The number of categories in each category set and the number of categories with at least 1 occurrence and 20 occurrences	74
Table 5.2: Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the number of unique terms in each category	77
Table 5.3: Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the WN approach	79
Table 5.4: Top 20 terms for category "Interest" for WLS and WN	80
Table 5.5: Macro-averaged f-measure for the Reuters-21578 dataset	82
Table 5.6: Micro-averaged f-measure for the Reuters-21578 dataset	82

Table 5.7:	Macro-averaged f-measure for the Reuters-21578 top ten categories	84
Table 5.8:	Micro-averaged f-measure for the Reuters-21578 top ten categories	84
Table 5.9:	F-measure for each category for Chi2, IG and WLS for the Reuters-21578 top ten categories	86
Table 5.10:	Top 20 terms for category “Acq” for Chi2, IG and WLS	87
Table 5.11:	Top 20 terms for category “Earn” for Chi2, IG and WLS	87
Table 5.12:	Top 20 terms for category “Interest” for Chi2, IG and WLS	87
Table 5.13:	Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 as well as IG and WLS-IG for the Reuters-21578 top ten categories	92
Table 5.14:	F-measure for each category for Chi2 and WLS-Chi2 as well as IG and WLS-IG at term size 10 for the Reuters-21578 top ten categories	94

Chapter 1 Introduction

1.1 Introduction to Text Categorization

Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns suggested by a training set of categorized documents (Yang and Liu, 1999). Text categorization has been actively pursued since the early 1960s. Classifiers are used to classify documents in text categorization. Previously, classifiers were manually built by knowledge engineering effort. This knowledge engineering approach requires experts to manually build classifiers by defining rules for categorization (refer to Section 2.3.1). Even today, categorization task is still largely carried out manually by human experts. With the increase in digital documents over the years, manual categorization by the experts becomes overwhelming. Therefore, an automated approach to text categorization is very much needed to assist in the categorization tasks.

With the escalating growth of new applications, more researches are now being carried out in the field of automated text categorization. Automated text categorization has been applied in many areas such as document filtering, automatic document indexing and search engines.

There are three approaches to automated text categorization. They are the machine learning approach, the similarity measurement approach and the knowledge-based approach. The machine learning approach uses learning schemes such as rule inducers, decision trees, support vector machines (SVM) and so on to perform

categorization. The similarity measurement approach uses a similarity measurement algorithm to determine the degree of similarity of a new document with the documents in a category to do categorization. The knowledge-based approach obtains knowledge from human expert to categorize documents. In this research, the focus is on the machine learning approach. According to Sebastiani (1999), this approach is the fastest and most effective approach to text categorization.

1.2 Machine Learning Approach to Text Categorization

Machine learning is the study of computer algorithms that automatically progress towards improvement of performance by learning experience. Applications of machine learning include data mining programs that discover general rules in large datasets, information filtering systems that automatically learn users' interests and categorization systems that classify documents into categories.

Machine learning has been applied to the field of text categorization since the early 1990s. In this approach, the classifier for a category is automatically built by learning from the documents in the training set, which has been classified manually by experts. The learner will generate a classifier by studying the characteristics of the documents from the training set and then the classifier will be used to classify new and unseen documents from the testing set. This type of learning is called supervised learning because the learning process is coupled with the knowledge of the categories to which a document belongs in the training set. There is also an unsupervised learning method where the process of learning is performed without any prior category knowledge or information. It is called unsupervised because the learning

process is not guided by category labels attached to the documents. This research adopts the supervised machine learning framework to exploit the available manually-classified documents.

1.3 Framework for Automated Text Categorization

The framework for automated text categorization is shown in Figure 1.1. The dataset to be used for text categorization is divided into the training set and the testing set. The training set is used for constructing a classifier through machine learning while the testing set is used for evaluating the effectiveness of the trained classifier. Indexing is the process of cataloging the documents in the training and testing set into a suitable representation of the documents' content. Before indexing is done, document preprocessing is first performed on the dataset. The general processes involved in document preprocessing are stop words removal and stemming. Stop words removal is performed to remove non-significant terms, such as articles and prepositions while stemming is a process of normalizing terms by removing prefixes and suffixes to derive root terms. There are two kinds of stemming; one is based on inflectional morphology while the other is based on derivational morphology. Stemming based on inflectional morphology deals with the change of a word form, usually by removing a suffix for plural words while stemming based on derivational morphology concerns with removing prefixes and suffixes of words to obtain the root form.

Stemming based on derivational morphology was not employed in this research, as this results in stemmed words losing their original meaning. For example, the word

“international” will give the stemmed form of “intern”. As such, we have explored the use of stemming based on inflectional morphology, whereby we make use of the WordNet’s dictionary to derive root terms from plural forms. The use of stemming in text categorization has been questioned by Sebastiani (1999) as there are studies (Baker and McCallum, 1998) reporting that it decreases the effectiveness in categorization. Moreover, stemming on large and full text database is very time-consuming as pointed out by Korfhage (1997). According to Korfhage, stemming only improves document representation by 5%, and thus, has a lesser impact on full text documents. The result of inflectional stemming as used in our research is shown in Appendix A. Although the number of terms is reduced when plural forms of words are excluded in the feature space, however, the results of using these features is slightly worse than that having the plural forms. This is because no stemming was performed to the test documents prior to the categorization task. Even with a minor reduction in performance, the reduction in feature space is considered significant.

Dimensionality reduction is performed to reduce the number of features in the feature space. Two popular approaches for dimensionality reduction are feature selection and feature extraction. These approaches will be discussed in Section 1.5.

After dimensionality reduction, the selected features will be used to train a classifier. The classifier will learn to categorize documents based on the features provided. After the learning process, the learned classifier could then be used to classify new and unseen documents. The testing set represents new and unseen documents, as this dataset is not involved in the induction of the classifier. In this way, the performance of the classifier is evaluated.

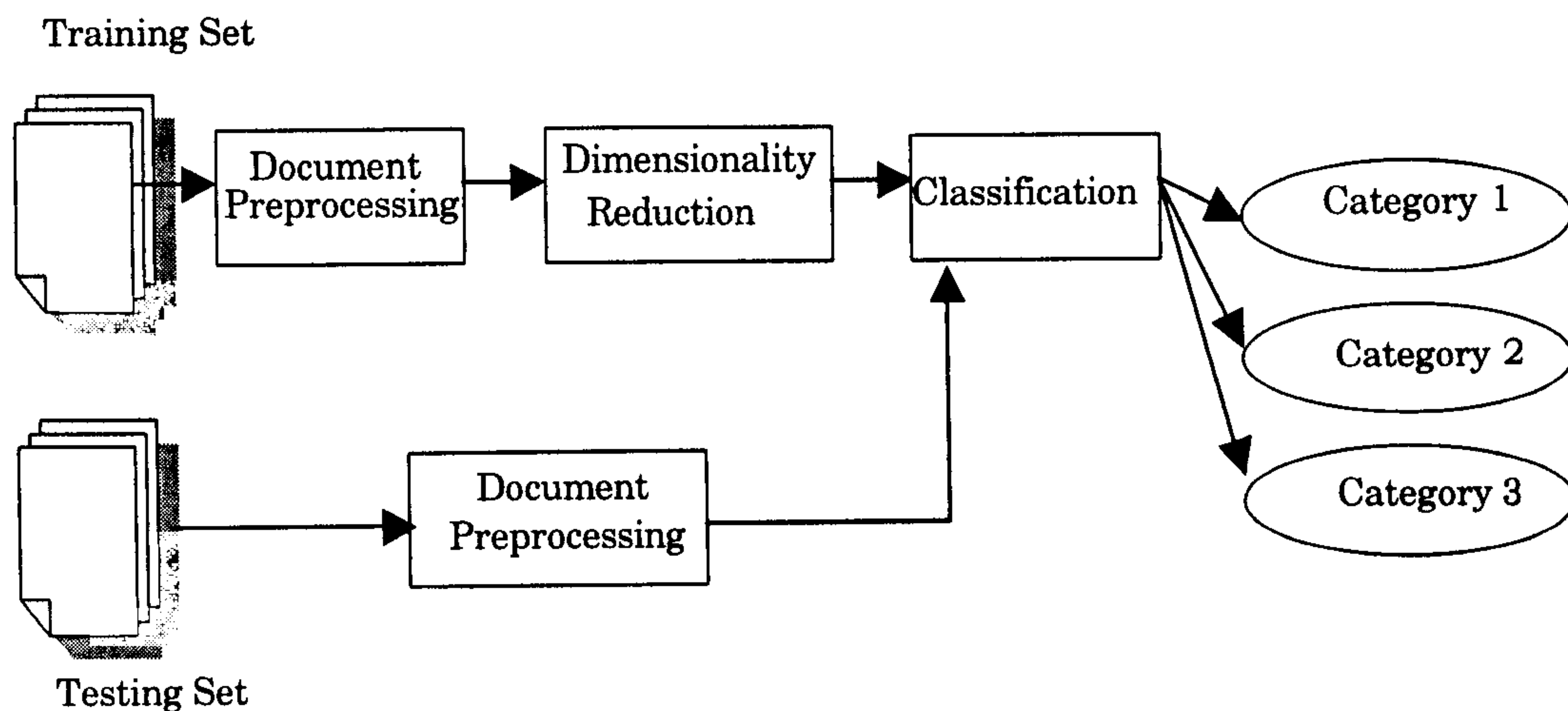


Figure 1.1: Framework for automated text categorization

1.4 Dimensionality Problem

The large dimensionality of the feature space is a well-known problem in text categorization. One document can easily contain hundreds to thousands of words and this number increases as the number of documents increases. This creates a problem for the learning process by the induced classifiers, as these classifiers are unable to handle the enormously high number of terms derived from the documents. Therefore, dimensionality reduction has to be performed where the number of terms, t is reduced to t' , where $t' < t$.

One advantage of dimensionality reduction is that it reduces the problem of overfitting. Overfitting describes the characteristics of a classifier that is particularly tuned towards the characteristics of the training data. In other words, the classifier is extremely good in classifying the data that it is trained on but is not as good at classifying other data. Overfitting can be avoided when training captures constitutive

features rather than contingent ones. Constitutive features are the terms that are generally representative of a category. On the other hand, contingent features are based on specific patterns that occur in the training set that need not actually be representative of the category and this usually causes the problem of overfitting. Sebastiani (1991) gave an example that, if a classifier for category “Cars for sale” were trained on just three positive examples where two of the examples contained the sale of a yellow car, then the classifier will consider the contingent feature “yellow” as a feature for the category. To avoid overfitting, experiments by Fuhr and Buckley (1991) suggested a ratio of 50 to 100 training examples per feature. However, if dimensionality reduction is performed, overfitting may be avoided even with a smaller set of training examples.

1.5 Feature Selection vs Feature Extraction

Both feature selection and feature extraction are approaches for dimensionality reduction. Feature selection is a process of selecting a subset of features from the feature space. It is performed to select meaningful features, representing the thousands of words in the training set. These selected features, when used in categorization task, should yield the smallest reduction in effectiveness when compared to using the whole feature set.

Feature extraction, on the other hand, is the creation of a new set of terms from the original set. This means that the new set of terms is not contained in the documents but is created from the existing terms to represent a category. Two popular feature extraction approaches are term clustering and latent semantic indexing (LSI). LSI is

a method that compresses the original features in the vector space into a new lower dimensional vector space that consist of the combinations of the original features based on their patterns of co-occurrence in the original vector space. Term clustering tends to deal with the problem of terms redundancy by grouping strongly related words into clusters. LSI has been promising but it is difficult to see why it works well as the features are latent in the semantic structure of the vocabulary used in the corpus (Sebastiani, 1999). A discussion on term clustering and LSI is presented in Section 2.5. In this research, we adopted feature selection to reduce the dimensionality of the vector space. We chose feature selection, as the features derived are made explicit for further analysis.

1.5.1 Statistical Feature Selection

Mathematical formulas are derived from the statistical information of documents and the terms and frequencies in a dataset are used to calculate a weight for each term in a document. The mathematical formulas, ranging from the simple Document Frequency to the more complex Chi-Square (Chi2) and Information Gain (IG) formulae, calculate the weight of each term to determine whether the term should be added to the feature space. Each term that surpasses a threshold value will be added as a feature in the feature space to represent a category. Both Chi2 and IG are feature selection approaches that are purely based on statistics of terms in the dataset. Therefore, these two approaches do not consider the semantics of each term to determine the relevance of the category it represents. Chi2 and IG are further discussed in Section 2.6.

1.5.2 Natural Language Processing (NLP) Approach to Feature Selection

Natural language processing (NLP) entails the use of computers to understand human language. NLP is used in a number of applications, among them, machine translation, information retrieval, text categorization, word sense disambiguation and human computer interaction. NLP approaches allow the user to exploit the lexical relations and semantics of each word in the documents.

Lexical information concerns with definitions and part-of-speech (POS) information for words while semantics is the study of word meanings. NLP employs digital dictionaries and word databases to handle the linguistics aspects of text documents. A commonly used resource in the NLP approach is WordNet (Richardson et al, 1994; Miller, 1995; Scott and Matwin, 1998; Masuyama and Nakagawa, 2002).

WordNet is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University by a group led by Professor George A. Miller (Miller et al, 1990). WordNet is a valuable resource that contains lexical and semantic information that can be exploited in many research works. At present, WordNet contains over 140,000 different word forms that are categorized into four different part-of-speech (POS): nouns, verbs, adjectives and adverbs. These have been organized into about 110,000 word meanings or set of synonyms.

We will make use of WordNet in this research as it contains common and general terms in the English language. Besides that, as mentioned before this, it also contains over 140,000 terms that are categorized into different POS. In addition,

WordNet has lexical and semantic links that depict the relationships between terms. These relationships can be further exploited for other research such as concepts identification.

1.6 Problem Statement

The growth of the computing technology has encouraged the introduction of new applications in text categorization. As text categorization becomes widely used in application areas, two research issues that need to be addressed are the absence of semantics in the text categorization tasks and also the large dimensionality of the vector space. To select features that best represent a particular category, semantic relationships needs to be effectively incorporated into the text categorization process.

The datasets or corpus used in automated text categorization varies from one research work to another. Standard test collections are lacking and only a few datasets have been used by more than one researcher. Thus, this makes the results hard to compare (Lewis, 1997).

The large dimensionality of the feature space is a challenge, as each of the documents contains thousands of terms that can be used as features to uniquely discriminate a category. Although stop words have been removed in the document preprocessing stage, the feature space still contains thousands of words. Furthermore, machine learning schemes like the Naïve Bayes learner and neural networks, are unable to support the large dimensionality space, as this would require a high computational power (Yang, 1999).

Feature selection aims to reduce the feature space by selecting a subset of the original set of terms and at the same time, yielding the smallest reduction in performance compared to using the full set of terms. Statistical feature selection approaches do not consider the meaning of the terms and choose terms based on statistical information. We believe that semantics plays an important role in choosing a better set of features to represent a category. In choosing features to represent a category, it is ideal to know the meaning of a term to see the relevance of the term to the category. Therefore, there is a need to incorporate semantics into feature selection to choose a better set of features. Although NLP resources are widely available, the potential to enhance the feature selection process has not been fully explored. There are a number of approaches for feature selection, as mentioned in Section 1.5.1. Generally, statistical approaches such as Chi2 and IG make use of the frequency information of each term for calculation and selection of terms. Therefore, these approaches do not consider the lexical and semantic relationships of the terms in a category. Lewis (1996) mentioned that the features used to represent a document must be able to capture the semantics of the natural language to some extent. We believe that these statistical approaches can be improved with the integration of a semantic component. The benefit of combining statistical approaches and NLP approaches is that this combined approach will be able to make use of the statistical information from the statistical approaches and also integrate a semantic value from the NLP approach.

1.7 Research Objectives

This research will thus explore the use of WordNet to determine the possibilities of employing a lexical database to select more constitutive terms. It also aims to provide

insights into the use of NLP approaches in feature selection by comparing directly with the statistical approaches. We will explore the possibility of combining the statistical feature selection approaches with the WordNet based approach to improve feature selection.

Therefore, the objectives of this research are to:

- Integrate WordNet as a lexical database into text categorization to obtain a better set of features to represent a category.
- Explore the use of lexical semantics as a feature selection method for dimensionality reduction.
- Compare the effectiveness of using statistical feature selection approaches with the WordNet based approach for feature selection.
- Explore the possibility of combining a statistical approach with the WordNet based approach for effective feature selection to determine constitutive features.

1.8 Scope

The incorporation of lexical semantics into the text categorization task is explored to improve the current feature selection approaches in text categorization by incorporating linguistic models into the existing approaches for dimensionality reduction. This will enable the task of text categorization, especially feature selection, to be more effective and accurate in discovering constitutive terms.

The experiments that are carried out in this research are done in a controlled environment so that comparisons can be made to other similar works conducted in the same environment.

This research will cover the following:

- The Reuters-21578 dataset will be used in this research as other works (Masuyama and Nakagawa, 2002; Rogati and Yang, 2002; Zheng and Srihari, 2003) on feature selection had used the same dataset. Thus, comparisons can be made with other works that uses this dataset.
- The Waikato Environment for Knowledge Analysis (WEKA) will be employed as the machine learning tool. The machine learning algorithm that will be used in this research is the multinomial Naïve Bayes classifier. This classifier is chosen for its simplicity and fast processing. Please refer to Section 2.4.1 for a discussion of this machine learning algorithm.
- In order to evaluate the performance of feature selection approaches, we will benchmark with two well-known feature selection approaches, namely, Chi2 and IG. We will also experiment on combining these statistical approaches with the WordNet-based approach.

1.9 Chapters Overview

In this dissertation, this chapter has given an overview of the research work. The following chapters will discuss the completed research in more detail. Chapter 2 will discuss the reviews of related works in automated text categorization and a comparison on the various approaches used in their researches. The machine

learning tool and scheme used will also be discussed. Chapter 3 will elaborate the design of the WordNet based approach for feature selection that is proposed for the task of text categorization. Subsequently, Chapter 4 discusses the implementation issues of the WordNet based feature selection approach. Chapter 5 will then discuss the experimental design and the results and analysis of the results. Finally, Chapter 6 will wrap up the dissertation with the contribution of this research, the limitations and conclusion as well as a discussion on future works.

Chapter 2 Literature Review

2.1 Introduction

This chapter describes a review of related works in the areas of text categorization. It also includes discussion on the integration of a WordNet based approach for feature selection in the task of text categorization. WordNet will be used to identify synonymous nouns and word sense in the English language so that a more effective set of features can be obtained. The task of text categorization especially feature selection will be discussed in detail as it is the focus of this research.

2.2 Text Representation for Text Categorization

In text categorization, there is a need to represent the texts in a way that is interpretable by the machine. The “bag of words” model or term vector model is the most popular text representation model (Cunningham et al., 1997). In this model, a document is represented by a set of terms in an n -dimensional vector space where n is the number of terms. In general, a term is usually a word while a term vector can be a Boolean vector that represent a set of words that appear in the document or a numeric vector representing the frequency of each term in the document. This model does not take into consideration the order and position of the words. Therefore, the sequence of the words is irrelevant in this representation model. Alternatively, the n -gram model represents each feature as a sequence of contiguous typographic symbols.

2.3 Automatic Approach to Text Categorization

With the escalating growth of digital documents, there is a need for an automatic approach to text categorization. Relying on humans to categorize text is not possible with the number of documents growing larger everyday. Automatic approach to text categorization can be applied through the field of knowledge engineering and machine learning.

2.3.1 Knowledge Engineering

Knowledge engineering is an alternative approach that can be used for text categorization. Expertise is elicited from human experts to manually build a knowledge based system, which typically consists of a set of rules. These rules, in the form of *if <DNF Boolean formula> then <category>*, place a document in a category if the document satisfy the Disjunctive Normal Form (DNF) Boolean formula. This approach is used in the CONSTRUE system (Hayes and Weinstein, 1990), an automated news story categorization system based on a fuzzy rule-based text categorization shell, developed by the Carnegie Group for the Reuters dataset. The disadvantage of this approach to categorization is that it requires repetitive effort by the human experts in upgrading the system when the set of categories is updated or when the system is ported to a different domain. The knowledge acquisition bottleneck highlights the challenges faced in deriving the expertise of human experts. A machine learning approach is able to overcome this problem.

2.3.2 Machine Learning

Machine learning concerns the discovery of patterns in a collection of data and applying these patterns to another collection of new data with the intention of performing tasks such as categorization, prediction and robotics among others.

Among the machine learning techniques that are the best developed are the decision trees and the rule induction classifiers. In these techniques, the algorithms learn solely from the input data. Therefore, it is important to have quality input data so that the concept description is accurate.

In the case of instance-based algorithms, new examples are classified by comparing each one with those stored training examples. In the *nearest neighbour* technique, the training examples are stored in n dimensional space, where n is the number of attributes describing the concept. Comparison of the examples is done using Euclidean distance where all the attributes are given equal importance. The Euclidean distance measures the straight-line distance between two points.

Neural network is a concept that models the human brain and its learning process. Knowledge is represented as weighted links between nodes in a network. The learning process involves adjusting the weights on the links so that the outputs are correctly engineered according to the input given (Cunningham et al., 1997).

Support Vector Machine (SVM) is one of the newer machine learning approach introduced by Vapnik in 1995 for solving two-class pattern recognition problems

(Vapnik, 1995). This scheme is based on the Structural Risk Minimization principle from computational learning theory that aims to find a decision surface that best separates the data points in two classes. SVM is widely used in text categorization as seen in works by Dumais et al. (1998), Joachims (1998), Yang and Liu (1999) and Masuyama and Nakagawa (2002).

The probabilistic models apply the Bayes theorem to calculate the probability of a document belonging to a category. The Naïve Bayes classifier is one of the widely used machine learning scheme in text categorization. The Naïve Bayes classifier is presented in detail in Section 2.4.1.

2.4 Machine Learning Approach to Text Categorization

The application of machine learning in the field of text categorization only emerged in the 1990s. The concept behind supervised machine learning in the task of categorization is generally described as a learner that automatically builds a classifier by learning from a set of documents that has already been classified by experts (Sebastiani, 1999).

The machine learning approach to text categorization has become the paradigm in text categorization as many researchers employ this approach in their works. In the works of Cohen (1995), the FOIL system is used on text categorization problems. FOIL is a first order learning system that is capable of creating classifiers with lower error rates and higher rates of precision and recall. RIPPER (Cohen, 1995), on the other hand, is a rule-learning algorithm revised from incremental reduced error pruning (IREP), which was

introduced by Fürnkranz and Widmer (Cohen, 1995; Cohen and Singer, 1996; Basili et al., 2000).

Other machine learning techniques that are applied to text categorization include nearest neighbour classifiers (Yang and Pedersen, 1997; Hall and Smith, 1998; Yang, 1999; Yang and Liu, 1999), probabilistic Bayesian models (Lewis and Ringuette, 1994; Hall and Smith, 1996; Dumais et al., 1998; McCallum and Nigam, 1998; Yang and Liu, 1999; Hall, 2000), decision trees (Lewis and Ringuette, 1994; Dumais et al., 1998; Holmes and Trigg, 1999), neural networks (Crestani, 1993; Yang and Liu, 1999) and support vector machines (SVM) (Joachims, 1998; Dumais et al., 1998, Yang and Liu, 1999; Masuyama and Nakagawa, 2002).

The advantage in applying machine learning in text categorization is that it is faster and cheaper to employ an automated text categorization system in categorizing documents rather than having human experts doing manual categorization (Sebastiani, 1999). By using supervised learning, a set of training data that have been previously classified manually by domain experts is provided to the machine learning tool to train the classifier using a learning scheme. This learnt classifier would then be used to determine whether documents in the test set are correctly classified. The machine learning approach to text categorization is able to overcome the difficulty of capturing knowledge, which is also known as the knowledge acquisition bottleneck.

One point to note is that the training data used must be of quality so that the construction of the classifier is precise and accurate. If the data is noisy and unreliable, then the learning process for the classifier induction will not be accurate. However, the

problem of inadequate number of documents and human errors in classifying are always found in datasets. In view of the absence of a “perfect” dataset, we will have to make use of the available datasets despite the problems mentioned.

2.4.1 Naïve Bayes (NB) Classifier

The Naïve Bayes (NB) classifier originates from works in pattern recognition (Duda and Hart, 1973). The NB classifier is widely used because of its computational simplicity. The NB approach is based on the mathematical theorem known as the Bayes Theorem. The Bayes Theorem is defined below with H as the hypothesis and D as an input document that bears on that hypothesis.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

where,

$P(H)$ is the probability of hypothesis H being correct.

$P(D)$ is the probability of document D being observed.

$P(D|H)$ is the probability of observing document D under the assumption of H being correct.

$P(H|D)$ is the probability that hypothesis H is correct, where document D is observed.

$P(H|D)$ is read as “the probability of H given D”. For example, if the hypothesis H is “input document belong to category C” and D is the input document, then this $P(H|D)$ value represents the probability of an input document D belonging to category C.

In the Naïve Bayes (NB) approach, the basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document (Yang and Liu, 1999). The term “naïve” is based on the assumption that the characteristic word patterns found in a document is independent of all other word patterns. This is of course not the true case in real world problems. Based on this assumption however, document structure and relationship of word positions can be ignored. Therefore, it makes the computation of the NB classifier more efficient as it does not use word combinations as predictors (Yang, 1999).

In the works of Clark and Niblett (1989), the NB classifier is compared to a decision tree learner and two rule learners. They found that it did surprisingly well despite its simplicity. Langley, Iba and Thompson (1992) compare the NB classifier with a decision tree learner and find that for four out of five of the University of California, Irvine (UCI) database collections, the NB classifier performs as well or better than the decision tree learner.

In the medical domain of diagnosing the cause of acute abdominal pain, Provan and Singh (1996) compare the NB classifier with the more sophisticated Bayesian network. The NB classifier outperforms the more complex Bayesian network in domains that are more complex using a smaller set of features. With only 169 features, the NB model is more accurate than the Bayesian network that uses 590 features. Friedman, Geiger and Goldszmidt (1997) also compare the NB classifier with Bayesian networks and found that the more complex Bayesian networks not only produce no improvements but also sometimes cause reductions in accuracy. This shows that the NB classifier is able to achieve good results in different domains.

Pazzani, Muramatsu and Billsus (1996) apply a few learning algorithms to predict interesting websites and find that the NB classifier was the most accurate in its prediction. Domingos and Pazzani (1997) compare the NB classifier with a decision tree learner, a nearest neighbour algorithm and a rule induction system. The results of their experiments show that for 22 out of 28 of the University of California, Irvine (UCI) database collection, NB classifier is at least as accurate as the nearest neighbour algorithm and the rule induction system. It is also at least as accurate as the decision tree learner for 20 out of 28 datasets.

There are two different NB models used in text categorization; the multi-variate Bernoulli model and the multinomial event model. The Bernoulli model specifies that a document is represented by a vector of binary attributes that indicate whether a word is present or absent in a document. There is no information on the number of times a word occurs. In calculating the probability of a document, the probability of all the attribute values including the probability of non-occurrence of words that do not occur in the document are multiplied. On the other hand, the multinomial event model specifies that a document is represented by a set of word occurrences. This model captures the number of times each word occurs in a document. In calculating the probability of a document, the probability of the words that occur is multiplied. Many researches had used this model in their text categorization works (Joachims, 1997; Nigam et al., 1998; McCallum and Nigam, 1998; Bong, 2001). For a more detailed discussion on the multinomial Naïve Bayes classifier, please refer to Section 3.2.3.

2.4.2 Single-label vs Multi-label Categorization

There are two different kinds of categorization; the *single-label* and the *multi-label* categorization. The single-label categorization, which is also known as the non-overlapping categorization, is when a single category is assigned to each document. Multi-label categorization, on the other hand, allows any number of categories to be assigned to a single document. The former approach employs a simpler algorithm, which can be adopted for multi-label categorization by taking each category assigned to the document and treating each one individually as a single-label categorization. On the other hand, adopting a multi-label categorization algorithm for a single-label categorization task is more cumbersome.

2.5 Dimensionality Reduction by Feature Extraction

Feature extraction, which is also known as reparameterisation (Sebastiani, 1999) create a new set of terms from the original set to maximize effectiveness. This method aims to solve the problem of polysemy and synonymy, where terms may not be useful for document content representation. Polysemy refers to the ambiguity of an individual word or phrase that can be used in different contexts to express two or more different meanings. Synonymy, on the other hand, refers to the semantic relation between two words that express the same meanings in a given context. Therefore, in term extraction, artificial terms that do not suffer from the language problem are created. Two common feature extraction approaches in text categorization are term clustering and latent semantic indexing (LSI).

2.5.1 Term Clustering

Term clustering is the process of grouping terms that are semantically related together into clusters so that the centroids of the clusters can be used as the feature in the vector space instead of the terms. The problem of synonymous terms can thus be addressed. In any term clustering method, there must be a technique to group the terms into clusters and also a technique to convert the original document representation into a new representation that consists of the synthesized features.

One of the earliest works of term extraction in text categorization is pursued by Lewis (1992a), where he employs a method called reciprocal nearest neighbour clustering. This method clusters together two terms that are most similar to each other based on some similarity measurement. This work reported that the clustering approach performed poorly as compared to single-word indexing.

Unlike the unsupervised clustering by Lewis where no category tags are attached to the documents, Baker and McCallum (1998) carried out experiments on supervised clustering. They cluster terms that point towards the same category using a Naïve Bayes classifier and their results show only a 2% loss of effectiveness with a feature set that is much smaller than the original feature set. However, with a lesser reduction of the feature set size, they record no improvements in effectiveness. This is a promising approach and can be used in combination of other approaches.

2.5.2 Latent Semantic Indexing (LSI)

Latent semantic indexing (LSI) (Deerwester et al., 1990) is a feature extraction method that combines the original features in the vector space and represents them in a new lower dimensional vector space based on their patterns of co-occurrence in the original vector space. The dependence among the original features in the vector space is deduced. This dependence information is then used to create a new independent vector space. The transformation of the original vector space to the new vector space is encapsulated in a function obtained from applying a singular value decomposition to the incidence matrix formed by the original document vectors. In the text categorization perspective, the mapping function is derived from the training set and is applied to each test documents to produce a new lower dimensional representation of the documents (Sebastiani, 1999).

Schütze et al. (1995) report that in one of the positive test instances, the neural network classifier gives an incorrect decision when a feature selection approach, Chi-Square, is used but is correct when LSI is applied to the task. Using the top 200 most relevant terms selected by Chi-Square, the neural network classifier is not able to give a correct decision for the positive test instance as none of the 200 terms occur in the document. However, LSI was able to come up with one or more terms that coincide with the terms in the document. LSI is able to capture semantic information implicitly, making it difficult to explain why it works well. NLP based approach on the other hand, allows a better interpretation of results. LSI has been applied successfully in a number of domains. The success of LSI shows the importance of exploiting information beyond word occurrences.

2.6 Dimensionality Reduction by Feature Selection

Feature selection is a process of selecting a subset of terms from the feature space. It is performed to reduce the dimensionality of the feature space. It includes the removal of non-informative terms according to corpus statistics. In this process, a subset of features is selected based on the following criteria (Dash and Liu, 1997).

1. The categorization accuracy does not significantly decrease.
2. The resulting class distribution, given only values for the selected features, is as close as possible to the original class distribution, given all features.

There are a number of feature selection methods, which over the years, are used in a wide range of text categorization tasks. Among the more widely used statistics based ones are Information Gain (IG) (Yang and Pedersen, 1997; Mladenić, 1998; Sebastiani, 1999; Debole and Sebastiani, 2003) and Chi-Square (Chi2) (Schütze et al., 1995; Yang and Pedersen, 1997; Sebastiani, 1999; Debole and Sebastiani, 2003). Table 2.1 shows the formula for each of the mentioned feature selection methods. In Table 2.1, probabilities are interpreted on an event space of documents, for example, $P(t_k, c_i)$ indicates the probability that, for a random document x , term t_k occurs in x and x belongs to category c_i and are estimated by counting the occurrences in the training set. Term absence is also taken into consideration in the calculation of the probabilities.

Feature Selection Method	Formula
Information Gain	$\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}$
Chi-Square	$\frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$

Table 2.1: Mathematical formula for feature selection methods (extracted from Debole and Sebastiani, 2003)

2.6.1 Information Gain (IG)

Information Gain or more popularly known as InfoGain is a feature selection approach that makes use of the presence and absence of a term in a document to select its features. It is frequently used as a term-goodness criterion in the field of machine learning. The number of bits of information is measured for category prediction by using the knowledge of the presence and absence of a term. The amount of information term t_k contains about category c_i is measured and terms that are more indicative of a category based on their presence or absence are chosen.

From the IG formula shown in Table 2.1, we can further expand it mathematically as follows:

$$P(t, c) \log_2 P(t, c) / P(t) / P(c) + P(\bar{t}, c) \log_2 P(\bar{t}, c) / P(\bar{t}) / P(c) + P(t, \bar{c}) \log_2 P(t, \bar{c}) / P(t) / P(\bar{c}) + P(\bar{t}, \bar{c}) \log_2 P(\bar{t}, \bar{c}) / P(\bar{t}) / P(\bar{c})$$

where,

$P(t, c)$ is the probability of term t occurring in category c .

$P(t)$ is the probability of term t in the data collection.

$P(c)$ is the probability of category c in the data collection.

$P(t, \bar{c})$ is the probability of term t not occurring in category c .

$P(\bar{t}, c) = 1 - P(t, c)$ is based on Debole and Sebastiani (2003).

Based on the rule in probabilities,

$$P(\bar{t}) = 1 - P(t)$$

$$P(\bar{c}) = 1 - P(c)$$

$$P(\bar{t}, \bar{c}) = 1 - P(t, c)$$

For each unique term in the training set, information gain is computed and those terms that are above a predetermined threshold are selected as features. A term with a high information gain indicates that it is a good feature for category prediction. Yang (1997) stated that IG is able to reduce the dimension of the vector space by a factor of 100 with no loss of categorization effectiveness.

2.6.2 Chi-Square (Chi2)

This method measures the degree of dependence between a term and a category. If a term is independent of a category, it will have a value of zero. A low value of Chi2 signifies a high degree of independence of term t_k and category c_i while a high value shows otherwise. A term with a high value of Chi2 shows that it is more dependant on a category and is therefore, more likely to be added to the feature space. These highly dependent features are selected because of their discriminating power.

From the Chi2 formula defined in Table 2.1, the probabilities are defined in the similar manner as the definition in IG (refer to Section 2.6.1 for definition of probabilities). For each category, the Chi2 value for each unique term is computed and those terms that are below a predefined threshold are removed from the feature space. One point to note is that Chi2 does not work well for terms with low frequencies (Dunning, 1993). This is due to the fact that Chi2 is a normalized value and can be compared across the terms in the same category. However, this normalization is not accurate for low frequency terms. Like IG, Chi2 can also reduce the dimension of the vector space by a factor of 100 with no loss of categorization effectiveness (Yang, 1997). It is thus desirable to develop feature selection methods with a performance comparable to both IG and Chi2.

2.6.3 Discussion on Information Gain and Chi-Square

Yang and Pedersen (1997) find that common terms are certainly informative for text categorization tasks based on the excellent performance of IG and Chi2 in their works. However, in information retrieval, common terms are considered non-informative. Based on their experiments, Yang and Pedersen conclude that this assumption in information retrieval need not be true for text categorization.

Yang and Pedersen (1997) also point out some reasons for the good and bad performance of feature selection approaches in text categorization tasks. These criteria are based on learning using the nearest neighbour classifier and a regression method called linear least square fit mapping (LLSF) and is shown in Table 2.2.

1. Supporting common terms or rare terms
2. Task-sensitive (using category information) or task free.
3. Using the absence of a term to predict the category probability.

