Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, March 2012

Learning to Classify Documents According to Formal and Informal Style

Fadi Abu Sheikha and Diana Inkpen

Published by CSLI Publications

LiLT volume 8, issue 1 March 2012

Learning to Classify Documents According to Formal and Informal Style

Fadi Abu Sheikha and Diana Inkpen, *School of Electrical Engineering and Computer Science, University of Ottawa, Canada*

Abstract

This paper discusses an important issue in computational linguistics: classifying texts as formal or informal style. Our work describes a genre-independent methodology for building classifiers for formal and informal texts. We used machine learning techniques to do the automatic classification, and performed the classification experiments at both the document level and the sentence level. First, we studied the main characteristics of each style, in order to train a system that can distinguish between them. We then built two datasets: the first dataset represents general-domain documents of formal and informal style, and the second represents medical texts. We tested on the second dataset at the document level, to determine if our model is sufficiently general, and that it works on any type of text. The datasets are built by collecting documents for both styles from different sources. After collecting the data, we extracted features from each text. The features that we designed represent the main characteristics of both styles. Finally, we tested several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines, in order to choose the classifier that generates the best classification results.

LiLT Volume 8, Issue 1, March 2012.
Learning to Classify Documents According to Formal and Informal Style.

1 Introduction

The need to identify and interpret possible differences in the linguistic style of texts–such as formal or informal–is increasing, as more people use the Internet as their main research resource. There are different factors that affect the style, including the words and expressions used and syntactical features (Karlgren, 2010). Vocabulary choice is likely the biggest style marker. In general, longer words and Latin origin verbs are formal, while phrasal verbs and idioms are informal (Park, 2007). There are also many formal/informal style equivalents that can be used in writing.

The formal style is used in most writing and business situations, and when speaking to people with whom we do not have close relationships. Some characteristics of this style are long words and using the passive voice. Informal style is mainly for casual conversation, like at home between family members, and is used in writing only when there is a personal or closed relationship, such as that of friends and family. Some characteristics of this style are word contractions such as "won't", abbreviations like "phone", and short words (Park, 2007). We discuss the main characteristics of both styles, in Section 3.

In this paper, we explain how to build a model that will help to automatically classify any text or sentence as formal or informal style. We tested several classification algorithms, in order to determine the classifier that generates the best classification results.

2 Background and Related Work

In this section, we briefly introduce machine learning techniques and describe the three machine learning algorithms we used: Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machines (SVM). For more details about machine learning algorithms see (Witten and Frank, 2005). We then discuss related work of text classification by formality and genre.

2.1 Machine Learning

Machine learning is about designing computer models that can learn from examples. The models can be used for prediction, explanation and understanding data. Machine learning can be implemented as 'supervised' (using labelled training data) or 'unsupervised' (using unlabelled data). A brief explanation of supervised learning for automatic classification follows.

Supervised learning is a machine learning method in which a system receives a training dataset consisting of many instances, with each

instance represented by different parameter values (input) and a class (output). The parameter values are known as attributes or features, and can be represented by numeric values (e.g., term frequency) or nominal values (e.g., Yes, No). The system infers a mathematical function, which automatically maps an input signal to an output signal. Thus, the system can determine the class (output) for any new instances with new attributes (inputs).

There are many algorithms that can be used for supervised learning. For our experiments we chose three classifiers (Decision Tree, Naïve Bayes, and Support Vector Machine) in order to have a classifier that is interpretable to humans (Decision Tree), one that works well with text (Naïve Bayes), and another known to achieve very good performance overall (SVM).

2.1.1 Decision Tree Learning Algorithm

The decision tree algorithm produces a tree-structured set of nodes that leads to a decision, with each node being either a decision node or a leaf node (indicating the value of the target attribute class). Figure 1 shows an example of a decision tree for whether to play a game outdoors, based on weather forecasting.

