uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Fadi Abu Sheikha

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.C.S.

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Analysis and Generation of Formal and Informal Text

TITRE DE LA THÈSE / TITLE OF THESIS

Diana Inkpen

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Babak Esfandiari

Stan Szpakowicz

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Analysis and Generation of Formal and Informal Text

by

Fadi Abu Sheikha

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements for the degree of
Master in Computer Science

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

© Fadi Abu Sheikha, Ottawa, Canada, 2010

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-74178-8
Our file Notre référence
ISBN: 978-0-494-74178-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

In this thesis, we discuss an important issue in computational linguistics: distinguishing between formal and informal style of texts, in document classification and in text generation. There is a need to identify formal texts and informal texts automatically. In addition, there is a need of having a computer system that could generate correct English texts in formal or informal style. Therefore, we propose to use two main techniques in order to solve the two tasks. The first technique is to build a model that can be used to classify any text or sentence as having formal or informal style. The second technique is based on natural language generation (NLG) and it generates correct English sentences with formal or informal style.

In order to achieve our goals, we start by studying the main differences between formal and informal style and summarize their characteristics. In addition, we manually collect parallel lists of formal versus informal words, phrases, and expressions from different sources that will be used for our proposed work. Then, we build our model for the classification task by using machine learning technique in order to classify texts and sentences into formal and informal style. The evaluation results show that our model is able to predict a class of formal / informal for any text or sentence with high accuracy. After that, we build our system that can generate formal and informal sentences by using NLG techniques. The evaluation results on a sample of generated sentences show that our NLG system can produce high-quality sentences in formal or informal style.

The main contribution of this work consists in designing a set of features that led to good results for both tasks: text classification and text generation with different formality levels.

Contents

List of Figures v

List of Tables vi

1 Introduction 1
- 1.1 Motivation 1
- 1.2 Thesis Goals 2
- 1.3 Contributions 2
- 1.4 Outline 3

2 Formal and Informal Language Style 4
- 2.1 Characteristics of Informal versus Formal Style 4
  - 2.1.1 Main characteristics of the informal style 5
  - 2.1.2 Main characteristics of the formal style 5
- 2.2 Informal versus Formal Lists 6
  - 2.2.1 Informal/Formal list of words and phrases 6
  - 2.2.2 Contractions Lists 7
  - 2.2.3 Abbreviation Lists 8
  - 2.2.4 Summary 9

3 Background and Related Work 10
- 3.1 Text Classification 10
  - 3.1.1 Machine Learning 10
    - 3.1.1.1 Decision Tree Learning Algorithm 11
    - 3.1.1.2 Naïve Bayes Learning Algorithm 12
    - 3.1.1.3 Support Vector Machine Learning Algorithm 12
  - 3.1.2 Text Classification by Formality and Genre 12
- 3.2 Text Generation 15
  - 3.2.1 Natural Language Generation (NLG) 15
    - 3.2.1.1 NLG Stages 15
    - 3.2.1.2 NLG Architecture 16
  - 3.2.2 SimpleNLG Package 17
  - 3.2.3 NLG and SimpleNLG 18
  - 3.2.4 Some NLG Systems 19

4 Automatic Classification of Texts by Formal/Informal Style 20
- 4.1 Data Set 20
  - 4.1.1 Document Level 20

4.1.2 Sentence Level ... 22
4.2 Features ... 23
4.2.1 Bag-Of-Words Representation ... 23
4.2.2 Representation based on Text Properties ... 23
4.3 Classification Algorithms ... 26
4.4 Experiments and Evaluation ... 26
4.4.1 Document Level ... 27
4.4.1.1 BOW Features ... 27
4.4.1.2 Text Property Features ... 27
4.4.2 Sentence Level ... 38
4.4.2.1 BOW Features ... 38
4.4.2.2 Sentence Property Features ... 38
4.5 Discussion ... 41

5 Automatic Generation of Formal and Informal Sentences ... 42
5.1 Generating formal versus informal sentences ... 43
5.1.1 Formality Features ... 43
5.1.2 Implementation ... 44
5.2 Results and Evaluation ... 48
5.3 Discussion ... 52

6 Conclusion and Future Work ... 53
6.1 Conclusion ... 53
6.2 Future Work ... 54

Bibliography ... 55

A. Informal versus Formal Lists ... 61

B. Classification Results on Medical Documents ... 84

C. The Generated Sentences from the Proposed System ... 86

List of Figures

Figure 3.1: An example of Decision Tree (Witten & Frank, 2005) 11
Figure 3.2: Architecture of an NLG system (Reiter and Dale, 2000). 17
Figure 3.3: A sample code implemented in Java using SimpleNLG package to generate a correct English sentence 18
Figure 4.1: A sample of informal text file extracted from Corpus of Late Modern English (prose). 21
Figure 4.2: A sample of formal text file extracted from Reuters Corpus. 22
Figure 4.3: The confusion matrix of the Decision Tree which shows the distribution of the actual and the predicted classes on document level using 10-fold cross validation. 29
Figure 4.4: Decision Tree visualization for text classification using 10-fold cross validation without feature selection. 31
Figure 4.5: Decision Tree visualization for text classification using 10-fold cross validation after applying a feature selection. 34
Figure 4.6: Decision Tree visualization for text classification trained on the whole training data (we applied it on the test set). 37
Figure 5.1: The first screen in our proposed NLG system. 45
Figure 5.2: The second screen in our proposed NLG system. 45
Figure 5.3: The third screen in our proposed NLG system. 46
Figure 5.4: The fourth screen in our proposed NLG system. 46
Figure 5.5: The fifth screen in our proposed NLG system. 46
Figure 5.6: The sixth screen in our proposed NLG system. 47
Figure 5.7: The seventh screen in our proposed NLG system. 47
Figure 5.8: The eighth screen in our proposed NLG system. 47
Figure 5.9: The ninth screen in our proposed NLG system. 48
Figure 5.10: The last screen in our proposed NLG system that contains the generated sentence. 48
Figure 5.11: The annotation result for Judge1 with the precision for each class. 49
Figure 5.12: The annotation result for Judge2 with the precision for each class. 50

List of Tables

Table 2.1: Examples of informal vs. formal sentences for different characteristics. 6
Table 2.2: An example of formal versus informal list of words. 7
Table 2.3: An example of the contractions versus the full forms. 8
Table 2.4: An example of the abbreviations versus the full forms. 8
Table 4.1: Document classification results (using BOW features representation) of Decision Tree, Naïve Bayes, and SVM classifiers, using 10-fold cross validation. 27
Table 4.2: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation before applying feature selection. 32
Table 4.3: The features with their weights in our model based on InfoGain feature selection on document level classification. 33
Table 4.4: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after performing feature selection. 34
Table 4.5: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after removing the abbreviations feature. 35
Table 4.6: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after removing the phrasal verbs feature. 36
Table 4.7: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers on the training/test set. 37
Table 4.8: Sentence classification results (using BOW feature representation) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation. 38
Table 4.9: Sentence Classification results (using sentence property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation before applying feature selection. 39
Table 4.10: The features with their weights in our model based on InfoGain feature selection on sentence level classification. 40
Table 4.11: Sentence classification results (using sentence properties features) of Decision Tree, Naïve Bayes, and SVM classifiers using the training/test set. 41
Table 5.1: Samples of the generated sentences with the annotations from both judges. 49
Table 5.2: The accuracy for our system according to each judge. 50
Table 5.3: The number of sentences for which Judge1 and Judge2 agreed or did not agree. .. 51

Table A.1: Informal/formal parallel list of words, phrase, and expressions. 77
Table A.2: Parallel list of contractions and full forms. 80
Table A.3: Abbreviations and full forms parallel list. 83

Table B.1: Classification results of SVM, Decision Trees, and Naïve Bayes classifiers using 10-fold cross validation. 84
Table B.2: Detailed evaluation for each class, for the SVM classifier. 84

Table B.3: Our model's features with their InfoGain scores for the dataset of medical texts. 85

Table C.1: Informal and Formal sentences generated by our system using different parameters 92

Acknowledgments

I wish to express my sincere gratitude and appreciation to my supervisor Dr. Diana Inkpen for her years of patience and guidance of this thesis. Without her this work would not have been possible.

Also, I would like to thank Alistair Kennedy and Chris Fournier for their help in judging the generated sentences from our NLG system.

I want thank to my lab colleagues and all the members of the NLP research group for their helpful suggestions, support, and advice, especially Prof. Stan Szpakowicz, Fazel Keshtkar, Fernanda Caropreso, Oana Frunza, and Martin Scaiano. They were really nice and available when I needed their help.

I want to thank Soumaya Chaffar for her helpful comments on the draft version.

I want to thank my parents, brothers, and sisters for their love, emotional support, and constant encouragement during all stages of my life.

Last but not least, I would like thank my wife, Sadoof Jaber, for her unrivaled support and encouragement during the many long days of my Master's program. She has been, always, my pillar, my joy and my guiding light.

Chapter 1

Introduction

In this thesis, we introduce two main important techniques that take into account the differences between formal text style and informal text style. The first technique is automatic text classification that can classify any text or sentence into formal or informal style. The second technique is automatic text generation that can generate texts that are formal or informal, based on the user preferences.

There are linguistic studies that state that there are different levels of formality (Hayakawa, 1994). We focus on a coarse-grained level, formal and informal style, but finer-grained levels are possible (e.g., informal, less formal, formal, extremely formal).

Adopting formality in writing or distinguishing formal texts from informal texts is not an easy task. There are different characteristics which affect formality, such as use of pronouns, active and passive voice, and sentence/word length. The choice of words (known as lexical choice) might be the most important style marker. In general, using longer words and Latin origin verbs leads to formal text, while using phrasal verbs and idioms to informal text. However, there are also many words that could be used in writing for both styles at the same time.

1.1 Motivation

The first motivation for our work consists in the need of identifying and distinguishing possible differences between formal and informal styles of texts. This need has increased nowadays, especially since more and more people are using the Internet as a main resource

for their research. Whereas scientific research papers are written in formal style, most of blogs and non-scientific papers are written in informal style.

The second motivation to propose this work is the need for a software tool that helps people to generate formal or informal texts. One of the difficult things of writing in English is the knowledge of how to adapt to formal or informal situations. Some situations (such as applying for a job) are likely to be formal, whereas others (such as emailing a friend or family member) are more likely to be informal. The real problem in writing in English is to know what words, phrases, or expressions should be used. There are some words, phrases, and expressions that are either formal or informal; therefore, if the wrong word is chosen, then the reader may think we are being either too friendly or too formal.

1.2 Thesis Goals

In this thesis we have several goals:

- Provide a good explanation and understanding of the differences between the formal and the informal style.
- Build a model that can be used to classify any text or sentence as formal or informal style.
- Build a system that can help in producing correct English sentences using Natural Language Generation (NLG) techniques, with a specified style: formal or informal.

1.3 Contributions

In this thesis, we make the following contributions:

- We explain the main differences between formal and informal style by summarizing the main characteristics of both styles and providing examples for each characteristic. In addition, we provide very useful parallel lists of formal versus informal words, phrases, and expressions that help to differentiate between both styles.
- We use machine learning techniques to classify text documents and sentences into formal and informal style. We have designed a very good set of features for this task. Evaluation results show that our model is able to predict a class of formal and informal style for any text with high accuracy.

- We present a system of generating formal and informal sentences. Our system is based on natural language generation technique. We have designed a set of features that help to generate sentences in formal or informal style. Evaluation results show that our system can generate formal and informal style successfully, with high accuracy.

Our work could be a part of more complex natural language processing systems.

1.4 Outline

The rest of the thesis is organized as follows: Chapter 2 discusses the main characteristics of the formal and the informal style. We also present our parallel lists of formal and informal words, phrases, and expressions, collected from several sources. In Chapter 3, we review some existing methods for text classification by style and by genre. In addition, we explain NLG, the SimpleNLG package (Gatt and Reiter, 2009), and we discuss related work for text generation. Then, Chapter 4 presents our approach of classifying texts and sentences by formality using machine learning techniques. In Chapter 5, we propose our system that generates formal and informal sentences, using NLG techniques. Finally, we conclude in Chapter 6 with a brief discussion of future work.

Chapter 2

Formal and Informal Language Style

In this chapter, we explain the main characteristics of informal versus formal style. We also present parallel lists of words, phrases, and expressions for both styles, which we collected from different sources. The understanding of the main differences between the styles will help to build models based on some of these main characteristics, for document/sentence classification tasks. In addition, it will help the sentence generation task by implementing some of these characteristics in our NLG system.

2.1 Characteristics of Informal versus Formal Style

In this section, we briefly explain and summarize the main characteristics of informal style versus formal style, as we found them described in (Dumaine and Healey, 2003; Obrecht, and Ferris, 2005; Akmajian et al., 2001; Park, 2007; Zapata, 2008; Siddiqi, 2008; Redman, 2003; Rob S. et al., 2008; Pavlidis, 2009; Obrecht, 1999). We need to understand the main differences between the styles, in order to be able to distinguish between them, and to identify the style of any text. These characteristics are used, as it will be explained in Chapter 4 and 5:

- for automatic classification of texts by Formal/Informal style in order to extract the values of these features from texts and to predict the class for new text documents and sentences.
- for building templates to generate sentences based on those characteristics.

2.1.1 Main characteristics of the informal style

The informal style has the following characteristics:

- It uses a personal style, using the first and second person ("I", "you") and the active voice (e.g., "I have noticed that...").
- It uses short simple words and sentences.
- It uses Contractions (e.g., "won't").
- It uses many abbreviations (e.g., "TV").
- It uses many phrasal verbs within the text (e.g., "find out").
- The words that express rapport and familiarity are often used in speech, such as "brother", "buddy", and "man".
- It uses a subjective style, expressing opinions and feelings (e.g., "pretty", "I feel").
- It uses vague expressions, it uses personal vocabulary and colloquial (slang words are accepted in spoken not in written text (e.g., "wanna" = "want to")).

2.1.2 Main characteristics of the formal style

The formal style has the following characteristics:

- It uses an impersonal style, using the third person ("it", "he", and "she") and often the passive voice (e.g., "It has been noticed that....").
- It uses complex words and sentences to express complex points.
- It does not use contractions.
- It does not use many abbreviations. Though, here are some abbreviations used in formal texts, such as titles with proper names (e.g., Mr.) or short names of methods in scientific papers.
- It uses appropriate and clear expressions, precise education, business, and technical vocabulary (Latin origin).
- It uses polite words and formulas such as "Please", "Thank you", "Madam", "Sir".
- It uses an objective style, using facts and references to support an argument.
- It does not use vague expressions and slang words.

In Table 2.1, we show examples of sentences that characterize informal style versus formal style.

Feature	Informal Style	Formal Style
Contractions	Use Contractions: e.g., "Many patients don't listen to their doctors."	Avoid Contractions: e.g., "Many patients do not listen to their doctors."
Phrasal Verbs	Use phrasal verbs: e.g., "I looked up information about nursing positions."	Avoid phrasal verbs: e.g., "I researched information about nursing positions."
Personal/Impersonal Pronouns	Use personal pronouns: e.g., "I think this is an effective plan."	Use impersonal pronouns: e.g., "This could be an effective plan."

Table 2.1: Examples of informal vs. formal sentences for different characteristics.

2.2 Informal versus Formal Lists

In this section, we present our parallel lists of informal versus formal words, phrases, and expressions. These lists were collected manually from different sources: the first list is for formal versus informal words and phrases, the second list is for most of the contractions in English, and the third list is for some of the common abbreviations in English. These lists are considered as important features for our models, for the text/sentence classification task or for the sentence generation task.

2.2.1 Informal/Formal list of words and phrases

This is a parallel list for informal versus formal words and phrases. We collected this list manually from different sources: (Gillett et al., 2009; Park, 2007; Redman, 2003: Rob et al., 2008). In addition, we obtained a new list that was extracted manually by Brooke et al. (2010) from the dictionary of synonyms *Choose The Right Word* (Hayakawa, 1994). There are some words in each list (formal/informal) that might be used for both styles. In order to

see the importance of the list for both styles, we counted the frequency of each word from the formal list in our dataset of formal texts (mentioned in section 4.1). We also counted the frequency of each word from the informal list in our dataset of informal texts. The words with very low or zero frequency might not be important for our model, in case CTRW states that they are both formal (but one is more formal than the other) or both informal (but one is more informal than the other). We manually detected them and removed them from the list. Other words with zero frequency (as long as they were pairs consisting of an informal word and its formal correspondent) were kept in the list because they might occur in other test data on which we will run our system in the future.