Criteria	Information Gain	Chi-Square
Supporting common terms	Y	Y
Task sensitive	Y	Y
Using term absence	Y	Y

Table 2.2: Criteria for performance of feature selection approaches (extracted from Yang and Pedersen, 1997)

From Table 2.2, it is noted that both IG and Chi2 support common terms and are task sensitive, where category information is used. Both IG and Chi2 also make use of the term absence criteria in predicting the category probability.

2.7 Word Sense Disambiguation

Word sense disambiguation (WSD) refers to the process of disambiguating words to determine the sense of an ambiguous word. For instance, the English word 'bond' can have at least two different senses as in "*The bond between a mother and her baby begins in the womb.*" and "*A government bond amounted to ten million dollars was issued to finance the underwater tunnel project.*" The former is most likely to be found in the field of biological science while the latter is in all likelihood a financial statement. Therefore, it is a WSD task to decide the correct sense in the sentence: "*The twin sisters have an exceptional bond since the day they were born.*"

Natural language ambiguity is one of the linguistic problems to be resolved in text categorization. WSD is included in the task of text categorization where indexing document by word senses (Mihalcea and Moldovan, 2000) is carried out instead of using word units as index terms. As reported by Mihalcea and Moldovan (2000), they achieved an accuracy of 92% in disambiguation with a semi-complete disambiguation (55% of nouns and verbs are disambiguated). In one of the experiments by Gonzalo et al. (1998), their manually disambiguated collection based on synset indexing produced an amazing improvement of 29% over word based indexing. The use of an aggregated form of WSD has the potential of enhancing the quality of features in text categorization. However, a complete form of disambiguation can be computationally expensive.

In general, WSD involves two steps (Ide and Véronis, 1998):

- 1) The determination of all the different senses that a word has.
- 2) A way to assign each occurrence of a word to the correct sense.

Step 1 can be achieved by having a list of senses from a dictionary, a group of associated words from a thesaurus and also an entry in a transfer dictionary that has translation to another language. Step 2 can be accomplished by relying on two sources of information. Firstly, the context of the word is identified by looking at the information contained within the text in which the word appears and secondly, by using external knowledge sources, which includes lexical and encyclopedic resources.

2.8 Introduction to WordNet

WordNet is an online thesaurus and an online dictionary. It can be considered as a dictionary based on psycholinguistics principles. WordNet contains nouns, verbs, adjectives and adverbs as parts-of-speech (POS). Function words are omitted based on the notion that they are stored separately as part of the syntactic component of language (Miller et al., 1990).

WordNet is organized by relations such as synonym, antonym, hyponym/hypernym and holonym/meronym. While synonym and antonym are lexical relations between word forms, both hyponym/hypernym and holonym/meronym are semantic relations between word meanings. Generally, synonyms are words having the same meaning and antonyms are words having opposite meanings. On the other hand, hyponym/hypernym is more commonly known as the subset/superset relation. It is also called the “IS A” relation. Holonym/meronym is known as the “PART WHOLE” relation.

The information in WordNet is organized into sets of words called synsets. Each synset in WordNet has a unique signature that differentiates it from other synsets. Each of the synset contains a list of synonymous words and semantic pointers that illustrate the relationships between it and other synsets. There are different types of semantic pointers that include “IS-A”, “PART-OF/HAS-PART”, “MEMBER-OF/HAS-MEMBER” and “SUBSTANCE-OF/HAS-SUBSTANCE” (Miller et al., 1990, Richardson et al., 1994). The “IS-A” relationship is also known as the hyponym/hypernym relationship where hyponym is the subset and hypernym is the superset. On the other hand, “PART-OF/HAS-PART”, “MEMBER-OF/HAS-MEMBER” and “SUBSTANCE-OF/HAS-

SUBSTANCE” is also known as the holonym/meronym relationship. Holonym is the inverse of meronym where, if x is a holonym of y , then y is a meronym of x . Table 2.3 shows some examples of the semantic relationships in WordNet.

WordNet Semantic Relationship	Examples
Hyponym/Hypernym ("IS-A")	- house IS A dwelling - bungalow IS A house
Holonym/Meronym ("PART-OF/HAS-PART")	- trunk PART OF tree - tree HAS PART trunk
Holonym/Meronym ("MEMBER-OF/HAS-MEMBER")	- tree MEMBER OF forest - forest HAS MEMBER tree
Holonym/Meronym ("SUBSTANCE-OF/HAS-SUBSTANCE")	- sapwood SUBSTANCE OF tree - tree HAS SUBSTANCE sapwood

Table 2.3: The semantic relationship in WordNet

2.9 Incorporating WordNet in the Task of Text Categorization

Many researches have been carried out in trying to achieve more effective results in text categorization by introducing the semantics of the natural language in the text categorization tasks. For this purpose, many have chosen WordNet as the semantic database because of its vast collection of words. Another reason why many choose WordNet is because the collection of words is more general and not based on a specified domain. Domain specific categories may require the use of domain specific dictionaries to provide the technical terms. Therefore, this gives the indication that WordNet might not be a good choice for those doing domain specific categorization.

WordNet synsets have been used as the indexing space in the works done by Gonzalo et al. (1998). Scott and Matwin (1998) use WordNet hypernyms for their new text

representation. Apart from that, WordNet is also exploited for text retrieval purposes. WordNet can be used for expanding queries with semantically-related terms (Voorhees, 1994) and comparison of queries and documents using conceptual distance similarity measures (Richardson and Smeaton, 1995). In the works of Masuyama and Nakagawa (2002), WordNet synonyms are used in their feature selection process called the cascaded feature selection. The usage of WordNet in these researches will be explored further in this section.

2.9.1 WordNet in Indexing

A classifier is unable to understand the basic structure of a document without any processing being done to it in a text categorization system. Therefore, indexing is done to map the document into a suitable representation of the document contents. Van Rijsbergen (1979) describes indexing as, “An index language is the language used to describe a document. The elements of an index language are index terms, which may be derived from the text of the document to be described”.

WordNet is used in the task of indexing (Gonzalo et al., 1998), as it is able to distinguish word senses in documents and queries and that match semantically related words. The corpus used in this research is SEMCOR, a subset of the Brown corpus of about 100 documents. They make use of WordNet synsets as the indexing space instead of using individual terms. The experiments in Gonzalo et al. (1998) are performed using a standard vector-model based text retrieval system, SMART (Salton, 1971). The experiments use three different indexing spaces:

- i) The original terms in the documents
- ii) Manually disambiguated version of the documents
- iii) WordNet synsets corresponding to the document terms

Gonzalo et al. (1998) reported encouraging results, showing a 29% of improvement if WordNet synsets are chosen as the indexing space instead of word forms. However, this result was achieved after having manually disambiguated test collection for both the documents and the queries. WordNet synsets were chosen over word forms because they offered the possibility to discriminate word senses and to match semantically related words. This work highlights the importance of WSD in ensuring the correct meaning of terms is preserved. The manual process of WSD is however time consuming and may not be practical for large datasets.

2.9.2 WordNet in Text Representation

In order to effectively categorize a set of documents, a system must be able to extract sufficient information from the documents to distinguish one category from another. A model of the documents' content must be represented accurately by capturing the fundamental information that tells us what the documents are about.

Scott and Matwin (1998) explored improvements in categorization accuracy when linguistic knowledge was incorporated into text representation. In their works, a new text representation based on WordNet hypernyms was proposed. There are three corpora used in their research; Reuters-21578, USENET and Digital Tradition (DigiTrad). Both Reuters and USENET have been commonly used in other researches.

DigiTrad is a public domain collection of 6500 folk song lyrics. All the datasets are split into two and each split section contains different categories; Reuter1, Reuter2, USENET1, USENET2, Song1 and Song2.

Their new *hypernym density* representation requires three passes through the corpus. During the first pass, they make use of the Brill tagger (Brill, 1992) to identify every word in the corpus. In the second pass, all nouns and verbs are looked up in WordNet and a list of all synonym and hypernym synsets is assembled. Synsets with a frequency of occurrence that is less than $0.05N$, where N is the number of documents in the corpus, is ignored. Lastly, in the third pass, the density of each synset is calculated using the formula:

$$\text{Density} = \frac{\text{Number of occurrences of a synset in the WordNet output}}{\text{The number of words in the document}}$$

One point to note is that during the computation of hypernym density, no word sense disambiguation is performed. Therefore, all senses of the words are included regardless of their senses. A parameter h that monitors the *height of generalization* is used to limit the number of steps upward through the hypernym hierarchy for each word. Therefore, at the height of $h=0$, only the synsets that contain the words in the corpus will be included. However, at $h>0$, all the hypernym synsets of up to h steps above in the hierarchy will be taken into consideration. There is also a value of $h=\text{max}$ where all the hypernym synsets will be counted no matter how far up they are in the hierarchy.

In their experiments, Scott and Matwin (1998) used two representations; the bag-of-words using only nouns and verbs and the filtered normalized density vectors for nouns and verbs. This is to highlight the improvement in accuracy derived from the use of WordNet synsets.

From the results of the experiments, they find that in the case of the Reuters dataset, no improvements over bag-of-words is observed using hypernym density. Both USENET1 and Song1 have shown substantial decrease in error rates compared to the bag-of-words with USENET1 registering a 34% drop in number of errors for $h=9$ and Song1 having a 47% drop for $h=9$. USENET2 and Song2 produced error rates that were comparable to bag-of-words. The experiments show that the hypernym density representation works best only when the appropriate value for the parameter h can be pre-determined for each hypernym path.

2.9.3 WordNet in Semantic Distance Measurement

In their work, Richardson and Smeaton (1995) use the nouns from WordNet as a knowledge base to compute a semantic distance measurement between concepts or words and use this word distance to compute a similarity between a query and a document.

The initial knowledge base consists of a number of hierarchical concept graphs (HCGs) constructed from WordNet data files. The advantages of constructing a knowledge base in this manner are that it contains a comparatively rich set of semantic link types and it is a comprehensive starting point in obtaining HCGs that contain all relevant concepts

in an information domain (Richardson and Smeaton, 1995). The corpus used in their experiments is the Text Retrieval Conference (TREC) collection.

In their work, Richardson and Smeaton (1995) employ a conceptual distance approach and an information based approach for estimating semantic similarity. The results of their experiments are not as encouraging as predicted. They compared their semantic based system with the traditional Term Frequency Inverse Document Frequency (TFIDF) ranking. Weights of the terms in TFIDF are calculated based on the frequencies of the terms in the document and the number of documents that contain that particular term over the whole document collection. They found that TFIDF ranking outperforms their systems at all levels of recall with an average precision of 0.2072 compared to conceptual distance approach with an average precision 0.1062 and information based approach, 0.1151.

Due to the broad domain of the HCGs, the conceptual distance measures are less accurate than expected. There are also the irregular densities of links between concepts that result in unexpected distance measures. This is due to the fact that in WordNet, some of the concepts have a deeper hierarchy as compared to other concepts. As word sense disambiguation is not taken into consideration in this experiment, all senses of words are considered. Therefore, the concept graphs also expanded on senses of terms not implied in the original documents. This gives an exaggerated information content value to polysemous words and multi-worded synsets. Figure 2.1 is a HCG, which shows the concept “car”.

The use of hypernym provides a means of identifying deeper concepts for terms found in a document. The difficulty in the use of hypernyms is in determining the actual depth in the hierarchy that will yield the best results. The use of hypernyms is discussed as future works in this research.

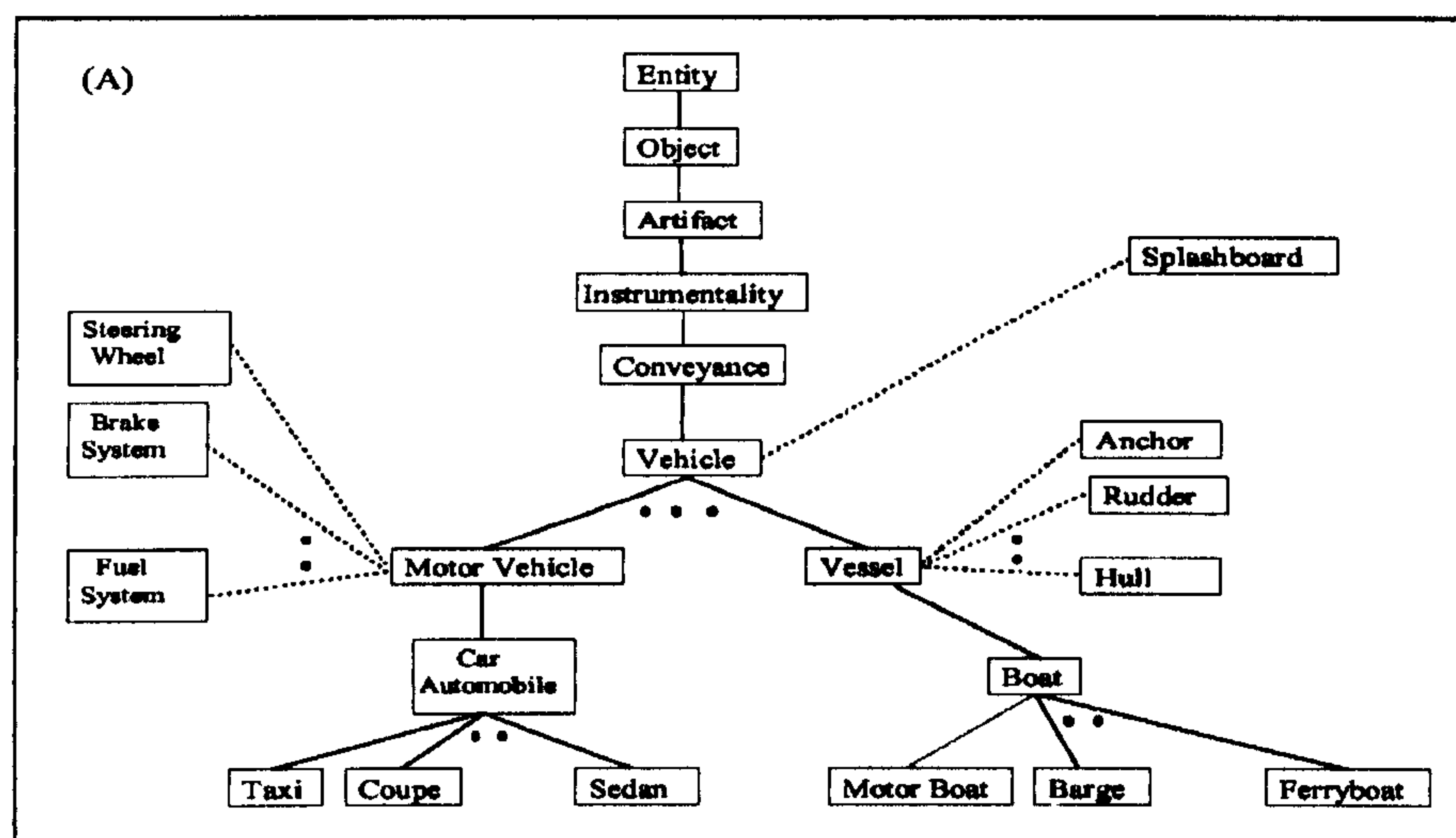


Figure 2.1: A HCG showing the concept “car” (extracted from Richardson and Smeaton, 1995)

2.9.4 WordNet in Feature Selection

Masuyama and Nakagawa (2002) use WordNet for cascaded feature selection (CFS). They employ the support vector machine (SVM) learning scheme in their text categorization task for the Reuters-21578 dataset. Their CFS approach consists of two steps. In Step 1, SVM classifies the test documents into a positive or negative set. In Step 2, SVM once again classifies the test documents categorized into a positive set in Step 1 into either a positive or negative set. Step 1 uses all types of parts of speech (POS) while Step 2 uses only nouns. The advantage of this work is that one can make use of the characteristics of each feature. Apart from that, the unnecessary test

documents of a category that should be rightfully classified in the negative set can be removed when Step 1 is performed.

The synonyms of the nouns in the documents are obtained from WordNet 1.7 and added as features for representing the documents. For each category, only 300 terms are extracted based on the threshold set for high mutual information. The word frequency is also used as an attribute value for each feature. The precision of the CFS method is 2.1% higher than the precision of the non-CFS method with a constant value of f-measure at 81.8%. With the CFS method, 75% of the categories with more than 100 documents in the dataset recorded an increase in precision while for categories with less than 100 documents, 22.9% of the categories recorded an increase in precision.

This work highlights the potential of employing the use of synonyms for feature selection. Their approach applied these synonyms twice in a cascaded manner to enhance the quality of features. Feature selection is employed purely syntactically. Besides that, their approach was not benchmarked with other statistical approaches.

2.10 Comparison and Discussion

WordNet is used more in information retrieval works compared to text categorization. As text categorization is closely related to information retrieval, it is worth reviewing to find out if WordNet can do any better for the text categorization tasks.

All the works reviewed in the previous section uses WordNet to add more information to the text analysis tasks. Synsets in WordNet are used in indexing in Gonzalo et al (1998),

hypernyms in WordNet are used in text representation (Scott and Matwin, 1998), WordNet data files are used to build hierarchical concept graphs (Richardson and Smeaton, 1995) and WordNet synonyms are added into the document representation vector (Masuyama and Nakagawa, 2002).

We find that by adding in words from WordNet into the feature space, as done by Masuyama and Nakagawa, there is a possibility that terms added does not contribute towards the effectiveness of automated text categorization, as the terms that are added are originally not in the documents. Also, with no WSD done, words that are added are from all word senses and this introduces unrelated terms into the feature space. While we are attempting to tackle the large dimensionality of the feature space by trying to reduce the number of terms, introducing more words into the feature space defeats this purpose. Table 2.4 compares the different works that makes use of WordNet.

Authors	Corpus Used	Machine Learning Scheme	WordNet's Usage	Considers Word Senses	Tasks
Gonzalo et al (1998)	SEMCOR	SMART	Synsets	Yes, manual	Indexing
Scott and Matwin (1998)	Reuters-21578, USENET, Digital Tradition	RIPPER	Synonyms, hypernyms	None	Text representation
Richardson and Smeaton (1995)	TREC Collection	—	Nouns	None	Build HCGs to be used for semantic distance measurement
Nakagawa and Masuyama (2002)	Reuters-21578	Support Vector Machine (SVM)	Synonyms	None	Cascaded feature selection by adding synonyms

Table 2.4: Comparison between works using WordNet

2.11 Evaluation Measures

As the aim of this research is intended to contrast the WordNet based approaches with statistical feature selection methods, two forms of evaluation has been proposed:

1. Standard metrics for performance evaluation of the classifier built using features selected.
2. Qualitative analysis to evaluate whether the selected features are representative of a category.

An analysis was also done to evaluate the performance of the classifier across the different categories in the dataset.

2.11.1 Performance Measures

There are various methods to judge the effectiveness of the text categorization classifier built using the machine learning approach. In binary categorization, the contingency table is used to evaluate the classifier for each category. From a contingency table, which is also known as a confusion matrix, performance measures that can be computed are in the form of precision (P), recall (R), accuracy (Acc), error (Err) and f-measure (F_1). These measures have been used in most of the previous researches. Prior to carrying out the experiments, the appropriate metrics were determined.

Precision is defined as the number of documents retrieved that are relevant over the total number of documents retrieved. Recall is defined as the number of documents

retrieved that is relevant from the total number of documents that are relevant (Witten and Frank, 2000). Accuracy is defined as the number of documents correctly retrieved over the total number of documents while error is the number of documents incorrectly retrieved over the total number of documents.

Both accuracy and error are not widely used as measures in text categorization (Scott and Matwin, 1998). The number of instances being tested for each category is always small as compared to the total number of instances in the test set. With a large denominator in accuracy and error, any variations in the correct instances will not have a significant impact on the value of accuracy and error. Therefore, it is not a good evaluator for a classifier.

Category Set		Classifier Judgments	
		Yes	No
Expert Judgments	Yes	TP	FN
	No	FP	TN

Table 2.5: Contingency table for binary categorization (extracted from Witten and Frank, 2000)

From Table 2.5, TP, FP, FN and TN represent the number of true positives, false positives, false negatives and true negatives respectively. True positives and true negatives are correctly classified instances. False positives are instances that are wrongly classified as positive when it is actually negative while false negatives are instances that are wrongly classified as negative when it is actually positive. Referring to Figure 4.8 in Chapter 4, the values for TP, FN, FP and TN are 12, 32, 36 and 334 respectively. These values can then be used to calculate the performance metrics based on the formula shown in Table 2.6.

Performance Metrics	Computation Formula
Precision	$P = TP / (TP + FP)$
Recall	$R = TP / (TP + FN)$
Accuracy	$Acc = (TP + TN) / (TP + FP + FN + TN)$
Error	$Err = (FP + FN) / (TP + FP + FN + TN)$
F-measure	$F_{\beta} = (\beta^2 + 1) \cdot P \cdot R / \beta^2 \cdot P + R$

Table 2.6: Computation for the performance metrics (extracted from Sebastiani, 1999)

In binary categorization, precision and recall are the common measures used in most researches (Sebastiani, 1999). However, neither precision nor recall makes an effective measure by itself. To have a high value of precision would mean that there is a trade-off of low recall and vice-versa. Often, the combination of both precision and recall is used to evaluate the effectiveness of a classifier. There are three main combined measures that are frequently used (Sebastiani, 1999):

- 1) 11-point average precision
- 2) Breakeven point
- 3) F-measure

In the 11-point average precision measure, 11 different recall values (0.0, 0.1, ..., 0.9, 1.0) are used with precision computed and averaged over these values. This average value is used as the measure for effectiveness. This method of evaluation is widely used in ranking-based information retrieval systems (Yang and Pedersen, 1997; Esteban et al., 1998; Yang, 1999).

In the breakeven point measure, the value for precision and recall are computed for different threshold and plotted on the precision-recall graph. Linear interpolation is

done to find the point where precision is equal to recall on the precision-recall graph. The point where precision is equal to recall is called the breakeven point. This single value is used as a measure for the classifier. Precision and recall breakeven point is described in detail in Lewis (1992b). This method is widely used in evaluation for text categorization (Scott and Matwin, 1998; McCallum and Nigam, 1998; Joachims, 1998, Dumais et al, 1998). There are, however, a few drawbacks in this effectiveness measure (Lewis, personal communication, 2003).

1. Lewis stated that by plotting the computed average value of recall on the x-axis and the computed averaged value of precision on the y-axis, a smooth decreasing curve could be obtained. However, when the value of recall and precision for individual categories are plotted, the curve is skewed and non-monotonic. Lewis noted that it is true for classifiers like Naïve Bayes that uses numeric threshold values and also discrete classifiers like decision trees and rules. Therefore, there may be no threshold that allows the classifier to achieve the breakeven point.
2. Also noted by Lewis is that the single value of precision equals to recall is not an informative target. If the breakeven point is optimized, it means that the user is trying to optimize the minimum value of recall and precision. Therefore, it is not a desirable target for categories that are diverse and it is not desirable for works with extensive dataset.
3. Categories in a dataset are often very diverse. Therefore, if the pair of precision and recall for each category is plotted on a graph, the points will be quite

dispersed. This means that an average over these varied values is not very informative and therefore the breakeven point will be even less informative.

F-measure is another metric that is commonly used in text categorization studies. It combines both precision and recall to obtain a single value to measure the performance of a classifier. F-measure is used in the works of Cohen (1995), Yang (1999), Yang and Liu (1999) and Lam et al. (1999). The general formula for f-measure is:

$$F_{\beta} = (\beta^2 + 1) \cdot P \cdot R / \beta^2 \cdot P + R$$

where, β is the parameter for differential weighting of precision and recall. The value for f-measure is maximized when the value of both precision and recall are close to each other. Typically, the value 1 is used for β to give equal weight to both precision and recall, thus giving f-measure the formula:

$$F_1 = 2 \cdot P \cdot R / P + R$$

For evaluating performance average across categories, there are two conventional methods, namely, macro-averaging and micro-averaging, used in all researches reviewed. Macro-averaged scores are computed by first computing the scores for each category using the category's contingency table and then taking an average of the values for all the categories. Micro-averaged scores is obtained by first summing up all the corresponding cells in the contingency table for all the categories and then compute the values from the global contingency table. The difference between these two is that micro-averaged scores give equal weighting to every document while macro-averaged

scores give equal weighting to every category. This implies that these two methods may give different results when the categories are unevenly populated.

In this research, we have opted to use the f-measure with both micro-averaged and macro-averaged scores across all the categories. F-measure combines both the value of precision and recall to give a more accurate result to indicate the effectiveness of the classifier's performance. By employing f-measure, it is easier to benchmark the works presented here with other related works.

2.11.2 Qualitative Analysis

There will also be a qualitative analysis on the set of features chosen by each feature selection scheme to see how descriptive and reflective the terms are to the category. This study is to observe the goodness of each term chosen to represent a category in a categorization task and also to identify noise terms. By doing qualitative analysis, constitutive and contingent terms can also be identified, which are not obvious in the statistical analysis.

2.12 Summary

Machine learning promises to allow classified text corpus to be used effectively to automate the task of text categorization. Many machine learning schemes have been applied in the text categorization task, which have, over the years been improved on to ensure a more effective and accurate categorization. Feature selection approaches are employed in text categorization to reduce the large dimensionality of the feature space.

Most methods of feature selection are statistical while the documents are in the natural language. As a result, the incorporation of semantic information in the task of text categorization can improve effectiveness and accuracy to some point. WordNet provides a valuable resource for incorporating and automating the tasks of text categorization. It is potentially capable of determining features that capture the semantics of terms. Although WordNet can be used to identify deeper concept of terms, little success has been achieved so far (Richardson and Smeaton, 1995).

Chapter 3 Design

3.1 Introduction

The automated text categorization framework and its components are described in this chapter. The design of the proposed feature selection methods will also be discussed. Typical statistical approaches, Chi-square (Chi2) and Information Gain (IG) will be used as a benchmark to evaluate the WordNet-based approaches. As both WordNet-based and statistical approaches have their own advantages and disadvantages, a combined approach is also explored.

In this research work, we will pursue the use of WordNet in the feature selection process in the task of automated text categorization using the machine learning approach. We explore further a single-phased approach to feature selection as opposed to a two-phased approach of Masuyama and Nakagawa (2002). We will explore the effectiveness of adding additional lexical and semantic information to select features to be added in the feature space. We will also explore the effectiveness of combining the statistical feature selection approaches with the WordNet-based approach to evaluate the effectiveness of incorporating a semantic component into the process of feature selection. We will employ single-label categorization as each document is classified as belonging to one single category and our categories are fixed to a number that is dependent on the corpus used. We will also employ the bag-of-words text representation model. Our research uses the multinomial Naïve Bayes model as the machine learning scheme as it has been seen to be better than the traditional Naïve Bayes model (McCallum and Nigam, 1998). Bong (2001) has also shown that term occurrence representation performs better as

compared to the Boolean representation for text categorization. In our experiments, we employ only positive examples for training as it has been found to be effective (Forman, 2003).

Therefore, our research attempts to use WordNet to identify existing synonyms in a category as features to represent each category instead of adding in terms discriminately or synsets, as are done by the other researchers. By that, we are also doing categorical sense determination to identify the senses for the selected synonyms, whereas in the work of the other researchers, they either disregard the word senses by adding in all senses of a term or they perform manual disambiguation. Words that are synonymously related give the indication that they are indeed from the same concept and thus, can be strong representatives of a category.

3.2 Text Categorization Framework

Figure 3.1 illustrates the machine learning framework (Bong, 2001) for automated text categorization. Document preprocessing will be carried out on both the training and testing set. Feature selection will then be performed on the training set to select features for category representation. These features will then be used in the process of machine learning to categorize documents into their respective categories. The design of these components will be discussed in detail in the following sub-sections. This framework will be adopted in our research, whereby we focus on the enrichment of the feature selection module.

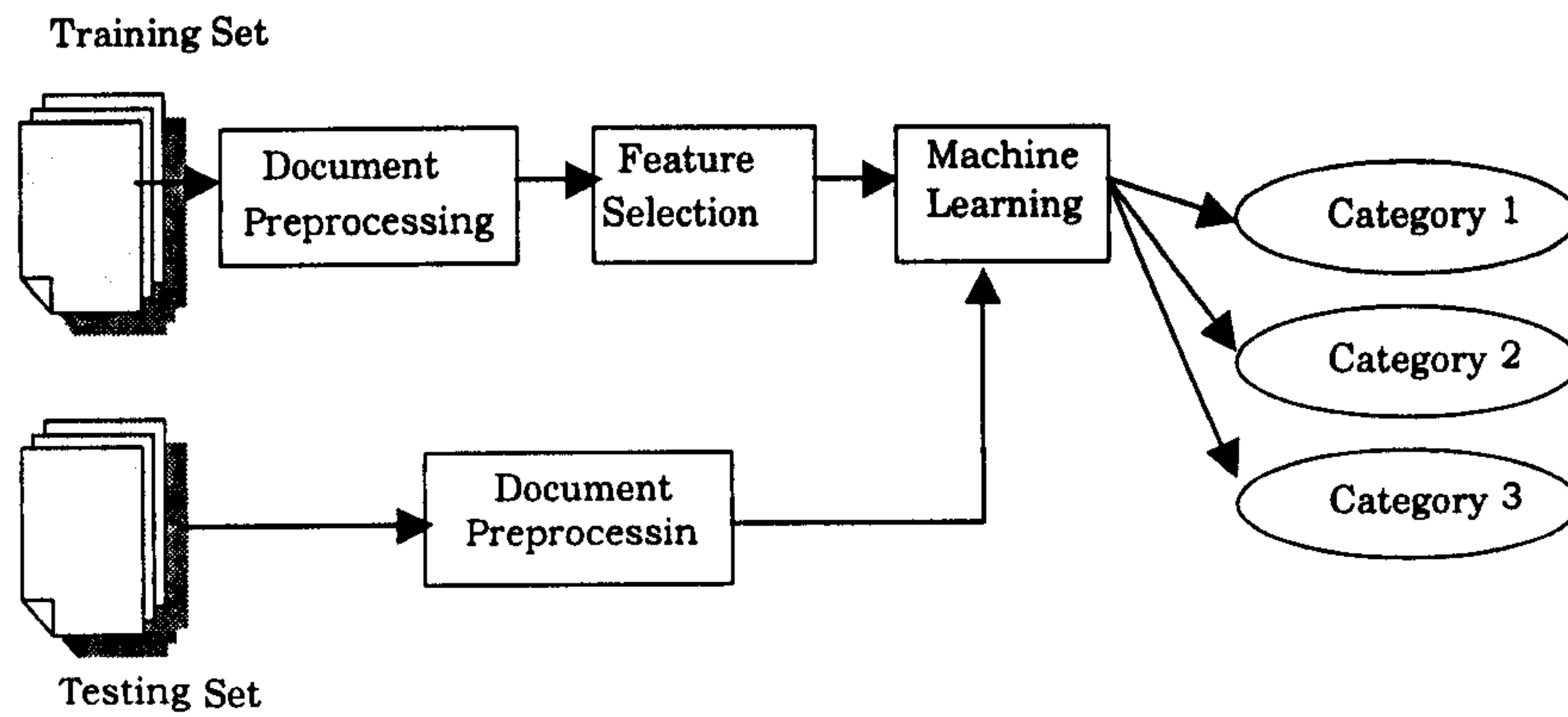


Figure 3.1: Text categorization framework

3.2.1 Document Preprocessing

In document preprocessing, the dataset used is first converted into plain text format. Therefore, we transform the input documents, which are in HTML format, by removing the HTML tags. Subsequently, stop word removal is performed to remove non-informative words. The stop list of 574 words used is the combination of stop words from Frakes and Baeza-Yates (1992) containing 425 words and that of the SMART project (Salton, 1971) with 571 words. Frakes and Baeza-Yates (1992) stated that stop words removal is able to reduce the documents' length between 30% to 50% to increase inter-document dissimilarity. The list of stop words used in this research is shown in Appendix B. After stop words removal is performed, an index file is created. Figure 3.2 shows the processes involved in document preprocessing.

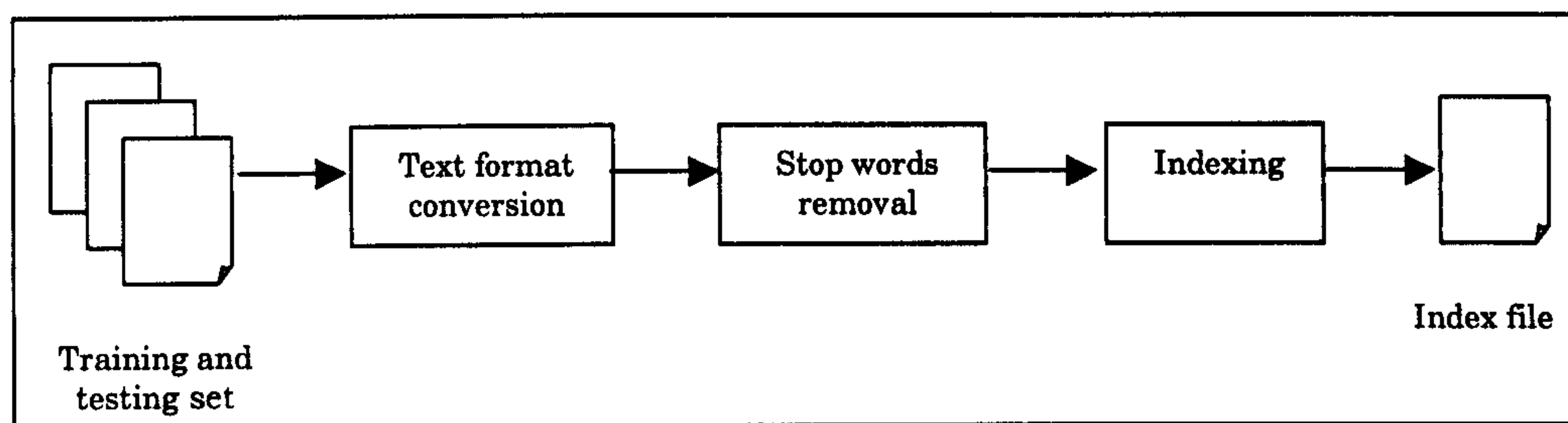


Figure 3.2: Document preprocessing

The general format of the index file is shown below.

<path\filename: term1=frequency1, term2=frequency2, term3=frequency3, ...>

Figure 3.3 is an extract from the index file used in this research. Each line represents a file from a category in the dataset derived from the Reuters-21578 dataset. An example of an index file is shown in Appendix C.

```

c:\temp\reuters-21578-10\train\acq\reut2_00010.txt:company=4,terminal=6,computer=7, .....
c:\temp\reuters-21578-10\train\acq\reut2_00012.txt:quarter=5,said=5,acquisitions=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00044.txt:lines=4,inc=3,transfer=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00045.txt:share=3,mIn=3,inc=3,company=7, .....
c:\temp\reuters-21578-10\train\acq\reut2_00068.txt:inc=3,cofab=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00096.txt:shares=4,cyclops=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00110.txt:company=5,reflect=3,market=9, .....
c:\temp\reuters-21578-10\train\acq\reut2_00125.txt:stake=3,said=3,wrather=4,shares=3,pct=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00128.txt:liebert=4,merger=3, .....
c:\temp\reuters-21578-10\train\acq\reut2_00135.txt:said=3,shares=3,management=3,robeson=5, .....
.
.
.

```

Figure 3.3: An extract of the Reuters-21578 index file

3.2.2 Feature Selection

Instead of using purely statistical methods, WordNet will be used to add in lexical and semantic information in the task of feature selection. In this research, three approaches will be explored in employing WordNet for feature selection: the WordNet nouns approach, the lexical semantics approach and the combination of the lexical semantics approach with statistical approaches.

In the WordNet nouns approach, all WordNet nouns found in each category are used as features representing that category. This approach is discussed in Section 3.2.2.1. In the WordNet lexical semantics approach, feature selection selects terms with overlapping synset signatures co-occurring in a category. The co-occurrence of synonyms is considered significant in identifying terms to represent a category. This approach is illustrated in Section 3.2.2.2.