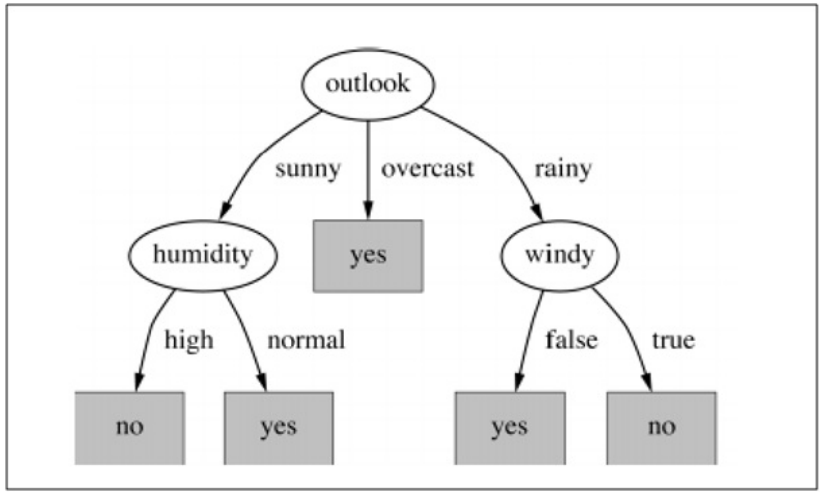

FIGURE 1 An example of Decision Tree (Witten and Frank, 2005).

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

The resulting Decision Tree is simple to understand and interpret (people are generally able to understand decision tree models after a brief explanation). This algorithm is fast, and it tends to perform well with large datasets.

2.1.2 Naïve Bayes Learning Algorithm

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

2.1.3 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 divides the data into two categories. It selects a small number of critical boundary instances (support vectors) from each class, and builds a linear discriminant function that separates them as widely as possible (Witten and Frank, 2005). After training, the SVM algorithm can predict if a new example will be in one class or the other.

2.2 Text Classification by Formality and Genre

Though there seems to be little research on automatic text classification by formal and informal style, some that has been done on automatic text classification by genre is relevant to our work. There is much research on classifying texts by topic, but this does not apply in our case since the texts can have different styles and be about the same topic, or, similarly, they can be about different topics and have the same style. In addition to classification by topic, there is research on classifying texts by author (from a set of possible authors), by the gender of the author, by opinion (positive, negative, neutral), or by emotion classes (happy, sad, angry, etc.). These are also not directly relevant to our work. Here we discuss related work on formal/informal classification and genre classification.

Heylighen and Dewaele (1999) proposed a method to determine the degree of formality for any text using a special formula, 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 according to an increase in formality. In our work, we want to build a model based on the main characteristics of formal and informal style so we can classify any text

into formal or informal, rather than creating a model based on the frequency of the words in both classes.

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

Brooke, Wang, and Hirst (2010) conducted detailed tests on several corpus-based methods for deriving real-valued formality lexicons. They quantified the formality of individual lexical items, assigning each word a formality score (FS) in the range of -1 to 1. They compiled two lists of words, one formal and one informal, to use as seeds for their dictionary construction method, and as a test set for evaluation. They used the Brown corpus and the Switchboard corpus to collect documents that represent both styles, and evaluated their lexicons using relative formality judgments between word pairs. The results of their evaluation suggest that the problem is tractable but not trivial. Though they achieved good accuracy in distinguishing the difference in formality using a small diverse corpus, they believe that larger corpora and more sophisticated methods are required to capture the full range of linguistic formality.

Qu, La Pietra, and Poon (2006) designed a natural language processing (NLP) experiment to categorize 120 blogs into four topic groups: personal diary, politics, news and sports. These classes are related to the topics of the documents. They reported that machine learning methods cannot accurately categorize blogs; the task is too difficult because a blog is written in an informal style. The authors used standard statistical measures to classify blogs. Their baseline feature is term document frequency (unigrams) excluding stop words. They also tried two other linguistic features: the title text of every blog post, and the anchor text over inbound links to the blog. Their proposed method performed reasonably well in classifying blogs, achieving 84% accuracy with unigrams features. This work is not directly relevant to our work, however, as it is classification by topic. We only mention it here to demonstrate that our model can deal with informal texts (e.g., blogs) using machine learning techniques, despite what is claimed in this work.

Kennedy and Shepherd (2005) discussed one type of web page classification by genre: how to distinguish home pages from non-home pages, and then classify the home pages as personal, corporate or organization. Their dataset consists of 321 web pages, and they focus on the difficult task of subgenre discrimination. Their best accuracy is 71.4% on per-

sonal home pages with a single classifier and manual feature selection, and without noisy pages.