Table 2.2 below shows samples of this parallel list; also Table A.1 in Appendix A shows the full list for both styles, with the frequencies for each word in the list, as counted in each dataset.

Informal	**Formal**
about	approximately
anybody	anyone
ask for	request
boss	employer
buy	purchase
enough	sufficient
get	obtain
go up	increase
have to	must

Table 2.2: An example of formal versus informal list of words.

2.2.2 Contractions Lists

This is a parallel list for most of the contractions in English (short forms) that represent the informal style versus the full forms of the contractions that represent the formal style. We obtained this list manually from (Redman, 2003; Garner, 2001; Pearl Production, 2005; Woods, 2010). In Table 2.3 below, we show samples of the parallel list of the contractions versus the full forms, also Table A.2 in Appendix A shows the full list of the contractions.

Informal	Formal
aren't	are not
can't	cannot
didn't	did not
hadn't	had not
hasn't	has not
I'm	I am

Table 2.3: An example of the contractions versus the full forms.

2.2.3 Abbreviation Lists

This is a parallel list for some of the most common abbreviations in English that represent the informal style versus the full forms that corresponds to these abbreviations as used in formal style. However, there are some abbreviations that are acceptable in formal texts (Obrecht, 1999). We will test in our experiments the importance of the abbreviations as a feature for informal style in our model.

We collected this list manually from (Redman, 2003; Gibaldi, 2003; Pearl Production, 2005). In addition, we extracted manually more pairs from *A Corpus of Late Modern English* prose (Denison, 1994); we will use a part of this corpus as our data set for the classification task in Chapter 3. Table 2.4 below shows samples of pairs from the parallel list of the abbreviations and the full forms; Table A.3 in Appendix A shows the list of the abbreviations that we have collected from different sources.

Informal	Formal
asap	as soon as possible
Corp.	Corporation
grad.	Graduate
Tue.	Tuesday
HR	Human Resources
USA	United States of America

Table 2.4: An example of the abbreviations versus the full forms.

2.2.4 Summary

We presented in this chapter the main characteristics of formal and informal style. In addition, we introduced three parallel lists for both styles: a list of informal versus formal words, a list of contractions versus full forms, and a list of abbreviations versus full forms. These characteristics of the styles and the parallel lists will be used to build the main features for document/sentence classification models and for the sentences generation system.

Chapter 3

Background and Related Work

In this chapter, we present the background information and related work concerning text classification and text generation. We discuss different methods and systems that are related to our work.

3.1 Text Classification

In this section, we briefly introduce the machine learning techniques, and briefly define three of machine learning algorithms which are Decision Tree, Naïve Bayes (NB), and Support Vector Machines (SVM). For more details about machine learning algorithms, see, for example, (Witten & Frank, 2005). After that, we discuss related work of text classification by formality and genre.

3.1.1 Machine Learning

Machine learning is concerned with designing computer models that can learn from examples. These models can be used for prediction, explanation, and understanding or the data. Machine learning can be implemented as supervised (using labeled training data) or unsupervised learning (using unlabeled data). We briefly explain here the idea of supervised learning, (automatic classification).

Supervised learning is a machine learning method where the system receives a training dataset consisting of many instances; each instance is represented with different parameter values (input) and a class (output). The parameter values are called attributes or features, and can be represented by numeric values (e.g., term frequency) or by nominal

values (e.g., Yes, No), Then, the system infers a mathematical function, which automatically maps an input signal to an output signal. Therefore, the system could discover the class (output) for any new instances with new attributes (inputs).

There are many algorithms used for supervised learning. We explain here three of them (Decision Tree, Naïve Bayes, and Support Vector Machine), which we used in our experiments to perform the classification tasks. We choose these three classifiers, in order to have a classifier that is interpretable to humans (Decision Tree), one that is known to work well with text (Naïve Bayes), and one that in known to achieve very good performance in general (SVM).

3.1.1.1 Decision Tree Learning Algorithm

The decision tree algorithm produces a tree-structured set of nodes which leads to a decision, where each node is either a decision node or a leaf node (indicates the value of the target attribute class). In Figure 3.1 below, we show an example of decision tree of playing a game outside based on the weather forecasting.

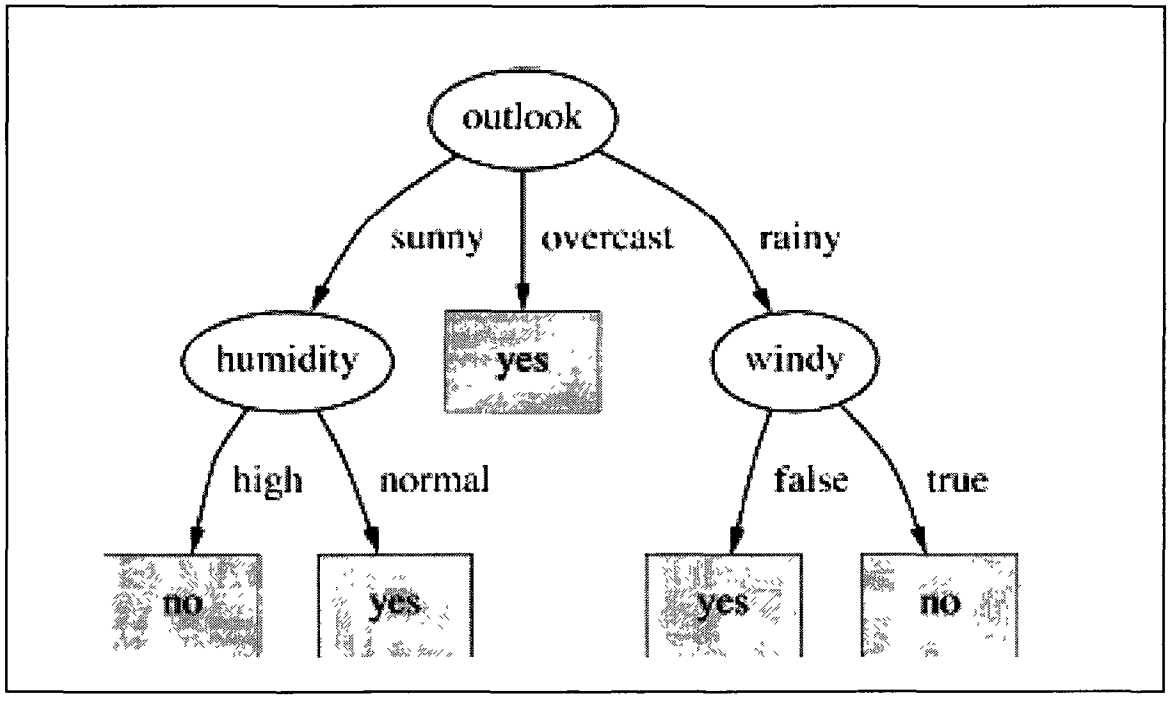

Figure 3.1: An example of Decision Tree (Witten & Frank, 2005)

Decision Tree learning algorithms can infer decision trees from datasets that contain examples of instances with different conditions (attributes) and outputs (classes or decisions).

The resulting Decision Tree is simple to understand and interpret; people are able to understand decision tree models after a brief explanation. This algorithm tends to perform well with large datasets, in a short time.

3.1.1.2 Naïve Bayes Learning Algorithm

The Naive Bayes method generates rules based on Bayes's rule of conditional probability. It uses the method of maximum likelihood. Naïve Bayes is a simple probabilistic classifier, it classifies based on the probability of each attribute independently. It can be trained very efficiently in a supervised learning setting. All attributes are used to make the decisions, and they are considered to be independent from each other and of equal importance.

3.1.1.3 Support Vector Machine Learning Algorithm

SVM is a set of related supervised learning methods that perform classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. It selects a small number of critical boundary instances called support vectors from each class and builds a linear discriminant function that separates them as widely as possible (Witten & Frank, 2005). After training, the SVM algorithm can predict if a new example falls into one class or the other.

3.1.2 Text Classification by Formality and Genre

As far as we are aware, there is little research on automatic text classification according to formal and informal style, however there is some work on automatic text classification by genre which is relevant to our work. Of course, there is a lot of research on classifying texts by their topic, but this does not apply in our case, since the texts can have different styles and be about the same topic. Similarly the texts can be about different topics and have the same style. Besides the classification by topic, there is research in classifying texts by author (from a set of possible authors), or by the gender of the author (male, female), by opinion (positive, negative, neutral), or by emotion classes (happy, sad, angry, etc.). These are also not directly

relevant to our work. We discuss here the related work on formal/informal classification, and also the work on genre classification.

Heylighen and Dewaele (1999) propose a method to determine the degree of formality for any text using a special formula. There formula is the F-score measurement which is based on the frequencies of different word classes (noun, adjective, preposition, article, pronoun, verbs, adverbs, interjection) in the corpus. The F-score increases with an increase of formality. In our work, we want to build a model based on some of the main characteristics of formal and informal style in order to classify any text into formal or formal, rather than a model based on the frequency of the words in both classes.

Dempsey, McCarthy, and McNamara (2007) propose to use phrasal verbs as text genre identifiers. Their results indicate that phrasal verbs significantly distinguish between both the spoken/written and the formal/informal dimensions. Their experiments are performed on the frequency of occurrence of phrasal verbs in spoken versus written text and in formal versus informal texts.

Brooke, Wang, and Hirst (2010) test in some detail several corpus-based methods for deriving real-valued formality lexicons. They quantify the formality of individual lexical items, assigning a formality score (FS) in the range −1 to 1 to each word. They collect two lists of words, one formal and one informal to use them as seeds for their dictionary construction method, and as test set for evaluation. They use the Brown corpus and the Switchboard corpus to collect documents that represent both styles. They evaluate their lexicons using relative formality judgments between word pairs. The results of their evaluation suggest that the problem is tractable but not trivial; they reached a good accuracy in distinguishing the difference in formality, using a small diverse corpus; however, they mention that they need both larger corpora and more sophisticated methods to capture the full range of linguistic formality.

Qu, Pietra, and Poon (2006) design a natural language processing (NLP) experiment for categorizing 120 blogs into four topic groups: personal diary, politics, news, and sports. These classes are related to the topic of the documents. They mention in their paper that machines learning methods cannot accurately categorize blogs because blog are written in an informal style; therefore it is a challenging task. The authors use standard statistical measures to classify blogs. Their baseline feature is term document frequency (unigrams) excluding

stop words. They try two more linguistic features which are the title text of every blog post and the anchor text over inbound links to the blog. Their proposed method has performed reasonably well in classifying blogs, they achieve 84% accuracy with unigrams features. This work is not directly relevant to our work, because it is about classification by topic. However, we mention this work here to show that our model can deal with informal texts such as blogs, using machine learning techniques, contrary to what is stated in this work.

Kennedy and Shepherd (2005) discuss one type of web page classification by genre: how to distinguish home pages from non-home pages as noise, and then to classify the home pages as personal, corporate, or organization. Their dataset consists in 321 web pages. They focus on the hard task of subgenre discrimination. The best accuracy they achieve is 71.4% is on personal home pages with a single classifier, manual feature selection, and without noisy pages.

Lim, Lee, and Kim (2005) suggest sets of features to detect the genre of web pages. They use 1224 web pages for their experiments. They investigated the efficiency of several feature sets to discriminate across 16 genres. These genres are: personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, official materials, FAQs, discussions, product specification, and informal texts (Internet fiction, private diaries, and so on). The classification efficiency has been tested on different parts of the web page space (title and meta-content, body, and anchors). The best accuracy they achieve is 75.7% with one of their feature sets, when applied only to the body and anchors.

Mason, Shepherd, and Duffy (2009) propose to use n-gram representation to automatically identifying web page genre. They use a corpus known as 7genre or 7-web-genre collection[1]. These genres are blogs, e-shops, FAQs, online newspaper front pages, listings, personal home pages, and search pages. The n-gram size ranges from 2 to 7 in increments of 1, and the number of the most frequent n-grams ranges from 500 to 5000 in increments of 500. Comparing to other approaches, they achieve better performance, with 94.6% accuracy on the web page genre classification task.

[1] http://www.itri.brighton.ac.uk/~Marina.Santini

3.2 Text Generation

In this section, we briefly explain the natural language generation techniques (Reiter and Dale, 2000), the SimpleNLG package (Gatt and Reiter, 2009), and we discuss the existing NLG systems that included stylistic variations.

3.2.1 Natural Language Generation (NLG)

Natural language generation is the process of constructing a natural language text from non-linguistic representation of information in order to meet specified communicative goals (McDonald, 1987). There are still many researches can be done in this domain, especially for who has the aim to build computer systems that can automatically produce correct texts in English, and other human languages (Reiter et al., 1995).

3.2.1.1 NLG Stages

By copy-paste of the canned text is possible to generate simple text, but the result might not be useful. Therefore, an NLG system which carries out all stages of such as planning and merging information is able to generate text which looks more natural and the stages for an NLG system can be consider as follows (Reiter and Dale, 2000):

1. Content determination: deciding what information to mention in the text.
2. Discourse planning: deciding the ordering and the structure of the information that should be conveyed.
3. Sentence aggregation: merging of similar sentences to improve readability and naturalness; e.g., the sentences "*The next train is the Ottawa Express*" and "*The next train leaves Montreal at 10am*" can be merged to: "*The next train, which leaves at 10am, is the Ottawa express*".
4. Lexicalization (Lexical choice): deciding which words, expressions, or phrase should be selected to express the concepts that will be constructed in the sentences. For example, deciding to select "go" or "depart"; this is the point when pragmatic issues are taken into consideration (e.g. should the text be formal or informal).
5. Referring expression generation: linking words in the sentences by introducing pronouns and other types of means of reference; For example, deciding to use "*in the*

Northern Isles and far northeast of mainland Scotland" to refer to a certain region in Scotland.

6. Linguistic realisation: adding all the above information, producing the surface document or text according to the syntactic rules, the morphological rules, and the orthographic rules (Dealing with casing, punctuation, and formatting). For example, adding full stop at the end of the sentence.

3.2.1.2 NLG Architecture

The NLG stages are usually combined in some way, to achieve most efficiency; this combination is called a pipelined architecture, introduced by Reiter and Dale (2000). We briefly explain the most important components of an NLG system, also shown in Figure 3.2: Architecture of an NLG system (Reiter and Dale, 2000). :

- Text Planner: this component gathers the information from the content determination and the discourse planning and then chooses an appropriate text structure.
- Sentence Planner: this component combines sentence aggregation, lexicalization, and referring expression generation.
- Linguistic Realiser (Surface Realiser): this component produces the actual text based on the lexical, grammatical, and orthographic rules and resources.