Combined approach, integrating the statistical feature selection approaches with the WordNet lexical semantics approach will also be explored. This approach is discussed in Section 3.2.2.4. There is a need to explore the strengths and weaknesses of the WordNet-based approaches as compared to existing statistical approaches.

3.2.2.1 WordNet Nouns Approach

WordNet's nouns that appear in each category are taken as features to represent the category. The process of applying the WordNet nouns approach is shown in Figure 3.4. The words in the index file for each category are looked up in WordNet. Nouns are

identified and added as features to represent each category. As opposed to prior works done, where synsets are chosen, our approach aims to identify the nouns. Only WordNet's nouns will be taken into consideration as preliminary experiments (refer to Appendix D) indicate that the use of other parts-of-speech (POS) does not significantly enhance performance.

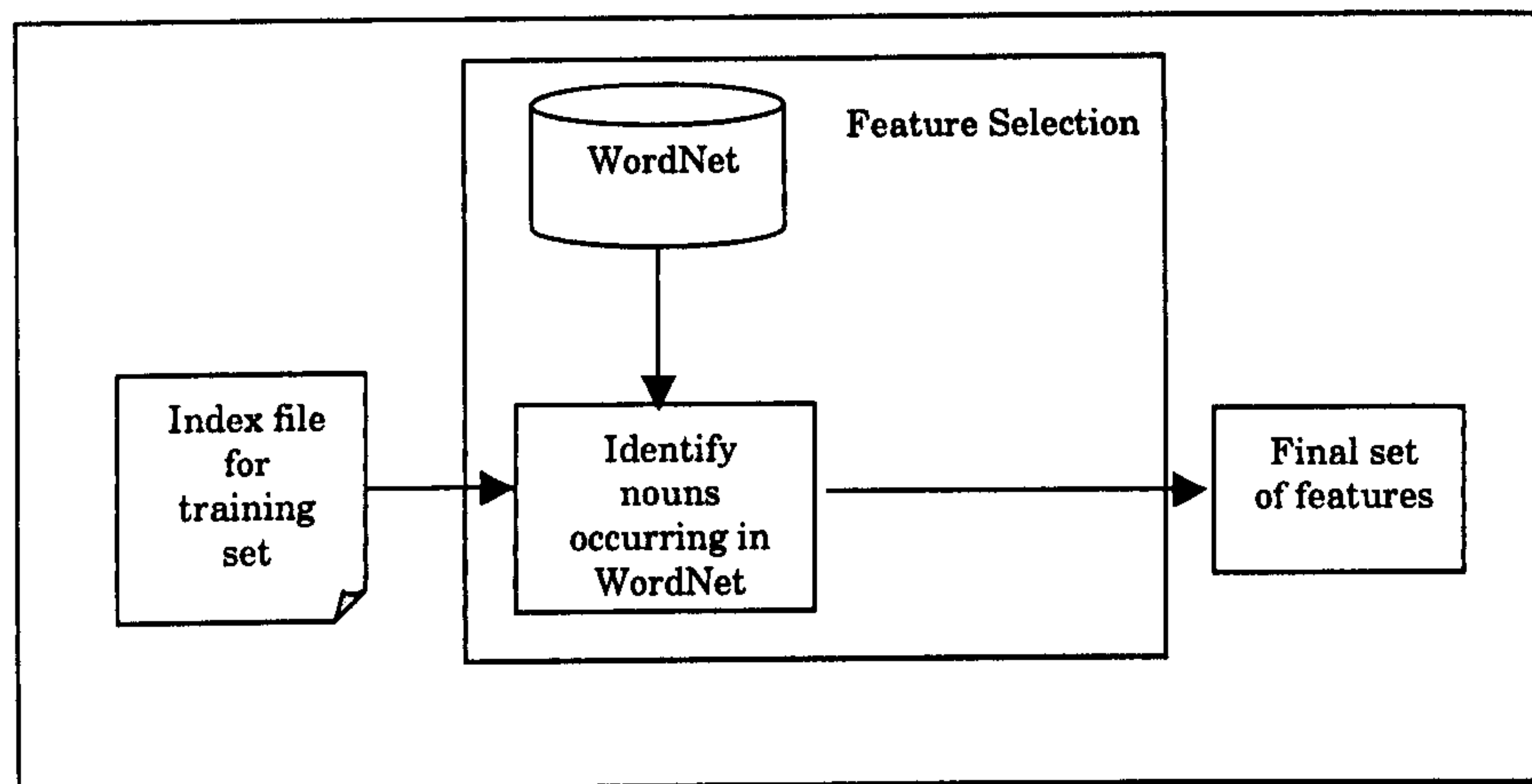


Figure 3.4: The WordNet nouns approach

3.2.2.2 WordNet Lexical Semantics Approach

In this approach, the words in the index file for each category are looked up in WordNet. This approach explores the use of synsets and word senses to capture the actual context of a word's usage. For terms co-occurring in a same category, cross-referencing is done to identify overlapping synset signatures. Cross-referencing is done by checking the list of noun synsets for all senses for similarity in their signatures. This is done by choosing the sense of the target word whose synset signature overlaps with the synset signature of another word in a pool of words from a particular category. Synonymous terms with

overlapping synsets will be chosen as features to be added in the feature space. As all these synonymous terms occurs in the testing documents, taking a single term reduces the performance measures based on the bag-of-words representation in machine learning. Although this approach tends to be simplistic, it allows the addressing of the synonymy problem and has a capacity to capture only terms with lexical semantic relationship that is related to the category.

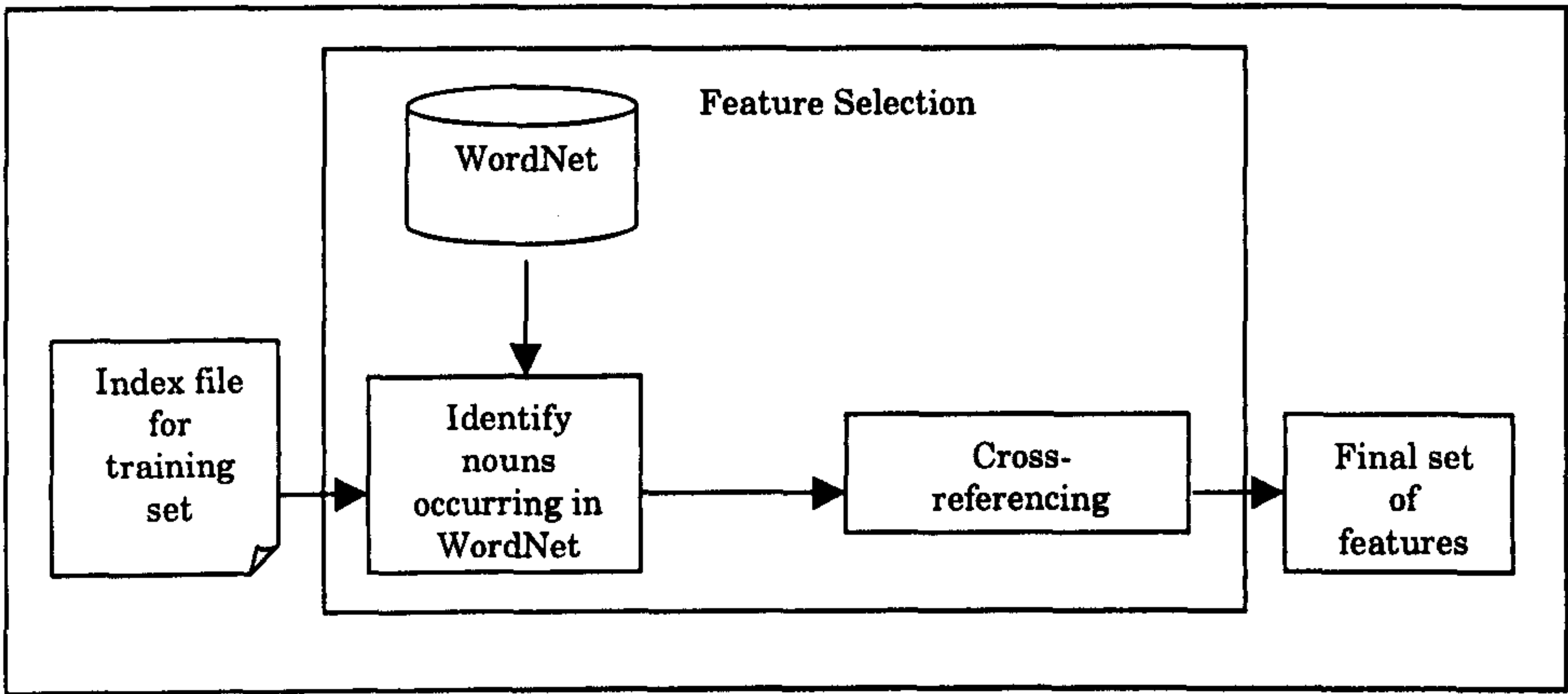


Figure 3.5: The WordNet lexical semantics approach

The following example illustrates the approach used. Let us look at all the senses for four nouns; ‘corn’, ‘maize’, ‘acquisition’ and ‘ship’. Each sense has a different signature, which is referred to as a synset, containing synonyms to reflect a sense.

Terms	Synsets for all senses
Corn	Sense 1: {corn, maize, Indian corn, Zea mays}
	Sense 2: {corn}
	Sense 3: {corn, edible corn}
	Sense 4: {corn, clavus}
	Sense 5: {wheat, corn}
	Sense 6: {corn whiskey, corn whisky, corn}
Maize	Sense 1: {corn, maize, Indian corn, Zea mays}
	Sense 2: {gamboge, lemon, lemon yellow, maize}

Acquisition	Sense 1: {acquisition} Sense 2: {learning, acquisition} Sense 3: {skill, accomplishment, acquirement, acquisition, attainment}
Ship	Sense 1: {ship}

Table 3.1: List of terms and their synsets in all senses

From Table 3.1, we can see that there is an identical synset. It is found in:

Corn (Sense 1) and Maize (Sense 1) - {corn, maize, Indian corn, Zea mays}

As sense 1 of corn and sense 1 of maize have identical synsets of the same signature, the original terms, “corn” and “maize” will be chosen from these synsets as features to represent the category.

Semantically related terms that are found in each category are listed out from the list of overlapping synsets. Their frequencies are then obtained from the index file. Therefore, from the list of similar synsets in the example shown, there are two features selected to represent the category. They are “corn” and “maize”. These terms have a stronger discriminative power in representing a category, as they are conceptually close. By selecting these terms, this approach is aimed at highlighting common terms to describe a category. Terms representative of a common theme will have at least a single synset in common with co-occurring term belonging to the same category.

3.2.2.3 Categorical Sense Determination

In text categorization, both step 1 and 2 of WSD described in Section 2.7 can be accomplished using the lexical semantics approach. WordNet has a list of senses for

each of the words in its dictionary. Therefore, WordNet is able to provide each word with a list of senses that it has. By looking at the context of a word in a category, WordNet can be used to provide the sense of the word. This method of sense determination is known as knowledge-driven sense determination, where information from an external knowledge source, in this case, WordNet, is used to identify the sense of a word. Our approach applies categorical sense determination to automatically disambiguate semantically related terms. Although we identify categorical word senses in this research, we do not make use of the sense information. We merely use the sense information to identify terms with dominant senses by finding the overlapping synset sense signatures. Further processing is required to incorporate the actual word sense of each noun. We applied categorical sense determination to determine the sense of the synonymous terms for a document. We do not focus on WSD per se in this research.

3.2.2.4 Combined Approach

The use of the combined approach allows us to determine whether the lexical semantics approach described in the previous section can enhance the quality of features for automated text categorization. The difference between the lexical semantics approach and the combined approach is that, in the combined approach, statistical feature selection approaches are first performed before using the lexical semantics approach to select features. Our research makes use of the Chi2 and IG formulae from the works of Debole and Sebastiani (2002).

In this combined approach, WordNet's nouns will be selected from the list of terms already selected by a statistical feature selection approach. Cross-referencing is then

done as in the lexical semantics approach and a final set of features is obtained. This combined method will benefit from both the statistical and semantic information. The combined approach process is shown in Figure 3.6.

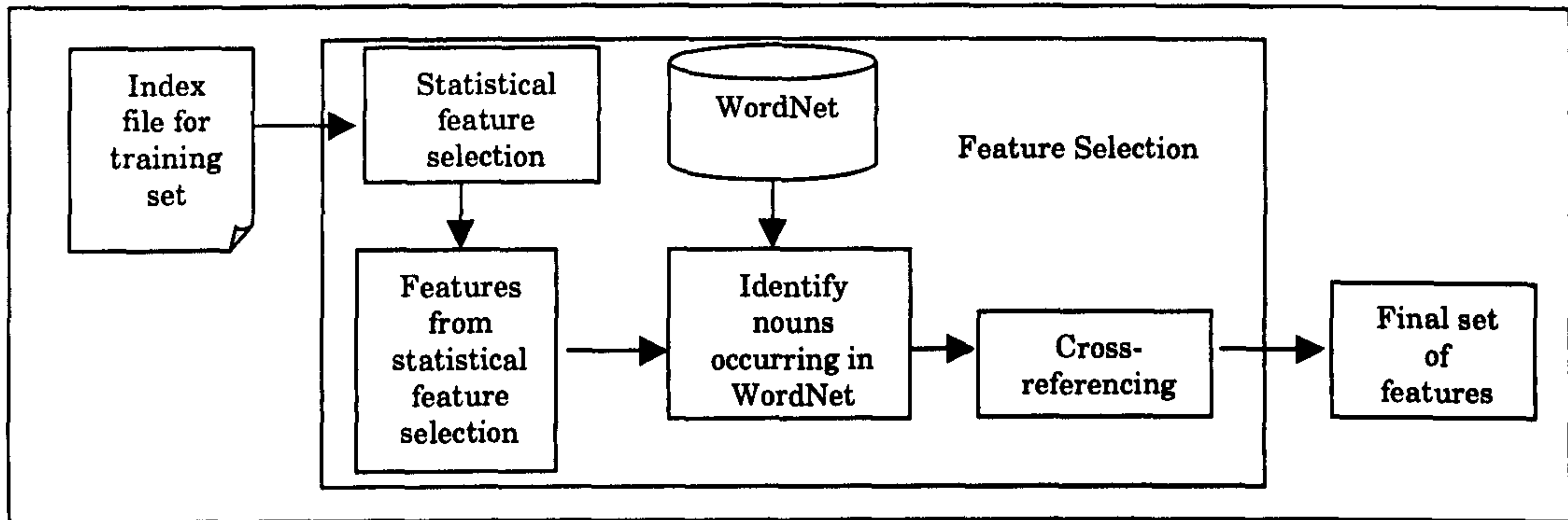


Figure 3.6: The combined approach

3.2.3 Machine Learning for Automated Text Categorization

The algorithm used for categorization in this research is the multinomial Naïve Bayes classifier from WEKA. WEKA is chosen because it has a readily implemented multinomial Naïve Bayes algorithm and also automated computation of effectiveness measures, such as precision, recall and f-measure. We chose the multinomial Naïve Bayes classifier over other types of classifiers as it has been widely used in text categorization tasks and it is simple and straightforward in its implementation. The core equation for the multinomial Naïve Bayes classifier is based on the Bayes theorem, which is described in Section 2.4.1. It is based on the naïve assumption that words in a document occur independently of each other given the class. The classifier parameters consist of the class prior probabilities and the class-conditioned word probabilities. Each class c_j has a prior probability, $P(c_j)$. For every word w_t in the vocabulary V , $P(w_t | c_j)$

indicates the probability that the classifier expects word w_t to occur in documents in class c_j . Learning of these parameters is achieved through the training documents, D , which are labeled. To count the word probability parameters, $P(w_t | c_j)$, the frequency with which word w_t occurs is counted among all word occurrences for documents in class c_j . Laplace smoothing is applied by adding a value of 1 to avoid probabilities of zero. Let $N(w_t, d_i)$ be the count of the number of times word w_t occurs in document d_i and $P(c_j | d_i) \in \{0, 1\}$. Using formula (1), the probability of word w_t in class c_j is

$$P(w_t | c_j) = \frac{1 + \sum_{d_i \in \mathcal{D}} N(w_t, d_i) P(c_j | d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{d_i \in \mathcal{D}} N(w_s, d_i) P(c_j | d_i)} \quad (1)$$

The class prior probability parameters are set in the same way, where $|C|$ indicates the number of classes.

$$P(c_j) = \frac{1 + \sum_{d_i \in \mathcal{D}} P(c_j | d_i)}{|C| + |\mathcal{D}|} \quad (2)$$

If we denote $w_{d_i, k}$ to be the k th word in document d_i , then classification is done based on formula (3).

$$\begin{aligned} P(c_j | d_i) &= P(c_j) P(d_i | c_j) \\ &= P(c_j) \prod_{k=1}^{|d_i|} P(w_{d_i, k} | c_j) \end{aligned} \quad (3)$$

The dataset will be represented in the Attribute Relation File Format (ARFF), which is a relational data file format. The ARFF file will store the list of features after feature selection is performed. A sample set of features is shown below. The features are preceded by the attribute tag “@attribute”. This format is further discussed in Section 4.2.3.

@attribute company numeric

@attribute shares numeric

@attribute pct numeric

@attribute corp numeric

@attribute offer numeric

The multinomial Naïve Bayes learning scheme will learn from the training document representation and induce a classifier. This classifier will then be tested on the testing document representation to evaluate its effectiveness in classification. Figure 3.7 shows the machine learning approach to text categorization.

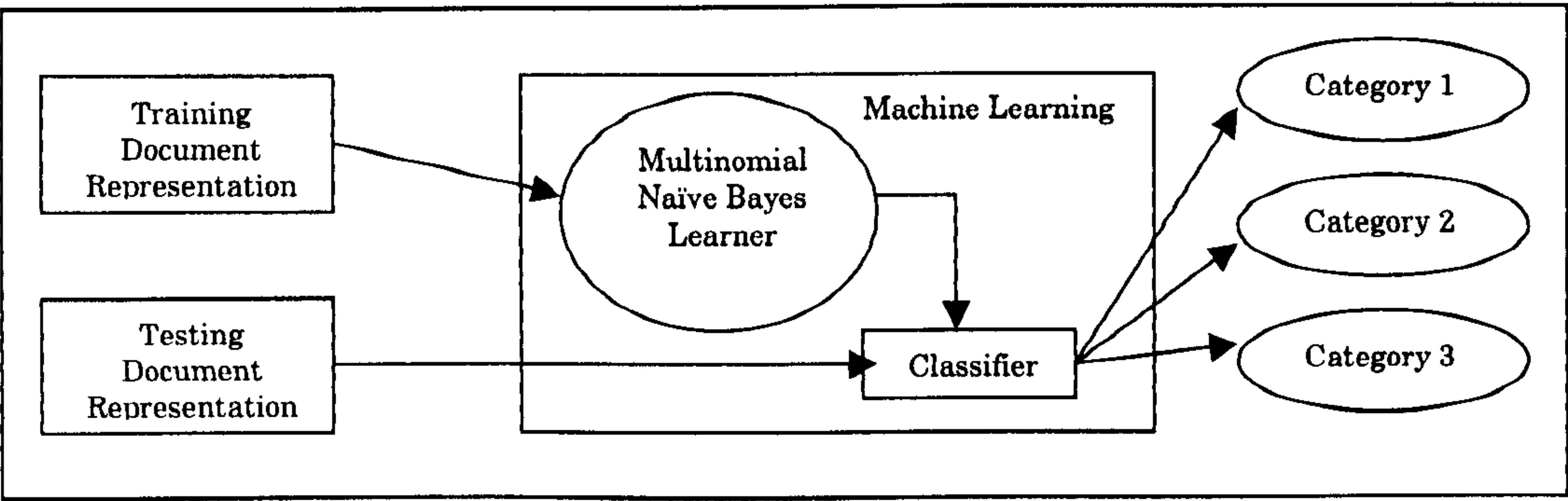


Figure 3.7: The machine learning approach for text categorization

3.3 Summary

The machine learning framework describing the automated text categorization system has been discussed in this chapter. We have also discussed the design of the WordNet nouns approach, the lexical semantics approach and the combined approach for feature selection. The use of WordNet for feature selection in automated text categorization allows the incorporation of lexical and semantic information in selecting features for categorization. In the combined approach proposed, we explore the potential of enhancement to existing statistical feature selection methods.

Chapter 4 Implementation

4.1 Introduction

This chapter gives an overview of the implementation of the machine learning based automated text categorization system of the proposed WordNet nouns and lexical semantics approaches to feature selection. The implementation of the combined approached is also discussed here.

4.2 Implementation

Figure 4.1 shows the context diagram for the automated text categorization process. The context diagram shows the overall processes involved in the text categorization task. Generally, the whole process is separated into three phases; document preprocessing, feature selection and lastly, machine learning.

In the document preprocessing phase, tasks of plain text conversion and index file creation are carried out. In the feature selection phase, three different approaches are applied to select features. The WordNet nouns approach is shown as Process 5 followed by Process 8 while the WordNet lexical semantics approach is shown by Process 5 and followed by Process 6 and 7. The combined approach is shown by Process 4 and followed by Process 5, 6 and 7. The design of these approaches were discussed in Section 3.2.2. In the machine learning stage, the creation of ARFF file is explained and the process of machine learning is carried out using the Waikato Environment for Knowledge Analysis

(WEKA) workbench. The results are then obtained and analysed. The following sub-sections will discuss in detail each of these three phases separately.

The implementation was done using the Java programming language. The Java 2 Standard Development Kit version 1.4.1 is used to perform java programming. A Java application programming interface (API) for accessing WordNet called the Java WordNet Library (JWNL) version 1.3, is also used to access the dictionary files in WordNet.

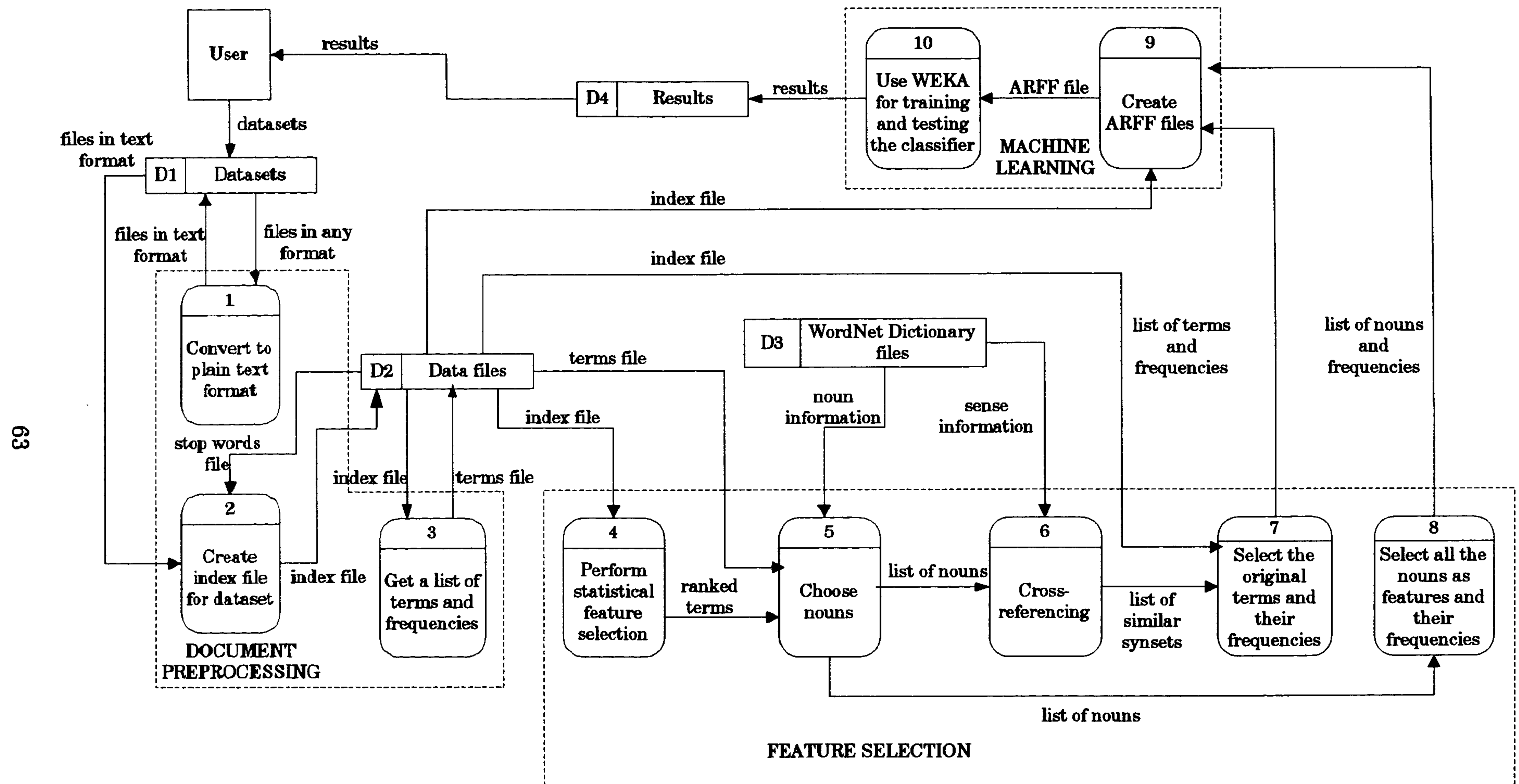


Figure 4.1: Context diagram for the text categorization process

4.2.1 Document Preprocessing

In document preprocessing, the dataset obtained by the user will need to be formatted and indexed before feature selection can take place. The dataset can be in any format, therefore, conversion of the dataset from its original format to text format is first done. Conversion to text format is done so that indexing of the dataset can be performed to ensure a more efficient retrieval in the future. When the dataset is already converted into text format, indexing will be done on the dataset.

Figure 4.2 shows the indexing process in detail. The path of the file is first obtained and that particular file will be accessed. The contents of the file will be listed out and compared against a list of stop words. Stop words are non-informative words such as articles, prepositions and conjunctions. Examples of stop words are 'and', 'after', 'the', 'a' and so on. These words do not provide any discriminative value and therefore does not help in the categorization task. Stop words will be removed from the list of words and a new list without the stop words will be produced. In this new list, the occurrence of each term in the list is counted and appended to the respective term. Lastly, the file path, terms and their frequencies will be written into the index file.

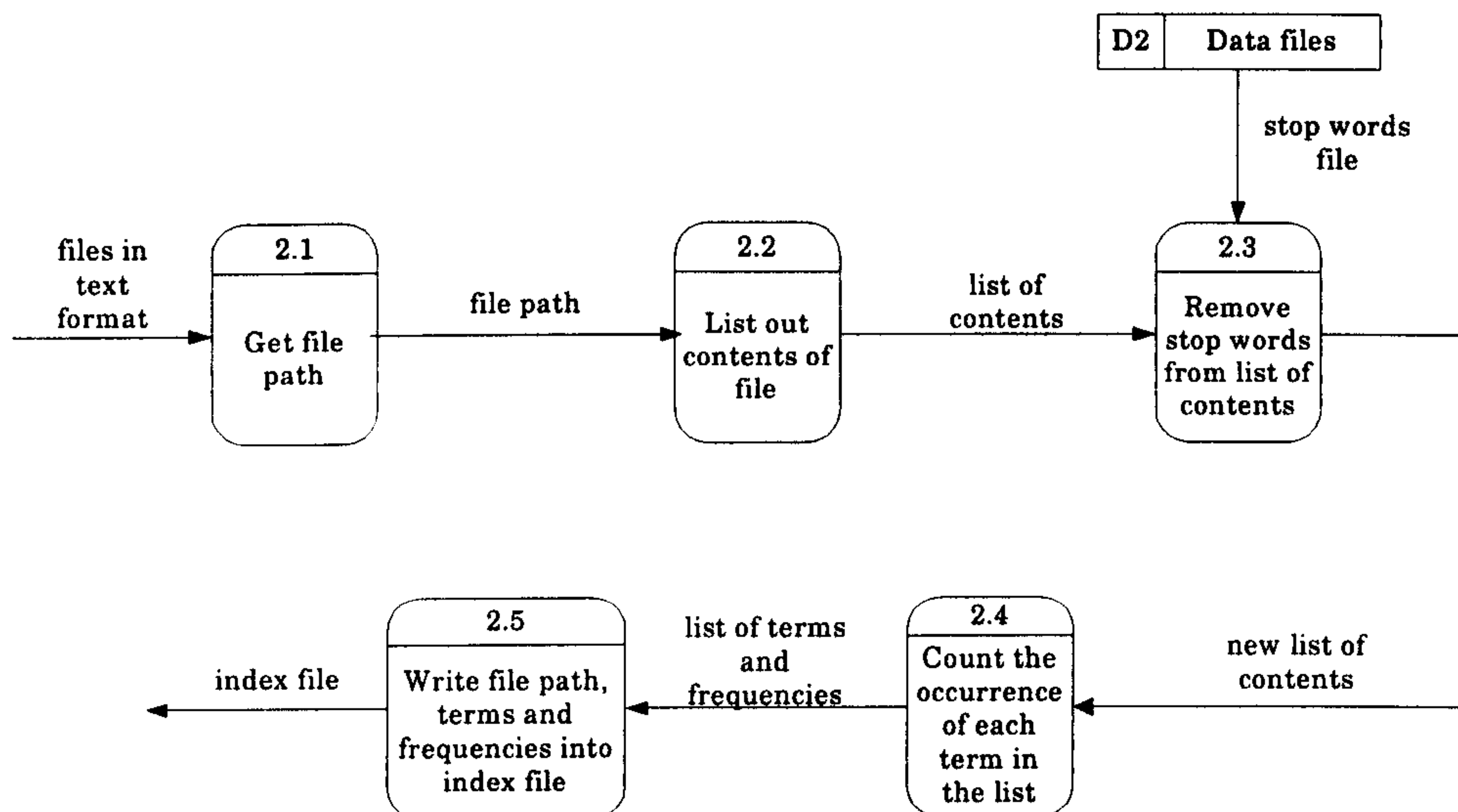


Figure 4.2: Child diagram for process 2 – Create index file for dataset

4.2.2 Feature Selection

Three WordNet based feature selection approaches are employed in this research, namely the WordNet nouns approach, the lexical semantics approach and the combination of the lexical semantics approach with statistical approaches. We will further discuss the implementation of each of these three approaches in the following sections.

4.2.2.1 WordNet Nouns Approach

The WordNet nouns approach is highlighted by Process 5, followed by Process 8 in Figure 4.1. In this approach, the terms file, which is produced in the preprocessing stage is used. This terms file contains a list of terms and their frequencies found in a category. WordNet's nouns will be chosen using the nouns information from the WordNet's

dictionary. The list of terms is checked against WordNet's nouns and terms that are not found in WordNet's nouns file will be omitted. The chosen nouns will then be used as features to represent the category.

4.2.2.2 WordNet Lexical Semantics Approach

The WordNet lexical semantics approach is illustrated by Process 5, 6, and 7 in Figure 4.1. In this approach, we also make use of the terms file. From this list of terms, only WordNet's nouns will be chosen. This is done through acquiring the noun information from WordNet's dictionary. Terms that are not found in WordNet's nouns file will be omitted as in the WordNet nouns approach.

The lexical semantics approach to feature selection in this research is based on synonymous terms that co-occur in a category. For terms co-occurring in a category, the sense assigned is selected based on the synset signature reference. Cross-referencing is done by checking the list of noun synsets in all senses for similar synsets or overlapping synsets. If there are any synsets that are similar, the similar synsets will be listed down while all the other synsets will be discarded. Figure 4.3 shows the processes involved in cross-referencing. Figure 4.4 shows five examples of synsets.

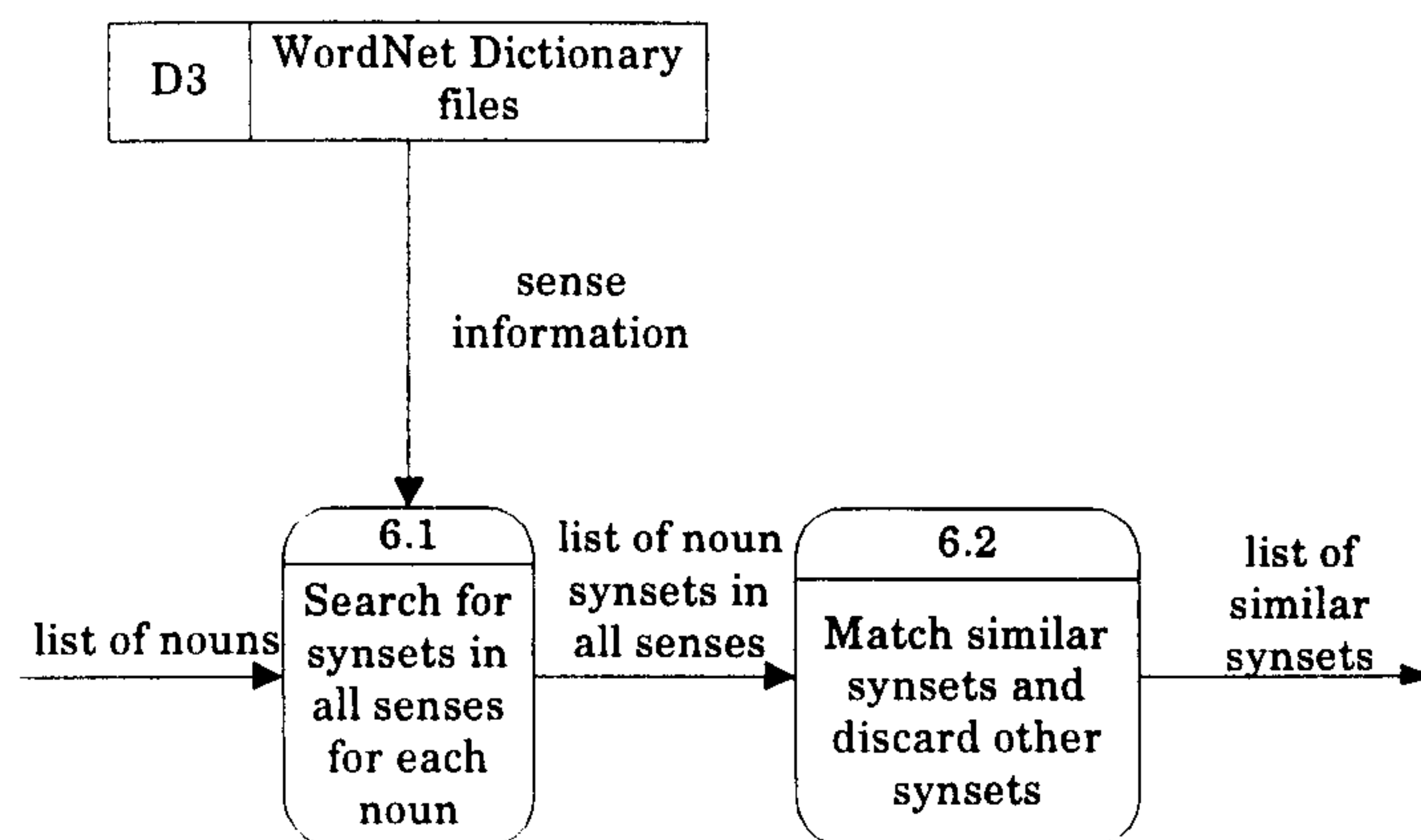


Figure 4.3: Child diagram of process 6 – Cross-referencing

[Synset 6952517@[POS noun]: "national, subject"]
 [Synset 517535@[POS noun]: "retention, keeping, holding"]
 [Synset 4745080@[POS noun]: "radio, radiocommunication, wireless"]
 [Synset 10426280@[POS noun]: "security"]
 [Synset 7354565@[POS noun]: "king, male monarch"]

Figure 4.4: Examples of synsets

4.2.2.3 Combined Approach

For the combined approach, statistical feature selection is first performed to select and rank terms from each category. This is shown as Process 4 in Figure 4.1. Firstly, terms and frequencies are obtained from the index file. Then, calculations for each term in a category will be performed based on the statistical feature selection approach used. Each term will be assigned a value from the calculations and will be ranked in descending order according to the value. The terms will then be used as an input to cross-referencing in the lexical semantics feature selection.