Lim, Lee, and Kim (2005) suggested sets of features to detect the genre of web pages. They used 1,224 web pages for their experiments, and investigated the efficiency of several feature sets to discriminate across 16 genres. The genres were: personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, informative materials, FAQs, discussions, product specification and others (informal texts). The classification efficiency was tested on different parts of the web page space (title and meta-content, body and anchors). The best accuracy they achieved was 75.7% with one of their feature sets, when applied only to the body and anchors.

Mason, Shepherd, and Duffy (2009) proposed n-gram representation to automatically identify web page genre. They used 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 increments of 1, and the number of most frequent n-grams ranges from 500 to 5,000 in increments of 500. They achieved better results than other approaches, with 94.6% accuracy on the web page genre classification task.

3 Learning Formal and Informal Style

In this section, we explain the main characteristics of the formal and informal styles. We also present parallel lists of words, phrases and expressions for both styles, which we collected from different sources. Understanding the differences between the styles facilitates building models based on the main characteristics for text classification tasks.

3.1 Characteristics of Formal versus Informal Style

Here we explain and summarize the main characteristics of formal style versus informal style, as they are 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 these differences in order to:

- distinguish between the styles;
- identify each style from texts;
- build features based on the characteristics; and,
- predict a class for new documents or sentences.

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

Explanations, examples and the characteristics of each style follow.

3.1.1 Main Characteristics of Informal Style Text

The informal style has the following characteristics:

1. It uses a personal style: the first and second person ("I" and "you") and the active voice (e.g., "I have noticed that...").
2. It uses short simple words and sentences (e.g., "latest").
3. It uses contractions (e.g., "won't").
4. It uses many abbreviations (e.g., "TV").
5. It uses many phrasal verbs in the text (e.g., "find out").
6. Words that express rapport and familiarity are often used in speech, such as "brother", "buddy" and "man".
7. It uses a subjective style, expressing opinions and feelings (e.g., "pretty", "I feel").
8. It uses vague expressions, personal vocabulary and colloquialisms (slang words are accepted in spoken text, but not in written text (e.g., "wanna" = "want to")).

3.1.2 Main Characteristics of Formal Style

The formal style has the following characteristics:

1. It uses an impersonal style: the third person ("it", "he" and "she") and often the passive voice (e.g.,"It has been noticed that...").
2. It uses complex words and sentences to express complex points (e.g., "state-of-the-art").
3. It does not use contractions.
4. It does not use many abbreviations, though there are some abbreviations used in formal texts, such as titles with proper names (e.g., "Mr.") or short names of methods in scientific papers (e.g., "SVM").
5. It uses appropriate and clear expressions, precise education, and business and technical vocabularies (Latin origin).
6. It uses polite words and formulae, such as "Please", "Thank you", "Madam" and "Sir".
7. It uses an objective style, citing facts and references to support an argument.
8. It does not use vague expressions and slang words.

Table 1 shows examples of sentences that characterize the informal style versus the 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 1 Examples of informal versus formal sentences for different characteristics.

3.2 Formal versus Informal Lists

Here we present parallel lists of formal versus informal words, phrases, and expressions. The lists were compiled manually from different sources: the first list is formal versus informal words and phrases, the second list is most of the contractions in English, and the third list is some of the common abbreviations in English. These lists are considered important features for our models for classifying text into formal or informal style.

3.2.1 Informal/Formal list of words and phrases

This is a parallel list for informal versus formal words and phrases. We populated this list manually from different sources: (Gillett et al., 2009; Park, 2007; Redman, 2003; Rob S. 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). The Informal/Formal list was very useful in our model. Table 2 shows an example of pairs of words from the Informal/Formal list. It is a general-domain list, which refers to the most common formal and informal words used in the English language, and is applicable to any domain. However, we could add more specific words to this list, to meet the requirements of certain domains (e.g., the legal domain).

Informal	Formal
about	approximately
and	in addition
anybody	anyone
ask for	request
boss	employer
but	however
buy	purchase
end	finish
enough	sufficient
get	obtain
go up	increase
have to	must

TABLE 2 An example from the lists of formal versus informal words.

3.2.2 Contractions List

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). Table 3 shows samples of the parallel list of the contractions versus the full forms. This list was a good feature of our model for text classification of formal versus informal style.

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 3 An example of the contractions versus the full forms.