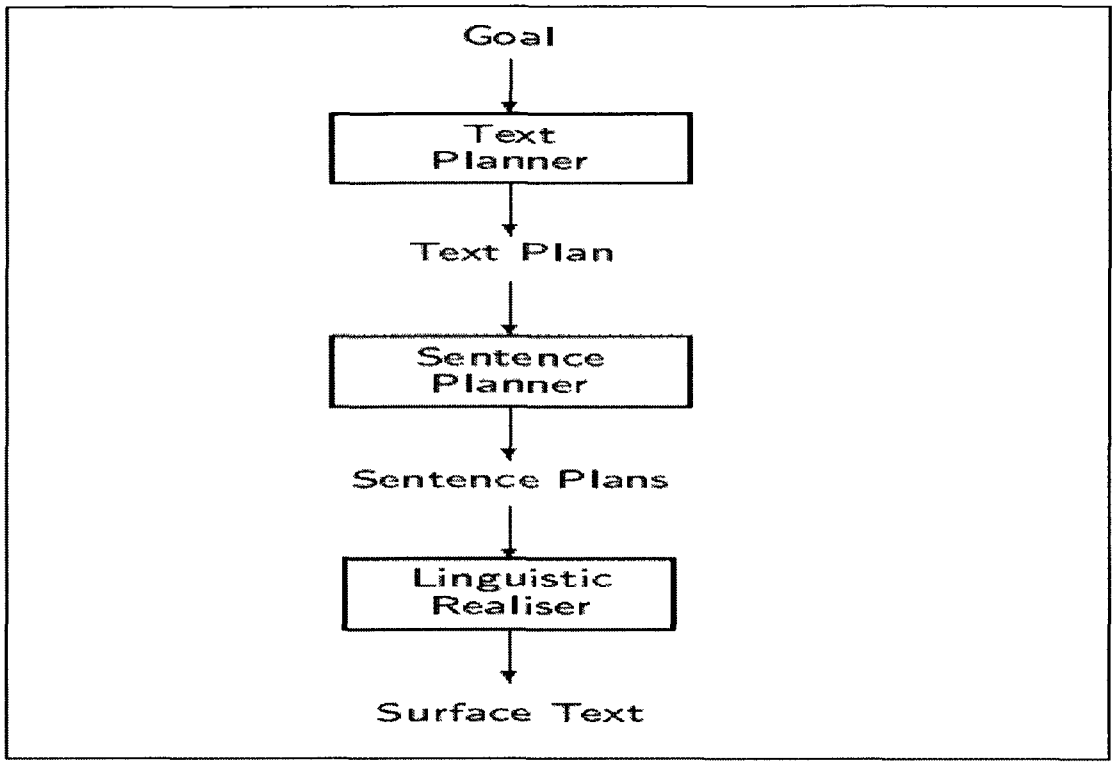

Figure 3.2: Architecture of an NLG system (Reiter and Dale, 2000).

3.2.2 SimpleNLG Package

The SimpleNLG[2] package (Gatt and Reiter, 2009) can be used to write a program which generates grammatically correct English sentences. It is a library, written in Java, which performs simple and useful tasks that are necessary for natural language generation.

SimpleNLG classes allow programmers to specify: the subject, the exact verb of a sentence that you want to appear in your sentence, and the object. e.g., subject: "my dog", verb: "chase", and object: "George". Once you have specify the content of the sentence it will be showed this information in SimpleNLG terms, and realize the parts of the sentence into a grammatical form. In our example, the resulting output would be: "*My dog was chasing George*", Figure 3.3 shows the code of SimpleNLG package that could generate that sentence.

Here are some of the tasks that SimpleNLG executed:

[2] http://www.csd.abdn.ac.uk/~ereiter/simplenlg/

- Capitalized the first letter of the sentence.
- Added the auxiliary "was" and made it agree with the subject.
- Added "ing" to the end of the verb (the progressive aspect of the verb is desired).
- Put all the words together in a grammatical form.
- Inserted the appropriate white space between the words of the sentence.
- Put a period at the end of the sentence.

SimpleNLG automates some of the more mundane tasks that all NLG systems need to perform. The main task that SimpleNLG performs is sentence realisation, which includes orthography, morphology, and simple grammar.

```
SPhraseSpec p = new SPhraseSpec();
p.setSubject(my dog);
p.setVerb(chase);
p.addComplement(George);
p.setTense(Tense.PAST);
p.setProgressive(true);
Realiser r = new Realiser();
System.out.println(r.realiseDocument(p));
```

Figure 3.3: A sample code implemented in Java using SimpleNLG package to generate a correct English sentence.

3.2.3 NLG and SimpleNLG

In section 3.2, we explained what NLG is and how the structure and stages of NLG look like. Now, we show what SimpleNLG does. Based on (Reiter and Dale, 2000) and SimpleNLG package, SimpleNLG perfectly performs Surface Realisation[3] which is one of the components of an NLG System. The Surface Realiser does the following tasks:

- Linguistic realisation: This component uses the grammar rules to convert abstract representations of sentences into actual text.
- Structure realisation: converts sentences and paragraphs into mark-up symbols and display the text.

[3] http://www.ling.helsinki.fi/kit/2008s/clt310gen/docs/simplenlg-tutorial-v37.pdf

3.2.4 Some NLG Systems

There are many NLG systems implemented to generate texts for specific purposes. Many of them are commercial systems. For example, the Forecast Generator (FOG) system was designed in 1992 by CoGenTex[4] to generate weather reports in English and French; the inputs of the system were graphical and numerical weather depictions (Goldberg et al., 1994).

We discuss here related work in NLG systems that take in consideration generating text under pragmatic constraints, especially according to style. As far as we are aware, there are only a few researchers who investigated producing text with varied styles.

Hovy (1988, 1990) introduced an NLG system called PAULINE, which is considered as one of the earliest examples of Natural Language Generation systems. Hovy proposed to generate text under pragmatic constraints, including formality. Although small scale, his experiments generated the same text in different styles, to achieve different effects on the reader, and incorporated some pragmatics into language generation. He suggested using different words to generate different styles.

Reiter et al. (2003) present the STOP system that was developed in University of Aberdeen for the British Health Services; it generates tailored letters to help people stop smoking. The STOP system makes the text friendlier by adding more empathy; it also makes the text easier to read for people with poor reading skills (Reiter et al. 2003).

[4] http://www.cogentex.com/

Chapter 4

Automatic Classification of Texts by Formal/Informal Style

In this chapter, we present the first part of our work which deals with an important issue in computational linguistics: classifying sets of documents into formal or informal style. We used machine learning techniques in order to classify texts into formal and informal style. We performed the classification experiments at the documents level and at the sentence level. We studied the main characteristics of each style, in order to train a system that can distinguish between them. We built our data set by collecting documents for both styles (formal and informal) from different sources. We performed pre-processing tasks on the collected documents to extract features that represent the main characteristics of both styles; these features were very useful for differentiating the two styles. We tested several classification algorithms, namely Decision Tree, Naïve Bayes, and Support Vector Machines. We performed feature selection for the three classifiers, to compare their performance, in order to choose the classifier that leads to the best classification results.

4.1 Data Set

4.1.1 Document Level

The data set that we collected consists of 1000 documents: 500 are informal texts and 500 are formal texts.

- **Informal Texts**

  We collected 500 texts that represent the informal style from the following sources:

  1. Corpus of Late Modern English Prose[5] (Denison, 1994): This corpus contains a set of annotated texts by David Denison[6]; most of these texts are informal texts (personal letters). Figure 4.1 shows one of these texts, as an example.
  2. Enron Email Dataset/Corpus[7]: This corpus contains email texts; most of them are personal letters (informal texts), as stated by Yu-shan & Yun-Hsuan (2005).
  3. Part of the Open American National Corpus[8]: this corpus contains some categories that are informal texts such as spoken language transcribed texts.

I have quite recovered and have polished off a lot of things that had got into arrears. We have all moved into winter quarters in the Office, out of dark, cold rooms into sunny ones. It is strange to welcome the sun again. My room is charming, warm and comfortable, with some delightful rugs which I've bought here on the floor, and all the new maps of Mesopotamia pinned up on the walls. Maps are my passion; I like to see the world with which I'm dealing, and everyone comes round to my room for geography.

Figure 4.1: A sample of informal text file extracted from Corpus of Late Modern English (prose).

- **Formal Texts**

  We collected 500 texts that represent the formal style from the following sources:

  1. Collection of news wire articles from the Reuters corpus[9]: This corpus contains a set of news texts; most of these texts are formal texts, as stated by Yu-shan & Yun-Hsuan (2005). Figure 4.2 shows a text from this set.
  2. Part of the Open American National Corpus, written technical texts: This corpus contains some categories that are formal texts such as formal publications.

[5] http://ota.ahds.ac.uk/catalogue/index-id.html
[6] http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/lmode-prose/#Summary
[7] http://www.cs.cmu.edu/~enron/
[8] http://www.anc.org/OANC/
[9] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

3. Corpus of Late Modern English Prose (Denison, 1994): This corpus contains a set of annotated texts by David Denison; some of these texts are formal texts. We extracted those formal texts manually, based on their annotation.

NEW YORK, March 16 - U.S. roastings of green coffee in the week ended March 7 were about 325,000 (60-kilo) bags, including that used for soluble production, compared with 290,000 bags in the corresponding week of last year and about 315,000 bags in the week ended February 28, George Gordon Paton and Co Inc reported.

It said cumulative roastings for calendar 1987 now total a 3,295,000 bags, compared with 3,620,000 bags by this time last year.

Figure 4.2: A sample of formal text file extracted from Reuters Corpus.

4.1.2 Sentence Level

We built our data set for sentence-level classification by splitting all the texts that we collected into sentences. We wrote a Perl program that splits text into sentences based on punctuation and on several simple rules (Manning and Schütze, 1999). . Each sentence is placed in a separate text file. We excluded from the data set the informal texts that we collected from OANC, because we could not split them into sentences. These texts are conversation (spoken) texts without any punctuation, especially full stops that are needed to detect the end of the sentences. In this case the whole text would have been considered as one sentence, and it would be too long as a sentence.

We performed sentence splitting on all texts for both classes (formal and informal). We excluded the empty files that contain no sentences.

We generated the following new data set:

1. 500 formal texts are split into 5158 sentences, which will represent formal sentences.
2. 452 informal texts are split into 5373 sentences, which will represent informal sentences.

In order to balance the number of the instances for each class and to have a baseline of 50:50, we randomly removed some sentences from each class to keep 5000 sentences for

each class. Therefore the final dataset for the sentence classification task consists of 5000 formal and 5000 informal sentences.

4.2 Features

We used several properties of the texts to encode documents or sentences as vectors of features. Before we present these features, we also present another text representation method that is commonly-used: the bag-of-words (BOW) representation.

4.2.1 Bag-Of-Words Representation

We encoded each text (sentence) into BOW features as one instance in our dataset. This task has done after cleaning these texts from all punctuations and stop words to reduce the size of the feature vector for our dataset; in addition, punctuation and stop words have high frequencies in almost any kind of text, which make them useless for predicting a class. The aim from this task is to build a bag-of-words model that help to see how the classifiers will do with BOW features before we build our model that is based on some of the main characteristics of formal and informal style.

This method of representing a text works very well when the target of classification is the topic of the text; but in our case, the classification task needs to be independent of the topic of the text, therefore we should not depend on this method for style classification.

4.2.2 Representation based on Text Properties

We extracted features from each text, based on some of the main linguistic characteristics of each style (formal or informal) that we had discussed earlier in Chapter 2. As mentioned in our related work chapter, these characteristics might represent each style. So, we hypothesize that these features might be good indicators to differentiate between both styles. Our experiments will be the ones left to decide which features help and which do not. We applied several statistical methods in order to extract the values of these features for each text in our dataset. Some of the features required us to parse each text. We parsed all the texts with the Connexor parser[10] (Tapanainen & Järvinen, 1997), which is known to produce high-quality

[10] http://www.connexor.com/

results (Pyysalo et al, 2006). We wrote UNIX scripts in order to run the parser for each text file and to generate files with parse trees for each text.

The features that we extracted are as follows:

1. **Formal words list**: This feature might characterize formal texts. It is based on the formal list that we had mentioned earlier in Chapter 2. The value of this feature is based on the sum of the frequencies of any words from this list. The word frequencies are normalized by the length of the text for each document, because the size of each document is different. We noticed that there are some words have very low frequency or zero frequency. As mentioned in Chapter 2, a few of these words that came from CTRW might not be really formal; therefore, these words were manually detected and excluded from our list. Other formal words with zero frequency should be kept because they might occur in other test data on which we will run our system in the future.
2. **Informal words list:** This feature might characterize informal texts. It is based on the informal list that we had mentioned it earlier in Chapter 2. The value of this feature is based on the sum of the frequencies of any words from this list. The word frequencies are normalized by the length of the text. Again, as in the formal list, we discovered that there are a few words that are not really informal (they came from the automatic acquisition from CTRW). These words were manually detected and excluded from our list.
3. **Formal pronouns:** This feature might characterize formal texts. We use this feature because formal style often uses an impersonal style, using third person pronouns. We extracted this feature from the parse trees returned by the Connexor parser; we counted how many times each text has impersonal pronouns. The count is normalized by the length of the text for each document.
4. **Informal pronouns:** This feature might characterize informal texts. Informal style often uses a personal style, using first and second person pronouns. We extracted this feature from the parse trees returned by Connexor parser. We counted how many times the texts have personal pronouns, and normalized by the length of the text for each document.

5. **Contractions:** This feature might characterize informal texts. As mentioned in Chapter 2, informal style tends to use contractions. We extracted this feature by counting the number of contractions (words that have apostrophe) in each text. The count of contractions is normalized by the length of the text for each document.
6. **Abbreviations:** This feature might characterize informal texts. This feature might not be a good feature for our model, because abbreviations might be used in formal texts in some cases (Obrecht, 1999). Therefore, we will experiment with using abbreviations as a feature to see if it is useful or not. We extracted this feature by counting the abbreviations in each text. The count of abbreviations is normalized by the length of the text for each document.
7. **Passive voice:** This feature might characterize formal texts. As mentioned in Chapter 2, formal style often uses the passive voice. We extracted this feature from the parse trees returned by Connexor parser. We counted how many times the text has a passive voice sentence structure. The count is normalized by the length of the text for each document.
8. **Active voice:** This feature might characterize informal texts. We extracted this feature from the parse trees returned by Connexor parser. The count of the active voice sentence structure is normalized by the length of the text for each document.
9. **Phrasal verbs:** This feature might characterize informal texts as stated by Dempsey, McCarthy, & McNamara (2007). It is based on one of the main characteristics of informal style. We will experiment with using phrasal verbs as a feature to see if it is useful or not. We extracted this feature from the parse trees returned by Connexor parser. We counted how many times each text has phrasal verbs. The count of the phrasal verbs is normalized by the length of the text for each document.
10. **Word length's average:** We hypothesized that this feature might characterize formal texts if the value is large (complex words), and it might characterize informal texts if the value is small (simple words). This hypothesis is based on the main characteristics of both styles as discussed in Chapter 2.
11. **Type/Token Ratio (TTR):** This feature refers to how many distinct words are in a text comparing to the total number of words in the text. The TTR in formal texts is lower than in informal texts as stated by Renkema (1984).

4.3 Classification Algorithms

We used Weka (Hall et al., 2009), a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a certain dataset or through a Java API. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited to develop new machine learning schemes.

We chose three machine learning algorithms: Decision Tree (J48[11]) because it allows human interpretation of what is learnt, Naïve Bayes because it is known to work well with text, and Support Vector Machines[12] (SMO[13]) because it is known to achieve high performance. We trained the three classifiers on all the features described above. Then we trained the three classifiers after removing the abbreviations feature and the phrasal verbs feature from our model to see if it will affect the accuracy of the classification. We also applied feature selection, in order to increase the performance of the classifiers and to determine the best features for our model. After performing the feature selection method on our model, we reduced the number of features from 11 features to 5 features, as discussed in section 4.4.1.2. We used the InfoGainAttributeEval[14] method from Weka, which shows the weight for each feature and that help us to know the weakest features that should be removed. In the next section, we will present more details about our experiments.

4.4 Experiments and Evaluation

The evaluation of the results is based on standard measures such as Precision and Recall, F-measure, Accuracy, and AUC-ROC[15] as implemented in Weka.

[11] J48 implements Decision Trees algorithm C4.5 in Weka.
[12] The kernel used is Linear Kernel, we tried other kernels but they are very slow to generate the classification result and they did not generate better performance than the Linear Kernel.
[13] Support Vector Machines algorithm is implemented in Weka by SMO.
[14] InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class.
[15] ROC-AUC: The area under the receiver operating characteristic curve.