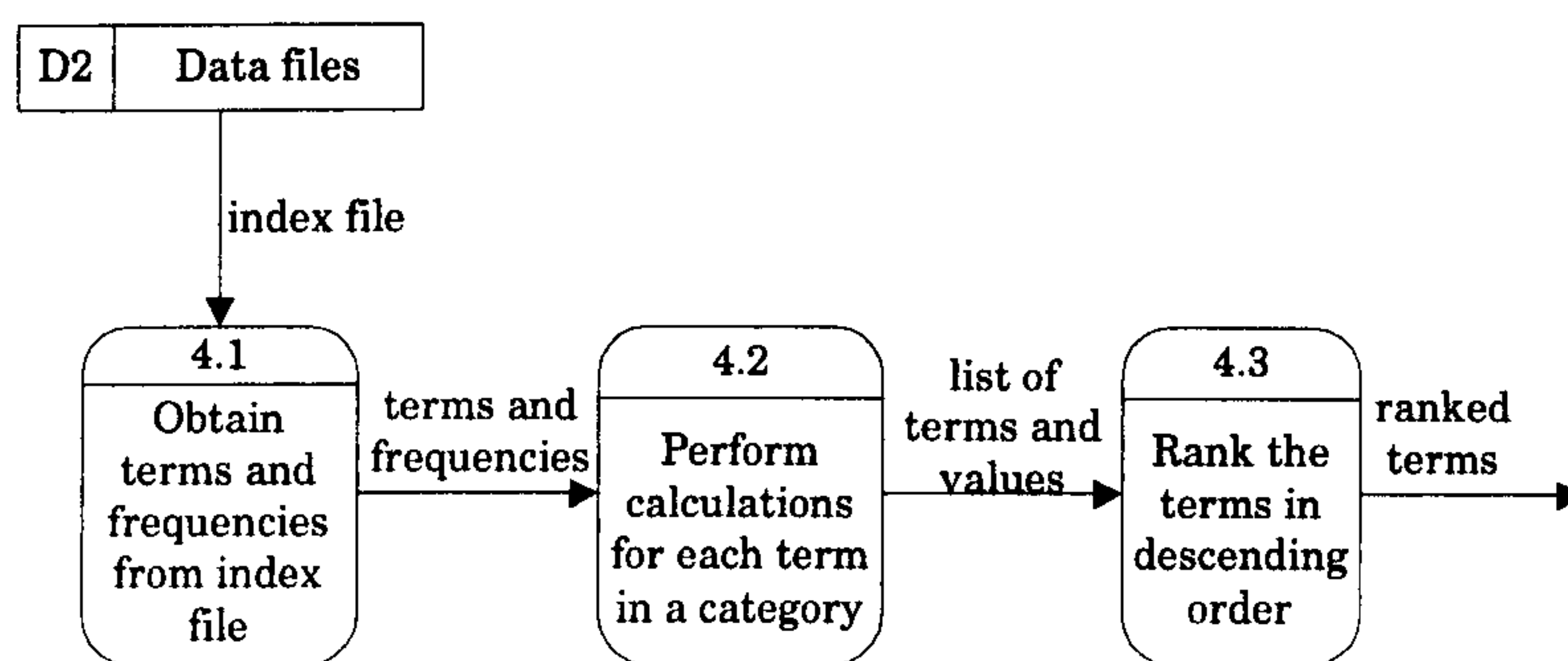


Figure 4.5: Child diagram of process 4 – Perform statistical feature selection

4.2.3 Machine Learning

The final stage is the employment of machine learning in the text categorization task. WEKA is used as the machine learning tool in this research. The input file for this machine learning tool has to be formatted into the Attribute Relation File Format (ARFF).

An ARFF file consists of the name of the relation, attributes and instances. Lines beginning with a “%” are comments. The name of the relation comes first. This is then followed by the attributes, which can be either nominal or numeric. Nominal values are specified after each attribute in curly braces while numeric ones are followed by the keyword *numeric*. The last attribute is usually the class where instances are categorized into these classes. The class attribute is however, not differentiated in any way in the data file. Therefore, any attributes can be used as the prediction attribute. The instances of the dataset are written after the “@data” line that separates the attributes and the instances. An instance represents a single document in a category.

From the feature selection process, a list of terms and their frequencies is produced. This list of terms is used in this stage as attributes in the ARFF file. The next step is to compare the attributes with each line in the index file of the same category. As mentioned in Section 3.2.1, each line in the index file represents a file from a category in the dataset. If the attribute exists in the line, the attribute index is written into the file and the attribute's frequency is recorded as the value of the instance. However, if the attribute does not exist in the line, meaning the frequency of occurrence for that attribute is 0 for that instance, then the attribute index will not be written onto the file. Each instance is written in the following format,

```
{2 1,3 1,40 1,42 1,66 1,115 1,165 1,189 1,196 1,212 1,258 3,278 1,293 1,660 2,714 1,1053
1,1129 3,1609 1,1669 2,1934 5,2000 0}
```

where, a pair of numbers separated by a white space, indicates <index> <space> <value>. Index represents the attribute index with values 0 to n, where n is a positive integer value. As shown in the example above, the first number pair (2 1) indicates that attribute number 3 occurs once in this instance. The last number pair (2000 0) indicates that attribute number 2001, which is the class for this instance, has value 0, indicating that this instance does not belong to the class. A value '1' will signify that the instance belongs to a particular class. An ARFF file is created after all the required information is written into the file. Figure 4.6 shows the process of creating an ARFF file. Figure 4.7 shows an extract of an ARFF file. An example of an ARFF file is shown in Appendix E.

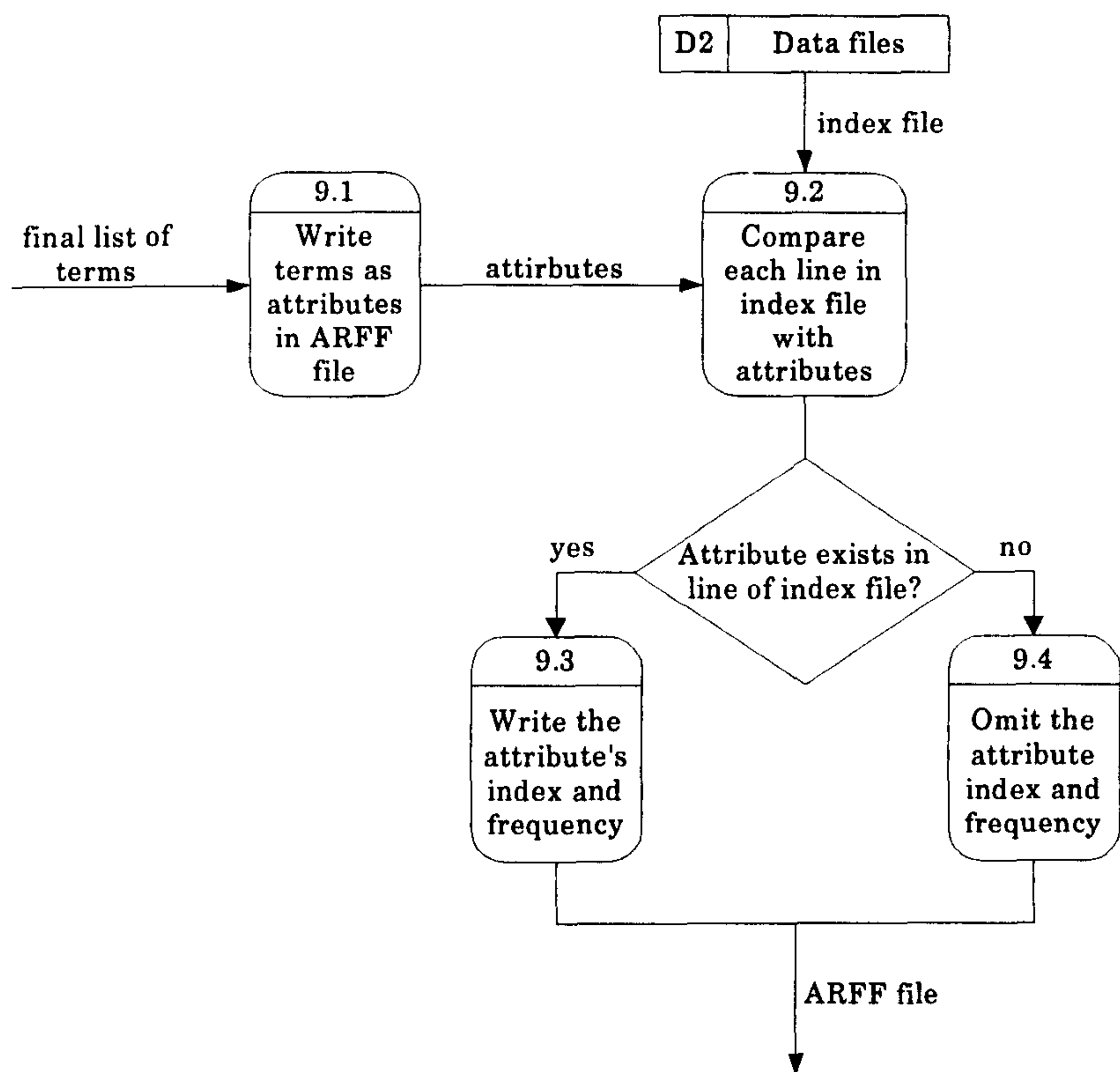


Figure 4.6: Child diagram for process 9 – Create ARFF files

```

%This is an ARFF file
@relation acq
@attribute company numeric
@attribute shares numeric
@attribute pct numeric
@attribute corp numeric
@attribute offer numeric
@attribute share numeric
@attribute stock numeric
@attribute stake numeric
.
.
.
@attribute class {1,0}
@data
{2 1,3 1,40 1,42 1,66 1,115 1,165 1,189 1,196 1,212 1,258 3,278 1,293
1,660 2,714 1,1053 1,1129 3,1609 1,1669 2,1934 5,2000 0}
{26 1,54 2,87 1,101 2,103 1,109 1,139 1,200 1,205 1,222 2,255 1,270
1,400 2,417 1,441 1,552 1,599 1,646 2,807 2,1609 3,1668 2,1782
1,2000 0}
{35 1,38 2,87 1,109 2,222 3,243 1,270 2,327 1,784 1,1623 1,1668
3,1758 1,1782 3,2000 0}
  
```

Figure 4.7: An extract of an ARFF file

4.2.3.1 Machine Learning Tool: Waikato Environment for Knowledge Analysis (WEKA)

The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning tool that consists of a complete set of state-of-the-art machine learning and data mining algorithms (Witten and Frank, 2000). WEKA was developed in the University of Waikato, New Zealand using the Java programming language in an object-oriented approach.

WEKA provides tools for pre-processing data, running the data through various machine learning schemes and analyzing the classifiers and their performance. WEKA uses classifiers to induce a rule set or decision tree that models the data. From a single representation model, the user can apply several learning schemes and compare their performance in order to choose the best learning scheme. In this research, we use the multinomial Naïve Bayes classifier, as described in Section 3.2.3.

The output from WEKA is shown as a confusion matrix, also known as a contingency table with its precision, recall and f-measure values calculated by WEKA. A screenshot of the output is shown in Figure 4.8. The values used for performance measures (i.e. precision, recall and f-measure) as well as the contingency table have been discussed in Section 2.11.1.

```

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   Class
  0.273     0.097     0.25       0.273     0.261       1
  0.903     0.727     0.913      0.903     0.908       0

=== Confusion Matrix ===

  a    b  <-- classified as
12  32 |   a = 1
36 334 |   b = 0

```

Figure 4.8: Screenshot of the output from the machine learning process

4.3 Summary

In this chapter, we present the details of the implementation of the automated text categorization system and the proposed feature selection approaches using WordNet. The experiments that are carried out to evaluate the classifier are discussed in Chapter 5.

Chapter 5 Results and Analysis

5.1 Introduction

This chapter will describe the experimental setup, the dataset used as well as the results of the experiments and the analysis of the results. The metrics to be used for evaluation will also be discussed here.

5.2 Dataset

Conducting experiments with a standard dataset is important to ensure that comparison can be done between different researches. The dataset used in conducting the experiments is from the benchmarked Reuters-21578 corpus.

The Reuters-21578 Distribution 1.0 collection is one of the most widely used benchmarked dataset in text categorization researches (Hull, 1994). This corpus contains 21578 documents that appear on the Reuters newswire in 1987. There are five category sets in this collection, namely, “Exchanges”, “Orgs”, “People”, “Places”, and “Topics”. The number of categories with at least 1 occurrence of documents and at least 20 occurrences of documents are shown in Table 5.1. The set of category that has been used in almost all previous researches is the “Topics” set (Lewis, 1997). This category set contains general economic subject categories “corn”, “grain”, “interest” and so on. Each document in the collection is assigned to at least one category by human indexers.

Category Set	Number of Categories	Number of Categories with at least 1 occurrence of documents	Number of Categories with at least 20 occurrences of documents
Exchanges	39	32	7
Orgs	56	32	9
People	267	114	15
Places	175	147	60
Topics	135	120	57

Table 5.1: The number of categories in each category set and the number of categories with at least 1 occurrence and 20 occurrences

In this research, the categories from “Topics” will be used so that we can benchmark with other research works. The training and testing set is split using the Mod Apte split (Apte et al., 1994). This split assigns documents on April 7, 1987 and before to the training set and documents on April 8, 1987 and after to the test set. Only documents with an occurrence of at least one “Topics” category are used. Many researchers use Mod Apte split so that benchmarking can be done with other similar works (Scott and Matwin, 1998; Cohen and Singer, 1998; Dumais et al., 1998; Joachims, 1998; Yang and Liu, 1999; Masuyama and Nakagawa, 2002).

5.3 Experiments

The experiments that are carried out are to test and compare the effectiveness of each feature selection method. The corpus text is represented using the bag-of-words representation, applying the frequency model, where the frequency of occurrence of the attribute is used as the value to represent each attribute. The experiments are carried out using WEKA, applying the multinomial Naïve Bayes machine learning scheme. The

experiments were conducted using the entire “Topics” set from Reuters-21578 and subsequently, the top 10 categories from the dataset to explore the results in more detail. Experiments were carried out to compare and contrast the following six feature selection methods:

1. Chi-square (Chi2)
2. Information Gain (IG)
3. WordNet nouns
4. WordNet lexical semantics
5. Combination of Chi-square and lexical semantics
6. Combination of Information Gain and lexical semantics

There were three sets of experiments. Each set of experiments consists of 7 different values for term size: 10, 20, 50, 100, 200, 500, and 1000. These different term sizes were chosen to evaluate the effectiveness of the classifier to see which term size can give optimal performance for the classifier.

1. WordNet nouns vs WordNet lexical semantics.
2. Comparison between WordNet nouns, WordNet lexical semantics, Chi2 and IG.
3. Comparison between Chi2, IG and the combined approaches.

Experiment 1 was carried out to first compare the WordNet nouns approach with the WordNet lexical semantics. In Experiment 2, statistical approaches were compared with the WordNet based approaches to evaluate the performance of a non-statistical approach with statistical approaches. Lastly, in Experiment 3, a combination of the lexical semantics approach and the statistical approaches is done to see whether the

lexical semantics approach is able to enhance the performance of the statistical approaches.

5.4 Results and Analysis for Reuters-21578 Dataset

The results and analysis for the Reuters-21578 dataset is reported in this section. It covers the f-measure analysis, where the f-measure value is used, as well as the qualitative analysis. Qualitative analysis was performed to complement the statistical results to indicate the reasons why some results are better or worse than the others. The extent of dimensionality reduction is also analyzed to see how much reduction is achieved through the proposed WordNet based feature selection methods. The percentage of reduction is calculated based on the number of terms in each category and the number of terms selected by the WordNet lexical semantics approach and the WordNet nouns approach. In the sections that follow, we denote the WordNet lexical semantics approach as WLS, the WordNet nouns approach as WN, Chi-square as Chi2 and Information Gain as IG.

5.4.1 Dimensionality Reduction

This section looks at the extent of dimensionality reduction that is achieved in this research by using the WLS approach. Table 5.2 shows the percentage of terms reduction when the WLS approach is used compared to the number of unique terms in each category. Generally, this approach to feature selection is able to reduce the number of terms by more than 67%.

Category	Number of unique terms in each category	Number of terms selected by WLS	Percentage of terms reduction (%)
Acq	10760	2793	74.0
Corn	3089	955	69.1
Crude	5890	1834	68.9
Earn	10152	2342	76.9
Grain	5231	1716	67.2
Interest	3679	1147	68.8
Money-fx	5323	1667	68.7
Ship	3902	1258	67.8
Trade	5569	1823	67.3
Wheat	3358	1109	67.0

Table 5.2: Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the number of unique terms in each category

5.4.2 Experiment 1: WN vs WLS

This experiment was carried out to see the performance of the WN approach as compared to the WLS approach.

5.4.2.1 F-measure Analysis

Figure 5.1 shows that the WLS approach works better than the WN approach across all the categories in the Reuters-21578 dataset. It, however, performed marginally lower than the WN approach at term size 10. From the micro-averaged results shown in Figure 5.2, the WLS approach has performed better than the WN approach at all term size thresholds. From the two graphs, we note that although the performance of the WLS approach is generally better than the WN approach, the margin of effectiveness is rather small for smaller term sizes, especially for 10 and 20 terms.

When we look into the individual category results, we find that for some categories, the WN approach did marginally better than the WLS approach while in some categories, the two approaches performed equally well and for some categories, the WLS approach did better, although, overall, the WLS approach still performed better than the WN approach.

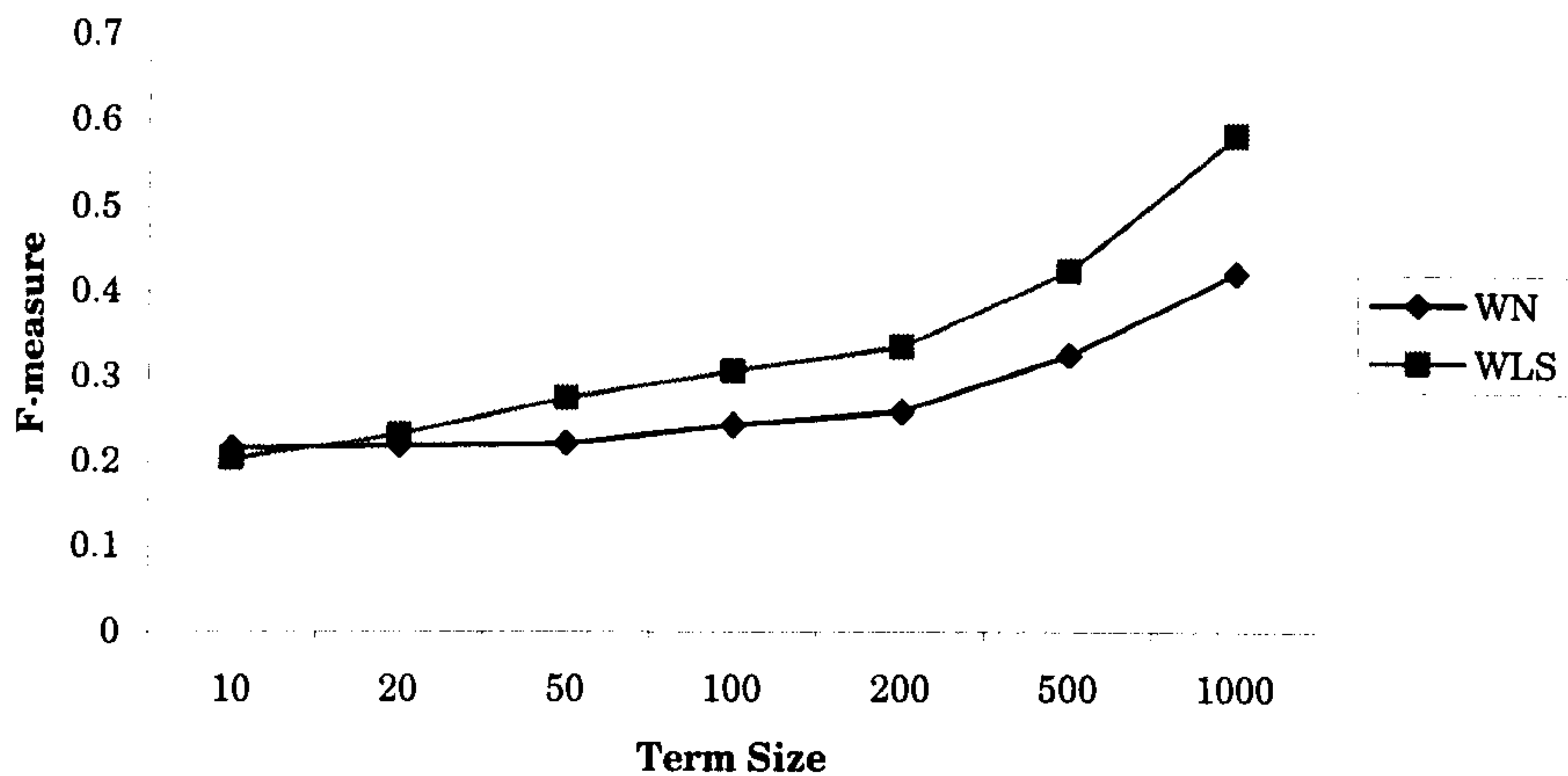


Figure 5.1: Macro-averaged f-measure comparison between WLS and WN for the Reuters-21578 dataset

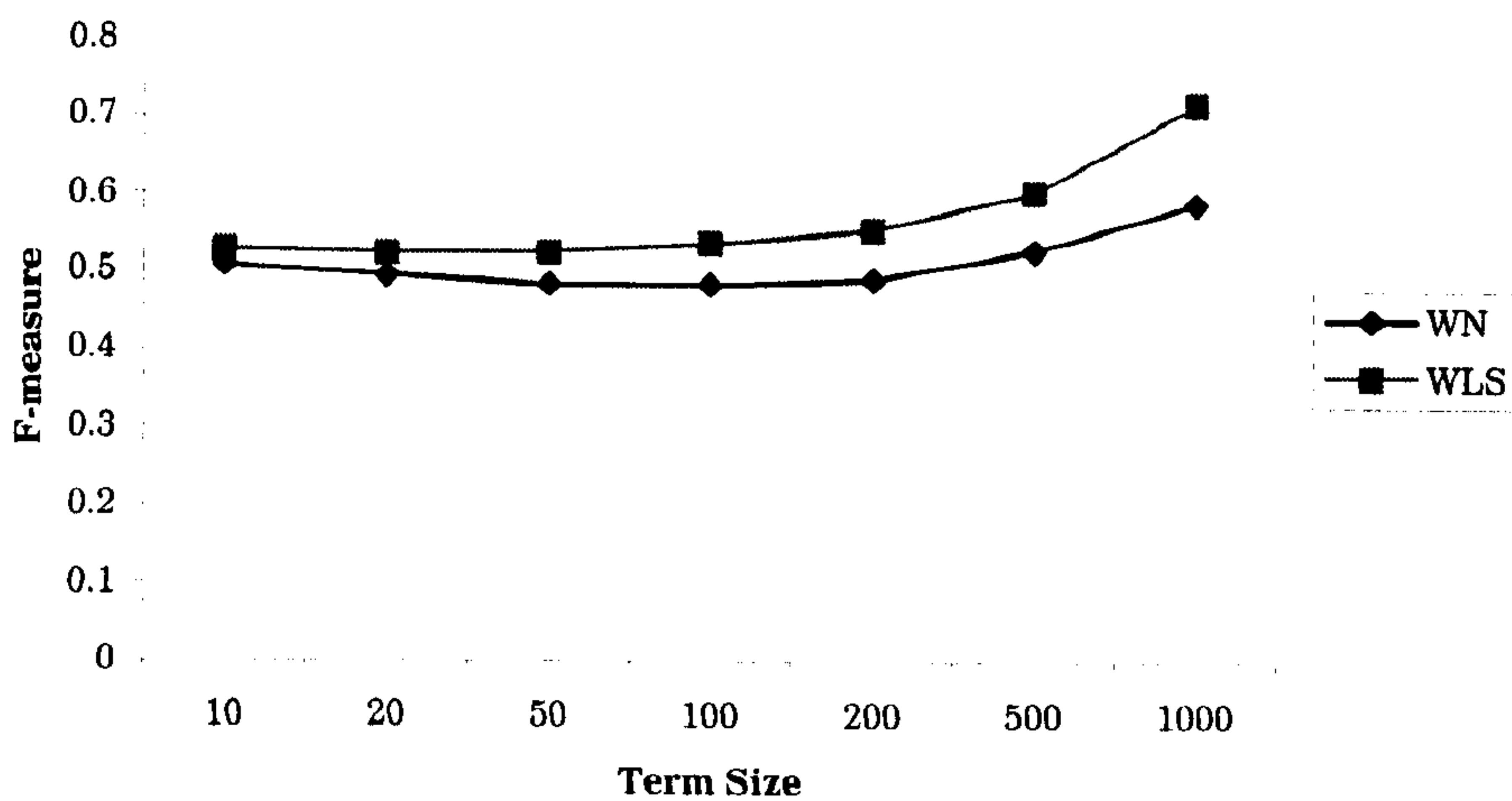


Figure 5.2: Micro-averaged f-measure comparison between WLS and WN for the Reuters-21578 dataset

5.4.2.2 Dimensionality Reduction

When compared to the WN approach, the WLS approach was also able to significantly reduce the number of terms chosen in each category by the WN approach. For example, in the category “Ship”, the WN approach is able to discover 2344 nouns. This number is reduced to 1258 when the WLS approach is applied to the same category. This results in a 46.3% reduction in the number of terms selected in feature selection. In another category “Wheat”, the WN approach found 2001 nouns. The WLS approach reduced this number to 1109, which produced a reduction of about 44.6%. This approach was able to reduce the number of features selected up to 40% and more and at the same time, perform better across all categories. Table 5.3 shows the percentage of terms reduction when the WLS approach is used compared to the WN approach.

Category	Number of terms selected by WN	Number of terms selected by WLS	Percentage of terms reduction (%)
Acq	4944	2793	43.5
Corn	1797	955	46.9
Crude	3206	1834	43.0
Earn	4327	2342	45.9
Grain	2934	1716	41.5
Interest	2039	1147	43.7
Money-fx	2832	1667	41.1
Ship	2344	1258	46.3
Trade	3108	1823	41.3
Wheat	2001	1109	44.6

Table 5.3: Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the WN approach

5.4.2.3 Qualitative Analysis

In category “Interest”, the WN approach was able to perform better than the WLS approach. The qualitative analysis of the top 20 terms selected by the two approaches is shown in Table 5.4.

Feature selection scheme	Top 20 terms for category “Interest”
WLS	pct, rate, market, fed, cut, prime, government, week, funds, federal, term, base, dealers, central, month, bills, reserve, february, growth, dollar
WN	pct, title, text, bank, rate, rates, market, money, banks, fed, cut, prime, government, week, funds, federal, term, base, dealers, central

Table 5.4: Top 20 terms for category “Interest” for WLS and WN

There were 6 unique terms selected by each method. The WLS approach has terms “month”, “bills”, “reserve”, “february”, “growth” and “dollar” while the WN approach has terms “title”, “text”, “bank”, “rates”, “money” and “banks”. As human evaluators, we can say that both “money” and “bank” are strong terms in representing the category “Interest”. The WLS approach however, failed to select “money” because in the WordNet dictionary files, this term exists as a single term in its synset. It has no synonyms and therefore, based on the simplistic WLS approach, this term was not selected. In the case of the term “bank”, the WordNet dictionary has phrases as synonyms for “bank”. As this research makes use of single words in the bag-of-words representation, we do not consider phrases or collocation as it is called in WordNet, as terms to represent our categories.

Overall, the WLS approach was able to perform better than the WN approach across all categories. The WLS approach has worked well as there were not many categories with words that do not have synonyms that are considered as important features for categorization.

5.4.3 Experiment 2: Comparison between WN, WLS, Chi2 and IG

In this experiment, the WordNet based approaches, WN and WLS, were compared against statistical approaches, Chi2 and IG, to see which of these approaches can perform best.

5.4.3.1 F-measure Analysis

In comparing the results of the WordNet based approaches with the statistical feature selection approaches, the macro-averaged and micro-averaged f-measure is reported for the experiments carried out on the Reuters-21578 “Topics” dataset, as well as the top ten categories from the dataset. The macro-averaged experiments on the full set of Reuters-21578 has shown that the WLS performed significantly better than Chi2 and IG while the WN performed as well as the statistical approaches. We will analyze why this is so in the following section. The results for macro-averaged and micro-averaged f-measure are shown in Table 5.5 and 5.6 and plotted in Figure 5.3 and 5.4 respectively. The results printed in bold shows the highest results for each of the term size.

Term Size	Chi2	IG	WN	WLS
10	0.192	0.228	0.217	0.205
20	0.208	0.224	0.220	0.234
50	0.240	0.213	0.222	0.275
100	0.270	0.209	0.243	0.306
200	0.271	0.228	0.259	0.335
500	0.271	0.283	0.325	0.423
1000	0.352	0.363	0.419	0.581

Table 5.5: Macro-averaged f-measure for the Reuters-21578 dataset

Term Size	Chi2	IG	WN	WLS
10	0.507	0.500	0.507	0.526
20	0.513	0.473	0.492	0.521
50	0.547	0.438	0.479	0.521
100	0.521	0.430	0.479	0.531
200	0.493	0.441	0.486	0.550
500	0.464	0.473	0.524	0.598
1000	0.519	0.529	0.584	0.713

Table 5.6: Micro-averaged f-measure for the Reuters-21578 dataset

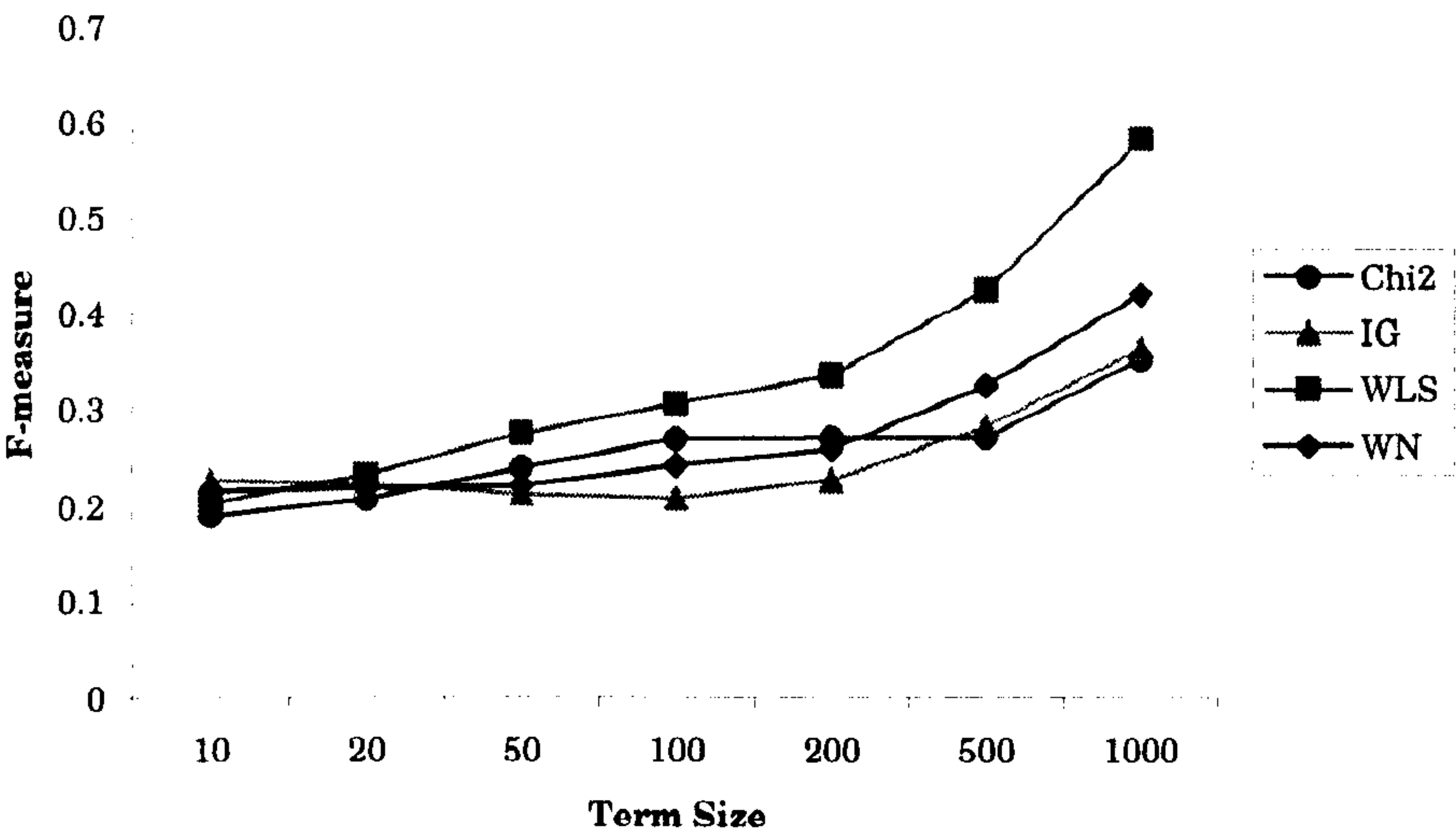


Figure 5.3: Macro-averaged f-measure for the Reuters-21578 dataset

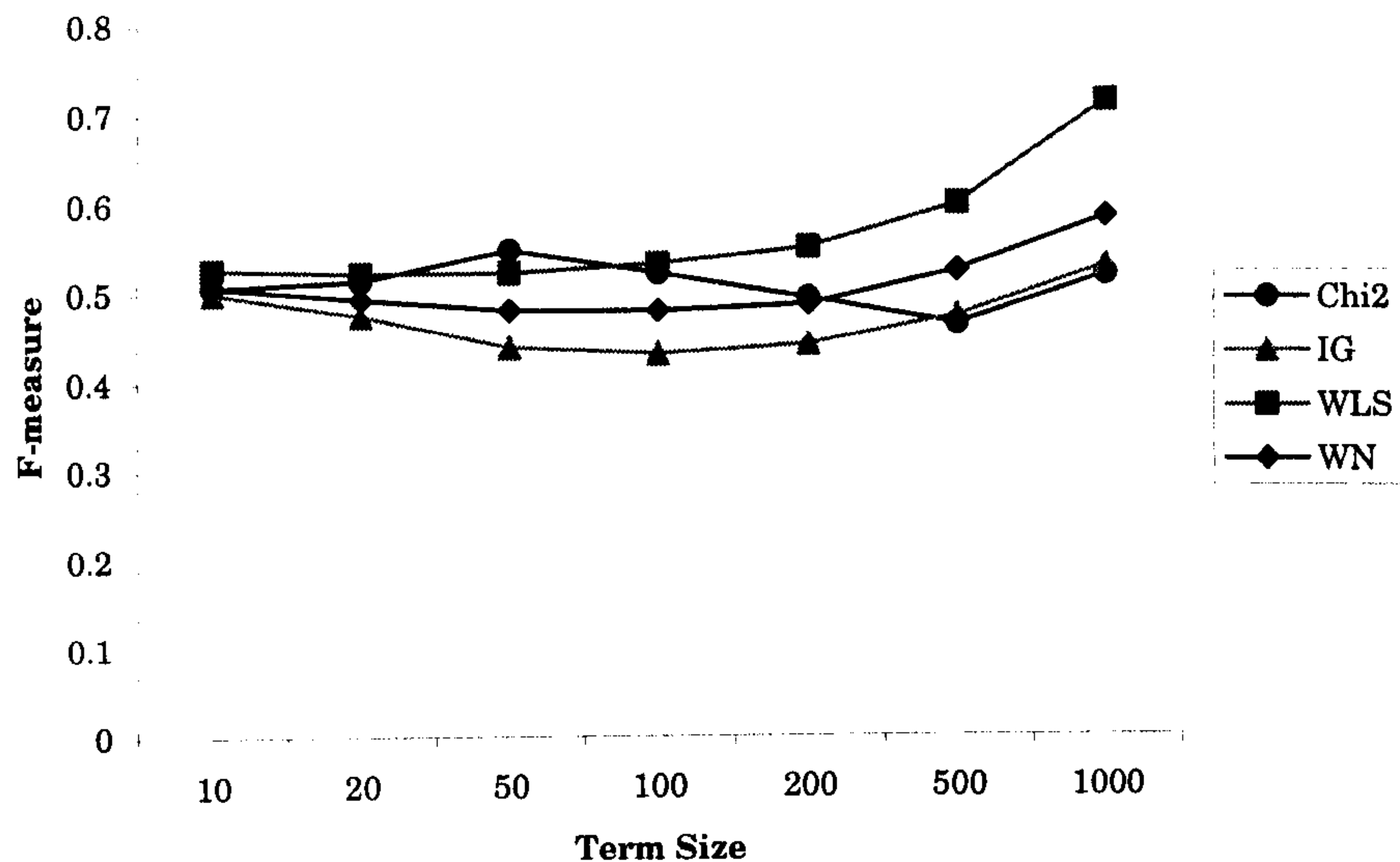


Figure 5.4: Micro-averaged f-measure for the Reuters-21578 dataset

As highlighted by the results of the Reuters-21578 dataset, the WLS approach for feature selection works better than the WN approach and the other statistical approaches. The WLS worked well across all the categories, taking into consideration, categories with small number of documents.