3.2.3 Abbreviations List

This is a parallel list for some of the most common abbreviations in English that represent the informal style, versus the full forms that correspond to these abbreviations used in formal style. However, there are some abbreviations that are acceptable in formal texts (Obrecht, 1999). We collected this list manually from (Redman, 2003; Gibaldi, 2003; Pearl Production, 2005). In addition, we manually extracted more pairs from *A Corpus of Late Modern English Prose*(Denison, 1994). Table 4 shows samples of pairs from the parallel list of abbreviations and their corresponding full forms.

Informal	Formal
asap.	as soon as possible
grad.	graduate
HR	Human Resources
Feb.	February
Lab	laboratory
temp.	temperature

TABLE 4 A sample of the abbreviations versus the full forms.

4 Data Sets

We built three data sets: the first represents genera-domain texts at the document level, the second represents general-domain texts at the sentence level, and the third represents medical documents that will be

used to determine if our classification model works well with specific kinds of text.

4.1 General-Domain Texts (Document Level)

The size of this dataset is 1,000 documents: 500 informal texts and 500 formal texts.

4.1.1 Informal Texts

We manually compiled 500 texts that represent the informal style for general-domain texts from the following sources:

1. *A Corpus of Late Modern English Prose*[2] (Denison, 1994) This corpus contains a set of annotated texts by David Denison[3]; most of the texts are informal texts (personal letters). Figure 2 shows an example of one of these texts.
2. Enron Email Dataset/Corpus[4] : This corpus contains email texts; most are personal letters (informal texts), as reported by Yu-shan and Yun-Hsuan (2005).
3. Part of the Open American National Corpus[5] : This corpus contains some categories that are informal texts, such as spoken language transcribed texts.

I'm getting over the attack of softening of the brain of which I told you, at least getting over it a little. I ride pretty regularly in the mornings, going out soon after dawn.

I get back to the office about 9 o'clock in better heart, and above all in a better temper. War is very trying to that vital organ, isn't it? I've been doing some interesting bits of work with Sir Percy which is always enjoyable. To-day there strolled in a whole band of sheikhs from the Euphrates to present their respects to him, and incidentally they always call on me.

FIGURE 2 A sample informal text file extracted from *A Corpus of Late Modern English Prose.*

[2]http://ota.ahds.ac.uk/catalogue/index-id.html

[3]http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/lmode-prose/#Summary

[4]http://www.cs.cmu.edu/∼enron/

[5]http://www.anc.org/OANC/

4.1.2 Formal Texts

We manually collected 500 texts that represent the formal style of general-domain texts from the following sources:

1. A collection of newswire articles from the Reuters corpus[6] : This corpus contains a set of news texts; most of these texts are formal texts, as reported by Yu-shan and Yun-Hsuan (2005). Figure 3 shows a text from this set. Most of our dataset for formal texts was extracted from this corpus.
2. Part of the Open American National Corpus, written technical texts: This corpus contains some categories that are formal texts, such as formal publications.
3. *A 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. There were very few formal texts in this corpus (only 3 letters).

ICO PRODUCERS TO PRESENT NEW COFFEE PROPOSAL

LONDON, Feb 26 - International Coffee Organization, ICO, producing countries will present a proposal for reintroducing export quotas for 12 months from April 1 with a firm undertaking to try to negotiate up to September 30 any future quota distribution on a new basis, ICO delegates said.

Distribution from April 1 would be on an unchanged basis as in an earlier producer proposal, which includes shortfall redistributions totalling 1.22 million bags, they said.

Resumption of an ICO contact group meeting with consumers, scheduled for this evening, has been postponed until tomorrow, delegates said.

FIGURE 3 A sample of formal text file extracted from Reuters Corpus.

4.2 General-Domain Texts (Sentence Level)

We built our dataset for sentence-level classification by splitting all the documents that we collected in Section 4.1 into sentences. We used our

[6]http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

own tool to divide text into sentences, based on punctuation marks and several simple rules (Manning and Schutze, 1999). Each sentence is in a separate text file (we excluded the informal texts that we collected from OANC, because we could not divide them into sentences). These texts are conversation (spoken) texts without any punctuation; in particular full stops, which are required to determine the end of sentences. For such a document, the whole text would have been incorrectly considered one very long sentence.

We performed sentence splitting on all the texts, for both classes (formal and informal). We generated the following new data set:

1. 500 formal general-domain documents are divided into 5,158 sentences, representing the class of formal sentences.

2. 452 informal general-domain documents are divided into 5,373 sentences, representing the informal sentences.