4.4.1 Document Level

4.4.1.1 BOW Features

When using the BOW feature representation, we have noticed that the Accuracy was high especially for SVM and Naïve Bayes, but in fact we cannot rely on BOW features because they are based on word vectors, so they can obtain good accuracies only when the topics of the texts in the two sets are distinct. We want to avoid the use of the topic of the texts in classification. The formal and informal style should not depend on the topic. From this point we conclude that we have to use specific features that work for any kind of text, independent of topic. Nonetheless, we present the comparison among the result of the three classifiers (Decision Tree, Naïve Bayes, and SVM). We used 10-fold cross validation to evaluate each classifier. This comparison is based on many measures such as: the number of correctly and incorrectly classified instances, the F-measure, and the AUC-ROC. Table 4.1 shows that best classifier among the three classifiers is Naïve Bayes, then SVM comes second, and Decision Tree comes last. Based on a paired t-test (significance level 0.05), both the NB and the SVM accuracies are significantly better than that of the Decision Tree. We can notice this difference if we look at the values of F-measure, success rate (Accuracy), and AUC-ROC. Higher values of these variables mean better performance. We notice that the classifiers have achieved high accuracy on BOW, which means that the features are successful in distinguishing the two classes.

Classifier	F-measure	Accuracy	AUC-ROC
SVM	0.994	0.994	0.994
Naïve Bayes	0.994	0.994	0.998
Decision Tree	0.977	0.977	0.972

Table 4.1: Document classification results (using BOW features representation) of Decision Tree, Naïve Bayes, and SVM classifiers, using 10-fold cross validation.

4.4.1.2 Text Property Features

We built our model with specific features that we extracted from the texts based on some of the main characteristics of both style. We used two methods to evaluate the classifiers which are 10-fold cross validation and training/test data set. In 10-fold cross validation, we applied

InfoGain feature selection. Using both methods, the three algorithms showed different classification results. The comparison between these algorithms is based on many measures such as: the correctly and incorrectly classified instances, the Accuracy and Precision of each algorithm in classifying each class, the Recall value, and the F-measure. We compare these values in order to find the best classifier among these algorithms.

The general formulas for Recall and Precision in information retrieval are (Witten & Frank 2005):

$$\text{Precision} = \frac{\text{Number of documents classified that are relevant}}{\text{Total number of documents that are classified}}$$

$$\text{Recall} = \frac{\text{Number of documents classified that are relevant}}{\text{Total number of documents that are relevant}}$$

As an example, we look at the confusion matrix of the Decision Tree classifier in Figure 4.3, based on the results produced by 10-fold cross validation; we can apply the formulas for Precision and Recall for both classes in Decision Tree result as follows:

$$\text{Recall (informal)} = \frac{TP}{TP + FN} \quad \text{Recall (formal)} = \frac{TN}{TN + FP}$$

$$\text{Precision (informal)} = \frac{TP}{TP + FP}$$

$$\text{Precision (formal)} = \frac{TN}{TN + FN}$$

Where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of the false positives predicted for the considered class.

After applying the different formulas, we obtained the following results:

- Recall (informal) = 0.986.
- Recall (formal) = 0.984.
- Precision (informal) = 0.984.
- Precision (formal) = 0.986.

		Predicted Class	
		Informal	**Formal**
		(a)	**(b)**
Actual Class	**Informal** = a	**TP** = 493	**FN** = 7
	Formal = b	**FP** = 8	**TN** = 492

Figure 4.3: The confusion matrix of the Decision Tree which shows the distribution of the actual and the predicted classes on document level using 10-fold cross validation.

Here we noticed the following:

1. The higher the Recall the better the classifier is, because the Recall value shows how close the number of predicted instances is to the original number of instances in that class (how good the classifier is in predicting the instances).
2. The higher value of Precision is preferred. This could be explained by examining the higher Precision value of the formal class. Only seven instances from the total of 500 in the class formal were incorrectly classified (i.e., FN= 7).
3. Recall and precision go against each other; while Recall is a good indicator for the performance of the classifier to learn the informal class, the Precision is lower for this class.

Based on this information, we choose as the best indicator the **F-measure,** which combines Recall and Precision. The higher value of F-measure, the better the classifier is in predicting a class. The **F-measure** is the weighted harmonic mean of Precision and Recall. The formula of F-measure is as follows:

$$\text{F-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Calculating the F-measure for the Decision Tree for both classes in our previous example gives the following:

- F-measure (informal) = 0.985.
- F-measure (formal) = 0.985.

The result shows that the classifier has the same predictive power (performance) for the first class and the second class.

To measure the goodness of a classifier in general, many evaluation measures can be used. The Accuracy (success rate) shows the instances from both classes that are classified correctly, using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly classified documents}}{\text{Total number of documents}}$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

In our case, the success rate (Accuracy) is equal to 0.985.

The ROC curve (Receiver Operating Characteristic) on the other hand, depicts the performance of a classifier without regard to class distribution or error costs. ROC is a term used in signal detection to characterize the trade-off between hit rate and false alarm rate over a noisy channel. ROC curves plot the number of positives included in the sample on the vertical axis, expressed as a percentage of the total number of positives, against the number of negatives included in the sample, expressed as a percentage of the total number of negatives, on the horizontal axis. The bigger the area under the curve (AUC-ROC), the better the classifier is.

Relying on the above analysis, we can measure the performance of the three algorithms in predicting the formal class and the informal class.

We notice the difference between the classifiers, if we look at the values of F-measure measure, the Accuracy, and AUC-ROC. Higher values of these measures mean better performance. We compare the overall values of the measures, to choose the best classifier.

1. Text classification results using 10-fold cross validation without feature selection

The best classifier among the three algorithms is the Decision Tree, SVM comes second, and Naïve Bayes comes last. Despite the fact that value of AUC-ROC for NB is high, as shown in Table 4.2, based on a paired t-test (significance level 0.05), both the Decision Tree and the SVM accuracies are significantly better than that of NB.

Figure 4.4 shows the visualization for the Decision Tree of text classification using 10-fold cross validation without attribute selection.

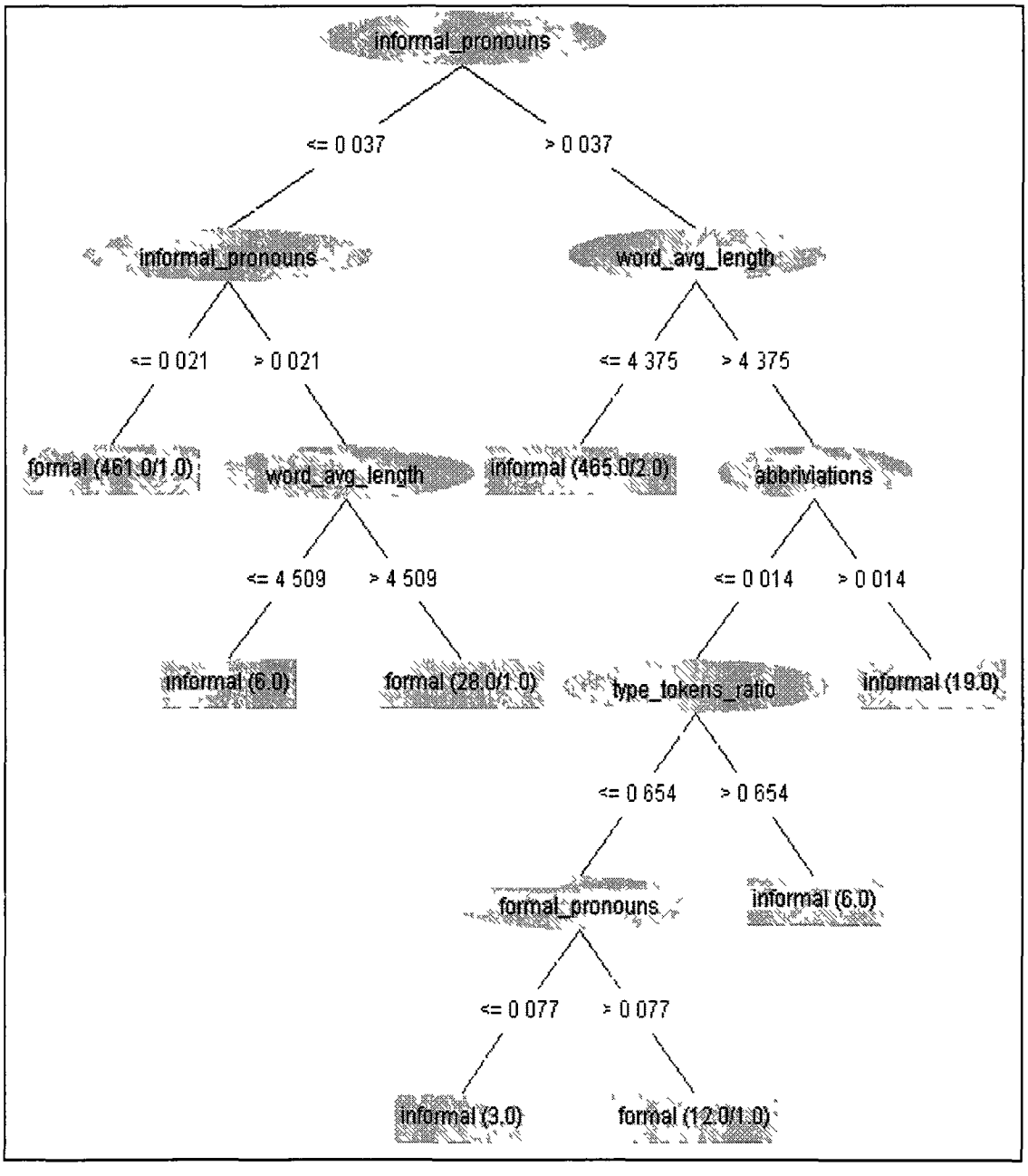

Figure 4.4: Decision Tree visualization for text classification using 10-fold cross validation without feature selection.

In Table 4.2 below, we show the comparison between the three algorithms using 10-fold cross validation before applying feature selection.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.985	0.985	0.982
Naïve Bayes	0.970	0.970	0.993
SVM	0.983	0.983	0.983

Table 4.2: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation before applying feature selection.

The results of the classifiers that use our proposed features are very good. We wanted to test on other kinds of texts to see if the features are still as good. We obtained again similarly good results, on a dataset of formal and informal medical texts. Classifying sets of medical documents into formal or informal style might be important for patient safety. Formal documents are more likely to have been published by medical authorities; therefore, the patients could trust them more than they can trust informal documents. The dataset consists of 990 medical texts characterizing the informal style, from a Medical newsgroups collection (Yu-shan & Yun-Hsuan 2005), and 990 medical texts characterizing the formal style from a medical abstracts collection (Joachims, 1997). Appendix B shows the results for our classification task on medical documents, using the text property features (Abu Sheikha & Inkpen, 2010). These results show that our classification model generalizes to other kinds of texts.

2. Text classification results using 10-fold cross validation with feature selection

We examined all the features by performing feature selection using InfoGain feature selection from Weka. We started by removing the weakest feature (which has the lowest weight) manually, until we obtained the features that lead to the best performance. Table 4.3 shows each feature with its weight according to the InfoGain feature selection, ranked in descending order from the strongest feature to the weakest feature.

Features	Weight
Informal pronouns	0.9031
Word length's average	0.7729
Informal list	0.4153
Active voice	0.3159
Contractions	0.2697
Type Tokens Ratio (TTR)	0.1523
Passive voice	0.1174
Abbreviations	0.0967
Phrasal verbs	0.0735
Formal list	0.057
Formal pronouns	0.0183

Table 4.3: The features with their weights in our model based on InfoGain feature selection on document level classification.

We removed 6 features: TTR, Passive voice, Abbreviations, Phrasal verbs, Formal list, Formal pronouns, in order to obtain the best performance. In addition, we noticed here that Formal list has a very low weight in our model because the list has many words with very low or zero frequency in our dataset.

The accuracy has improved from 0.985 to 0.988 for Decision Tree after applying feature selection. The best classifier among the three algorithms is still the **Decision tree**, SVM comes second, and Naïve Bayes comes last. Based on paired t-test (significance level 0.05), the Decision Tree accuracy is significantly better than the accuracy obtained by SVM and Naive Bayes. Figure 4.5 shows the visualization for the Decision Tree of text classification using 10-fold cross validation after applying feature selection.

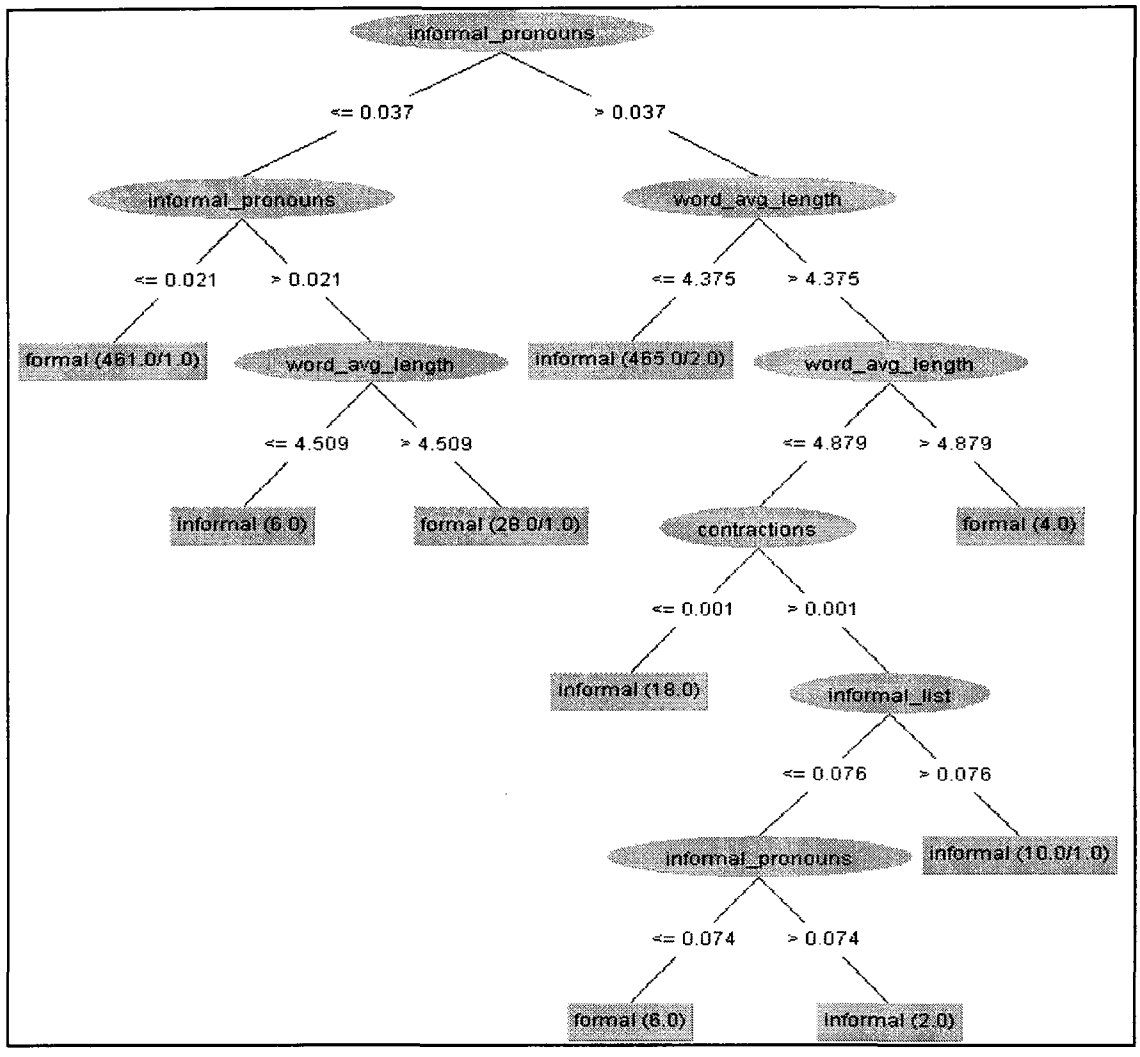

Figure 4.5: Decision Tree visualization for text classification using 10-fold cross validation after applying a feature selection.

In Table 4.4 below, we show the comparison between the three algorithms using 10-fold cross validation after applying feature selection.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.988	0.988	0.989
Naïve Bayes	0.969	0.969	0.995
SVM	0.973	0.973	0.973

Table 4.4: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after performing feature selection.