As the document distribution in the Reuters-21578 collection is uneven and there are a number of categories containing a small number of documents, further experiments were carried out on the top ten most populated categories of the Reuters-21578 dataset. This was also aimed at exploring the actual reasons why the WLS approach worked better. The macro-averaged and micro-averaged f-measure for Reuters-21578 top ten categories are shown in Table 5.7 and 5.8 and plotted in Figure 5.5 and 5.6 respectively. The results printed in bold shows the highest results for each of the term size.

Term Size	Chi2	IG	WN	WLS
10	0.356	0.604	0.588	0.564
20	0.470	0.628	0.609	0.607
50	0.567	0.631	0.629	0.637
100	0.616	0.627	0.647	0.662
200	0.624	0.644	0.658	0.668
500	0.633	0.653	0.666	0.663
1000	0.644	0.662	0.669	0.694

Table 5.7: Macro-averaged f-measure for the Reuters-21578 top ten categories

Term Size	Chi2	IG	WN	WLS
10	0.707	0.737	0.704	0.706
20	0.756	0.751	0.740	0.750
50	0.746	0.766	0.765	0.773
100	0.753	0.762	0.778	0.792
200	0.761	0.779	0.791	0.799
500	0.774	0.786	0.797	0.798
1000	0.782	0.796	0.801	0.813

Table 5.8: Micro-averaged f-measure for the Reuters-21578 top ten categories

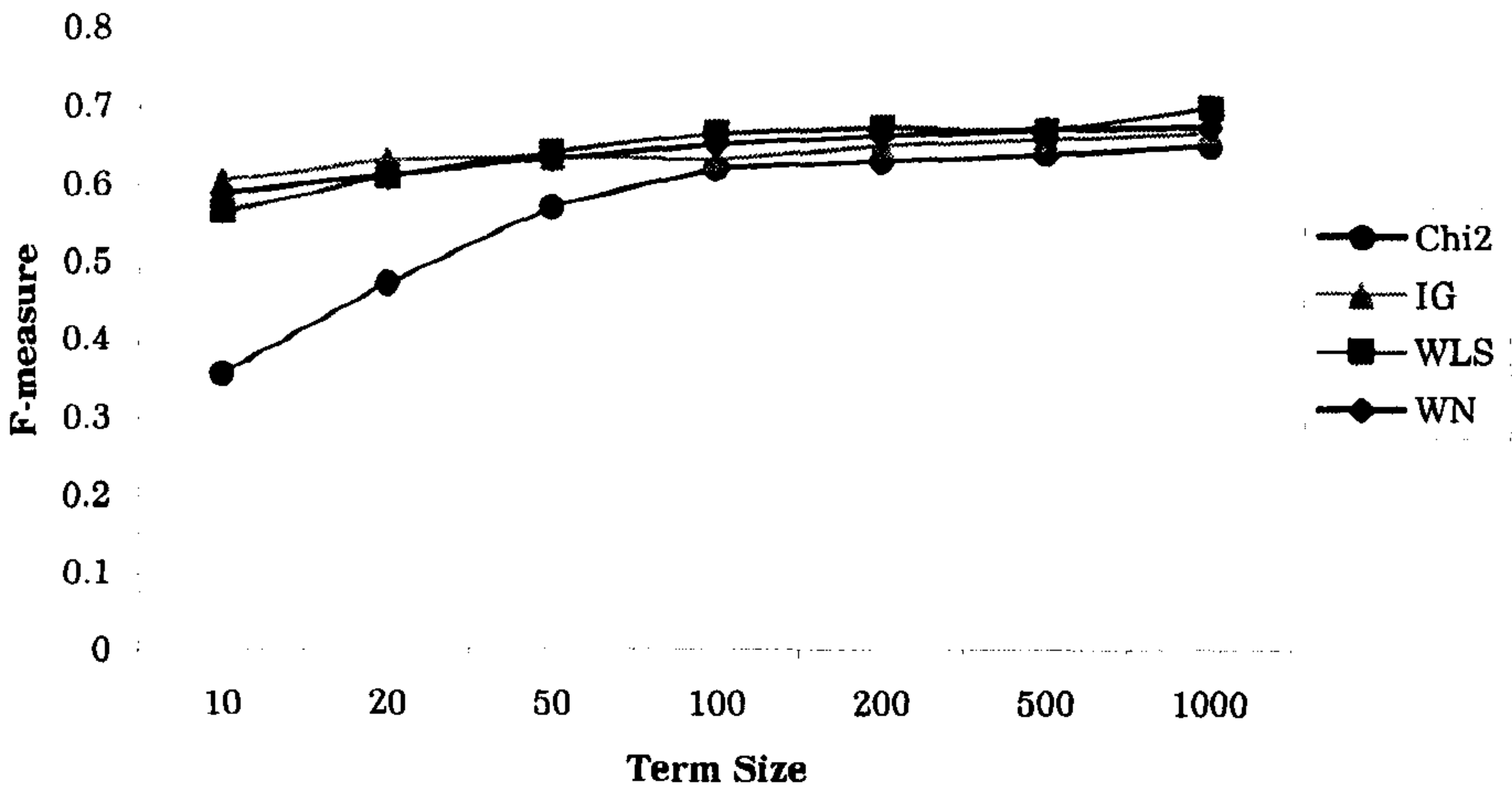


Figure 5.5: Macro-averaged f-measure for the Reuters-21578 top ten categories

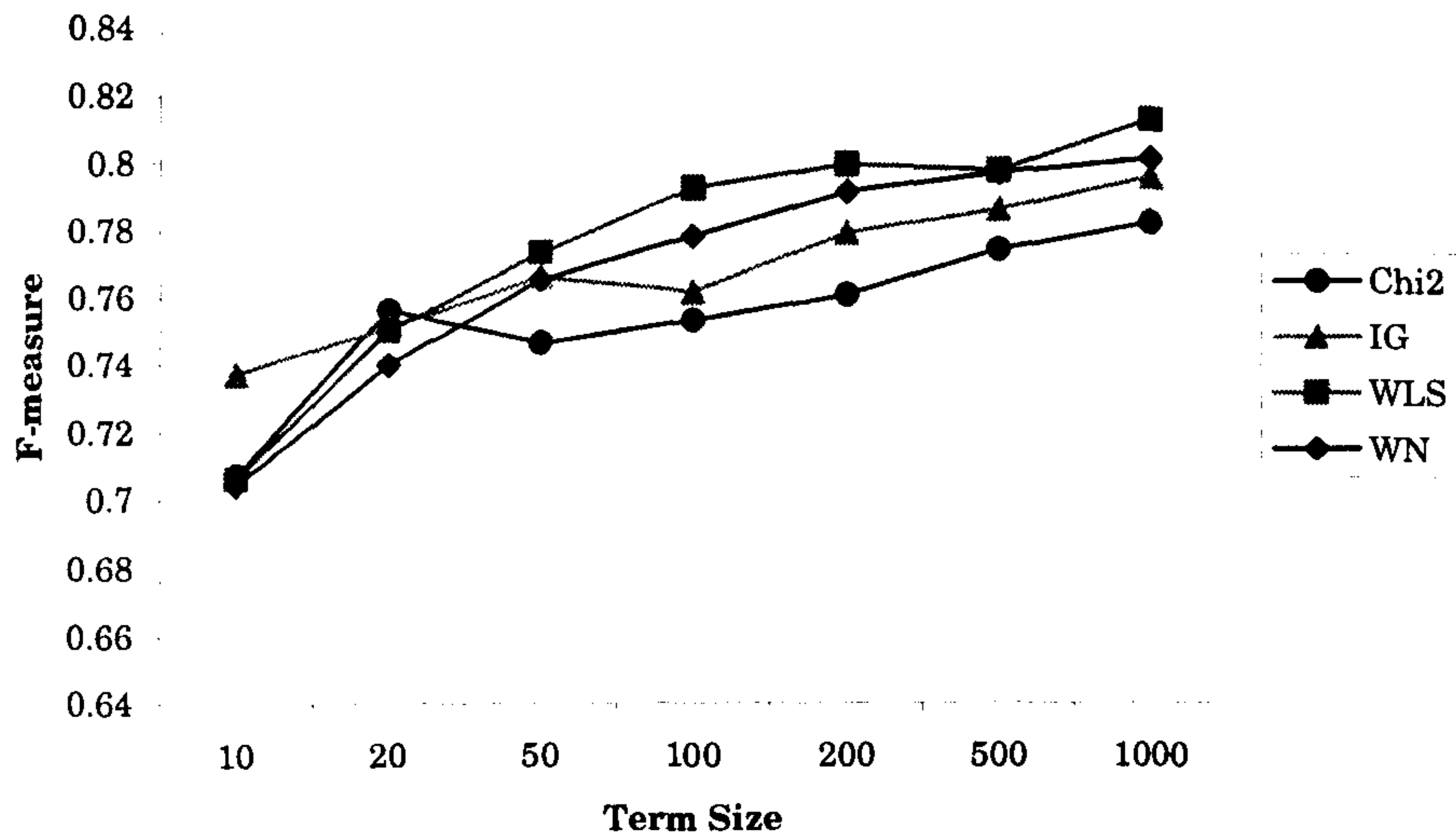


Figure 5.6: Micro-averaged f-measure for the Reuters-21578 top ten categories

From Figure 5.5 and 5.6, we can see that the WLS approach generally performs better than the WN approach and the statistical feature selection methods. The WN approach by itself performed comparatively to statistical approaches. However, Chi2 and IG registered better results at feature size 10 and 20, as shown in Figure 5.6. We investigated the reason for this by looking at the performance of Chi2, IG and WLS for each category at term size 10 and 20. Table 5.9 shows the f-measure for each category for Chi2, IG and WLS at term size 10 and 20. The results printed in bold show the highest results for each of the term size.

Category/Term Size		Chi2	IG	WLS
Acq	10	0.760	0.790	0.581
	20	0.847	0.859	0.731
Corn	10	0.034	0.400	0.385
	20	0.094	0.346	0.408
Crude	10	0.342	0.723	0.718
	20	0.620	0.716	0.715
Earn	10	0.910	0.908	0.938
	20	0.915	0.922	0.942
Grain	10	0.013	0.581	0.529

	20	0.603	0.620	0.556
Interest	10	0.410	0.404	0.402
	20	0.444	0.561	0.410
Money-fx	10	0.306	0.579	0.517
	20	0.474	0.608	0.595
Ship	10	0.022	0.649	0.442
	20	0.063	0.661	0.590
Trade	10	0.646	0.576	0.688
	20	0.562	0.593	0.702
Wheat	10	0.120	0.429	0.439
	20	0.075	0.396	0.421

Table 5.9: F-measure for each category for Chi2, IG and WLS for the Reuters-21578 top ten categories

From Table 5.9, it is noted that IG generally performs better in categories “Acq”, “Crude”, “Grain”, “Money-fx” and “Ship” while WLS performs better in categories “Earn”, “Trade” and “Wheat”. IG only performs better in category “Interest” at 20 terms while Chi2 did better for 10 terms.

5.4.3.2 Qualitative Analysis

We will take a look at the top 20 terms for category “Acq”, where IG performs better, category “Earn”, where WLS performs better and also category “Interest”, where Chi2 perform best. Listed in Table 5.10, 5.11 and 5.12 respectively, are the top 20 terms for category “Acq”, “Earn” and “Interest” for Chi2, IG and WLS. For all the other categories, the performance of WLS, Chi2 and IG in each category for term size 100 is shown in Figure 5.7. Note that at term size 100, WLS performs better in six categories compared to only three categories at term sizes 10 and 20.

Feature selection scheme	Top 20 terms for category “Acq”
Chi2	shares, offer, lt, stake, merger, cts, acquisition, company, inc, acquire, net, loss, corp, usair, common, mln, unit, shr, stock, sell
IG	cts, shares, net, loss, lt, shr, offer, stake, company, merger, tonnes, wheat, acquisition, trade, inc, mln, profit, qtr, corp, acquire
WLS	title, company, shares, pct, corp, offer, share, stock, stake, acquisition, merger, common, unit, buy, agreement, board, shareholders, sell, american, investment

Table 5.10: Top 20 terms for category “Acq” for Chi2, IG and WLS

Feature selection scheme	Top 20 terms for category “Earn”
Chi2	vs, cts, net, loss, mln, shr, profit, revs, qtr, oper, dlrs, note, dividend, th, shrs, avg, div, earnings, prior, lt
IG	vs, cts, net, loss, mln, shr, profit, revs, qtr, dlrs, oper, note, dividend, trade, th, shrs, avg, tonnes, div, said
WLS	vs, text, title, cts, net, loss, profit, revs, share, company, pct, note, corp, quarter, dividend, record, th, earning, stock, tax

Table 5.11: Top 20 terms for category “Earn” for Chi2, IG and WLS

Feature selection scheme	Top 20 terms for category “Interest”
Chi2	rate, rates, bank, money, fed, prime, pct, banks, stg, lending, base, funds, market, discounts, cut, bills, sterling, bundesbank, england, band
IG	rate, rates, bank, pct, money, fed, prime, market, stg, banks, loss, net, lt, cut, company, funds, dlrs, base, lending, discount
WLS	pct, rate, market, fed, cut, prime, government, week, funds, federal, term, base, dealers, central, month, bills, reserve, february, growth, dollar

Table 5.12: Top 20 terms for category “Interest” for Chi2, IG and WLS

In Table 5.10, we observe that both Chi2 and IG have one term that strongly represents the “Acq” category, which is “acquire”. With the WLS approach however, “acquire” was not chosen because it is a verb and our approach only considers nouns. Together with other terms chosen by Chi2 and IG, their performance are better than WLS, as abbreviations such as “lt”, “cts”, “mln” and “shr” are not chosen by WLS although they are statistically strong terms.

For category “Earn, out of the 11 unique terms, Chi2 has 9 abbreviations and IG has 8 abbreviations, which are not proper English abbreviations. The only word that is reflective of the category “Earn” is “earnings” for Chi2 and “trade” for IG. In the case of WLS, it has words like “share”, “company”, “earning”, “stock” and “tax” that are reflective of the category “Earn”. Therefore, in this category, WLS is able to perform better.

For category “Interest” as seen in Table 5.12, Chi2 has terms like “sterling”, “bundesbank” and “England”. IG has 5 terms similar to Chi2 and also two abbreviations “lt” and “stg” that are not chosen by WLS. This explains why they have better performance when compared to WLS, as WLS does not contain the abbreviations.

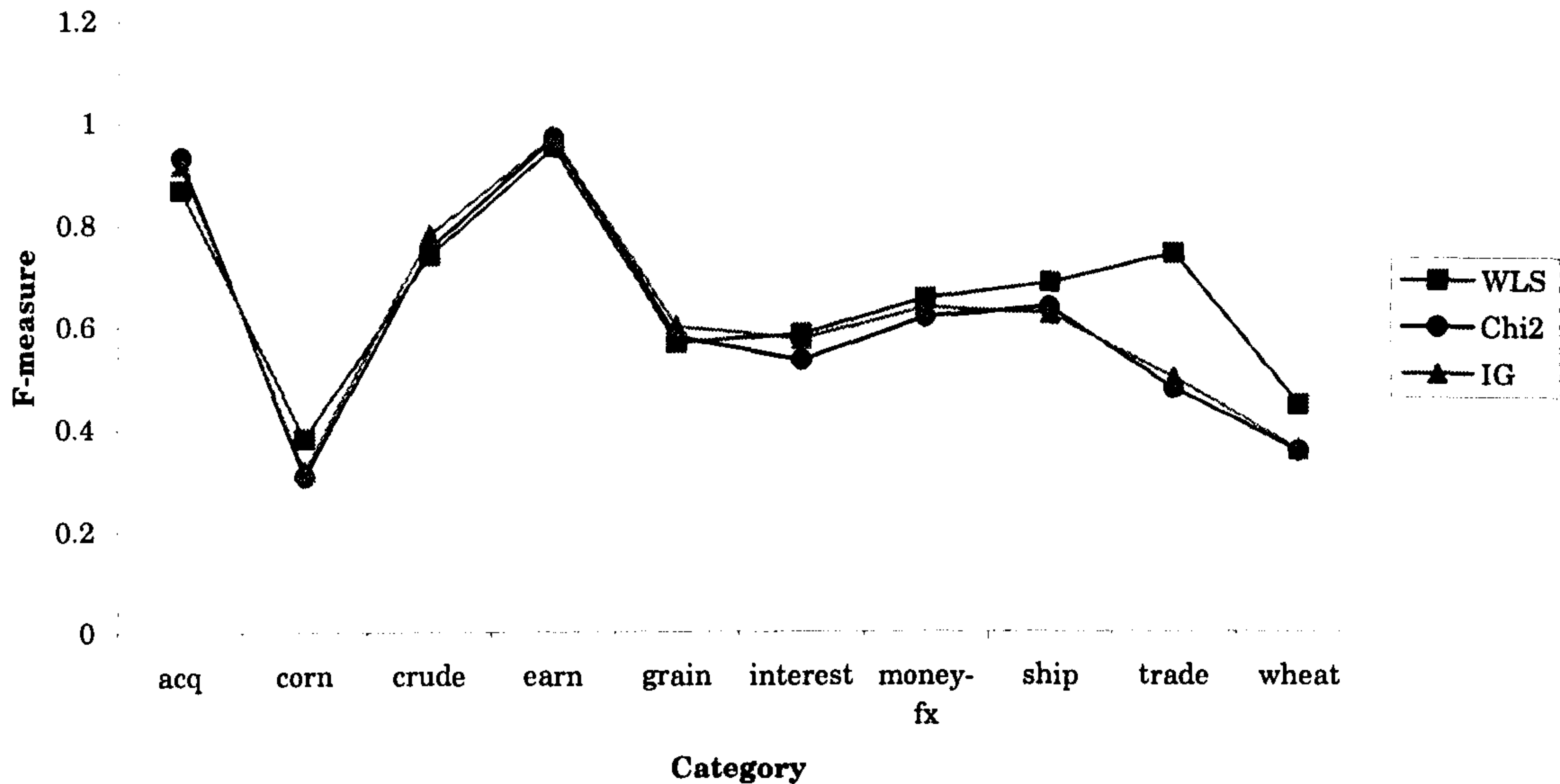


Figure 5.7: Comparison of the performance of WLS, Chi2 and IG in different categories at term size 100 for the Reuters-21578 top ten categories

5.4.3.3 Discussion

From the qualitative analysis, we can see that WLS has a set of terms that better describe category “Earn”. Chi2 outperformed WLS for category “Interest” because it chose unique features for its representation of the category. Words that are chosen by Chi2 are mostly contingent words that are found in that particular category. Thus, it can categorize the documents from that category very well. We call this phenomenon “semantic overfitting”, as the chosen features fits very well with one particular category. However, if the chosen features were used for a same category in a different dataset, then the results would not be as appealing. The advantage of the WLS approach is that, unlike Chi2, it captures a subset of the constitutive keywords that reflects the general content of the category and that knowledge can be applied to another dataset with a

similar category. Chi-square is incapable of doing this as it relies dominantly on unique terms of a category in a dataset.

Generally, the WLS approach performs better from 50 terms onwards because it is capable of choosing constitutive terms that are more reflective of a category. For a small number of features, the WLS approach is not able to capture adequate representative categorical features. This approach stabilizes as the number of features exceeds 50 features, as a larger number of features are needed to overcome the effects of overfitting that arises due to selecting statistically unique terms with a lesser semantic significance. With fewer terms at 10 and 20 terms, IG has a lot of terms that has strong statistical values. Although those terms are not in general, representative of the category concerned, it is able to perform well in machine learning because of its statistical patterns.

In the individual category evaluation, we found that both category “Acq” and “Earn” have the highest number of training examples, thus producing the two highest values of f-measure. On the other hand, categories with a low number of training examples perform less effectively. This is proven in the correlation test that was carried out to see whether the number of training examples actually affects the performance of the classifier. We observed a strong correlation with a positive value of 0.8136 between the number of training examples and the corresponding f-measure values (refer to Appendix F).

Furthermore, categories with more synonyms found in WordNet performed better. The correlation for the number of synonyms found and the performance of the classifier has

the positive value of 0.8789 (refer to Appendix F). Thus, both categories “Acq” and “Earn” again outperformed the other categories with category “Corn” having the lowest performance. We observed the same positive results for the number of nouns found in WordNet and the corresponding f-measure values. The correlation for the number of nouns found and the performance of the classifier has the positive value of 0.8643 (refer to Appendix F).

With the WLS approach to feature selection, insignificant words and noise can be filtered. These insignificant words consist of non-English words, wrongly-spelt words, insignificant abbreviations and names. Terms like “lt”, “govodi” and “pik” are actually meaningless in representing a category. If human experts were asked to choose a set of words to represent a category, they would not have chosen those words that are mentioned. However, statistical feature selection approaches like Chi2 and IG, will include such words if the statistical data show that the words are statistically significant, which means that the words have high statistical values. The Chi2 and IG values are calculated based on the formula shown in Table 2.1. These statistical approaches do not take into consideration whether a term is misspelt or is reflective of a category. By using the WLS approach in feature selection, we can actually tackle this problem by filtering these terms and at the same time, make use of the available synonym relationship and word senses in WordNet to identify semantic features in a category. The WLS approach is able to choose a set of terms that are more reflective of a category’s content. This is in line with inducing a classifier that will act more like a human expert rather than having a classifier rely only on statistical findings. This also highlights the potential of the WLS approach to be used for areas such as spam filtering.

5.4.4 Experiment 3: Comparison between Chi2, IG and Combined Approaches

The WLS approach generally performs better than the WN approach as seen in Experiment 1 and 2. Therefore, only the WLS approach is employed in combining WordNet based approach with statistical approaches. This is to test whether the semantic component does indeed make a difference in feature selection. We focus on whether the combination of WLS and a statistical approach does actually improve on the results of the statistical approach itself. Therefore, experiments were carried out on the combination of WLS and Chi2 as well as WLS and IG. We denote the combination of WLS and Chi2 as WLS-Chi2 and the combination of WLS and IG as WLS-IG.

5.4.4.1 F-measure Analysis

Table 5.13 shows the micro-averaged f-measure of Chi2 and WLS-Chi2 as well as IG and WLS-IG for the Reuters-21578 top ten categories. The results in bold show the higher of the two values being compared.

Term Size	Chi2	WLS-Chi2	IG	WLS-IG
10	0.707	0.705	0.737	0.724
20	0.756	0.782	0.751	0.760
50	0.746	0.766	0.766	0.767
100	0.753	0.769	0.762	0.778
200	0.761	0.780	0.779	0.786
500	0.774	0.782	0.786	0.789
1000	0.782	0.817	0.796	0.817

Table 5.13: Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 as well as IG and WLS-IG for the Reuters-21578 top ten categories

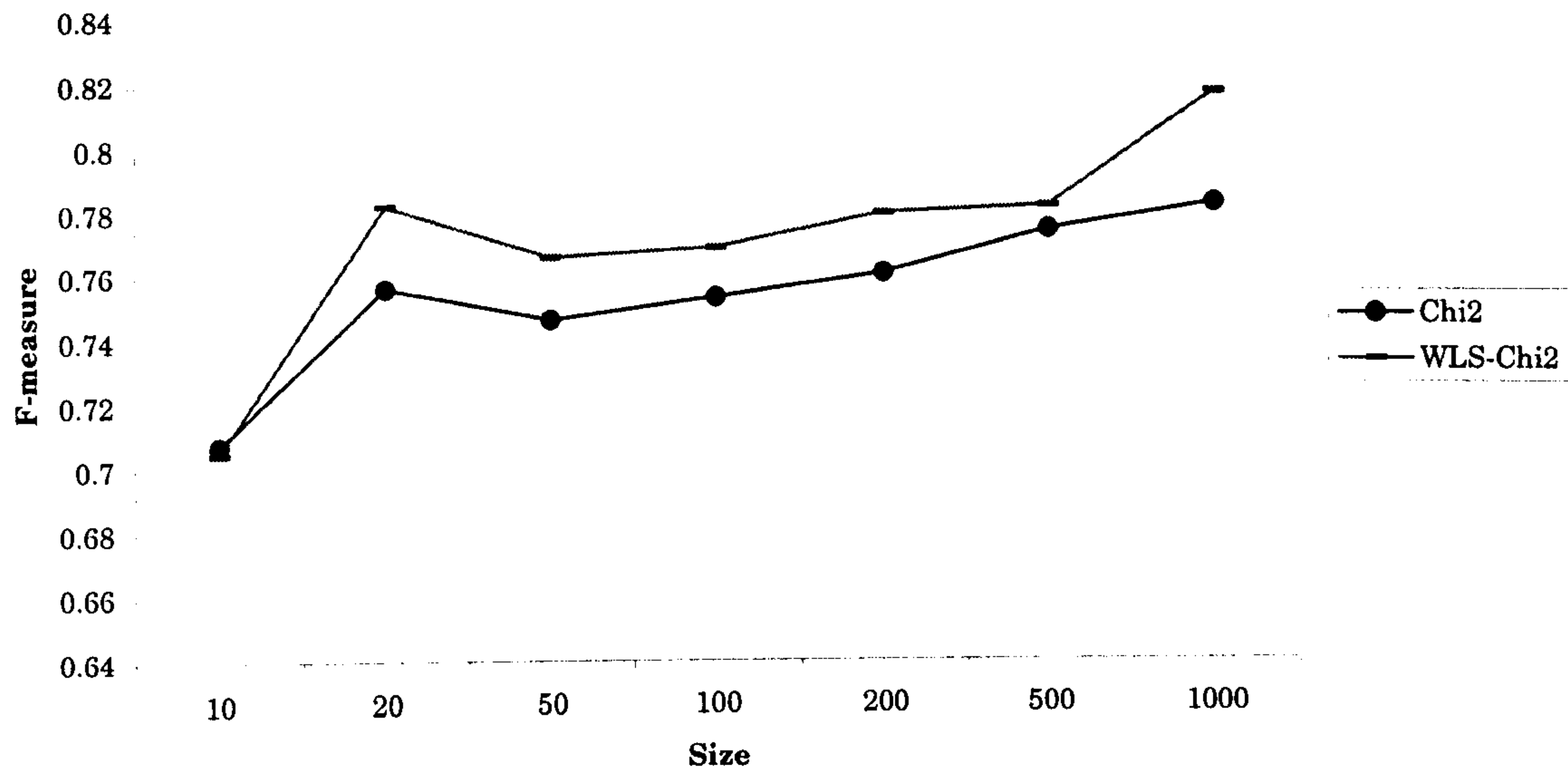


Figure 5.8: Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 for the Reuters-21578 top ten categories

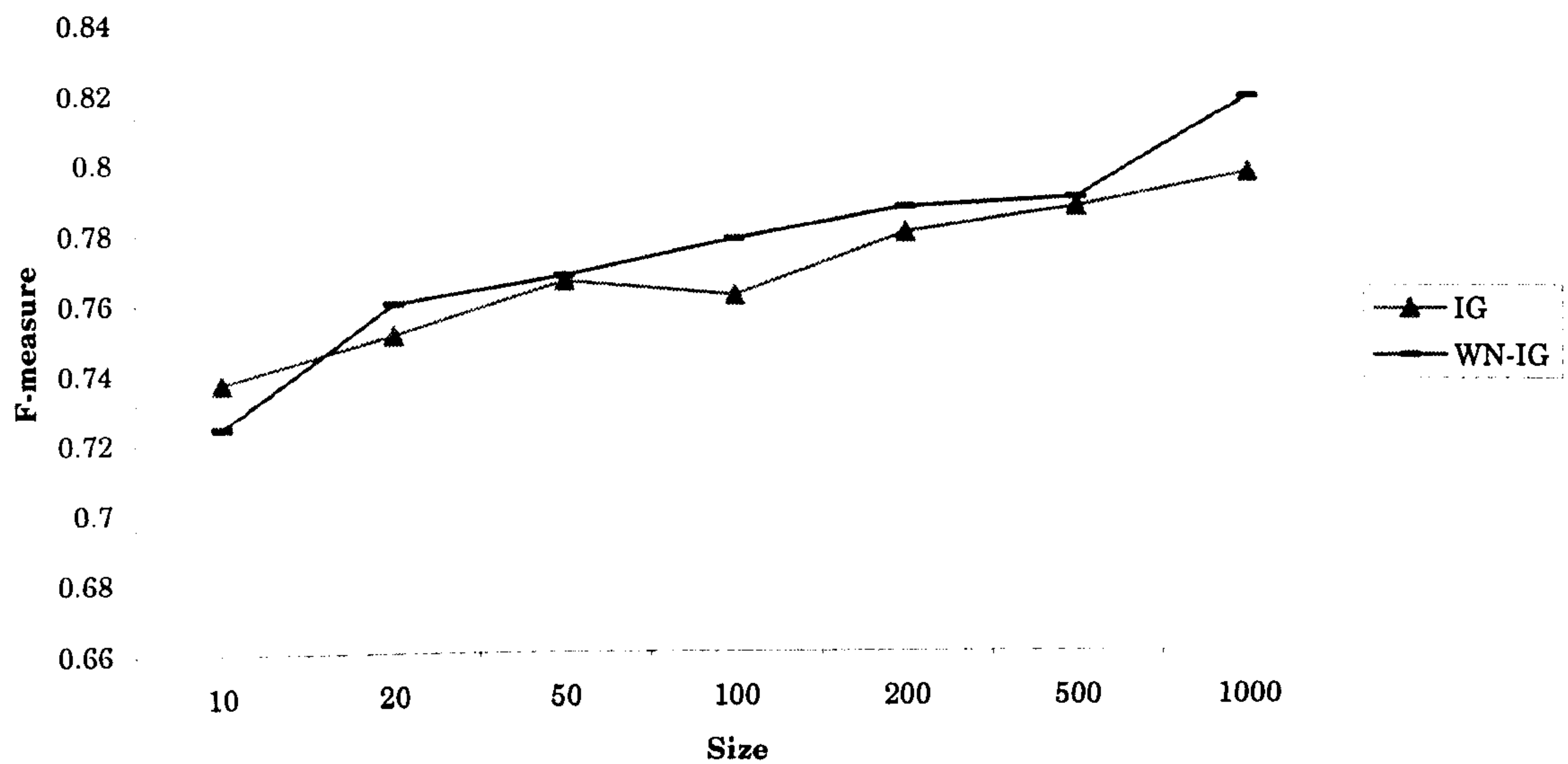


Figure 5.9: Comparison of micro-averaged f-measure between IG and WLS-IG for the Reuters-21578 top ten categories

From Figure 5.8 and 5.9, it is noted that both WLS-Chi2 and WLS-IG is able to perform better than the statistical approach itself, with the exception at term size 10. At all

other term size thresholds, there is a slight increase in f-measure when the WLS approach is combined with a statistical approach.

Table 5.14 shows the f-measure for each category for Chi2 and WLS-Chi2 as well as IG and WLS-IG at term size 10. The results in bold shows the higher of the two values being compared. It is noted that Chi2 performed better in five categories while WLS-Chi2 only performed better in four. The result for category “Corn” is the same for both approaches. Meanwhile, IG performed better in four categories while WLS-IG performed better in five categories. Again, category “Corn” gives the same results for both approaches.

Category/Term Size		Chi2	WLS-Chi2	IG	WLS-IG
Acq	10	0.760	0.769	0.790	0.689
Corn	10	0.034	0.034	0.400	0.400
Crude	10	0.342	0.478	0.723	0.715
Earn	10	0.910	0.908	0.908	0.914
Grain	10	0.013	0.039	0.581	0.599
Interest	10	0.410	0.366	0.404	0.491
Money-fx	10	0.306	0.343	0.579	0.545
Ship	10	0.022	0.000	0.649	0.571
Trade	10	0.646	0.500	0.576	0.627
Wheat	10	0.120	0.119	0.429	0.461

Table 5.14: F-measure for each category for Chi2 and WLS-Chi2 as well as IG and WLS-IG at term size 10 for the Reuters-21578 top ten categories

5.4.4.2 Discussion

From the results, we can conclude that the statistical feature selection approaches only did marginally better than the combination methods for term size 10. The WLS approach needs a larger number of features to capture adequate representative

categorical features. This is the reason why it did not improve on the results of the statistical approaches at term size 10, while across all other term sizes, the combination method is able to produce some improvements over the statistical methods. This is also a significant contribution to prove that the choice of terms is now better representative of the documents.

5.5 Summary

The results of the experiments that were carried out using the WLS approach and the WN approach shows that the WLS approach is significant and promising as it is able to reduce the number of features and at the same time, gives better overall results. When the WLS approach is compared with the statistical approaches, namely Chi2 and IG, it also generally performed better at term size 50 onwards. The combination of the WLS approach with Chi2 and with IG, also gave positive results where the combination was able to improve on the results of pure statistical approaches alone.

From the qualitative analysis done by looking into features chosen for each category, we were able to see why certain categories did better than the others for each feature selection scheme. The WN and WLS approaches were able to identify certain constitutive features that were representative of a category's content as well as contingent features that were not significant. The WLS approach was able to capture a subset of the constitutive keywords that reflects the general content of the category. Therefore, it is able to select features that are reflective of a category instead of insignificant terms that Chi2 and IG tend to choose because of their high statistical values.

Our approach relies on the availability of a large number of terms found in the document that also exist in the lexical database. This was found to be true in the experiments conducted on the Reuters-21578 dataset. Therefore, our approach may not work as well in technical areas where, WordNet lacks the technical terms. Domain specific dictionaries are needed to provide the lexical and semantics information of technical terms.

To summarize, we have shown that WordNet can be used to enhance feature selection. The WN approach alone can be considered for feature selection, as it is comparable to the WLS approach, although the feature space is larger. The WLS approach needs a minimum term size to work well. It was also found that the combined approach performed as well as the WLS approach. However, the introduction of lexical semantics terms has been shown to enhance the performance of statistical approaches.