3. Text classification results using 10-fold cross validation after removing the abbreviations feature

We removed the abbreviations feature in order to see it is useful for our model or not. As we discussed in Chapter 2, informal texts tend to use many abbreviation, but formal style also uses abbreviations in some cases. The classification results after removing the abbreviations are very similar to the classification results of our model with the abbreviations feature. Therefore, the abbreviations feature might not be a good feature for our model. This is also confirmed by its low InfoGain score in Table 4.3.

In Table 4.5 we show the comparison between the three algorithms using 10-fold cross validation, after removing the abbreviations feature.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.985	0.985	0.982
Naïve Bayes	0.970	0.970	0.993
SVM	0.983	0.983	0.983

Table 4.5: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after removing the abbreviations feature.

4. Text classification results using 10-fold cross validation after removing the phrasal verbs feature

We removed the phrasal verbs feature in order to see it is important for our model or not. As we discussed in Chapter 2 that phrasal verbs might be used in some formal texts. The classification results after removing the phrasal verbs are almost the same as the classification results for our model with the phrasal verbs feature. Therefore, the phrasal verbs feature might not be a good feature for our model. This is also confirmed by its low InfoGain score in Table 4.3.

Table 4.6 shows the comparison between the three algorithms using 10-fold cross validation, after removing phrasal verbs feature.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.985	0.985	0.982
Naïve Bayes	0.968	0.968	0.993
SVM	0.983	0.983	0.983

Table 4.6: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation after removing the phrasal verbs feature.

5. Text classification results using training/test set

We divided our dataset randomly into two files, one to represent the training data set and the other one to represent the test data set.

- The training file includes the following instances:
  - Informal class: 350 instances, randomly chosen from our data set from the informal class.
  - Formal Class: 350 instances, randomly chosen from our data set from the formal class.
- The test file includes the following instances:
  - Informal class: the rest of 150 instances from our dataset, for the informal class.
  - Formal Class: the rest of 150 instances from our data set, for the formal class.

The best classifier among the three algorithms becomes **SVM**, Decision Tree comes second, and Naïve Bayes comes last. Figure 4.6 shows the visualization for the Decision Tree of text classification using supplied test set (therefore trained on the whole training data).

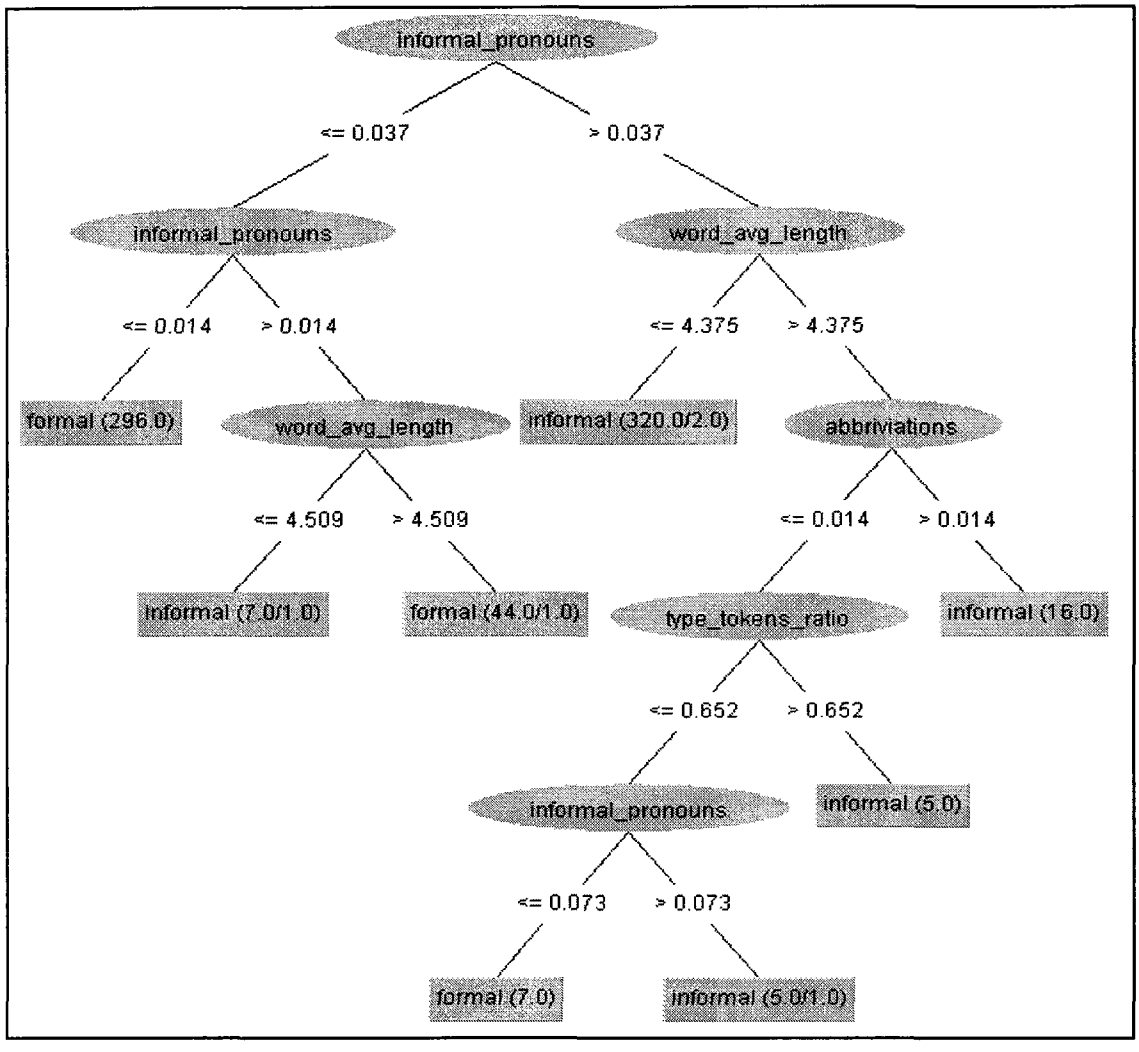

Figure 4.6: Decision Tree visualization for text classification trained on the whole training data (we applied it on the test set).

In Table 4.7 below, we show the comparison between the three algorithms using training/test set method.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.987	0.987	1.00
Naïve Bayes	0.970	0.970	0.997
SVM	0.990	0.990	0.990

Table 4.7: Document classification results (using text property features) of Decision Tree, Naïve Bayes, and SVM classifiers on the training/test set.

4.4.2 Sentence Level

4.4.2.1 BOW Features

We notice again that the accuracy is high for BOW feature representation, this time for the sentence-level classification, as it was at the document-level classification, especially for SVM and Naïve Bayes. As we have mentioned earlier in section 4.4.1.1, we cannot rely on BOW features because they are based on word vectors. However, we want to show the result for BOW, to compare it to our model. In Table 4.8, we show the comparison between the three algorithms using 10-fold cross validation for BOW. These results show that the best classifier is SVM, followed by Naïve Bayes and Decision Tree.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.862	0.864	0.880
Naïve Bayes	0.871	0.873	0.957
SVM	0.938	0.938	0.938

Table 4.8: Sentence classification results (using BOW feature representation) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation.

4.4.2.2 Sentence Property Features

We built our model to classify sentences based on the same features that we used to classify longer texts (whole documents) because it has achieved high accuracy in predicting the two classes. Also, we used the same two methods to evaluate the classifiers: 10-fold cross validation and training/test set. In 10-fold cross validation, we applied InfoGain feature selection. We did comparisons between the three algorithms based on the accuracy, the F-measure, and the AUC-ROC, as we did on the text classification section 4.4.1, in order to find the best classifier for predicting the formal class and the informal class at sentence level.

1. **Sentence classification results using 10-fold cross validation without feature selection**

The best classifier among the three algorithms is the **Decision Tree**, SVM comes second, and Naïve Bayes comes last. Based on paired t-test (significance level 0.05), both the Decision Tree and the SVM accuracies are significantly better than that of NB.

In Table 4.9 below, we show the comparison between the three algorithms using 10-fold cross validation before applying feature selection.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.865	0. 865	0.905
Naïve Bayes	0.784	0.786	0.888
SVM	0.843	0.843	0.844

Table 4.9: Sentence Classification results (using sentence property features) of Decision Tree, Naïve Bayes, and SVM classifiers using 10-fold cross validation before applying feature selection.

2. **Sentence classification results using 10-fold cross validation with feature selection**

We examined all the features by performing feature selection using InfoGain feature selection from Weka. InfoGain feature selection has ranked the features in descending order. We started by removing the weakest features and then we train the classifiers, but we discovered that this decreased the accuracy for the three algorithms. So, we decided to keep all the features in our model, as all the features at the sentence level might be important in order to achieve good performance. Table 4.10 shows each feature with its weight according to the InfoGainA feature selection.

Features	Weight
Informal pronouns	0.35444
Word length's average	0.25982
Formal pronouns	0.11995
Informal list	0.07730
Active voice	0.06262
Abbreviations	0.05403
Passive voice	0.04760
Type Tokens Ratio (TTR)	0.03796
Formal list	0.02126
Contractions	0.01652
Phrasal verbs	0.00368

Table 4.10: The features with their weights in our model based on InfoGain feature selection on sentence level classification.

3. Sentence classification results using training/test set

We divided our dataset randomly into two sets: the first one represents the training data and the second represents the test data.

- The training set includes the following instances:
  - Informal class: 3500 instances, randomly chosen from our data set that represent the informal class.
  - Formal Class: 3500 instances, randomly chosen from our data set that represent the formal class.
- The test file includes the following instances:
  - Informal class: the rest of the 1500 instances of the informal class.
  - Formal Class: the rest of the 1500 instances of the formal class.

The best classifier among the three algorithms became **Decision Tree**; SVM comes second, and Naïve Bayes comes last.

In Table 4.11 below, we show the comparison between the three algorithms using supplied test set method.

Classifier	F-measure	Accuracy	AUC-ROC
Decision Tree	0.867	0. 867	0.896
Naïve Bayes	0.822	0.823	0.903
SVM	0.856	0.856	0.856

Table 4.11: Sentence classification results (using sentence properties features) of Decision Tree, Naïve Bayes, and SVM classifiers using the training/test set.

4.5 Discussion

Our experiments show that it is possible to classify any text or sentence according to formal and informal style. We achieved reliable accuracies for all three classifiers, especially on Decision Tree and SVM classifiers in our model. This indicates that we selected high quality features to include in our model.

Our experiments showed that the abbreviations feature and the phrasal verbs feature might not be good features to characterize the informal style. In addition, we noticed that there are some words in the formal/informal list that have very low frequency in the dataset for each style, therefore they might not be important for our model and might not characterize the distinction between formal and informal style.

Our model can generate good results whether it is applied to a single topic or to different topics (as opposed to a BOW model).

Chapter 5

Automatic Generation of Formal and Informal Sentences

In this chapter, we present the second part of our work which consists in using a natural language generation technique to generate formal and informal sentences. In chapter 2, we listed the main characteristics of each style, in order to distinguish between them. This helps us choose features that can produce sentences in one of the two styles. We collected some ready-made parallel list of formal/informal words and phrases from different sources. In addition, we obtained a new list that was extracted manually by Brooke et al. (2010) from the dictionary of synonyms *Choose The Right Word* (Hayakawa, 1994). These lists are considered as important features of our system; they help to generate sentences which represent both styles. We obtained two more parallel lists for both styles: one contains most of the contractions in English (short forms), and the other one is some common abbreviations. These parallel lists might help to generate sentences in the preferred style by changing words or expressions for that style (Appendix A shows these lists).

We implement our NLG system in Java using the SimpleNLG package (Gatt and Reiter, 2009), which we use in order to implement templates from which we generate valid English texts with formal or informal style. We used human judges in order to evaluate the quality of the generated sentences and their level of formality.

5.1 Generating formal versus informal sentences

In chapter 1 and chapter 2, we stated the importance of distinguishing between formal and informal style. In addition, we stated the need to implement a system that can generate correct English sentences based on different styles. Therefore, we proposed to use NLG techniques to automatically generate formal and informal sentences. In this thesis, we used natural language generation technique from the SimpleNLG package (which is in fact a library of Java classes). It can be integrated in any system that needs to generate correct English sentences based on different criteria. We explained the NLG techniques and the SimpleNLG package in Chapter 3 (section 3.2).

In this section, we present our proposed features that we used in building a system that generates formal or informal sentences. Then, we describe our system that generates sentences using the SimpleNLG package with our proposed features. At the end of this section, we show an example of generating a formal sentence using the system. Our parallel lists of words and the use of the SimpleNLG package are key success factors and efficient features for our proposed system. SimpleNLG has some of the syntactic features that are needed to represent some of the characteristics of formal and informal style such as passive voice. Sometimes, we need to generate sentences in passive voice to represent the formal style, and this feature is already implemented in SimpleNLG.

5.1.1 Formality Features

We introduced the following two main features that will be used in constructing formal/informal sentences. We hypothesize that both features might help to produce sentences in both styles.

1. Phrase, expression, and word choice (lexical choice): This feature might help to generate sentences in both styles (Hovy, 1988). We implement this feature in our system based on the parallel lists (formal/informal words, the contraction list, and the abbreviation list) that we have described in Chapter 2 (section 2.1).
2. Passive/Active voice option: This feature is based on the characteristics of both styles which we mentioned in Chapter 2 (section 2.1). In addition, it was suggested by Hovy

(1988). We added this feature to our system and we let the user choose a sentence in the passive or the active voice.

5.1.2 Implementation

Our system can generate natural language sentences in a formal/informal style with different inputs of subject, verb, and complement (by complement, we mean one or more words including subordinate clauses). Therefore, the user might not worry about choosing any word that he is not very familiar with, whether the word is formal or informal, because the system will manage to replace some words with better words, based on the desired style.

In addition, our system might interact with the user directly, or it can be integrated with any system that has the ability to send and receive commands from Java programs.
In the following, we explain the main steps for our system to generate sentences using NLG. The steps are as follows:

1. Ask the user which style he likes to generate (Formal/ Informal). The user should choose one style.
2. The option will be sent to the main Java class, which is considered as a Java interface. This interface will call the class that represents formal or informal style, based on the user's preference.
3. The system then will ask the user to enter a template that represents the sentence in the form of a subject, a verb, and a complement.
4. If the user has chosen Formal, then the system will ask him if he would like to generate a sentence in a passive voice or not. This feature is implemented in our system based on the characteristics of formal style.
5. Then the system will ask the user about some syntactic features: the verb tense (present, past, future), progressive (yes, no), perfect (yes, no), and negation (yes, no).
6. The system then checks the verb in the formal/informal parallel list; if it is formal or informal, and the system will find a synonym of the verb in the list, it will replace it based on the preferred style.
7. After the sentence is constructed, the system will search for any word, phrase, or expression from the formal/informal list, the abbreviations list, and the contractions list, in order to replace it with a synonym, based on the preferred style.

8. Lastly, our system will generate a natural language sentence according to the preferred style.

For more explanation and clarification on how our proposed system works, we provide the following example with snapshots of our system, with different parameters to generate a formal sentence These snapshots are from a prototype for our proposed system, to show how our system is generating the sentences The user interface can be improved in future work

1. The system asks the user to choose the preferred style (formal/informal) of the sentence that will be generated (see Figure 5.1).

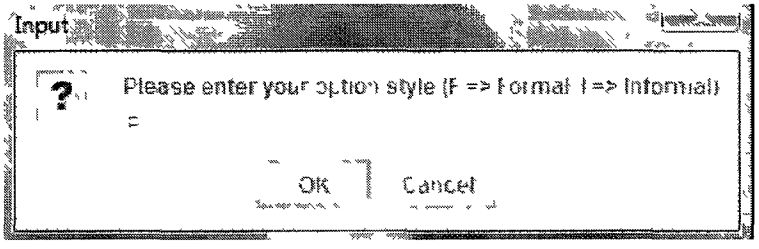

Figure 5.1: The first screen in our proposed NLG system.

2. After determining the preferred style, the system asks the user to enter the subject of the sentence (see Figure 5.2).