The last chapter, which is Chapter 6, will summarize the contribution of this research, as well as the limitations of the WLS approach. The possible future works on this approach will also be discussed.

Chapter 6 Summary and Future Works

6.1 Introduction

This chapter will highlight the contribution of this research, as well as the limitations of the proposed approach for feature selection. The future works to be carried out will also be discussed.

6.2 Research Contributions

In this research, three WordNet-based approaches for feature selection were explored. These approaches are fully automated with no human intervention. We present the WordNet nouns (WN) approach that selects all nouns that exist in the WordNet dictionary in a category to be used as features for category representation. We also extended the use of WordNet in feature selection through the WordNet lexical semantics (WLS) approach where synonyms that co-occur in a category are chosen as features to represent a category. The last approach presented was the combined approach, where the more effective lexical semantic approach was combined with statistical feature selection methods in an effort to enhance the performance of the statistical feature selection methods.

From the exploration of the three approaches to feature selection, we have gained insights into feature selection. We have discovered the strength of the WordNet-based approaches over the statistical approaches. By doing a qualitative analysis of the terms chosen for each scheme, we found that statistical approaches that chose terms purely

based on statistics of the terms are susceptible to choosing contingent terms, where non-standard abbreviations, wrongly spelt words and non-meaningful words are chosen to represent a category. These drawbacks are tackled using the WordNet-based approaches where these terms are omitted from the feature space. WordNet contains a large number of terms found in the Reuters-21578 dataset. The WN approach by itself is comparable in performance with statistical feature selection methods when used in a single pass. Our aim of choosing a set of constitutive features that are representative of a category is achieved using the lexical semantics approach. The WLS approach is able to capture common themes based on a wide range of categories in the Reuters-21578 dataset. This set of features mirror the set of features that a human expert will choose. Apart from that, we have also achieved a dimensionality reduction of over 65% from the full set of unique terms in each category by using the WLS approach. The WLS approach also reduced more than 40% of the feature space of the WN approach and is able to perform better than the WN approach. We also applied categorical sense determination to identify dominant senses in a category. This method can also be exploited further in the proposed future works.

Another contribution to be noted is the significant improvements based on the combination of the WLS approach and the statistical feature selection approaches. Here, the combination of Chi2 and the WLS approach, as well as IG and the WLS approach, have produced improvements in the effectiveness of the categorization tasks. Both combination approaches have performed better than the statistical approaches alone.

Therefore, we have achieved all the objectives of this research. Through the use of WordNet as a lexical database, we are able to obtain a better set of features in

representing a category as seen in the qualitative analysis of terms. We are also able to achieve a better dimensionality reduction without compromising the effectiveness of categorization. We have also benchmarked the WordNet-based approaches with statistical approaches in feature selection and obtain a better performance through the WLS approach. Combined approaches also registered positive results where the performance of statistical approaches were improved.

6.3 Limitations

Even though the WLS approach has produced positive results in the experiments that are carried out, during the analysis of the results, we have found some limitations to this approach. We found that, although WordNet has a large database of words, these words are general words and are not domain specific. As such, when a particular category has many domain specific words, the WordNet-based approach will not be able to identify synonyms from the category. Therefore, domain specific dictionaries are needed to complement WordNet in dealing with domain specific words. Apart from that, there is also the case where a word that can be found in WordNet does not have a synonym in WordNet. Thus, that particular word will not be chosen as a feature based on our lexical semantics approach. Words with no synonyms are not considered, as our hypothesis states that synonymous words give stronger discriminative power in representing a category as features. Therefore, this approach will work well only with words that have synonyms listed in WordNet. However, we will explore the effects of adding words without synonyms in future works. Though most of these words are not reflective of the particular category, we might still find a small number of words that actually relate to

the category. Thus, the WLS approach needs to be extended to be able to handle words without synonyms. This is highlighted in Section 6.4.1.

We chose WordNet despite these limitations, as WordNet is an available lexical database that has rich lexical relations and semantic relations. Based on the research carried out, we find that WordNet has a good potential as a lexical database especially when further extension to this research is done in the future. We will highlight the proposals to extend the current WordNet-based approaches in the following sections.

6.4 Future Works

As future works to this research, we have identified that WordNet's hypernyms can be used for concepts identification to identify concepts to be used as features for text categorization. There are a number of issues to be addressed in employing hypernyms. Determining the level of hypernyms to work on and derivation of word sense disambiguation for different levels of hypernyms can be major research works by themselves. The following section will explain the use of a single level hypernyms for concepts identification. Apart from that, another proposal is the synonymization of terms. An enhanced framework for automated text categorization using hypernyms and synonymization is also proposed.

6.4.1 Concepts Identification Based on WordNet's Hypernyms

WordNet organized its synsets in a hierarchical tree where there exists the hyponym/hypernym or subset/superset relation. At the top of the WordNet tree are more

general concepts while, traversing down the tree, more specific concepts can be found. Besides that, WordNet also has the meronym/holonym relation, also known as the “part-whole” relation where more components that are related to a word can be identified.

Based on WordNet nouns, there is a hierarchy of synsets that shows how the synsets are related with one another. Each group of synsets that are closely related can be pictured as a concept. Each word in the feature space can be looked up in WordNet to see its relationship with the other words. Words that relate closely to one another can be grouped into a similar concept by simply belonging to the same parent synset or hypernym in WordNet. Therefore, a document can be represented by having a few main concepts and the concepts can be used as features in categorizing the documents.

There are 25 unique beginners for WordNet nouns. A unique beginner is a synset with no superordinate. They are as listed in Figure 6.1 below.

{act, action, activity}	{feeling, emotion}	{plant, flora}
{animal, fauna}	{food}	{possession}
{artifact}	{group, collection}	{process}
{attribute, property}	{location, place}	{quantity, amount}
{body, corpus}	{motive}	{relation}
{cognition, knowledge}	{natural object}	{shape}
{communication}	{natural phenomenon}	{state, condition}
{event, happening}	{person, human being}	{substance}
		{time}

Figure 6.1: List of 25 unique beginners for WordNet nouns (extracted from Miller et al., 1990)

Figure 6.2 shows seven unique beginners for WordNet nouns being classified under {living thing, organism} and {non-living thing, object}. These two synsets are in turn the hyponyms or subset of {thing, entity}.

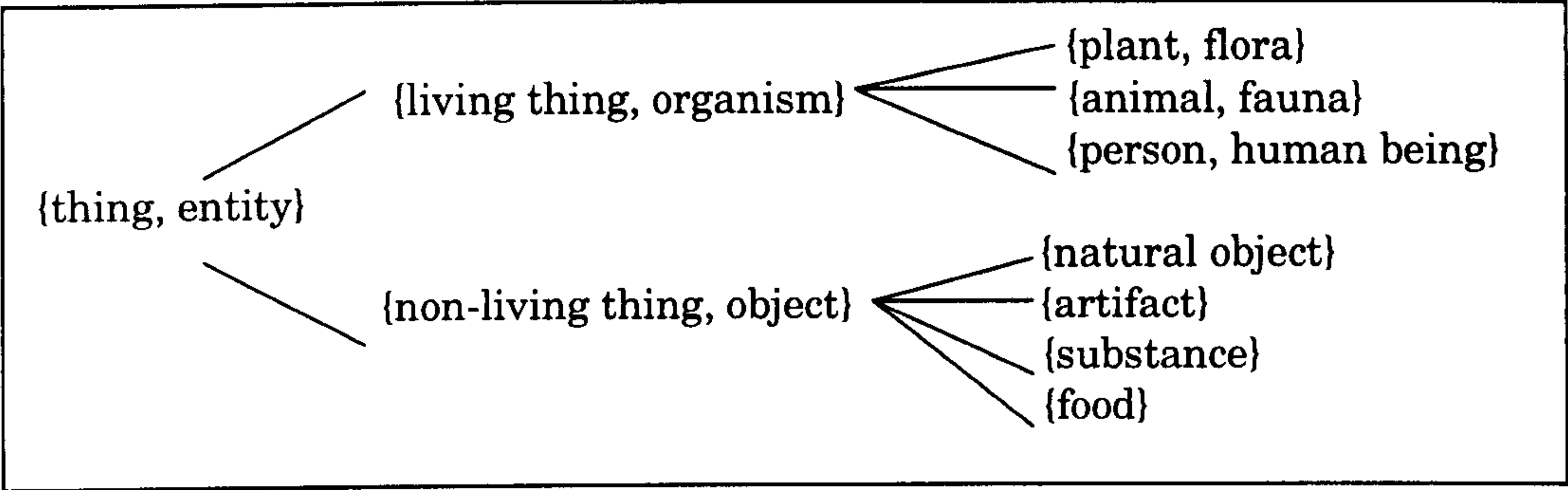


Figure 6.2 Diagrammatic representation of hyponymic relations among seven unique beginners denoting different kinds of tangible things (extracted from Miller et al., 1990)

We will illustrate how concept identification can be done based on WordNet’s hypernyms in the example that follows. Let us say that the category “AGRICULTURE” has ten features representing it as listed below. These features are looked up in WordNet to obtain their synsets. Categorical sense determination can also be applied here to identify the dominant senses in each category.

AGRICULTURE: corn, crop, export, grain, oat, maize, soybean, vegetable, wheat, tonne

Corn -> {corn, maize, Indian corn, Zea mays}

Crop -> {crop, harvest}

Export -> {export}

Grain -> {grain}

Oat -> {oat}

Maize -> {corn, maize, Indian corn, Zea mays}

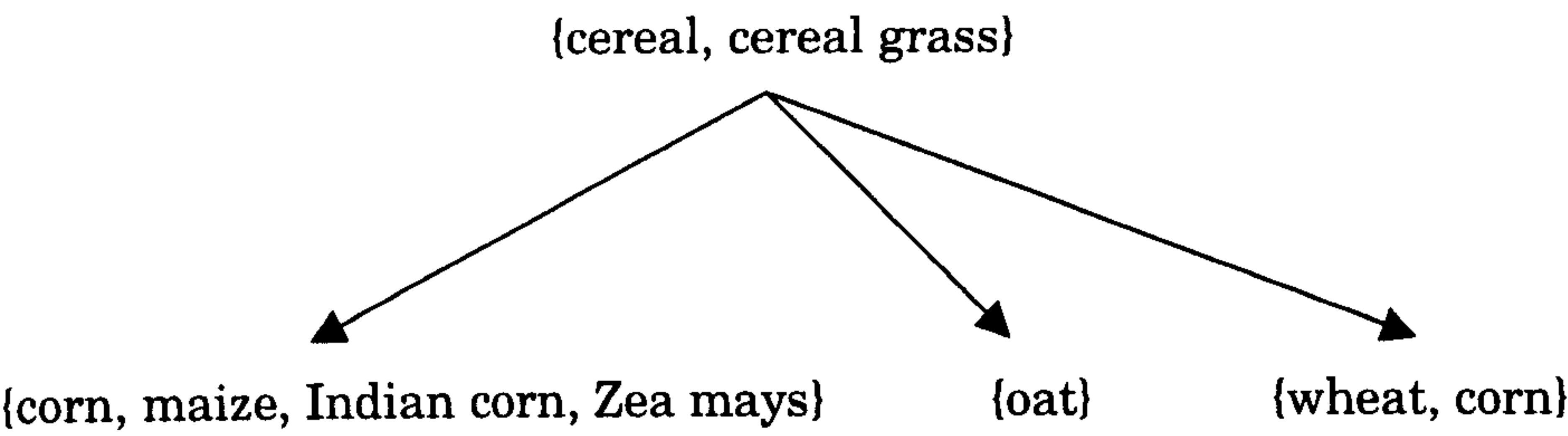
Soybean -> {soy, soybean, soyabean}

Vegetable -> {vegetable, veggie}

Wheat -> {wheat, corn}

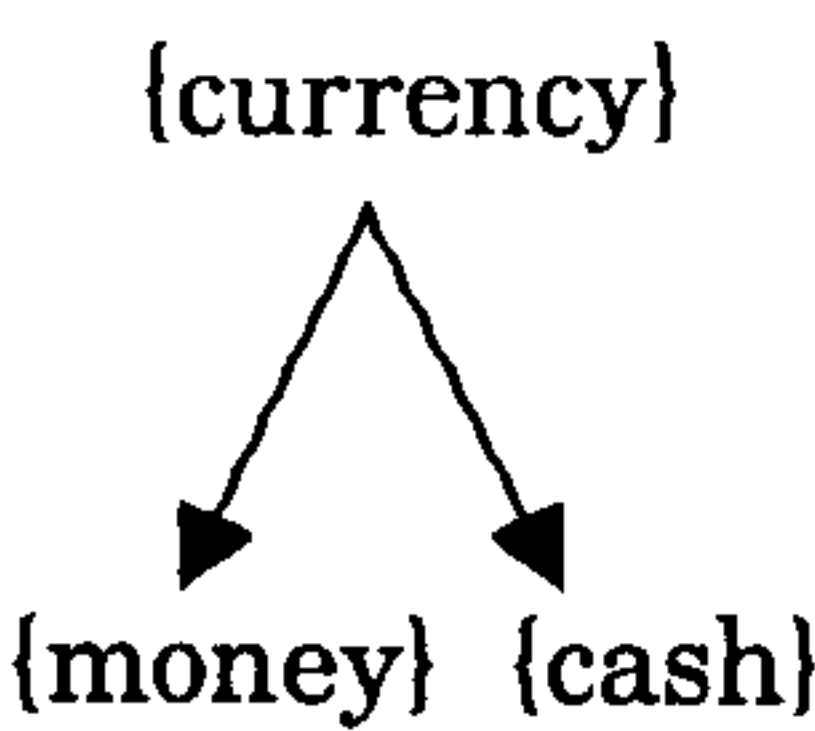
Tonne -> {metric ton, MT, tonne, t}

It is discovered from WordNet dictionary files that {corn, maize, Indian corn, Zea mays}, {oat} and {wheat, corn} belong to the same parent synset {cereal, cereal grass} as shown below.



Hence, “corn”, “maize”, “oat” and “wheat” can be grouped under the concept “cereal”, which is the superclass or parent class of those terms. Therefore, the concept “cereal” can be chosen as a feature to represent the category.

This method can be used to handle the limitation where words without synonyms are not chosen by the present lexical semantics approach as mentioned in Section 6.3. The following is an example to illustrate this point.



Based on the current lexical semantics approach, both “money” and “cash” were not chosen as features, as they do not have any synonyms in their synsets. One way to overcome this is to look at the hypernyms relationship, where both these terms have the same parent synset {currency}. Terms that have the same parent synset are considered to be under the same concept. Therefore, we have identified that both “money” and “cash” are closely related under the same concept and thus, are meaningful enough to be features representative of a category. Similarly, term clustering techniques can be used to identify features that are closely related. Term clustering has been discussed in Section 2.5.1.

6.4.2 Synonymization Based on WordNet’s Synsets

Synonymization is the process of replacing one synonym with another. Using the same set of terms in the example in Section 6.4.1, we obtain the synset for each term from WordNet and apply the categorical sense determination to identify the dominant senses in each category.

Corn -> {corn, maize, Indian corn, Zea mays}

Crop -> {crop, harvest}

Export -> {export}

Grain -> {grain}

Oat -> {oat}

Maize -> {corn, maize, Indian corn, Zea mays}

Soybean -> {soy, soybean, soyabean}

Vegetable -> {vegetable, veggie}

Wheat -> {wheat, corn}

Tonne -> {metric ton, MT, tonne, t}

It is observed that both “corn” and “maize” belong to the same synset {corn, maize, Indian corn, Zea mays}. Therefore, both the terms are said to be synonymous and the replacement of one term with the other will not change the meaning of the term. Both “corn” and “maize” will only be replaced by one common term between them. By synonymizing all terms in a category through replacing synonymous terms with a common term between them, the dimensionality of the feature space can thus be reduced. These terms can then be used as features for category representation.

6.4.3 Improved Framework for WordNet based Automated Text Categorization

It is noted that concepts identification using hypernyms and the synonymization process using synsets will not work well based on the current framework. This is due to the fact that in the current framework, a document is represented by all the terms in it after removing stop words. By introducing concepts identification and synonymization, we have to represent each document with concepts and synonyms respectively. We will therefore, propose an improved framework for automated text categorization using the WordNet-based approaches.

We propose to use only either hypernyms or synonyms that are identified in each document as representation of the document instead of using all terms in a document as done currently. A hypernyms/synonyms derivation process is applied to first identify the

hypernyms/synonyms in a document to be used for representation. Figure 6.3 shows the proposed framework.

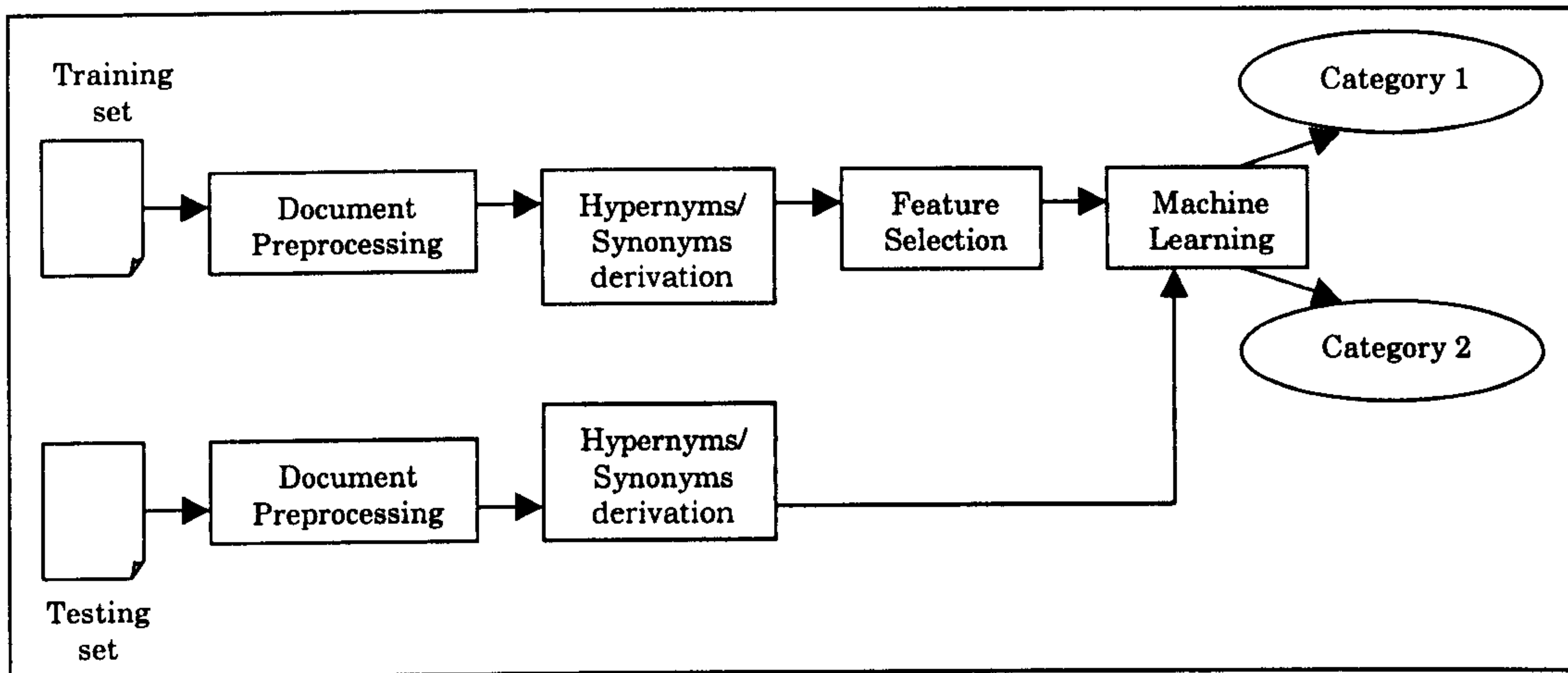


Figure 6.3: Proposed framework to improve automated text categorization using the WordNet-based approaches

6.5 Conclusion

This research has shown that, WordNet-based approaches, particularly the lexical semantics approach is a promising approach for feature selection. Not only has it performed better than Chi2 and IG, it has improved on Chi2 and IG themselves in the combination approach. A framework and proposals to overcome its limitations has been proposed in this chapter. We are confident that if the limitations are overcome, the lexical semantics approach can be further improved. The lexical semantics approach can also be further extended to cover hierarchical concepts identification using the semantic relation hypernyms as well as in the process of synonymization. The incorporation of linguistics in feature selection has addressed the problems of relying on contingent terms.

List of References

- Apte, C., Damerau, F. J. and Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12, 3: 233 – 251.
- Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text categorisation. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, Melbourne, AU, pp 96 – 103.
- Basili R., Moschitti A. and Pazienza M. T. (2000). Language sensitive text classification. In *Proceedings the Sixth International Conference of Recherche d'Information Assistee par Ordinateur (RIAIO)*, Paris.
- Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*,. pp 152 – 155.
- Bong, C. H. (2001). *A Machine Learning Framework for Automated Text Categorization*. MSc Thesis, Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak.
- Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning* 10 (1): 57 - 78.

Cohen, W. W. (1995). Learning to classify English text with ILP methods. In *Proceedings of the 5th International Workshop on Inductive Logic Programming*, pp 3 – 24.

Cohen, W.W. and Singer, Y. (1996). Context-sensitive learning methods for text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp 307 – 315.

Crestani, F. (1993). Learning strategies for an adaptive information retrieval system using neural networks. In *Proceedings of the IEEE International Conference on Neural Networks*, San Francisco, CA.

Cunningham, S. J., Littin, J. and Witten, I. H. (1997). Applications of machine learning in information retrieval. Working paper 97/6. C.S Department, University of Waikato, New Zealand.

Dash, M and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1: 131 – 156.

Debole, F. and Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, Melbourne, AS, pp 784 – 788.

- Deerwester, S., Dumais, S. T., Furnas, G. W. and Landauer. T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41: 391 – 407.
- Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29: 103 – 130.
- Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*. New York, NY: Wiley.
- Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*.
- Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19: 1, pp 61 – 74.
- Esteban, A. D., Buenaga, M., Ureña, L. and García-Vega, M. (1998). Integrating linguistic resources in an uniform way for text classification tasks. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, España.
- Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3: 1289 – 1305.

- Frakes, W. B. and Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-HALL, North Virginia.
- Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29: 131–163.
- Fuhr, N. and Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9, 3: 223 – 248.
- Gonzalo, J., Verdejo, F., Chugur, I. and Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp. 38 – 44.
- Hall M. A. and Smith, L. A. (1996). Practical feature subset selection for machine learning. In *Proceedings of the Australian Computer Science Conference* (University of Western Australia).
- Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the 17th International Conference on Machine Learning*.
- Hayes P. J. and Weinstein, S. P. (1990). CONSTRUE/TIS: a system for content-based indexing of a database of news stories. In *Proceedings of the 2nd Conference on Innovative Applications of Artificial Intelligence (IAAI-90)* (Rappaport A. and Smith, R. eds), pp 49 – 66, Menlo Park, AAAI Press.

- Holmes G. and Trigg L. (1999). A diagnostic tool for tree based supervised classification learning algorithms. In *Proceedings of the Sixth International Conference on Neural Information Processing (ICONIP)*, Perth, Western Australia, Volume II, pp 514 – 519.
- Hull, D. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the ACM SIGIR*.
- Ide, N. and Véronis, J. (1998). *Word sense disambiguation: The state of the art. Computational Linguistics*, 24: 1 – 40.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the Tenth European Conference on Machine Learning (ECML)*, pp 137 – 142.
- Korfhage, R. R. (1997). *Information Storage and Retrieval*. John Wiley & Sons.
- Lam, W., Ruiz, M. and Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*, Vol 11, No. 6.
- Langley, P., Iba, W. and Thompson, K. (1992). An analysis of bayesian classifiers. In *Proceedings of AAAI-92*, pp 223 – 228.

- Lewis, D. D. (1992a). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, pp 37 – 50.
- Lewis, D. D. (1992b). *Representation and learning in information retrieval*, PhD thesis, Graduate School of the University of Massachusetts.
- Lewis, D. D. and Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*.
- Lewis, D. D. (1996). Challenges in machine learning for text classification. In *Proceedings of the Ninth ACM Annual Conference on Computational Learning Theory*, Desenzano del Garda, Italy, p 1.
- Lewis, D. D. (1997). Reuters-21578 text categorization test collection. README file (V 1.3), available from:
<http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt>
- Masuyama, T. and Nakagawa, H. (2002). Applying cascaded feature selection to SVM text categorization. *DEXA Workshops*, pp 241 – 245.
- McCallum A. and Nigam, K. (1998). A comparison of event model for naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*.

- Mihalcea, R. and Moldovan, D. (2000). Semantic indexing using WordNet senses. In *Proceedings of the ACL Workshop on IR & NLP*, Hong Kong.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 4: 235 – 244.
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38, 11: 39 – 41.
- Mladeníć, D. (1998). Turning Yahoo into an automatic web-page classifier. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI)*, pp 473 – 474.
- Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. In *Proceedings of Fifteenth National Conference on Artificial Intelligence (AAAI)*.
- Pazzani, M., Muramatsu, J. and Billsus, D. (1996). Syskill & Webert: Identifying interesting websites. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp 54 – 61.
- Provan, G. and Singh, M. (1996). Data mining and model simplicity: A case study in diagnosis. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

- Richardson, R., Smeaton, A. and Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. In *Proceedings of AICS Conference*. Trinity College, Dublin.
- Richardson, R. and Smeaton, A. (1995). Using WordNet in a knowledge-based approach to information retrieval. *Technical Report CA-0395*, School of Computer Applications, Dublin City University, Dublin, Ireland.
- Rogati, M. and Yang, Y. (2002). High-performing feature selection for text classification. *CIKM'02*, pp 659 – 661.
- Salton, G. (editor) (1971). *The SMART retrieval system*. Prentice Hall, Englewood Cliffs.
- Schütze, H., Hull, D. A. and Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, Seattle, US, pp 229 – 237.
- Scott, S. and Matwin, S. (1998). Text classification using WordNet hypernyms. *Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp 45 – 51.

Sebastiani, F. (1999). A tutorial on automated text categorization. In *Proceedings of ASAI-99, 1st Argentinean Symposium on Artificial Intelligence* (Analia Amandi and Ricardo Zunino, eds), pp 7 – 35, Buenos Aires, AR.

Van Rijsbergen , C. J. (1979). *Information Retrieval*. Butterworths, London, 2nd Edition.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer: New York.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland.

Witten, I. H. and Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann.

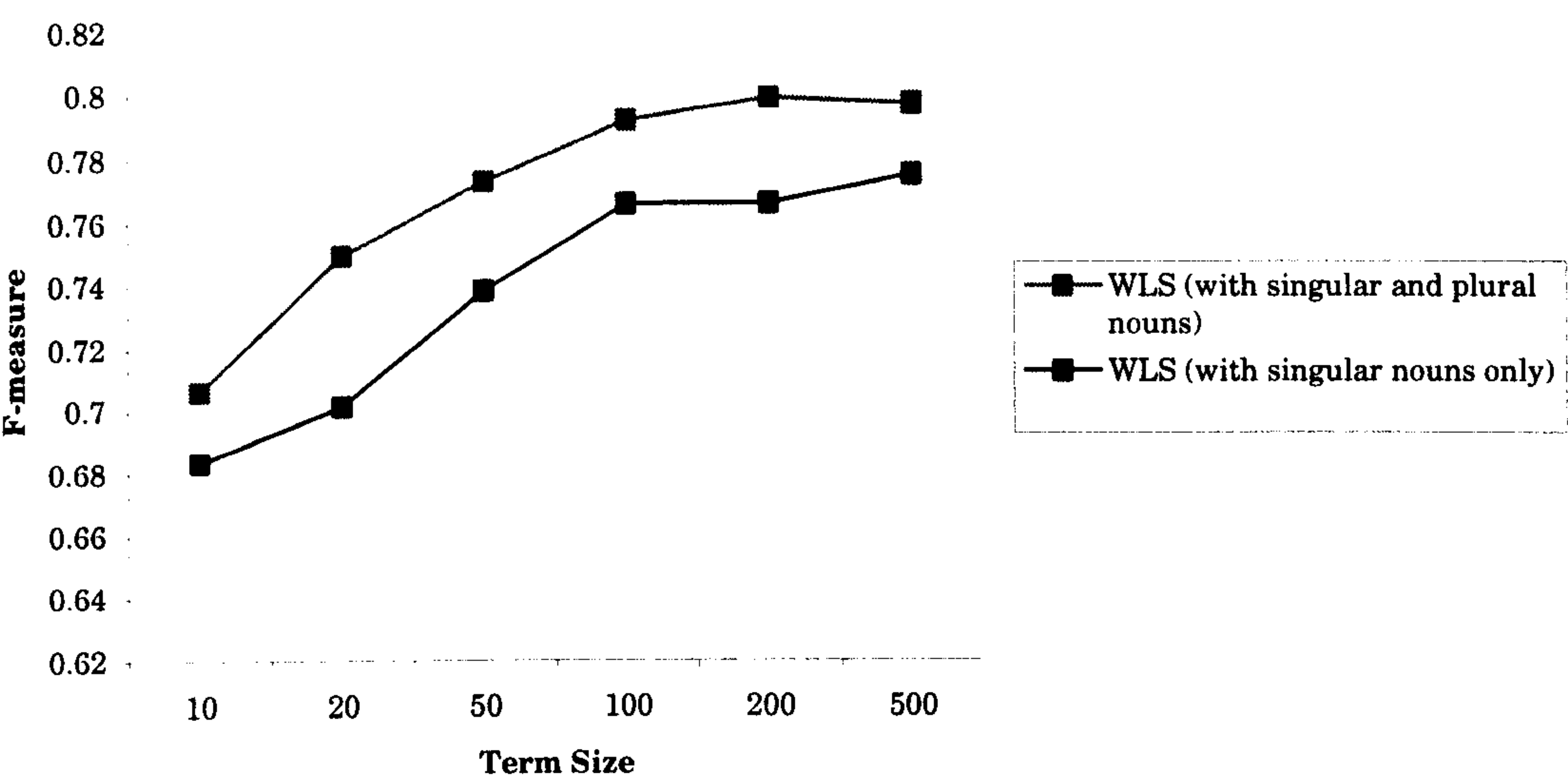
Yang Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML 97)*, pp 412 – 420, Nashville, TE, USA.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1/2(1): 67 – 88.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pp 42 – 49, Berkeley, CA.

Zheng, Z. and Srihari, R. (2003). Optimally combining positive and negative features for text categorization. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*.