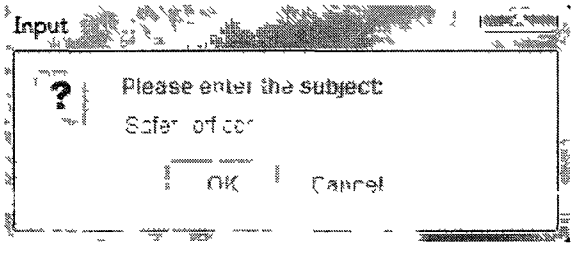

Figure 5.2: The second screen in our proposed NLG system.

3. Then, the system asks the user to enter the verb of the sentence (see Figure 5.3).

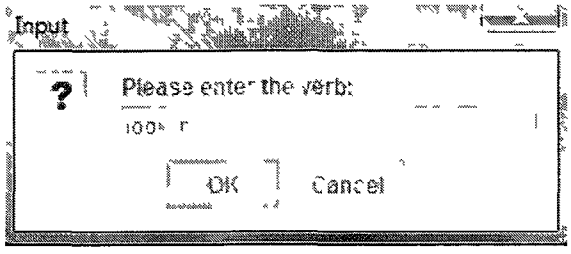

Figure 5.3: The third screen in our proposed NLG system.

4. After that, the system asks the user to enter the rest of the sentence (called "the complement" in SimpleNLG), (see Figure 5.4)

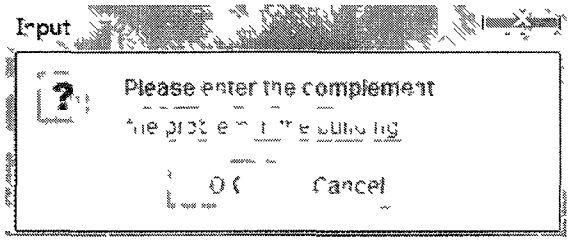

Figure 5.4: The fourth screen in our proposed NLG system.

5. After entering all the sentence components (subject, verb, complement), the system checks the style. If the chosen style is formal, then the user will be asked if he/she prefers to put the sentence in a passive voice If the chosen style is informal, then the user will not be asked about the passive voice (see Figure 5.5)

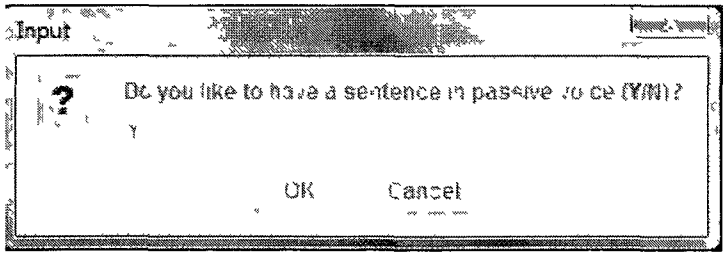

Figure 5.5: The fifth screen in our proposed NLG system.

6. Then, the system asks the user about the tense of the verb (see Figure 5 6)

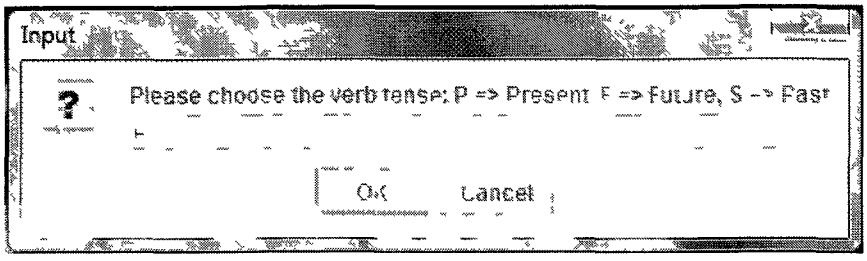

Figure 5.6: The sixth screen in our proposed NLG system.

7. After choosing the tense, the system asks the user if he wants the verb's tense to be progressive (see Figure 5.7).

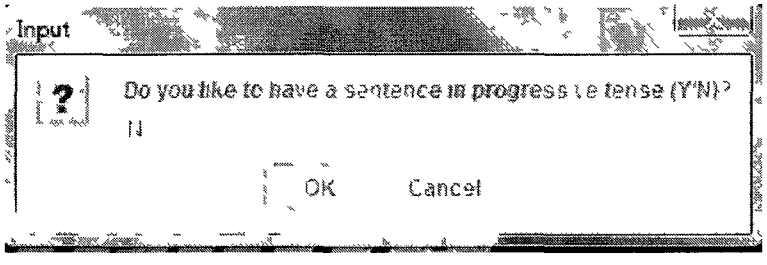

Figure 5.7: The seventh screen in our proposed NLG system.

8. Then, the system asks the user if he wants the verb's tense to be perfect (see Figure 5.8)

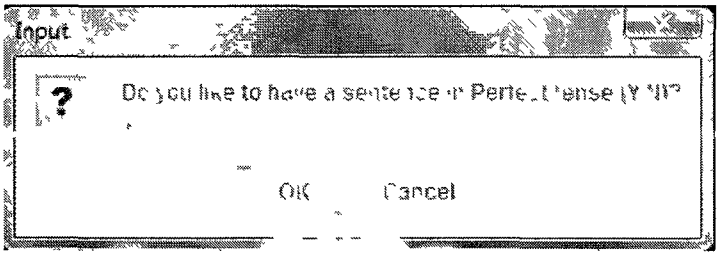

Figure 5.8: The eighth screen in our proposed NLG system.

9. After that, the system asks the user if he wants the verb's tense to be negated (see Figure 5.9).

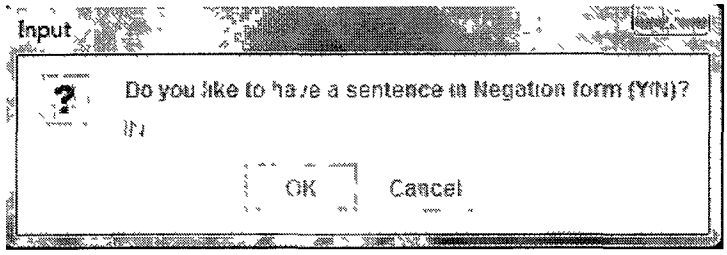

Figure 5.9: The ninth screen in our proposed NLG system.

10. Lastly, The user will get the generated sentence based on his preferences (see Figure 5.10:

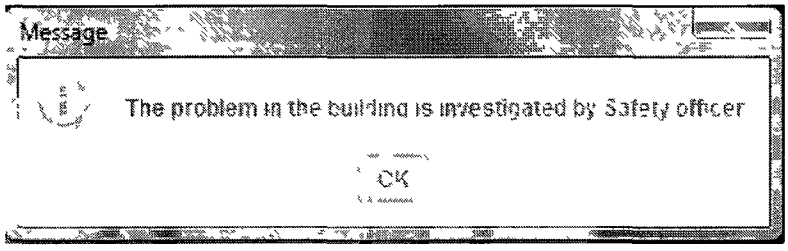

Figure 5.10: The last screen in our proposed NLG system that contains the generated sentence.

As we see, the system has replaced the phrasal verb "look in" with "investigate", to make the sentence formal. In addition, the system generated the sentence in passive voice to make it more formal, as required by the user's choice.

5.2 Results and Evaluation

Natural language generation is most often evaluated using scores given by human judges (Reiter and Belz, 2009). Our evaluation target was to measure the degree of formality (Formal/Informal) of the generated sentences. We asked two human judges (graduate students in computational linguistics, native speakers of English) to annotate the 100 generated sentences as having formal or informal style. The instructions that we gave to the judges were simple: we asked the judges to annotate each sentence by 0 if they think that the

sentence is Informal, and to annotate it by 1 if they think that the sentence is Formal. Table 5.1 shows samples of the generated sentences with the judges' annotations. In addition, Table C.1 in Appendix C shows all the generated sentences with the annotations of both judges. We estimate the correctness of our system by comparing the original class of the generated sentences (formal/informal) to the annotations of Judge1 and to the annotations of Judge 2. We calculated several evaluation measures, to see if our proposed system achieves good quality in producing English sentences in formal and informal style. These measures are the accuracy (correctness) of our system according to each judge, and the precision for each class according to each judge. The results of their annotations show high accuracy for the generated sentences. In Figure 5.11 and Figure 5.12, we show the summary results of their annotations. Table 5.2 shows that the accuracy of our system is 95% according to Judge 1 and 99% according to Judge 2.

Sentence	**Actual Class**	**Judge1 annotation**	**Judge2 annotation**
The plane is going to go on Jan. 5th.	Informal	Formal	Informal
She has wished to work as soon as possible.	Formal	Formal	Informal

Table 5.1: Samples of the generated sentences with the annotations from both judges.

		Predicted Class		
		Informal (a)	**Formal** (b)	**Precision**
Actual Class	**Informal** = a	**TP** = 45	**FN** = 0	0.90
	Formal = b	**FP** = 5	**TN** = 50	1.00

Figure 5.11: The annotation result for Judge1 with the precision for each class.

		Predicted Class		
		Informal (a)	**Formal** (b)	**Precision**
Actual Class	**Informal** = a	**TP** = 50	**FN** = 1	1.00
	Formal = b	**FP** = 0	**TN** = 49	0.98

Figure 5.12: The annotation result for Judge2 with the precision for each class.

Judge	**Accuracy**
Judge1	95%
Judge2	99%

Table 5.2: The accuracy for our system according to each judge.

The following equations are used to calculate the precision for each class and the accuracies for both classes together.

The precision is the proportion of the sentences that are relevant according to a judge's annotations.

$$\text{Precision} = \frac{\text{Number of sentences that are relevant according to a judges annotation}}{\text{Total number of sentences per class}}$$

$$\text{Precision (informal)} = \frac{TP}{TP + FP}$$

$$\text{Precision (formal)} = \frac{TN}{TN + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Now, we calculate the agreement between the two judges, and the kappa statistic that compensated for agreement by chance (Cohen, 1960) (Manning et al., 2008). In Table 5.3, we show the agreement between Judge1 and Judge2, and then we show the equations that we used in the calculation of the kappa statistic.

	Judge2 Relevance			
Judge1 Relevance		Informal	Formal	**Total**
	Informal	45	1	**46**
	Formal	5	49	**54**
	Total	**50**	**50**	**100**

Table 5.3: The number of sentences for which Judge1 and Judge2 agreed or did not agree.

The following equations are used to calculate the agreement between the two judges which will lead to calculate the kappa statistic:

- Observed proportion of the times the judges agreed

$$P(A) = \frac{((Informal_agree) + (Formal_agree))}{Total_of_Sentences}$$

In our case, P (A) is equal to 0.94

- Calculate of the probability of each event:

The probability for the Informal class is:

$$P(\text{Informal}) = \frac{((Total_Informal_rows) + (Total_Informal_columns))}{((Grand_Total_rows) + (Grand_Total_columns))}$$

In our case, P (Informal) is equal to 0.48

The probability for the Formal class is:

$$P(\text{Formal}) = \frac{((Total_Formal_rows) + (Total_Formal_columns))}{((Grand_Total_rows) + (Grand_Total_columns))}$$

In our case, P (Formal) is equal to 0.52

- Probability that the two judges agreed by chance (P(E))

$$P(E) = P(Informal)^2 + P(Formal)^2$$

In our case, P (E) is equal to 0.5008

Now based on the previous equation, we can calculate the kappa statistic as follows:

- Kappa statistic (k)

$$\text{Kappa(k)} = \frac{((Observed_agreement) - (Chance_agreement))}{(1 - Chance_agreement)}$$

$$\text{Kappa(k)} = \frac{(P(A) - P(E))}{(1 - P(E))}$$

In our case kappa statistic (k) is equal to 0.88.

Therefore, the agreement between the two judges is 94% and the kappa value is 0.88. This means a very good agreement for the task.

5.3 Discussion

Overall, the evaluation shows that it is possible to generate any text or sentence according to formal and informal style, according to the high accuracy for the generated sentences.

The accuracies for our system according to both judges were very high (the accuracy is 95% according to Judge1 and 99% according to Judge2). The strength of the inter-judge agreement is considered to be very good, reaching 94% and the kappa value is 0.88. We showed that our proposed features were selected successfully, and our proposed system is accurate.

Chapter 6

Conclusion and Future Work

In this chapter, we conclude our work and we discuss possible future work.

6.1 Conclusion

In this thesis, we addressed the tasks of analysis and generation of formal and informal texts. The main characteristics of formal and informal style that we identified are success factors for our work, because they helped us to build the features that lead to good classification and generation results. In addition, the parallel lists of formal versus informal words and phrases that we collected from different sources were very important in building our model for classification task and designing our system for the generation task.

The experiments and the evaluation results showed that it is possible to classify any text or sentence according to formal or informal style. In our model, we chose high quality features to characterize formal and informal texts. The evaluation results showed high accuracies in predicting the class. Classifying texts into formal and informal is very useful for different applications (e.g., evaluating the written texts in research papers, to check their degree of formality).

We developed an NLG system that can generate formal and informal sentences. We used NLG technique and the SimpleNLG package to implement our system. We proposed some important features that are used in generating formal and informal sentences. We think that these features were selected successfully because the evaluation with human judges showed a high accuracy in generating formal and informal sentences with our system in generating formal and informal sentences. Generating formal and informal sentences is very useful for

different applications (e.g., generating feedback for e-learning games, letters to clients, and other formal or informal documents).

6.2 Future Work

- Our immediate future work will be on extracting more formal and informal lists which should increase the accuracy of the classifiers and increase the possibility of generating more and more formal/informal sentences. We will apply different techniques, such as bootstrapping, which can be used in order to extract more lists of words, based on some seed words.

- We plan to do more experiments on formal/informal lists on other kinds of texts of both styles, in order to determine the most important words that we should keep in the list for each style.

- For the classification task, we will experiment by adding more features to our model such as a sentence length feature, in order to obtain a classifier with close to 100% accuracy.

- We plan to experiment our classification model on other kinds of texts, in order to reach the model that could classify any texts into formal and informal style.

- We plan to extend the implementation for our NLG system to cover generating longer texts (e.g., generating several sentences, by adding aggregation, or replacing some nouns with pronouns to avoid repetitions).

Bibliography

Abu Sheikha, Fadi, and Inkpen, Diana (2010). Learning to Classify Medical Documents According to Formal and Informal Style. Proceedings, Workshop on Intelligent Methods for Protecting Privacy and Confidentiality in Data, 23rd Canadian Conference on Artificial Intelligence, May 31st to June 2nd, University of Ottawa, Ontario, Canada, pp. 27-34.

Akmajian, Adrian, Demers, Richard A., Farmer, Ann K., and Harnish, Robert M. (2001). Linguistics: an introduction to language and communication. 5th Edition, MIT Press, Cambridge (MA), pp. 287-291.

Brooke, Julian, Wang, Tong, and Hirst, Graeme (2010). Inducing Lexicons of Formality from Corpora. Proceedings. Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods, 7th Language Resources and Evaluation Conference, 17-22 May, Valetta, Malta, pp. 605—616.

Cohen, Jacob (1960). Coefficient of agreement for nominal scales. Educational and Psychological Measurement. VOL. XX, No. 1, pp. 37-46.

Dempsey, K.B., McCarthy, P.M., & McNamara, D.S. (2007). Using phrasal verbs as an index to distinguish text genres. In D. Wilson and G. Sutcliffe (Eds.), proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference, Menlo Park, California: The AAAI Press, pp. 217-222.

Denison, David (1994). A Corpus of Late Modern English Prose. In Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March, 1993, edited by Merja Kytö, Matti Rissanen, and Susan Wright, Amsterdam, Rodopi, pp. 7-16.

Dumaine, Deborah, and Healey, Elisabeth C (2003). Instant-Answer Guide to Business Writing: An A-Z Source for Today's Business Writer. 2003 Edition, Writers Club Press, Lincoln, pp. 153-156.

Garner, Bryan A. (2001). A Dictionary of Modern Legal Usage. 2nd Edition, pp. xxv, Oxford University Press, US.

Gatt, Albert, and Reiter, Ehud (2009). SimpleNLG: A realisation engine for practical applications. In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), 30-31 March, Athens, Greece, pp. 90-93.

Gibaldi, Joseph (2003). MLA Handbook for Writers of Research Papers. 6th edition, Section 7.4, Modern Language Association of America, (ISBN: 0873529863 / 0-87352-986-3).

Gillett, Andy, Hammond, Angela, and Martala, Mary (2009). Inside Track to Successful Academic Writing. Pearson Education, ISBN: 978-0273721710.

Goldberg, E., Driedger, N., and Kittredge, R. I. (1994). Using Natural Language Processing to Produce Weather Forecasts. IEEE Expert: Intelligent Systems and Their Applications, 9(2): 45-53.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Hayakawa, S. I., editor. (1994). Choose the Right Word: A Contemporary Guide to Selecting the Precise Word for Every Situation. Second Edition, revised by Eugene Ehrlich. HarperCollins Publishers, NY, USA.

Heylighen, Francis, and Dewaele, Jean-Marc. (1999). Formality of language: definition and measurement. Internal Report, Center "Leo Apostel", Free University of Brussels.

Hovy, Eduard H. (1988). Generating Natural Language under Pragmatic Constraints. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, pp. 82 – 87.

Hovy, Eduard H. (1990). Pragmatics and natural language generation. Artificial Intelligence. vol. 43, pp. 153–197.

Joachims, Thorsten (1997). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. LS8-Report 23, Universitat Dortmund, LS VIII-Report.

Kennedy, A., and Shepherd, M. (2005). Automatic Identification of Home Pages on the Web. Proceedings of the 38th Hawaii International Conference on System Sciences.

Lim, Chul, Lee, Kong, and Kim, Gil (2005). Automatic Genre Detection of Web Documents. In Su K., Tsujii J., Lee J., Kwong O. Y. (eds.) Natural Language Processing, Springer, Berlin.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press, Chapter 8, pp. 165- 166.

Manning, Christopher D., and Schütze, Hinrich (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA,

Mason, J., Shepherd, M., and Duffy, J. (2009). An n-gram based approach to automatically identifying web page genre. Proceedings 41st Annual Hawaii International Conference on System Sciences (HICSS-42).

McDonald, David D. (1987). Natural language generation. In Stuart C. Shapiro, editor, Encyclopaedia of Artificial Intelligence. John Wiley and Sons, pp. 642-655.

Obrecht, Fred (1999). Minimum Essentials of English. 2nd Edition, Barron's Educational Series Inc., Los Angeles Pierce College, New York, page 13.

Obrecht, Fred, Ferris, Boak (2005). How to Prepare for the California State University Writing Proficiency Exams. 3rd Edition, Barron's Educational Series Inc., New York, page 173.

Park, David (2007). Identifying & using formal & informal vocabulary. IDP Education, the University of Cambridge and the British Council, the Post Publishing Public Co. Ltd.

Pavlidis, Mara (2009). Target Your Study Skills: Optimise Your Learning. Faculty of Health Sciences. La Trobe University, Victoria, Australia, Chapter 5, page 3.

Pearl Production (Ed) (2005). English Language Arts Skills & Strategies Level 5. Saddleback Publishing, Inc. USA, ISBN 1-56254-839-5.

Pyysalo, Sampo, Ginter, Filip, Pahikkala, Tapio, Boberg, Jorma, Jouni, Jarvinen, and Salakoski, Tapio (2006). Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. International Journal of Medical Informatics, 75(6):430-442.

Qu, Hong, La Pietra, Andrea, and Poon, Sarah (2006). Blog classification using NLP: Challenges and pitfalls. In Spring Symposium Series Technical Reports. AAAI Press.

Redman, Stuart (2003). English vocabulary in use: Pre-intermediate & intermediate. 2nd Edition, Cambridge University press, UK.

Reiter, Ehud and Belz, Anja (2009). An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. Computational Linguistics 25:529–558.

Reiter, Ehud and Dale, Robert (2000). Building Natural Language Generation Systems (Studies in Natural Language Processing). Cambridge University Press 2000. ISBN 0-521-62036-8.

Reiter, Ehud, Mellish, Chris, and Levine, John (1995). Automatic Generation of Technical Documentation. Applied Artificial Intelligence 9, pp. 259-287.

Reiter, Ehud, Robertson, Roma, and Osman, Liesl (2003) Lessons from a Failure: Generating Tailored Smoking Cessation Letters. Artificial Intelligence, 144, pp. 41-58.

Renkema, J. (1984). On Functional and Computational LSP Analysis: the Example of Officialese. In: Pugh, A.K., and Ulijm, J.M (eds), Reading for Professional Purposes: Studies in Native and Foreign Languages. London: Heinemann Educational, pp. 109 – 119.

Rob, S. et al. (2008). How to Avoid Colloquial (Informal) Writing. WikiHow.

Siddiqi, Anis (2008). The Difference Between Formal and Informal Writing. EzineArticles.

Tapanainen, P. and Timo Järvinen. (1997). A nonprojective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C. Association for Computational Linguistics, pp. 64–71.

Witten, Ian H., Frank, Eibe. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.

Woods, Geraldine (2010). English Grammar for Dummies. 2nd Edition, page 147, Wiley Publishing, Inc. NJ, USA.

Yu-shan, Chang and Yun-Hsuan, Sung (2005). Applying Name Entity Recognition to Informal Text. Ling 237 Final Projects.

Zapata, Argenis A. (2008). Inglés IV (B-2008). Universidad de Los Andes, Facultad de Humanidades y Educación, Escuela de Idiomas Modernos.

Appendix A

Informal versus Formal Lists

1. Informal versus Formal List of Words

Informal	Formal	Frequencies for Informal words	Frequencies for Formal words
a bit	a little	9	7
about	approximately	0	3
about	concerning	47	2
absentminded	abstracted	3	0
absorb	ingest	93	1
abundant	copious	93	0
add	affix	5	0
advise	counsel	0	16
again and again	repeatedly	0	3
aim	intend	0	3
allot	allocate	11	3
allow	permit	21	6
and	furthermore	15	0
and	in addition	0	0
anybody	anyone	134	7
anyplace	anywhere	0	2
apathy	anomie	47	0
appeal	petition	3	0
applause	acclamation	93	0

application	requisition	93	0
approval	commendation	5	0
around	approximately	0	3
arty	mannered	0	0
ask	enquire	0	0
ask about	enquire	11	0
ask for	request	21	5
assert	aver	15	0
assign	designate	0	2
assume	postulate	134	0
at first	initially	0	4
at once	immediately	47	23
attribute	characteristic	3	0
authority	jurisdiction	93	0
avoid	eschew	93	0
aware	cognizant	5	0
awful	ill	0	3
basic	fundamental	0	0
be going to	will	11	639
beach	littoral	21	0
beg	plead	15	0
beginner	novice	134	0
belittle	minimize	0	1
bellyache	whine	47	0
better	superior	3	5
better	ameliorate	93	0
big	major	93	73
big	large	5	44
bigger	greater	0	19
bitterness	acerbity	0	0

blabbermouth	informer	0	0
blame	reprehension	11	0
blessedness	beatitude	21	0
bloody	sanguinary	15	0
boozer	drunkard	134	0
boundless	illimitable	47	0
brag	vaunt	3	0
brilliant	resplendent	93	0
bring	convey	93	1
bring up	vomit	5	0
broad-minded	complaisant	0	0
broke	insolvent	0	0
bug	exasperate	0	0
build	construct	0	4
bully	terrorize	11	0
busy	occupied	21	0
but	however	15	58
buy	purchase	0	15
cancel	eradicate	134	0
carry	transport	0	17
catch	apprehend	47	0
catch on	understand	3	11
cease	desist	93	0
chance	opportunity	93	15
change	transform	5	1
change	alter	0	2
chat	dialogue	0	4
cheap	inexpensive	0	0
check up	investigate	0	0
chew	masticate	11	0

childish	immature	21	0
choose	select	15	0
chubby	portly	0	0
chump	fool	134	0
clean	immaculate	47	0
clear	transparent	3	0
clear	unmistakable	93	0
cleave	sunder	93	0
climb	ascend	5	0
clothing	apparel	0	4
comfort	condole	0	0
command	directive	11	0
conceit	vanity	21	0
concern	solicitude	15	0
conduct	deportment	0	0
conference	assembly	134	0
confusion	disarray	0	0
conscious	mindful	47	3
consider	deem	3	0
console	solace	93	0
control	govern	93	1
convert	transmute	5	0
copy	replica	0	0
cowardly	craven	0	0
coworker	associate	0	3
crony	friend	0	1
crowd	throng	11	0
cry	wail	21	0
cuddle	fondle	15	0
curse	anathema	0	0

cute	pretty	134	8
dad	father	0	1
daily	diurnal	47	0
deal with	handle	3	22
decay	decompose	93	0
decent	ethical	93	2
dedicate	consecrate	5	0
delete	expunge	0	0
delicious	flavorful	0	0
determine	ascertain	0	1
difficult	arduous	11	1
digest	imbibe	21	0
diligent	assiduous	15	0
dim	indistinct	0	0
disable	incapacitate	134	0
disapproval	aspersion	47	0
disaster	calamity	3	0
discussion	colloquy	93	0
disease	malady	93	1
disgusting	repugnant	5	0
do	perform	0	1
dog	hound	11	0
doubt	dubiety	21	0
doubter	skeptic	15	0
douse	submerge	0	0
dread	foreboding	134	0
drive	impel	0	0
drop	decline	47	49
drunk	intoxicated	3	0
dry	desiccated	93	0

dub	designate	93	2
dumb	obtuse	5	0
dunk	immerse	0	0
duplicate	facsimile	0	0
earlier	previous	0	53
eat	dine	11	0
edgy	restless	21	1
embarrassement	discomposure	15	0
empty	vacant	0	2
encourage	gladden	134	0
end	finish	0	0
endless	unending	47	0
enough	sufficient	3	4
erase	efface	93	0
everlasting	unending	93	0
every year	annually	5	7
everybody	everyone	0	9
everyday	quotidian	0	0
evil	nefarious	0	0
excuse	remit	0	0
explain	elucidate	11	0
facts	data	21	37
fair	disinterested	15	0
fall	decrease	0	4
famous	renowned	134	0
farming	agriculture	0	32
farsightedness	prescience	47	0
Fast	swift	3	2
fat	corpulent	93	0
feeling	emotion	93	0

fib	lie	5	2
field	specialization	0	0
filmy	diaphanous	0	0
find	locate	0	2
find	discover	0	0
fix	rectify	11	1
flabbergasted	astounded	21	0
flashy	gaudy	15	0
fleshy	mesomorphic	0	0
flimsy	unsubstantial	134	0
foretell	augur	47	1
forgive	pardon	3	0
fragile	frangible	93	0
Free	release	93	10
Free	exempt	5	1
fridge	refrigerator	0	0
friendly	amiable	0	0
frisky	sportive	11	0
funny	comic	21	0
gabby	talkative	15	0
gap	aperture	0	0
gardening	tillage	134	0
gather	convene	0	2
generous	magnanimous	47	0
get	obtain	3	7
get	acquire	93	8
get out	leave	93	8
get smaller	decrease	5	4
gist	crux	0	0
give	donate	0	2

give	contribute	0	2
give out	distribute	0	4
glasses	spectacles	11	0
gleaming	luminous	21	0
go	depart	15	0
go down with	contract	0	44
go through	perform	134	1
go up	increase	0	99
goal	objective	47	21
good	beneficial	3	0
goodwill	generosity	93	0
goof	mistake	93	3
gourmet	gastronome	5	0
great	reputable	0	0
greedy	avaricious	0	0
grill	interrogate	0	0
gripe	complain	11	0
grown-up	adult	21	0
guess	believe	15	20
guy	man	134	0
happiness	felicity	0	0
hard	laborious	47	0
harshness	acrimony	3	0
have to	must	93	55
heavy	burdensome	93	2
hefty	ponderous	5	0
help	assist	0	7
helper	assistant	0	8
high	elevated	0	0
hint	insinuation	0	0

hire	employ	11	1
hoard	preserve	21	4
hobby	avocation	15	0
home	residence	0	0
house	residence	134	0
house	dwelling	0	0
hug	caress	47	0
huge	enormous	3	2
humanity	humankind	93	0
humorous	jocular	93	0
hurry	expedite	5	2
illness	infirmity	0	0
imply	connote	11	0
important	consequential	21	0
improve	ameliorate	15	0
in charge	responsible	0	3
in the end	finally	134	2
inbred	innate	0	0
incidental	adventitious	47	0
include	comprise	3	5
indirectness	circumlocution	93	0
inhabit	reside	93	1
interject	interpose	5	0
jam	preserve	0	4
jaundiced	cynical	11	1
job	occupation	21	0
jolt	impact	15	22
keep	preserve	0	4
keep	retain	134	3
kid	child	47	14

kind of	somewhat	3	1
kind of	approximately	93	3
kindness	benevolence	93	0
lack	deficiency	5	0
lady	woman	0	7
lay back	relax	0	0
laziness	indolence	0	0
learn	detect	11	1
learner	scholar	21	0
learning	pedantry	15	0
leave	depart	0	0
leftover	residue	134	1
lessen	allay	0	0
let	permit	47	6
let	authorize	3	0
letter	correspondence	93	1
letup	respite	93	0
lighten	alleviate	5	1
like	such as	0	0
lit up	illuminate	0	0
live	reside	11	1
lively	energetic	21	0
loaded	prosperous	15	0
loneliness	disaffection	0	0
lonesome	lonely	134	0
look up	research	47	22
loud	clamorous	3	0
lucky	fortunate	93	3
lukewarm	tepid	93	0
mad	insane	5	0

mainly	principally	0	0
make sure	ensure	0	16
many	numerous	0	2
maybe	possibly	0	4
maybe	perhaps	11	2
mean	in other words	21	0
means	instrumentality	15	0
meant	denote	0	0
mend	repair	134	1
method	procedure	0	4
middle	midst	47	0
modify	transform	3	1
mom	mother	93	4
moral	virtuous	93	0
move	transfer	5	16
mushy	maudlin	0	0
nab	arrest	0	1
need	require	0	7
neighboring	contiguous	0	2
next	subsequently	11	0
nice	agreeable	21	0
nitpicking	pedantic	15	0
nobody	no one	0	0
nosy	prying	134	0
numb	anesthetized	47	0
numbskull	dullard	3	0
obscure	arcane	93	0
offer	proffer	93	0
ok	satisfactory	5	2
old	aged	0	1

older	senior	0	33
old-fashioned	archaic	0	0
on and off	intermittently	0	0
oppose	gainsay	11	0
optimistic	sanguine	21	0
originate	emanate	15	0
outcome	denouement	0	0
outstanding	paramount	134	1
own	possess	0	0
pale	wan	47	0
parched	dehydrated	3	0
participate	partake	93	0
pay	compensation	93	4
peak	summit	5	0
phone	telephone	0	6
photo	photograph	0	0
piddling	negligible	11	0
pigeonhole	categorize	21	0
plan	scheme	15	13
plane	aeroplane	0	0
Plucky	valiant	134	0
portion	passage	47	2
power	sway	3	0
praise	laud	93	0
preacher	minister	93	77
premonition	presentiment	5	0
present	gift	0	2
pretty much	essentially	0	2
prize	award	11	6
project	undertaking	21	2

promise	assure	15	0
prompt	motivate	0	2
promptness	alacrity	134	0
pushy	ambitious	47	0
put up	manage	3	1
quick	rapid	93	7
quit	resign	93	0
quotation	quote	5	0
raunchy	risque	0	0
really	quite	0	4
reasoning	ratiocination	11	0
rebirth	renaissance	21	0
redundant	pleonastic	15	0
relentless	inexorable	134	0
remain	abide	0	2
remember	recall	47	1
replace	supersede	3	0
request	appeal	93	7
resemble	parallel	93	2
resolution	fortitude	5	0
rest	repose	0	0
rile	annoy	0	1
ripen	mature	0	2
risk	jeopardy	11	1
rob	extort	21	0
rot	spoil	15	0
sanction	approbation	134	0
say no	reject	47	1
scanty	exiguous	3	0
scold	chide	93	0