APPENDIX A – Comparison of WLS (with singular and plural nouns) vs WLS (with singular nouns only)



Micro-averaged f-measure for Reuters-21578 top ten categories

APPENDIX B – Stop Words List

a	another	backs	billion
about	any	be	both
above	anybody	became	but
according	anyhow	because	by
across	anyone	become	c
actually	anything	becomes	came
adj	anywhere	becoming	can
after	are	been	can't
afterwards	area	before	cannot
again	areas	beforehand	caption
against	aren't	begin	case
all	around	beginning	cases
almost	as	behind	certain
alone	ask	being	certainly
along	asked	below	clear
already	asking	beside	clearly
also	asks	besides	co
although	at	best	co.
always	away	better	come
among	b	between	could
amongst	back	between	couldn't
an	backed	beyond	d
and	backing	big	did

didn't	ending	for	goods
differ	enough	former	got
different	etc	formerly	great
differently	even	forty	greater
do	evenly	found	greatest
does	ever	four	group
doesn't	every	from	grouped
don't	everybody	full	grouping
done	everyone	fully	groups
down	everything	further	h
downed	everywhere	furthered	had
downing	except	futhering	has
downs	f	furtheres	hasn't
during	face	g	have
e	faces	gave	haven't
each	fact	general	he
early	facts	generally	he'd
either	far	get	he'll
eg	felt	gets	he's
eight	few	give	hence
eighty	fifty	given	her
either	find	gives	here
else	finds	go	here's
elsewhere	first	going	hereafter
end	five	good	hereby

herein	interest	later	member
hereupon	interested	latter	members
hers	interesting	latterly	men
herself	interests	latest	might
high	into	least	million
higher	is	less	miss
highest	isn't	let	more
him	it	let's	moreover
himself	it's	like	most
his	its	likely	mostly
how	itself	long	mr
however	j	longer	mrs
hundred	just	longest	much
i	k	ltd	must
i'd	keep	m	my
i'll	keeps	made	myself
i'm	kinds	make	n
i've	knew	makes	namely
ie	know	making	necessary
if	known	man	need
important	knows	many	needed
in	l	maybe	needing
inc.	large	me	needs
indeed	largely	meantime	neither
instead	last	meanwhile	never

nevertheless	on	p	rather
new	once	part	really
newer	one	parted	recent
newest	one's	parting	recently
next	only	parts	reuters
nine	onto	per	reuter
ninety	open	perhaps	right
no	opened	place	room
nobody	opening	places	rooms
non	opens	point	s
none	or	pointed	same
nonetheless	order	pointing	says
noone	ordered	points	second
nor	ordering	possible	seconds
not	orders	present	see
nothing	other	presented	seem
now	others	presenting	seemed
nowhere	otherwise	presents	seeming
o	our	problem	seems
of	ours	problems	sees
off	ourselves	put	seven
often	out	puts	seventy
old	over	q	several
older	overall	quite	shall
oldest	own	r	she

she'd	somewhere	there'll	thousand
she'll	state	there're	three
she's	states	there's	through
should	still	there've	throughout
shouldn't	stop	thereafter	thru
show	such	thereby	thus
showed	sure	therefore	to
showing	t	therein	today
shows	take	thereupon	together
side	taken	these	too
sides	taking	they	took
since	ten	they'd	toward
six	than	they'll	towards
sixty	that	they're	trillion
small	that'll	they've	twenty
smaller	that's	thing	two
smallest	that've	things	u
so	the	think	under
some	their	thinks	unless
somebody	them	thirty	unlike
somehow	themselves	this	unlikely
someone	then	those	until
something	thence	though	up
sometime	there	thought	upon
sometimes	there'd	thoughts	us

use	well	whether	x
uses	wells	which	y
used	went	while	year
using	were	whither	years
v	weren't	who	yes
very	what	who'd	yet
via	what'll	who'll	you
w	what's	who's	you'd
want	what've	whoever	you'll
wanted	whatever	whole	you're
wants	when	whom	you've
wanting	whence	whomever	young
was	whenever	whose	younger
wasn't	where	why	youngest
way	where's	will	your
ways	whereafter	with	yours
we	whereas	within	yourself
we'd	whereby	without	yourselves
we'll	wherein	won't	z
we're	whereupon	would	
we've	wherever	wouldn't	

APPENDIX C – Index File Format

c:\temp\reuters-21578-10\test\acq\reut2_14843.txt:

individuals=1,goldman=1,scratch=1,entered=1,article=1,simon=1,strategy=1,crumbling=
1,optimistic=1,occur=1,city=1,efficiency=1,board=1,hands=1,gottardo=2,yen=1,st=1,disap
pearing=1,komatsu=9,decision=2,comparable=1,expect=1,heiwa=4,european=1,don=1,mo
ves=2,business=8,president=1,ministry=1,placed=1,soon=1,told=1,analysts=4,change=1,s
eparates=1,parent=1,gamble=1,allowed=1,benson=1,widely=1,sachs=1,result=1,waves=1
,expand=2,holds=2,kind=1,osaka=1,act=2,cautiously=1,divided=1,regulations=1,lt=1,koh
=1,analyst=2,ll=1,glass=1,forbidden=1,weak=1,government=1,geared=1,businesses=1,qu
ick=1,hard=1,devote=1,deregulation=2,walls=1,said=15,desposits=1,profitability=1,tokyo
=1,juggling=1,decade=1,home=1,reluctant=1,investment=2,initially=1,move=2,cost=1,la
me=1,confident=1,bond=1,changes=1,increase=1,licences=1,wisdom=1,despite=1,quickly
=2,struggling=1,agreed=2,hopes=1,regain=1,ve=1,mln=1,smithson=2,network=2,wrong=
1,merged=1,metropolitan=1,banking=1,swiss=1,ailing=1,grant=1,absorbing=1,recovery=
1,lending=1,entering=1,japan=2,fed=1,partnership=1,interview=1,added=1,create=1,link
=1,strengthening=1,dealing=1,wait=1,medium=1,status=1,question=1,including=1,may=
1,october=1,securities=9,finance=1,relatively=1,version=1,focus=1,enter=1,overseas=1,ce
ntury=1,lot=1,strategies=1,willing=1,aims=1,paid=1,universal=2,demand=1,subsidiaries
=2,reserve=1,expected=2,replace=1,text=2,related=2,position=2,pay=3,exchange=2,branc
h=2,digesting=1,profitable=2,limited=1,steagall=1,lose=1,tie=1,lucrative=1,del=1,pushin
g=1,patient=1,questioned=1,august=1,looked=1,sumi=1,relations=1,meiko=1,subsidiary=
2,various=1,similar=2,initiatives=1,argue=1,reduce=1,kleinwort=2,pct=3,share=2,sumito
mo=15,vastly=1,trust=1,title=2,time=3,commercial=2,routes=1,dllrs=2,sogo=4,future=1,d
ebts=1,chances=1,competition=1,estimated=1,bank=12,merger=5,expanded=1,duck=1,cr

eating=1,banks=3,strict=1,based=1,opportunities=1,banca=1,near=1,effective=1,view=1,american=1,personnel=1,domestic=2,federal=1,push=1,limits=1,following=1,financial=3,unrecoverable=1,bold=2,aggressive=1,loans=1

c:\temp\reuters-21578-10\test\acq\reut2_14852.txt:

repaid=1,company=4,life=1,central=2,accurate=1,loan=1,statement=3,operations=1,atlc=1,infusion=1,war=1,operation=1,ore=1,october=1,underground=1,restructuring=1,acquire=1,troubled=1,talks=1,bona=1,development=4,memorandum=1,cebu=1,half=1,properties=1,told=1,negotiations=2,reasonably=1,seriously=1,text=2,corp=5,accelerated=1,december=1,continuing=1,tonnage=1,exploration=1,loss=2,bail=2,compared=1,kong=1,copper=3,island=1,producer=1,world=2,losses=1,international=1,extending=1,lt=3,pesos=1,mines=1,consolidated=1,wilson=1,said=8,increased=1,gold=3,creation=1,able=1,scaled=1,title=2,reported=1,mining=4,mabate=2,prices=1,dlrs=1,package=1,fund=1,equity=1,total=1,holding=1,bond=6,details=1,cut=1,philippines=2,biggest=1,firm=1,bank=1,banks=2,province=1,mln=2,manager=1,understanding=1,pre=1,mn=1,net=2,signed=1,hit=1,holdings=1,acquisition=1,existing=1,atlas=10,considering=2,loans=2,discussing=1,depressed=1,started=1,investments=1,hong=1

c:\temp\reuters-21578-10\test\acq\reut2_14865.txt:

hold=2,nl=3,creek=4,consolidated=1,whim=4,pty=1,statement=1,mln=2,text=2,owns=2,croesus=1,disclose=1,said=2,pay=1,resources=1,consortium=2,pct=3,gold=5,producing=1,forest=3,project=1,title=2,sold=1,reported=2,western=1,cra=3,mining=1,craa=1,acquisition=1,austwhim=1,price=1,dlrs=2,ounces=1,lt=5,leading=1,combined=1,unit=1,undeveloped=1,yesterday=2,australia=1,mines=1

c:\temp\reuters-21578-10\test\acq\reut2_14888.txt:

beverage=1,bidders=1,diaz=3,ramon=1,bid=3,company=2,offered=2,filipino=1,economic=1,shares=7,ariadne=1,farmlands=1,bud=1,government=3,australian=3,subsidiary=1,nept

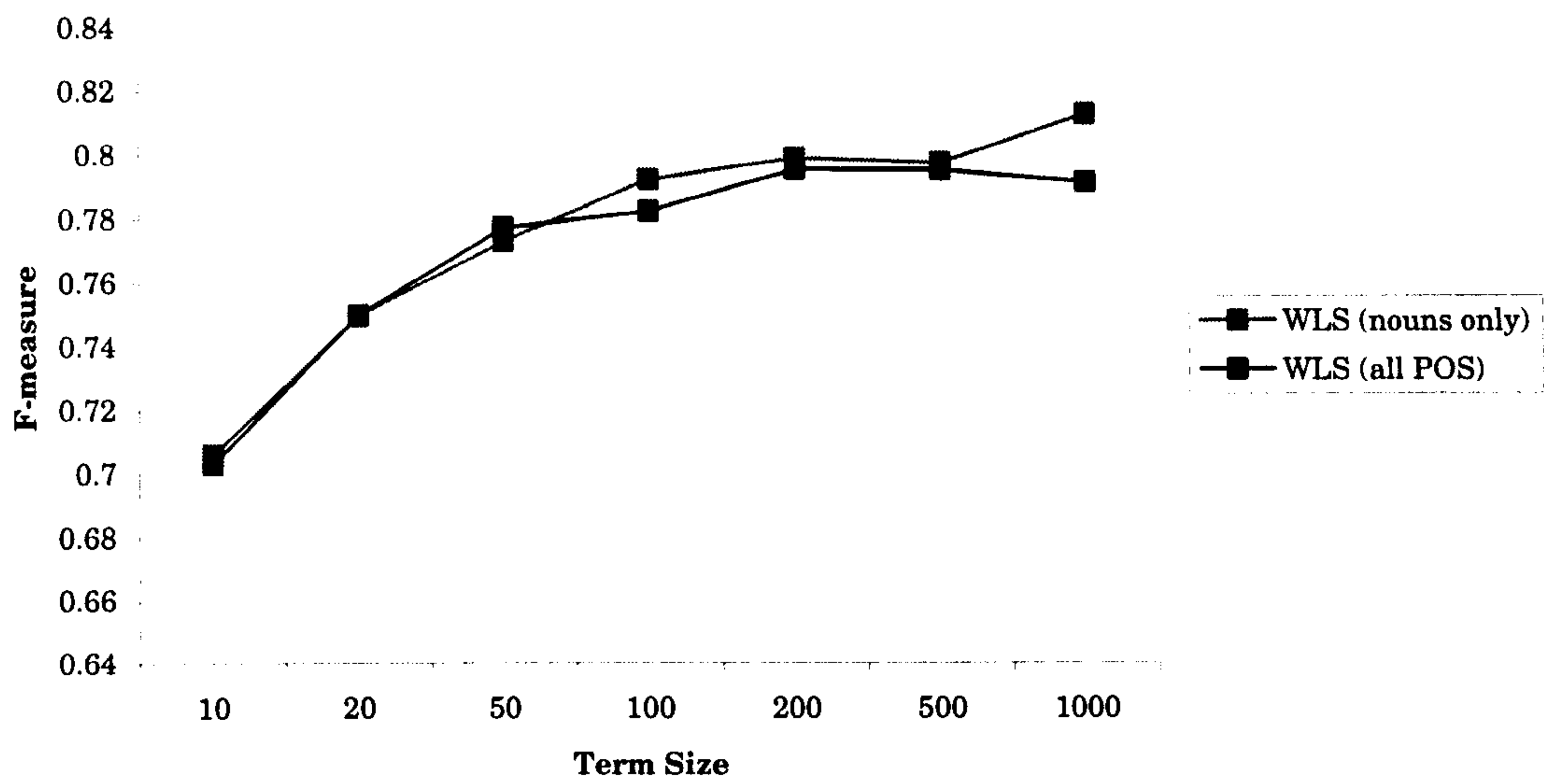
unia=1,commission=1,dlr=1,said=5,quoted=1,busch=4,pct=2,san=5,head=1,share=1,food
=1,inc=2,planning=1,sequestered=2,largest=1,companies=1,broker=1,title=2,foreign=1,in
vestment=1,elxa=1,york=1,price=1,joined=1,talks=1,sale=1,maker=1,bond=2,june=1,phil
ippines=1,earlier=1,firm=1,listed=1,bank=1,broke=1,secretary=1,journal=1,controls=1,m
elbourne=1,anheuser=4,told=3,based=1,miguel=5,mln=4,brewer=1,text=2,agree=1,brewe
ry=1,corp=4,disclose=1,elders=1,owned=2,barwon=1,branch=1,mn=1,brewing=1,kong=2,i
xl=1,buying=2,alan=1,saying=1,stock=1,presidential=1,spokesman=1,pcgg=1,entire=1,ho
ldings=1,sanm=1,block=1,allen=1,joins=1,lt=7,terms=1,buy=2,panel=1,pesos=1,month=2,
jacksons=1,australia=1,hong=2

c:\temp\reuters-21578-10\test\acq\reut2_14900.txt:

fell=1,bid=1,redland=3,advised=1,brief=1,close=1,previous=1,ash=1,diversified=1,chairm
an=1,shares=4,due=1,fly=1,australian=2,britain=2,said=4,marketer=1,plc=1,may=2,pct=
1,annual=1,peak=1,housing=1,largest=2,lead=1,offer=1,recommended=1,title=2,notice=1
,roof=1,foreign=1,low=1,reported=1,manufacturer=1,rose=1,discussions=1,monier=6,dllr
=3,talks=1,june=1,earlier=1,hold=1,soon=1,ended=1,report=1,building=1,half=1,mln=3,ti
le=1,text=2,materials=1,issued=1,rdld=1,earnings=1,exchange=1,trading=1,downturn=1,
net=1,locke=2,stock=1,holds=1,world=1,mnra=1,progressed=1,shareholders=1,lt=2,bill=1
,concrete=1,yesterday=1,zealand=1,australia=1

.
.
.
.
.
.
.

APPENDIX D – Comparison of WLS (nouns only) vs WLS (all POS)



Micro-averaged f-measure for Reuters-21578 top ten categories

APPENDIX E – ARFF File Format

@relation corn

@attribute tonnes numeric

@attribute corn numeric

@attribute title numeric

@attribute pct numeric

@attribute grain numeric

@attribute wheat numeric

@attribute maize numeric

@attribute agriculture numeric

@attribute export numeric

@attribute program numeric

@attribute week numeric

@attribute production numeric

@attribute department numeric

@attribute acres numeric

@attribute trade numeric

@attribute total numeric

@attribute ec numeric

@attribute crop numeric

@attribute farmers numeric

@attribute report numeric

@attribute prices numeric

@attribute soybean numeric

@attribute sources numeric
@attribute imports numeric
@attribute market numeric
@attribute price numeric
@attribute farm numeric
@attribute exports numeric
@attribute official numeric
@attribute union numeric
@attribute acreage numeric
@attribute bushels numeric
@attribute harvest numeric
@attribute producers numeric
@attribute tonne numeric

.
.
.

@attribute class {1,0}

@data

{2 2,3 3,58 2,64 4,72 2,85 1,91 1,98 1,103 1,113 2,114 1,118 1,142 2,153 1,157 1,164 1,165
1,170 8,182 1,203 2,209 1,254 1,257 1,280 1,297 2,319 1,323 2,357 1,396 1,405 1,428
3,437 1,511 1,522 1,599 1,608 1,632 1,644 1,705 1,768 1,789 2,797 2,828 1,835 2,860
1,893 9,955 0}

{2 2,15 1,20 1,61 2,86 1,99 1,115 1,146 1,153 1,218 1,245 3,251 5,288 4,298 2,331 2,335
4,357 2,434 1,508 1,515 1,516 1,632 6,710 1,773 2,778 1,796 2,806 1,854 1,955 0}

{2 2,3 3,25 1,245 1,355 1,382 1,413 2,428 1,849 1,955 0}

{2 2,3 2,25 1,51 1,65 2,78 2,104 2,194 1,203 1,218 1,237 1,238 1,250 1,251 4,277 3,316
1,335 2,337 1,388 1,412 1,516 1,564 7,632 2,691 1,710 1,955 0}

{2 2,3 1,19 1,46 1,61 1,97 1,142 1,145 1,190 1,218 1,277 1,279 1,413 1,521 1,564 4,689
1,796 1,955 0}

{2 2,3 2,59 1,212 1,228 1,245 1,253 1,279 1,332 2,335 2,515 1,564 1,589 1,640 1,893 1,955
0}

{0 1,2 2,8 1,9 1,11 1,20 1,22 1,28 1,51 1,53 6,58 1,59 1,85 1,94 2,104 1,126 1,129 2,132
1,216 1,238 1,251 1,267 1,335 1,361 1,394 1,523 1,529 5,536 1,539 2,711 2,955 0}

{2 2,3 4,10 2,18 1,19 1,22 3,25 1,38 1,39 4,40 2,51 2,66 2,78 1,85 1,86 2,90 1,98 5,104
2,115 1,126 1,142 1,155 1,161 1,171 3,182 2,190 1,194 1,198 1,200 2,203 3,212 2,216
1,232 2,237 2,245 3,250 1,251 5,252 1,253 1,277 2,279 2,281 1,332 1,335 5,412 1,413
1,453 1,468 1,481 2,508 1,541 1,563 1,564 21,614 1,633 1,671 1,688 2,689 1,796 1,893
1,898 1,941 2,955 0}

{2 2,25 1,46 2,71 2,98 1,150 1,203 1,277 1,335 1,361 1,564 1,777 1,955 0}

{2 2,63 1,65 1,93 1,109 1,115 1,161 1,194 1,216 1,250 1,277 1,471 1,474 2,540 1,562 1,747
1,955 0}

{2 2,218 1,251 1,955 0}

{2 2,3 1,38 1,109 1,157 1,203 1,218 2,251 2,277 1,298 1,335 1,359 1,434 2,564 2,571 1,746
1,747 1,847 1,941 1,955 0}

{2 2,40 1,85 1,200 1,243 1,271 1,316 1,335 2,380 1,597 1,746 2,941 2,955 0}

{2 2,3 5,41 3,46 2,51 1,65 1,68 2,71 3,77 1,97 7,98 1,102 2,103 2,122 2,142 1,170 2,174
1,222 1,240 2,244 1,273 1,279 1,323 3,329 1,335 2,355 1,391 1,402 1,405 2,429 1,435
1,466 1,481 1,531 1,548 1,549 1,589 1,603 1,708 1,754 1,779 1,818 1,878 1,893 1,955 0}

{2 2,109 1,251 2,316 1,955 0}

{2 2,3 2,41 2,108 1,250 1,316 1,335 1,955 0}

{2 2,128 1,206 1,251 1,335 1,404 3,611 1,955 0}
 {2 2,77 3,93 1,114 1,162 2,243 3,250 2,251 1,335 1,398 1,419 1,804 1,955 0}
 {2 2,39 1,251 1,564 1,955 0}
 {2 2,3 1,25 1,200 1,251 1,382 1,449 2,710 1,955 0}
 {2 2,38 1,40 2,98 1,252 1,329 1,380 1,557 1,646 1,941 1,955 0}
 {2 3,251 1,316 1,382 2,425 1,955 0}
 {2 2,3 7,9 1,15 1,19 1,20 4,24 3,27 2,29 1,33 1,48 2,61 1,64 1,65 1,79 4,87 1,95 2,102 1,104
 3,108 7,109 4,111 2,113 1,126 11,136 1,154 2,165 4,170 5,171 1,190 1,196 2,203 1,232
 1,238 1,250 3,251 3,253 1,257 1,276 1,279 3,282 1,297 1,301 1,335 1,342 3,350 1,375
 1,376 1,436 1,509 1,510 1,522 1,523 1,526 1,539 5,544 1,549 1,577 2,604 1,614 1,633
 1,639 1,662 1,711 1,738 1,779 2,815 2,847 3,868 1,913 1,942 1,943 1,946 1,955 0}
 {2 2,19 1,277 1,476 1,955 0}
 {2 2,19 4,24 1,46 1,104 1,182 1,216 1,229 1,250 1,256 1,273 1,277 2,279 1,307 1,332 2,335
 7,476 1,544 1,571 1,746 2,747 1,955 0}
 {2 2,69 1,142 1,254 1,316 1,746 2,777 1,955 0}
 {2 2,3 1,142 1,941 1,955 0}
 .
 .
 .
 .
 .
 .

APPENDIX F – Correlation Test Results

Data 1

Category	No of training documents	F-measure (WLS)
acq	1650	0.8211
corn	181	0.3762
crude	389	0.7392
earn	2877	0.9482
grain	433	0.5786
interest	347	0.5506
money-fx	538	0.6255
ship	197	0.628
trade	369	0.6805
wheat	212	0.4287

Correlation Test Results 1

	No of training documents	F-measure (WLS)
No of training documents	1	
F-measure (WLS)	0.8136	1

Data 2

Category	No of synonyms in WordNet	F-measure (WLS)
acq	2793	0.8211
corn	955	0.3762
crude	1834	0.7392
earn	2342	0.9482
grain	1716	0.5786
interest	1147	0.5506
money-fx	1667	0.6255
ship	1258	0.628
trade	1823	0.6805
wheat	1109	0.4287

Correlation Test Results 2

	No of synonyms in WordNet	F-measure (WLS)
No of synonyms in WordNet	1	
F-measure (WLS)	0.8789	1

Data 3

Category	No of nouns in WordNet	F-measure (WN)
acq	4944	0.8283
corn	1797	0.3588
crude	3206	0.7262
earn	4327	0.9515
grain	2934	0.5831
interest	2039	0.6054
money-fx	2832	0.6235
ship	2344	0.661
trade	3108	0.6529
wheat	2001	0.4258

Correlation Test Results 3

	No of nouns in WordNet	F-measure (WN)
No of nouns in WordNet	1	
F-measure (WN)	0.864290912	1

APPENDIX G – Extract of Codes

Extract of codes for cross-referencing

```
Vector vecsynset = new Vector();

Vector vecfreq = new Vector();

Vector vecword = new Vector();

Vector vecfrequency = new Vector();

Vector vecsingleword = new Vector();

Vector vecsinglefreq = new Vector();


try{

    POS n = POS.NOUN;

    IndexWord word = Dictionary.getInstance().lookupIndexWord(n, str);

    Synset[] synset= word.getSenses();


    for(int i=0; i<synset.length; i++) {

        String synsetString = synset[i].toString();

        //check for same term

        if(vecsynset.contains(synsetString)) {

            int index = vecsynset.indexOf(synsetString);

            vecsynset.setElementAt(synsetString, index);

            vecfreq.setElementAt(new Integer(CTFw), index);

            vecword.addElement(str);

            vecfrequency.addElement(new Integer(CTFw));

        }

    }

}
```

```

        vecword.addElement(vecsingleword.elementAt(index));
        vecfrequency.addElement(vecsinglefreq.elementAt(index));
    }

    else {

        vecsynset.addElement(synsetString);
        vecfreq.addElement(new Integer(CTFw));
        vecsingleword.addElement(str);
        vecsinglefreq.addElement(new Integer(CTFw));
    }

}

}

catch(Exception exp){

    System.out.println("Exception - " + exp.toString());

}

```

Text Categorization: Using WordNet To Improve Feature Selection

Stephanie Chua¹

Narayanan Kulathuramaiyer²

¹*Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.*

Tel: +60-8-267-1000, Fax: +60-8-267-2301, Email: hlchua@calm.unimas.my¹, nara@fit.unimas.my²

Abstract

This research explores the use of WordNet [1], a lexical database for performing feature selection for an automated text categorization system. The WordNet based approach employs noun synonyms and word senses for feature selection and this allows the selection of terms that are semantically representative of a category of documents, as opposed to statistical approaches traditionally used for feature selection. The approach that is adopted extends the use of WordNet, which has been typically used for Word Sense Disambiguation [2]. Our experiments on the Reuters-21578 dataset have shown that the proposed approach has performed better than Information Gain, Chi-Square and Category Term Frequency as a feature selection scheme.

Keywords

Text categorization, feature selection, WordNet, machine learning

Introduction

Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns suggested by a training set of categorized documents [3]. This categorization task used to be done manually by human experts in the past when the number of documents were not significantly large. With the increase in digital documents over the years, manual categorization by the experts proves to be a heavy workload. Therefore, an automated system employing machine learning is required to assist in the categorization tasks.

Approach and Methods

Feature Selection

One major difficulty in text categorization is the high dimensionality of the feature space. There are thousands of terms occurring in documents and it has been noted that learning algorithms cannot handle such a large number of terms. Feature selection is performed to reduce the dimensionality of the feature space. There are a number of feature selection methods, which over the years, are used in many text categorization tasks. Among the more

popular ones are Document Frequency (DF), InfoGain (IG), Chi-Square (Chi2) and Mutual Information [4], [5]. These approaches that are currently used for feature selection are purely statistical.

The works by [6] has shown that approaches considering semantics are promising. Their work on cascaded feature selection extracts two sets of documents, one with all parts of speech (POS) and another with only nouns from the Reuters-21578 dataset. Those terms are then looked up in WordNet and the synonyms of those terms are added for representing documents. Our work, however, will explore the use of synsets and word senses that represents the actual context of a word's usage. It is thus our aim to identify the existing synonyms in a category instead of adding in synonyms.

This research focuses on the feature selection process where we employ WordNet to discover synonymous terms based on cross-referencing. We will compare the WordNet synonyms approach with statistical techniques such as Chi2, IG and Category Term Frequency (CTF). IG is a feature selection technique that makes use of the presence and absence of a term in a document to select its features. Chi2 measures the degree of independence between a term and a category and select features with high dependency on a category while CTF is based on counting the number of occurrence of each term in a category. Terms with high frequencies of occurrence will be chosen as features. CTF is actually the extension of term frequency (TF) as used in TFIDF term weighting scheme. While TF is the frequency of terms occurring in a document, CTF is the frequency of terms occurring in a category.

Our experiments will explore the use of synonyms and word senses in feature selection to help derive a better set of features for category representation as well as better categorization effectiveness.

WordNet

WordNet is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University by a group led by Professor George A. Miller [1]. At present, WordNet contains more than 150,000

different unique terms. These are organized into some 115,000 word meanings or set of synonyms called synsets. The lexicon in WordNet is divided into nouns, verbs, adjectives and adverbs. In this research, only nouns are used.

Dataset: Reuters-21578

The Reuters-21578 Distribution 1.0 collection is one of the most popularly used benchmark in text categorization researches [7]. This corpus contains 21578 documents that appear on the Reuters newswire in 1987. There are five category sets in this collection, namely, “Exchanges”, “Orgs”, “People”, “Places”, and “Topics”. Each document in the collection is assigned to at least one category by human indexers.

In this research, the categories from “Topics” are used. This set of categories has been used in almost all previous researches that makes use of the Reuters-21578 dataset [8]. It contains economic subject categories such as “corn”, “grain”, “interest” and so on. The training and testing set is split using the Mod Apte split, which has been used in many past works [4], [6], [9], [10], [11], [12].

Table 1 – The number of categories in each category set and the number of categories with at least 1 occurrence and 20 occurrences

Category Set	Number of Categories	Number of Categories with at least 1 occurrence	Number of Categories with at least 20 occurrences
Exchanges	39	32	7
Orgs	56	32	9
People	267	114	15
Places	175	147	60
Topics	135	120	57

Using WordNet for Feature Selection

We have carried out two experiments on the WordNet based approach, where we used WordNet terms in the first and subsequently, we explored the use of WordNet synonyms and word senses. In the WordNet terms approach, all WordNet terms that occur in a category is chosen as features and ranked according to their frequencies. These features are words that occur in the category and contained in the WordNet dictionary. Therefore, insignificant terms, such as, “twa”, “shr”, “lme”, “natl” and other non-English terms will be filtered and omitted. The WordNet term approach is denoted as WN1 in Figure 1 in the Appendix section.

In the WordNet synonyms approach, feature selection is based on synonymous terms that co-occur in a category. The co-occurrence of terms belonging to the same synset is used as an indicator of significant terms to represent a category. For terms co-occurring in a category, the sense assigned is selected based on the synset signature

reference. Cross-referencing is done by checking the list of noun synsets for all senses for similar synsets. The sense of different terms with similar synsets highlights the actual semantic context of the terms. The original terms from the category that belongs to the similar synsets will then be identified and added as features for category representation. WordNet synonyms approach is denoted as WN in Figure 1 in the Appendix section.

Here is an example of how cross-referencing works. We will take a look at all the senses for five nouns; ‘corn’, ‘wheat’, ‘maize’, ‘acquisition’ and ‘business’. All these nouns are assumed to occur in the same category.

Table 2 – List of terms and their synsets in all senses

Terms	Synsets in all senses
Corn	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {corn} Sense 3: {corn, edible corn} Sense 4: {corn, clavus} Sense 5: {wheat, corn} Sense 6: {corn whiskey, corn whisky, corn}
Wheat	Sense 1: {wheat, corn} Sense 2: {wheat, wheat berry}
Maize	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {gamboge, lemon, lemon yellow, maize}
Acquisition	Sense 1: {acquisition} Sense 2: {acquisition} Sense 3: {learning, acquisition} Sense 4: {skill, accomplishment, acquirement, acquisition, attainment}
Ship	Sense 1: {ship}

From Table 2, we can see that there are two similar synsets. These synsets are found in:

1. Corn (Sense 1) and Maize (Sense 1) - {corn, maize, Indian corn, Zea mays}
2. Corn (Sense 5) and Wheat (Sense 1} - {wheat, corn}

Therefore, for the term ‘corn’, sense 1 and sense 5 will be selected, while for the term ‘maize’ and ‘wheat’, sense 1 will be selected. All the other synsets in the other senses will be discarded. The synsets that are chosen will be used in feature selection to select the original terms found in each category. Synonymous terms that are found in each category are listed out from the list of similar synsets. Their frequencies are then obtained from the index file. Therefore, from the list of similar synsets in the example shown, there are three features selected to represent the category. They are ‘corn’, ‘wheat’ and ‘maize’. The rationale behind choosing synonymous terms in representing a category is that the synonymous terms are semantically related. Therefore, there can be a stronger

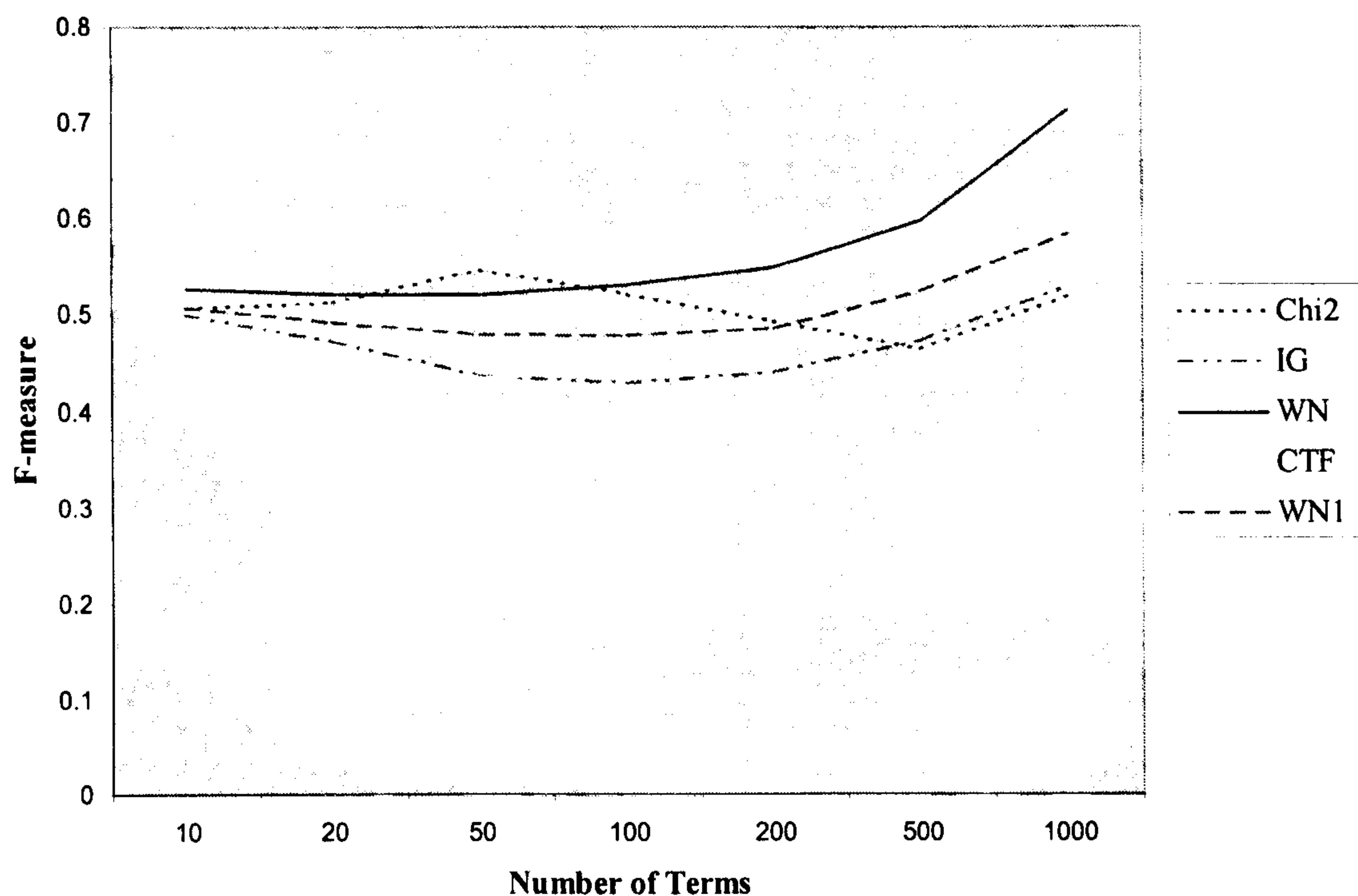


Figure 1 – Comparison of Chi2, IG, WN, CTF and WN1

discriminative power in those terms in representing a category, as they are conceptually close.

Word sense disambiguation (WSD) refers to the process of disambiguating words by telling which sense an ambiguous word belongs to. Literature has highlighted that WordNet has been used in the process of WSD [2]. WSD is a complicated process where the sense of every term in every sentence is important. Our focus in this research is not on WSD. However, we simplify the WSD process in this research by identifying significant synsets for a category by finding the overlapping synset sense signatures. We applied a simplified version of WSD because past works have reported the importance of WSD in improving categorization performance [13], [14].

Experiments

The experiments were conducted using the naïve bayes machine learning scheme. The effectiveness of each approach was measured using the F-measure value, which gave the same weight to both precision and recall. Statistical schemes like Chi2, IG and CTF were compared with the WordNet based approach to see how well a natural language approach performs as compared to a statistical approach.

Experiments were also carried out to see the significance of using synonymous words in WordNet as features to represent a category for text categorization. This is done by comparing the performance of the WordNet synonyms approach with the WordNet terms approach.

Results

The micro-averaged f-measure is reported for the experiments carried out on the entire Reuters-21578 “Topics” categories. The WordNet synonyms approach is able to outperform all the statistical feature selection techniques as well as the WordNet terms approach in all the experiments performed. The results are shown in Table 3 and plotted in Figure 1.

Table 3 – Micro-averaged F-measure for various feature selection techniques

Number of terms	Chi-Square (Chi2)	InfoGain (IG)	WordNet Synonyms (WN)	Category Term Frequency (CTF)	WordNet Terms (WN1)
10	0.506667	0.499688	0.525933	0.473122	0.506942
20	0.512587	0.472947	0.52078	0.443562	0.49185
50	0.546537	0.437541	0.52091	0.434923	0.47922
100	0.520722	0.430292	0.53139	0.442529	0.478796
200	0.49298	0.440706	0.549589	0.466821	0.485978
500	0.463756	0.472804	0.598316	0.50174	0.523821
1000	0.519098	0.529123	0.713126	0.548593	0.584407

Discussion

The WordNet terms approach surprisingly performed well when compared to the statistical techniques. Only Chi-Square did marginally better at the threshold of 200 terms and below. On the other hand, we can see that the

WordNet synonyms approach generally has better results than all the other feature selection techniques with the exception of Chi-Square at the threshold of 50 terms. For other thresholds, the WordNet synonyms approach was able to achieve the highest rate of effectiveness. The Chi Square feature selection technique tends to favor unique terms for category representation, which was significant only when 50 terms were chosen. Both IG and CTF also performed worse than the WordNet synonyms approach. This shows that statistical data by itself is not sufficient to select features that are significant and meaningful for category representation and some form of semantics can improve the feature selection process.

Conclusion

The terms in WordNet by itself can serve as an approach to identify generic terms that are representative of the categories. From the experiments conducted, we can see that the WordNet synonyms approach can improve the feature selection process in text categorization. It is able to provide a better set of features for category representation as the Reuters-21578 collection has many terms that are covered comprehensively by WordNet. Therefore, the WordNet synonyms approach has proven to be an effective feature selection approach.

However, WordNet based approaches do not work well for categories that are domain specific. Domain specific dictionaries would need to be incorporated in order to provide the technical and domain specific terms that are lacking in WordNet.

References

- [1] Miller G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Vol 3, No.4 (Winter 1990), 235-244.
- [2] Li, X., Szpakowicz, S., Matwin, S. 1995. A WordNet-based Algorithm for Word Sense Disambiguation. In *Proceedings of the IJCAI-95*, 1368-1374.
- [3] Sebastiani, F. 1999. A Tutorial on Automated Text Categorisation. In *Analia Amandi and Ricardo Zunino, editors, Proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, pages 7—35, Buenos Aires, AR.
- [4] Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 42—49, Berkeley, CA.
- [5] Yang, Y. and Pedersen, J. A. 1997. Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference (ICML 97)*, pages 412—420, Nashville, TE, USA.
- [6] Masuyama, T. and Nakagawa, H. 2002. Applying Cascaded Feature Selection to SVM Text Categorization. In *the DEXA Workshops 2002*, 241-245.
- [7] Hull, D. 1994. Improving Text Retrieval for the Routing Problem Using Latent Semantic Indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 282-291.
- [8] Lewis, D. D. 26 September 1997. Reuters-21578 Text Categorization Test Collection Distribution 1.0 README File (v 1.2). Citing Internet sources URL <http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt>
- [9] Scott, S. and Matwin, S. 1998. Text Classification Using WordNet Hypernyms. In *Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp. 45-51.
- [10] Cohen, W.W. and Singer, Y. 1996. Context-sensitive learning methods for text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, 307-315.
- [11] Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 148-155, Bethesda, Maryland, USA.
- [12] Joachims, T. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, pp. 137—142.
- [13] Mihalcea, R. and Moldovan, D. 2000. Semantic indexing using WordNet senses. In *Proceedings of the ACL Workshop on IR & NLP, Hong Kong*.
- [14] Gonzalo, J., Verdejo, F., Chugur, I. And Cigarran, J. 1998. Indexing with WordNet Synsets can Improve Text Retrieval. In *Proceedings of the Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp. 38-44.

Semantic Feature Selection Using WordNet

Stephanie Chua
*Faculty of Computer Science and
Information Technology,
Universiti Malaysia Sarawak.
hlchua@calm.unimas.my*

Narayanan Kulathuramaiyer
*Faculty of Computer Science and
Information Technology,
Universiti Malaysia Sarawak.
nara@fit.unimas.my*

Abstract

The web has caused an explosion of documents, requiring the need for an automated text categorization system. This paper explores the notion of semantic feature selection by employing WordNet [1], a lexical database. The proposed semantic approach employs noun synonyms and word senses for feature selection to select terms that are semantically representative of a category of documents. The categorical sense disambiguation extends the use of WordNet, which has been typically used for text retrieval and word sense disambiguation [2]. Our experiments on the Reuters-21578 dataset have shown that automated semantic feature selection is able to perform better than well known statistical feature selection methods, Information Gain and Chi-Square as a feature selection method.