seem	appear	93	14
send	transmit	5	0
send back	return	0	27
sentiment	affect	0	11
set up	request	11	5
shameful	dishonorable	21	0
sharp	acute	15	0
shining	effulgent	0	0
shiv	knife	134	0
shock	impact	0	22
shorten	decrease	47	4
shorten	reduce	3	17
show	demonstrate	93	2
show	evince	93	0
show up	appear	5	14
sickness	ailment	0	0
sight	vision	0	1
skimpy	meager	0	0
slander	aspersion	11	0
slapdash	cursory	21	0
slushy	mawkish	15	0
small	diminutive	0	0
smooch	kiss	134	0
snatch	seize	0	2
sneaky	underhand	47	0
so	therefore	3	2
so	consequently	93	0
sociable	social	93	24
somebody	someone	5	10
sort	type	0	5

sort of	Rather	0	0
so-so	mediocre	0	0
sot	alcoholic	0	0
sourness	asperity	11	0
speech	oration	21	0
speed	velocity	15	0
spread	propagate	134	0
spud	potato	0	0
stab	penetrate	47	0
start	begin	3	26
stick	adhere	93	1
stickup	larceny	93	0
stint	assignment	5	0
stipend	emolument	0	0
stop	cease	0	0
stop	halt	0	1
story	narrative	11	0
strong	stalwart	21	0
stuff	materials	15	6
stuff	things	0	19
stuff	items	134	3
surround	circumscribe	47	0
swamp	deluge	3	0
swap	barter	93	1
sweat	perspiration	93	0
swing	oscillate	5	0
tactful	diplomatic	0	2
take on	employ	0	1
takeoff	mimicry	0	0
tasty	palatable	0	0

teach	educate	11	0
tell	inform	21	0
tell	recount	15	0
thing	matter	0	15
thing	issue	134	57
think	cogitate	0	0
timeless	eternal	47	0
times	multiply	3	0
tip	gratuity	93	0
tired	fatigued	93	0
tomb	sepulcher	5	0
too	also	0	202
totally	completely	0	7
touching	poignant	0	1
trim	ornament	11	0
trip	voyage	21	0
truthful	veracious	15	0
try	endeavour	0	0
tune	song	134	0
unchangeable	immutable	47	0
understanding	comprehension	3	0
unhappy	dissatisfied	93	0
unruly	intractable	93	0
unselfishness	altruism	5	0
upset	disturb	0	0
uptight	nervous	0	2
use	consume	11	0
various	sundry	21	0
very	highly	15	5
want	desire	0	1

want	wish	134	3
watch	observe	47	0
wealthy	affluent	3	0
whole	complete	93	11
whole	entire	93	2
willing	compliant	5	0
wisecrack	joke	0	1
wordy	verbose	0	0
workable	feasible	0	0
worry	apprehension	11	0
worse	inferior	21	0
wrong	incorrect	15	1
yardstick	criterion	0	0

Table A.1: Informal/formal parallel list of words, phrase, and expressions with its frequencies in the dataset.

2. Contractions versus Full Forms List of Words

Contraction (Informal)	Full Form (Formal)
ain't	am not
aren't	are not
can't	cannot
couldn't	could not
didn't	did not
doesn't	does not
don't	do not
hadn't	had not
hasn't	has not
haven't	have not
he'd	he had
he'd	he would

he'll	he shall
he'll	he will
he's	he is
he's	he has
I'd	I would
I'd	I had
I'll	I shall
I'll	I will
I'm	I am
I've	I have
isn't	is not
let's	let us
mightn't	might not
mustn't	must not
shan't	shall not
she'd	she had
she'd	she would
she'll	she will
she'll	she shall
she's	she is
she's	she has
shouldn't	should not
that's	that is
that's	that has
there's	there is
there's	there has
they'd	they would
they'd	they had
they'll	they will
they'll	they shall

they're	they are
they've	they have
we'd	we would
we'd	we had
we're	we are
we've	we have
weren't	were not
what'll	what shall
what'll	what will
what're	what are
what's	what is
what's	what has
what've	what have
where's	where has
where's	where is
who's	who had
who's	who would
who'll	who will
who'll	who shall
who're	who are
who's	who has
who's	who is
who've	who have
won't	will not
wouldn't	would not
would've	would have
you'd	you had
you'd	you would
you'll	you will
you'll	you shall

you're	you are
you've	you have
that'll	that will
it's	it is
we'll	we will
we'll	we shall
it'd	it would
it'd	it had

Table A.2: Parallel list of contractions and full forms.

3. Abbreviations versus Full Forms List of Words

Abbreviations (Informal)	**Full Form (Formal)**
e.g.	for example
i.e.	that is
etc.	and so on
ok	Okay
Jan.	January
Feb.	February
Mar.	March
Apr.	April
May.	May
Jun.	June
Jul.	July
Aug.	August
Sep.	September
Oct.	October
Nov.	November
Dec.	December
Sat.	Saturday

Sun.	Sunday
Mon.	Monday
Tue.	Tuesday
Wed.	Wednesday
Thu.	Thursday
Fri.	Friday
plane	airplane
lab.	laboratory
asap	as soon as possible
USA	United States of America
tons	tonnes
undergrad.	undergraduate
grad.	graduate
HR	Human Resources
Prof.	Professor
AI	Artificial Intelligence
ur	your
&	and
abt	about
mt	might
mt.	might
shd	should
shd.	should
wch	which
wd	would
wd.	would
wh	which
wh.	which
yr	your
yr.	your

yrs	yours
yrs.	yours
2d	second
2n	second
3d	third
acct	account
acct.	account
advt	advertisement
aftern	afternoon
aftern.	afternoon
aftn	afternoon
aftn.	afternoon
AM	ante meridiem
A M	ante meridiem
anti-Cathc.	anti-Catholic
Arc.	Archaeological
Ass.	Association
B.M.	British Museum
bkfst	breakfast
bkfst.	breakfast
brkfst	breakfast
C.B.E.	Commander of the Order
cent	century
chap	chapter
chap.	chapter
chaps	chapters
depart.	department
edn	edition
Eng.	English
H. of Lds	House of Lords

Hist.	History
hrs.	hours
in.	inches
Ld.	Lord
Ldy	Lady
Ldy.	Lady
lovg	loving
Ly	Lady
Ly.	Lady
max.	maximum
mem.	memoires
min.	minutes
min:	minutes
Phys.	Physical
prob.	probably
recvd	received
Secty.	Secretary
temp.	temperature
to =	to equal
Trans.	Transactions
Univ.	University
vol.	volume
Xmas	Christmas
Xtian	Christian
yestdy	yesterday
yrs	years
acad.	academic
adm.	administration
bib.	bible

Table A.3: Abbreviations and full forms parallel list.

Appendix B

Classification Results on Medical Documents

The size of the data set is 1980 documents: 990 informal texts and 990 formal texts.

The results of our experiments are presented in Table B.1.

Machine Learning Algorithm	F-measure (Weighted Avg.)	Accuracy
Support Victor Machine (SMO)	0.977	0.977
Decision Trees (J48)	0.972	0.972
Naïve Bayes (NB)	0.965	0.965

Table B.1: Classification results of SVM, Decision Trees, and Naïve Bayes classifiers using 10-fold cross validation.

Class	Precision	Recall	F-Measure
Informal	0.991	0.963	0.976
Formal	0.964	0.991	0.977
Weighted Avg.	0.977	0.977	0.977

Table B.2: Detailed evaluation for each class, for the SVM classifier.

Features	Weight
Word length's average	0.745
Active voice	0.5719
Informal pronouns	0.5636
Contractions	0.4571

Passive voice	0.2192
Informal list	0.1913
Type Tokens Ratio (TTR)	0.1598
Formal pronouns	0.0913
Formal list	0.0815
Phrasal verbs	0.0748
Abbreviations	0.0168

Table B.3: Our model's features with their InfoGain scores for the dataset of medical texts.

Appendix C

The Generated Sentences from the Proposed System

No.	Sentence	Actual	Judge1	Judge2	Agreement
1	I'm choosing two books.	0	0	0	Informal - Yes
2	I'll have asked about the admission.	0	0	0	Informal - Yes
3	You didn't ask for a car & van.	0	0	0	Informal - Yes
4	I haven't been buying a new phone.	0	0	0	Informal - Yes
5	He won't catch on the idea	0	0	0	Informal - Yes
6	I didn't change the lock for my door.	0	0	0	Informal - Yes
7	We've been checking up the samples in the lab.	0	0	0	Informal - Yes
8	The guy is going to choose one helper for the new job.	0	0	0	Informal - Yes
9	I'd dealt with any problem that is going to happen.	0	0	0	Informal - Yes
10	I ain't finding my keys, so I'm unhappy.	0	0	0	Informal - Yes
11	You'll get a new job.	0	0	0	Informal - Yes
12	You hadn't gotten out early this Fri.	0	0	0	Informal - Yes

13	She isn't giving some money to poor guy.	0	0	0	Informal - Yes
14	The plane is going to go on Jan. 5th.	0	1	0	Informal - No
15	I'd given out my old clothes, & shoes.	0	0	0	Informal - Yes
16	The boss is giving $1000 to the older guy.	0	0	0	Informal - Yes
17	The kid brought up two times today. I'll see the doctor.	0	0	0	Informal - Yes
18	I'd gone up your salary. so You've to work hard.	0	0	0	Informal - Yes
19	You've been believing that is going to be changed asap.	0	0	0	Informal - Yes
20	They won't encourage cute girls to apply for the new job.	0	0	0	Informal - Yes
21	They didn't keep my options open.	0	0	0	Informal - Yes
22	Nobody has been letting to smoke inside the building.	0	0	0	Informal - Yes
23	The lady is going to live alone, but She's enjoying.	0	0	0	Informal - Yes
24	I'll make sure You'll happy if you listen to me.	0	0	0	Informal - Yes
25	He's fixing a bit stuff in my car.	0	0	0	Informal - Yes
26	She'll put up with her neighbour's noise.	0	0	0	Informal - Yes
27	My dad was needing special tools to fix the fence.	0	0	0	Informal - Yes
28	My son is saying no for my offer, it'd be nice if you talk with him.	0	0	0	Informal - Yes

29	Anybody is going to help my dad, I'll give him the prize.	0	0	0	Informal - Yes
30	It seems so-so problem, so don't worry.	0	0	0	Informal - Yes
31	You've been sending back my cheque last Sat., Jun. 2nd.	0	1	0	Informal - No
32	I'll be asking for at once 2 helper.	0	0	0	Informal - Yes
33	She shortened his working hours last Mon.	0	1	0	Informal - No
34	He's been showing the new system for everybody.	0	0	0	Informal - Yes
35	I won't tell the HR manager to take on somebody.	0	0	0	Informal - Yes
36	Prof. Tong is teaching AI course every year.	0	1	0	Informal - No
37	I'll go through a look up with somebody.	0	0	0	Informal - Yes
38	You didn't want to learn ur fault.	0	0	0	Informal - Yes
39	I don't want to change my schedule for this week.	0	0	0	Informal - Yes
40	Sara had notified us that She'll get out early today.	0	1	0	Informal - No
41	Our boss won't assign a new helper for the project.	0	0	0	Informal - Yes
42	He didn't build a house for his dad.	0	0	0	Informal - Yes
43	I don't feel very well since last Fri.	0	0	0	Informal - Yes
44	I won't let you down, I'll help you.	0	0	0	Informal - Yes
45	It's my fault.	0	0	0	Informal - Yes

46	We're really happy to see you.	0	0	0	Informal - Yes
47	It won't be easy, but We've a chance.	0	0	0	Informal - Yes
48	My coworker hasn't helped me to go through the task.	0	0	0	Informal - Yes
49	The helper didn't write the letter the our boss.	0	0	0	Informal - Yes
50	He doesn't worry so, keep going.	0	0	0	Informal - Yes
51	It appears repeatedly in the model.	1	1	1	Formal - Yes
52	Specific activities and goals had been determined by us.	1	1	1	Formal - Yes
53	He will be beginning a new occupation on next Monday.	1	1	1	Formal - Yes
54	She has wished to work as soon as possible.	1	1	0	Formal - No
55	The residence has required some works, We will repair it.	1	1	1	Formal - Yes
56	The employer will depart early today, therefore We will meet tomorrow.	1	1	1	Formal - Yes
57	The woman is consecrating to finish the task on time.	1	1	1	Formal - Yes
58	The salaries for the employees had not been decreased by the corporation.	1	1	1	Formal - Yes
59	The child was wailing much today.	1	1	1	Formal - Yes
60	We have gladdened anyone to obtain share in our corporation.	1	1	1	Formal - Yes

61	They were transmuting the raw materials to finished goods.	1	1	1	Formal - Yes
62	His father had transformed the new system, however it is still require more work.	1	1	1	Formal - Yes
63	Her mother will depart to United States of America on Sunday, June 21st.	1	1	1	Formal - Yes
64	The window that is completely was broken had been repaired by the man.	1	1	1	Formal - Yes
65	My record from the system was expunged by him.	1	1	1	Formal - Yes
66	Their reputation over the word is being propagated by them.	1	1	1	Formal - Yes
67	He had postulated that you will agree my proffer.	1	1	1	Formal - Yes
68	We will permit the students to dine only in the break.	1	1	1	Formal - Yes
69	They are not performing the undertaking in a reputable way.	1	1	1	Formal - Yes
70	Mike had petitioned a new corporation last year on December 23rd.	1	1	1	Formal - Yes
71	My invitation for the dinner is rejected by Natasha.	1	1	1	Formal - Yes
72	Two undergraduate courses, furthermore one graduate course is being educated by Professor Tomas.	1	1	1	Formal - Yes
73	The author had been apprehending his mistake, therefore he will repair	1	1	1	Formal - Yes

	it.				
74	3 employees in the Human Resources department are employed by the British council	1	1	1	Formal - Yes
75	We had transferred to new residence last Monday.	1	1	1	Formal - Yes
76	They will not transmit the correspondence to my college.	1	1	1	Formal - Yes
77	Twenty members were observing the voting procedure to select the president.	1	1	1	Formal - Yes
78	United States of America is constructing a major bridge in California.	1	1	1	Formal - Yes
79	It appears that the new traffic law will transform numerous rules.	1	1	1	Formal - Yes
80	She will purchase a used car subsequently week, I will assist her.	1	1	1	Formal - Yes
81	The new procedure to attract more people had been elucidated by him.	1	1	1	Formal - Yes
82	We transformed our corporation name last year on May 20th.	1	1	1	Formal - Yes
83	He will not enquire his associate how to perform the task.	1	1	1	Formal - Yes
84	Approximately 500 tonnes are transported by the vessel.	1	1	1	Formal - Yes
85	They had been endeavouring to obtain a superior deal.	1	1	1	Formal - Yes
86	My aunt has numerous children.	1	1	1	Formal - Yes
87	The problem in the building is	1	1	1	Formal - Yes

	investigated by Safety officer.				
88	The cost of cleaning services has increased by 25% over the last three years.	1	1	1	Formal - Yes
89	It had been suggesting that no action should be taken at this stage.	1	1	1	Formal – Yes
90	This result is dependent on factors such as weight, age, and so on.	1	1	1	Formal – Yes
91	We will play soccer on Friday, 23 of January.	1	1	1	Formal – Yes
92	My wife had begotten a pretty girl last Saturday.	1	1	1	Formal – Yes
93	We will be travelling this summer for vacation.	1	1	1	Formal – Yes
94	I had impacted from my Mark Yesterday.	1	1	1	Formal – Yes
95	It was a rapid decision to discharge Mike from his work.	1	1	1	Formal – Yes
96	The police officer had permitted the driver to depart after giving him a fine.	1	1	1	Formal – Yes
97	I would like to take pictures for my children.	1	1	1	Formal – Yes
98	The Sun will rise at 5:20 morning Tomorrow.	1	1	1	Formal – Yes
99	We designated someone to assist building the vessel.	1	1	1	Formal – Yes
100	They had located a procedure that will enhance the productivity.	1	1	1	Formal – Yes

Table C.1: Informal and Formal sentences generated by our system using different parameters