1. Introduction

Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns suggested by a training set of categorized documents [3]. One major difficulty in text categorization is the high dimensionality of the feature space. There are thousands of terms occurring in documents and it has been noted that learning algorithms cannot handle such a large number of terms. Feature selection is performed to reduce the dimensionality of the feature space. There are a number of feature selection methods, which over the years, are used in a wide range of text categorization tasks. Among the more popular statistics based ones are Document Frequency (DF), InfoGain (IG), Chi-Square (Chi2) and Mutual Information [4], [5].

The works by [6] has shown that approaches considering semantics are promising. Their work on cascaded feature selection extracts two sets of documents, one with all parts of speech (POS) and another with only nouns from the Reuters-21578 dataset. These terms are

then looked up in WordNet and the synonyms associated with those terms are used as features for representing documents. Our work, however, explores the use of synsets and word senses to capture the actual context of a word's usage.

2. WordNet

WordNet is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University by a group led by Professor George A. Miller [1]. At present, WordNet contains more than 150,000 different unique terms. These are organized into some 115,000 word meanings or set of synonyms called synsets. The lexicon in WordNet is divided into nouns, verbs, adjectives and adverbs. Only nouns are considered in our research as preliminary experiments indicate that the use of other part-of-speech does not significantly enhance performance.

3. Using WordNet for feature selection

This research focuses on the feature selection process where WordNet is employed to discover synonymous terms based on cross-referencing. We will compare the WordNet synonyms approach with statistical methods such as Chi2 and IG. IG is a feature selection technique that makes use of the presence and absence of a term in a document to select its features. Chi2 measures the degree of independence between a term and a category. It selects features with high dependency on a particular category. The WordNet synonyms approach will explore the use of synonyms and word senses to derive a better set of features for category representation to achieve better categorization effectiveness.

In the WordNet synonyms approach, feature selection is based on terms with overlapping word senses co-occurring in a category. The co-occurrence of terms with the same synset signature is used as an indicator of significant terms to represent a category. For terms co-occurring in a category, the correct sense is determined

based on the synset signature cross-referencing. Cross-referencing is done by checking the list of noun synsets for all senses for similarity in the signatures. The senses of different terms with overlapping synsets aggregate the semantic context of a category. The original terms from the category that belongs to the similar synsets will then be identified and added as features for category representation. The WordNet synonyms approach is denoted as WN in all the figures.

The example below illustrates the approach used. Let us look at all the senses for five nouns; ‘corn’, ‘maize’, ‘acquisition’ and ‘ship’. Each sense has a signature, which is referred to as a synset containing synonyms to reflect a sense.

Table 1. List of terms and their synsets for each sense

Terms	Synsets for all senses
Corn	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {corn} Sense 3: {corn, edible corn} Sense 4: {corn, clavus} Sense 5: {wheat, corn} Sense 6: {corn whiskey, corn whisky, corn}
Maize	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {gamboge, lemon, lemon yellow, maize}
Acquisition	Sense 1: {acquisition} Sense 2: {acquisition} Sense 3: {learning, acquisition} Sense 4: {skill, accomplishment, acquirement, acquisition, attainment}
Ship	Sense 1: {ship}

From Table 1, we can see that there are two identical synsets. Sense 1 of ‘corn’ and sense 1 of ‘maize’ have identical synsets with the same signatures. These synsets will be used in feature selection to select the original terms found in each category. Categorical sense disambiguation is performed here to automatically disambiguate semantically related terms. By finding identical synsets with the same signatures, the dominant senses can be determined for the synonymous terms in each category.

Semantically related terms that are found in each category are listed out from the list of similar synsets. Their frequencies are then obtained from the index file. Therefore, in the example shown, there are two features selected to represent the category. They are ‘corn’ and ‘maize’. The rationale behind choosing synonymous terms in representing a category is that these terms have a

stronger discriminative power in representing a category, as they are conceptually close.

Word sense disambiguation (WSD) refers to the process of disambiguating words by telling which sense an ambiguous word belongs to. WordNet has also been used in the process of WSD [2]. WSD is a complicated process where the sense of every term in every sentence is important. Our focus in this research is not on WSD. However, we simplify the WSD process in this research by identifying significant synsets for a category by finding the overlapping synset sense signatures. We applied a simplified version of WSD because past works have reported the importance of WSD in improving categorization performance [7], [8].

4. Experiments

The experiments were conducted using the multinomial naïve Bayes scheme from the Waikato Environment for Knowledge Analysis (WEKA) [9] machine learning tool. We chose this machine learning scheme because of its simplicity and fast processing. Although support vector machines (SVM) has been shown to outperform other schemes [10], our objective has not been to show which machine learning scheme is able to perform best. On the other hand, we are interested in the relative performance of the feature selection methods regardless of what machine learning schemes used.

The effectiveness of each feature selection method is evaluated using f-measure value, which gives the same weight to both precision and recall. Statistical methods like Chi2 and IG were compared with the WordNet synonyms approach to see how well a semantic approach performs as compared to a statistical approach.

Experiments were also carried out to see the significance of using the nouns from WordNet that occur in each category as features to contrast with the performance of the WordNet synonyms approach. Both experiments rank features according to their frequencies. WordNet nouns approach is denoted as WN1 in Figure 4 and 5.

5. Results and analysis

As a benchmark, we perform experiments on the Reuters-21578 dataset using the ModApte split [11]. This split assigns documents from April 7, 1987 and before to the training set and documents from April 8, 1987 and after to the test set. Only documents with an occurrence of at least one “Topics” category are used. Figure 1 shows the macro-averaged f-measure comparison between Chi2, IG and WN for the Reuters-21578 dataset. It shows that the WordNet synonyms approach is able to outperform the statistical feature selection methods, Chi2

and IG. However, this paper explores the performance of the WordNet synonyms approach on the top ten categories of the Reuters-21578 “Topics” dataset to provide an insight as to why this approach is able to outperform the statistical approaches.

The macro-averaged and micro-averaged f-measure is reported for the experiments carried out on the top ten categories of the Reuters-21578 “Topics” dataset. The

WordNet synonyms approach is able to outperform Chi2 and IG in all the experiments performed except at 10 and 20 terms. We will analyse why this is so in the following sections. The results for the macro-averaged f-measure and the micro-averaged f-measure for the Reuters-21578 top ten categories are shown graphically in Figure 2 and Figure 3 respectively.

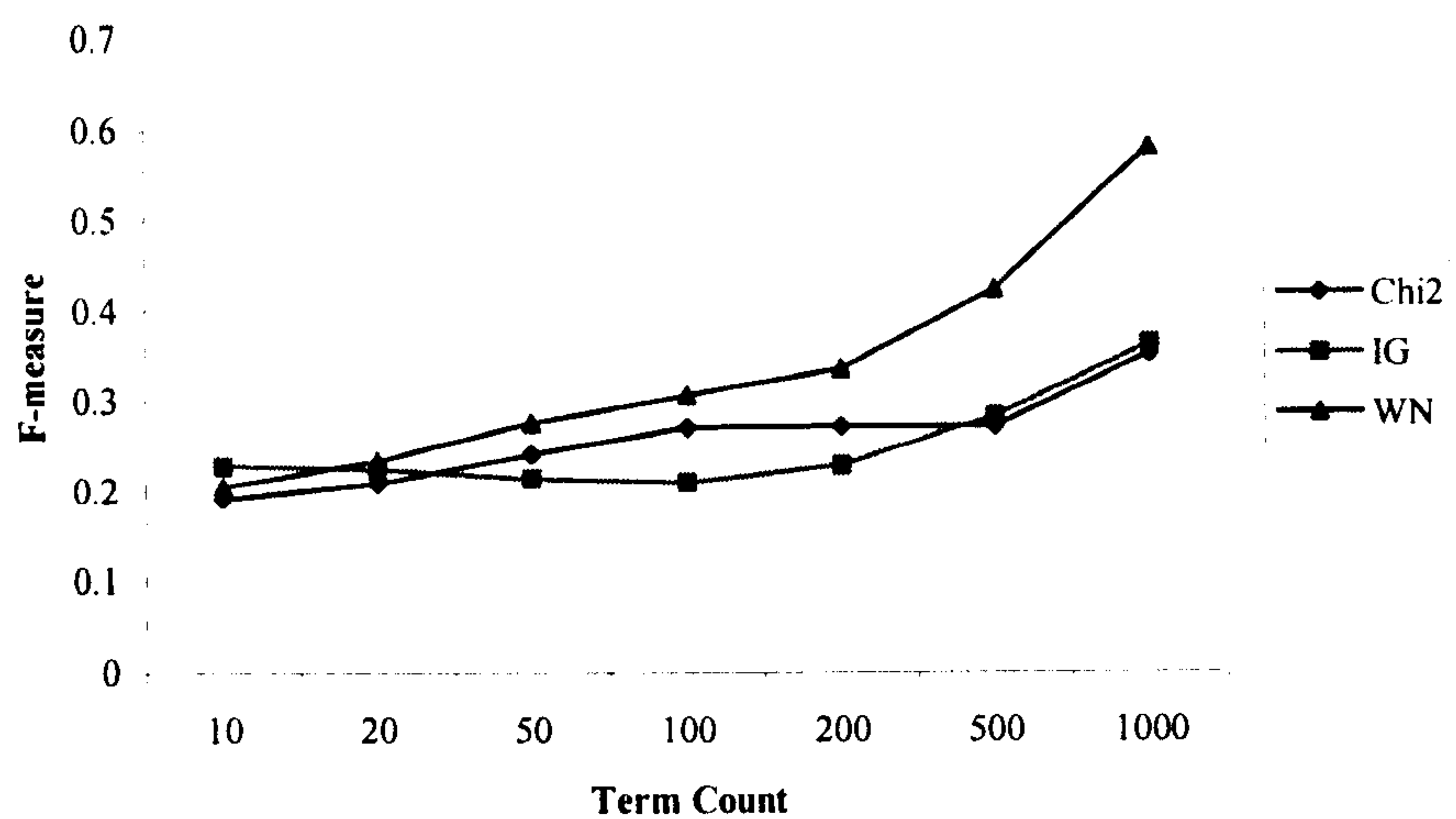


Figure 1. Macro-averaged f-measure comparison between Chi2, IG and WN for the Reuters-21578 dataset

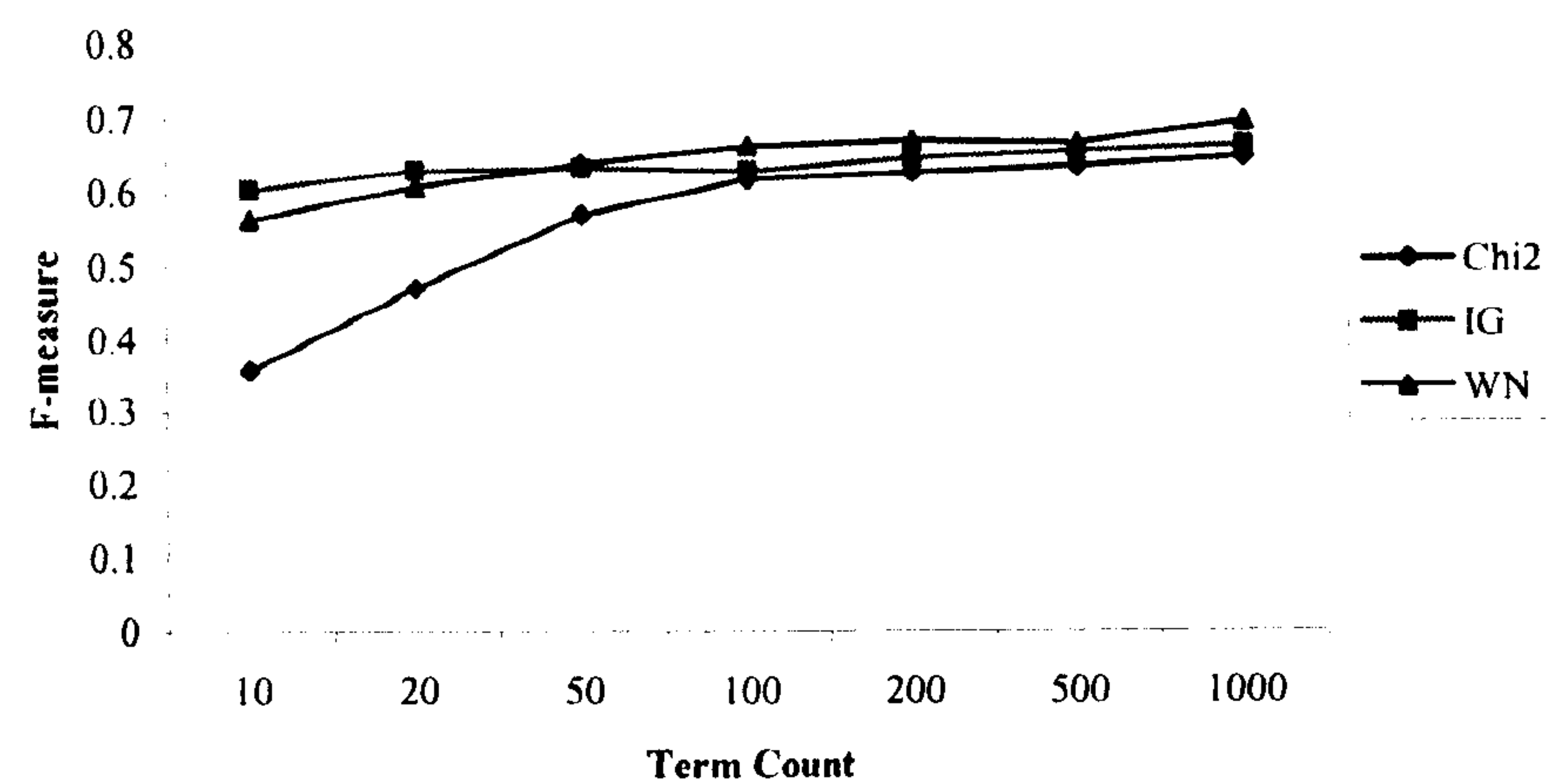


Figure 2. Macro-averaged f-measure comparison between Chi2, IG and WN for Reuters-21578 top ten categories

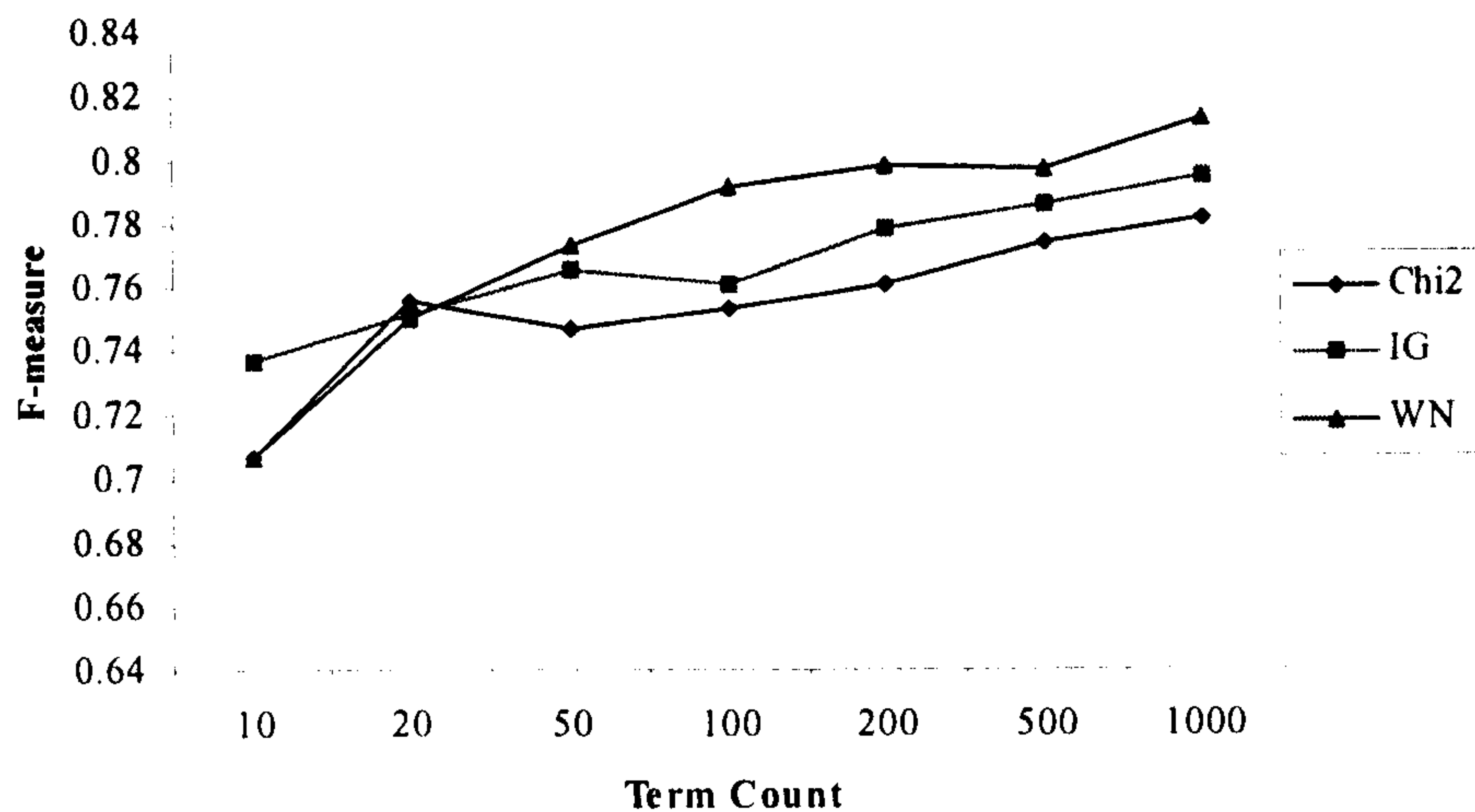


Figure 3. Micro-averaged f-measure comparison between Chi2, IG and WN for Reuters-21578 top ten categories

By choosing semantically related features, closer concepts can be identified within the category. This point will be illustrated in detail when we evaluate the performance of the categories individually. With the use of WordNet in feature selection, insignificant words can be filtered. These consist of non-English words, wrongly spelt words, insignificant abbreviations and names. Therefore, insignificant terms found in the dataset, such as, “twa”, “shr”, “lme”, “natl” and so on will be filtered and omitted. Terms like “lt”, “govodi” and “pik” are actually meaningless in representing a category. If human experts were asked to choose a set of words to represent a category, they would not have chosen those words that are mentioned. However, statistical feature selection techniques will include such words if the statistical data show that the words are significant statistically. These statistical techniques do not take into consideration whether a term is misspelt or is reflective of a category. By incorporating WordNet in feature selection, we can actually tackle this problem by filtering these insignificant terms and at the same time, make use of the available synonym relationship in WordNet to identify semantic features in a category.

From Figure 2 and 3, we can see that the WordNet synonyms approach generally performs better than the statistical feature selection methods, Chi2 and IG with the exception of results at term count 10 and 20. At term count 50 and above, the WordNet synonyms approach outperforms both Chi2 and IG.

In the individual category evaluation, we found that both category “Acq” and “Earn” have the highest number of training examples, thus producing the two highest values of f-measure. On the other hand, categories with a low number of training examples perform less effectively. This was proven in the correlation test that was carried

out to see whether the number of training examples actually affects the performance of the classifier. We observe a strong correlation with a positive value of 0.81. Furthermore, another factor is observed, where categories with more synonyms found in WordNet, performed better. The correlation for the number of synonyms found and the performance of the classifier has the positive value of 0.88. Thus, both categories “Acq” and “Earn” again outperformed the other categories with “Corn” having the lowest performance.

To look into why WN performs less effectively at 10 and 20 terms, the results for the individual categories are analyzed. Table 2 shows the results for each category at term count of 10 and 20. The results printed in bold shows the highest of three f-measures.

Table 2. F-measure for each category for Chi2, IG and WN

Category/Term Count		Chi2	IG	WN
Acq	10	0.760	0.790	0.581
	20	0.847	0.859	0.731
Corn	10	0.034	0.400	0.385
	20	0.094	0.346	0.408
Crude	10	0.342	0.723	0.718
	20	0.620	0.716	0.715
Earn	10	0.910	0.908	0.938
	20	0.915	0.922	0.942
Grain	10	0.013	0.581	0.529
	20	0.603	0.620	0.556
Interest	10	0.410	0.404	0.402
	20	0.444	0.561	0.410
Money-fx	10	0.306	0.579	0.517

	20	0.474	0.608	0.595
Ship	10	0.022	0.649	0.442
	20	0.063	0.661	0.590
Trade	10	0.646	0.576	0.688
	20	0.562	0.593	0.702
Wheat	10	0.120	0.429	0.439
	20	0.075	0.396	0.421

From Table 2, it is noted that IG generally performs better in categories “Acq”, “Crude”, “Grain”, “Money-fx” and “Ship” while WN performs better in categories “Earn”, “Trade” and “Wheat”. For category “Corn”, IG is better with 10 terms while WN performs better for 20 terms. IG only performs better in category “Interest” at 20 terms while Chi2 did better for 10 terms. Therefore, we will take a look at the top 20 terms for category “Acq” where IG performed better, category “Earn” where WN performed better and category “Interest” where Chi2 performed best at 10 terms. Listed in Table 3, Table 4 and Table 5 respectively, are the top 20 terms for category “Acq” , “Earn” and “Interest” for Chi2, IG and WN.

Table 3. Top 20 terms for category “Acq”

Feature selection scheme	Top 20 terms for category “Acq”
Chi2	shares, offer, lt, stake, merger, cts, acquisition, company, inc, acquire, net, loss, corp, usair, common, mln, unit, shr, stock, sell
IG	cts, shares, net, loss, lt, shr, offer, stake, company, merger, tonnes, wheat, acquisition, trade, inc, mln, profit, qtr, corp, acquire
WN	title, company, shares, pct, corp, offer, share, stock, stake, acquisition, merger, common, unit, buy, agreement, board, shareholders, sell, american, investment

Table 4. Top 20 terms for category “Earn”

Feature selection scheme	Top 20 terms for category “Earn”
Chi2	vs, cts, net, loss, mln, shr, profit, revs, qtr, oper, dlrs, note, dividend, th, shrs, avg, div, earnings, prior, lt
IG	vs, cts, net, loss, mln, shr, profit, revs, qtr, dlrs, oper, note, dividend, trade, th, shrs, avg, tonnes, div, said
WN	vs, text, title, cts, net, loss, profit, revs, share, company, pct, note, corp, quarter, dividend, record, th, earning, stock, tax

Table 5. Top 20 terms for category “Interest”

Feature selection scheme	Top 20 terms for category “Interest”
Chi2	rate, rates, bank, money, fed, prime, pct, banks, stg, lending, base, funds, market, discounts, cut, bills, sterling, bundesbank, england, band
IG	rate, rates, bank, pct, money, fed, prime, market, stg, banks, loss, net, lt, cut, company, funds, dlrs, base, lending, discount
WN	pct, rate, market, fed, cut, prime, government, week, funds, federal, term, base, dealers, central, month, bills, reserve, february, growth, dollar

The similar terms between Chi2, IG and WN for category “Acq” are removed to see the combination of words that are unique to each feature selection method. From the 13 unique terms obtained, both Chi2 and IG have one term that strongly represents the “Acq” category, which is “acquire”. With the WN approach however, “acquire” was not chosen because it is a verb and our approach only considers nouns. Together with other terms chosen by Chi2 and IG, their performance are better than WN, as abbreviations such as “lt” and “shr” are not chosen by WN although they are statistically strong terms.

For category “Earn”, 11 unique words are identified. Out of the 11 unique terms, Chi2 has 9 abbreviations and IG has 8 abbreviations, which are not proper English abbreviations. However, Chi2 and IG choose these terms because they give good statistical values. The only word that is reflective of the category “Earn” is “earnings” for Chi2 and “trade” for IG. In the case of WN, it has words like “share”, “company”, “earning”, “stock” and “tax” that are reflective of the category “Earn”. Therefore, in this category, WN is able to perform better and has a set of terms that better describe this category.

There are 12 unique terms for category “Interest”. From the 12 unique terms, Chi2 has terms like “sterling”, “bundesbank” and “england” that gives high statistical value. IG has 5 terms similar to Chi2 and also an abbreviation “lt” that is not chosen by WN. They also have one similar abbreviation that is not chosen by WN, “stg”. This explains why they have better performance when compared to WN, as WN does not contain the abbreviations that gave good statistical values in machine learning.

Generally, the WordNet synonyms approach performs better from 50 terms onwards because it is capable of choosing terms that are more reflective of a category. With fewer terms at 10 and 20 terms, IG has a lot of terms

that has strong statistical values. Although those terms in general, are not representative of the category concerned, it is able to perform well in machine learning because of its statistics. In the case of WordNet synonyms approach, for term size 10 and 20, WordNet perform worse than the statistical methods because it ignores the terms like “It” and “stg” that gives good statistical performance but are nonetheless meaningless in the language context. Therefore, the performance of the WordNet synonyms approach here is marginally reduced. However, the significance of these abbreviations becomes less significant as the number of terms increased, as we can see from both Figure 2 and 3 that WordNet synonyms approach performs better from 50 terms onwards.

To determine the significance of using WordNet synonyms approach for feature selection, we compare the

use of WordNet nouns. From Figure 4, we can see that WN1 performs slightly better than WN at term count 10, 20 and 500. We attribute this to the existence of nouns with high frequencies that gives good statistical values for machine learning. However, in Figure 5, WN outperformed WN1 at all term counts. Generally, the WordNet synonyms approach can perform better than the WordNet nouns approach. Another plus point of WN over WN1 is that WN, which chooses only synonyms, reduces more of the features for each category. We have observed that WN has a feature space that is more than 40% smaller than that of WN1 but still gives an overall result that is better than WN1. This again helps in the dimensionality reduction problem.

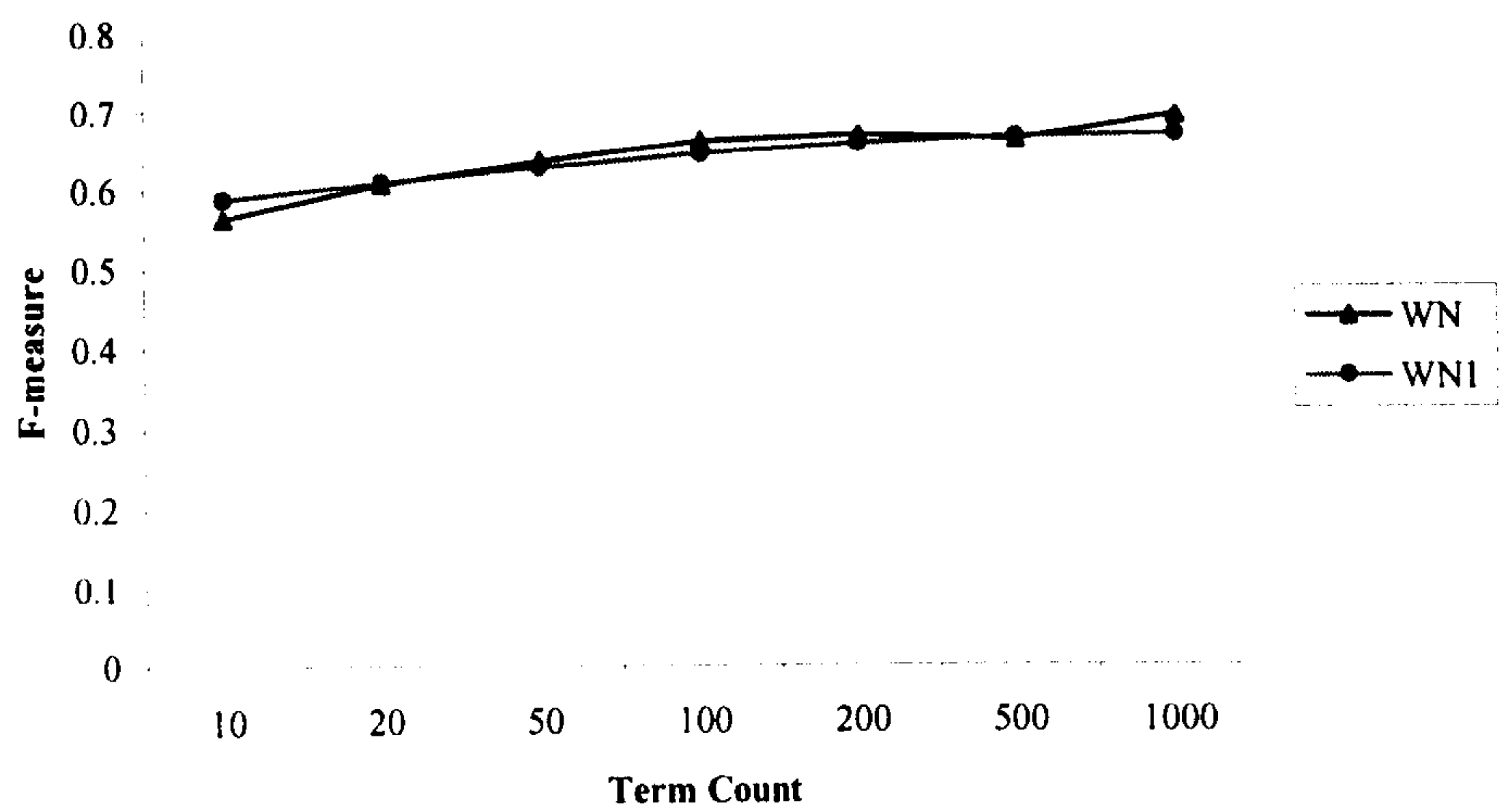


Figure 4. Macro-averaged f-measure comparison between WN and WN1 for Reuters-21578 top ten categories

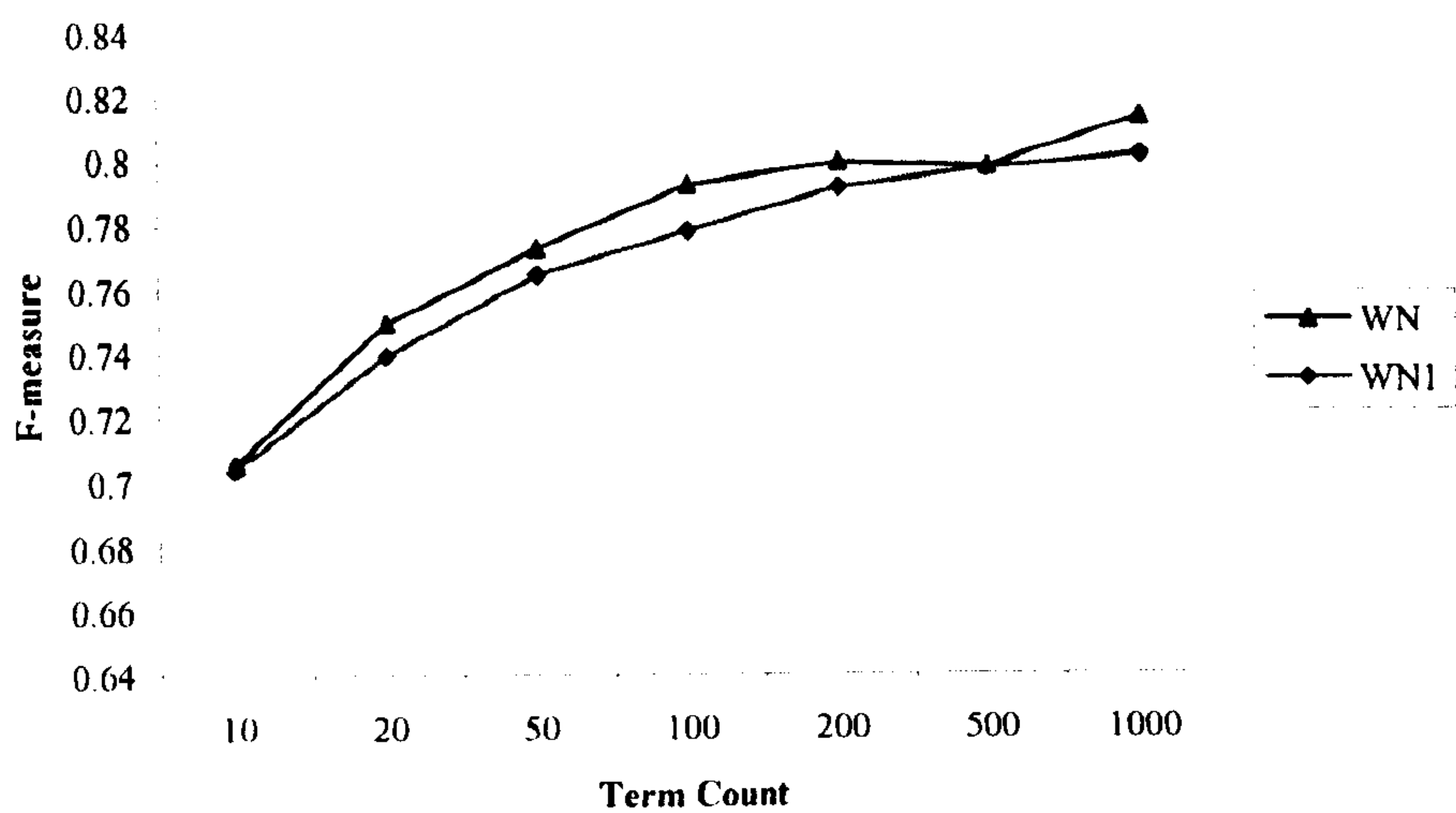


Figure 5. Micro-averaged f-measure comparison between WN and WN1 for Reuters-21578 top ten categories

6. Conclusion

This paper highlights that the WordNet based semantic feature selection is promising. From the experiments conducted, we can see that the WordNet based semantic approaches can improve the feature selection process in text categorization. It is able to provide a better set of features for category representation as the Reuters-21578 collection has many terms that are covered comprehensively by WordNet. The categorical sense disambiguation is promising as an effective approach to automatically disambiguate semantically related terms. Therefore, semantic feature selection is seen to be promising. Further works are being carried out to derive more meaningful features by employing hypernyms in WordNet.

However, for domain specific areas containing technical words, domain specific dictionaries need to be incorporated to complement the terms in WordNet.

7. Reference

- [1] Miller G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J., "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, Vol 3, No.4 (Winter 1990), pp. 235-244.
- [2] Li, X., Szpakowicz, S., Matwin, S., "A WordNet-based Algorithm for Word Sense Disambiguation", in proceedings of the IJCAI-95, 1995, pp. 1368-1374.
- [3] Sebastiani, F., "A Tutorial on Automated Text Categorisation", (Analia Amandi and Ricardo Zunino, editors), in proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR., 1999, pp. 7—35.
- [4] Yang, Y. and Liu, X., "A re-examination of text categorization methods", in proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, pp. 42—49.
- [5] Yang, Y. and Pedersen, J. A., "Comparative Study on Feature Selection in Text Categorization", in proceedings of the 14th International Conference (ICML 97), Nashville, TE, USA, 1997, pp. 412—420.
- [6] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", in the DEXA Workshops, 2002, pp. 241-245.
- [7] Mihalcea, R. and Moldovan, D., "Semantic indexing using WordNet senses", in proceedings of the ACL Workshop on IR & NLP, Hong Kong, 2000.
- [8] Gonzalo, J., Verdejo, F., Chugur, I. And Cigarran, J., "Indexing with WordNet Synsets can Improve Text Retrieval", in proceedings of the Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems, 1998, pp. 38-44.
- [9] Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.
- [10] Aas, K and Eikvil, L. "Text Categorization: a survey", Technical Report #941, Norweigan Computing Center, 1999.
- [11] Apte, C., Damerau, F. J. and Weiss, S. M., "Automated learning of decision rules for text categorization", *ACM Transactions on Information Systems*, 12, 3, 1994, pp. 233 – 251.