In order to balance the number of instances for each class, and to have a baseline of 50:50, we randomly removed some sentences in order to retain 5,000 sentences for each class. Therefore the final dataset for the sentence classification task consists of 5000 formal and 5000 informal sentences.

4.3 Medical Texts

The size of the dataset that we collected is 1,980 documents: 990 characterize informal text and 990 characterize formal text.

4.3.1 Informal Medical Texts

We chose 990 texts that characterize the informal style as reported by Yu-shan and Yun-Hsuan (2005) from the medical newsgroups collection. This corpus, known as '20 Newsgroups' [7] contains 20 topics, and each topic has 1,000 texts that characterize informal style. We used one of these topics (the medical texts), and we excluded 10 documents which had less than two words. Figure 4 shows a sample of these informal medical texts.

4.3.2 Formal Medical Texts

We randomly selected 990 texts that characterize the formal style from the medical abstracts collection. This collection contains 23 cardiovascular diseases categories, as reported in (Joachims, 1997). Figure 5 shows a sample of one of these formal medical texts.

[7] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

There were a few people who responded to my request for info on treatment for astrocytomas through email, whom I couldn't thank directly because of mail-bouncing probs (Sean, Debra, and Sharon). So I thought I'd publicly thank everyone. Thanks!
(I'm sure glad I accidentally hit "rn" instead of "rm" when I was trying to delete a file last September. "Hmmm...'News?' What's this?"....)
-Brian

FIGURE 4 A sample informal text file extracted from medical newsgroups corpus.

Gastrointestinal tuberculosis. Report of four cases.
Gastrointestinal tuberculosis is a rare disease in the United States. Correct identification is often delayed because it is not considered early on in the differential diagnosis.
Four patients with gastrointestinal tuberculosis and the symptoms, diagnosis, complications, and treatment of the disease are discussed.
Gastrointestinal tuberculosis should be considered in Asian immigrant patients who present with symptoms and signs of inflammatory bowel disease.

FIGURE 5 A sample informal text file extracted from the medical abstracts collection.

5 Features

We extracted features from each text, based on some of the main linguistic characteristics of the formal and informal styles discussed in Section 3. We hypothesize that these features could be good indicators to differentiate between the two styles. Most of the features are based on lexical choices and syntactic features, which are important for distinguishing between the two styles, as reported by (Karlgren, 2010).

We applied several statistical methods to extract the values of these features for each text in our dataset. Some of the features required parsing each text. Texts were parsed with the Connexor parser[8] (Tapanainen and Jarvinen, 1997), which is known to produce high-

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

quality results, as reported by (Pyysalo et al., 2006).

The features we extracted follow (see Table 6 for concrete examples of the values used in our model for the features):

1. **Formal words list:** This feature could characterize formal texts. It is based on the formal list in Section 3. Table 2 shows examples of the formal words. The value of this feature is based on the sum of the frequencies of any words from the list that appear in the text. The word frequencies are normalized by the length of the text for each document, since the size of each document is different.
2. **Informal words list:** This feature could characterize informal texts. It is based on the informal list in Section 3. Table 2 shows examples of the informal words. The value of this feature is based on the sum of the frequencies of any words from this list that occur in the text. The word frequencies are normalized by the length of the text.
3. **Formal pronouns:** This feature could characterize formal texts. We use this feature because formal style is often impersonal, using third person pronouns. We extracted this feature from the parse trees returned by the Connexor parser, counted how many times each text had impersonal pronouns, and normalized this by the length of the text for each document.
4. **Informal pronouns:** This feature could characterize informal texts. Informal style is often personal, using first and second person pronouns. We extracted this feature from the parse trees returned by Connexor parser, counted how many times the texts have personal pronouns, and normalized this by the length of the text for each document.
5. **Contractions:** This feature could characterize informal texts, since the informal style tends to use contractions, as discussed in Section 3. Table 3 shows examples of contractions. We extracted this feature by counting the number of contractions in each text. The count of contractions is normalized by the length of the text for each document.
6. **Abbreviations:** This feature could characterize informal texts, though it might not be a good feature for our model because, in some cases, abbreviations are used in formal texts, as reported by (Obrecht, 1999). Therefore, we let the machine learning algorithms determine if this feature will be used, based on its distribution in the training data. Table 4 shows examples of abbreviations. We extracted this feature by counting the abbreviations

in each text. The count is normalized by the length of the text for each document.

7. **Passive voice:** This feature could characterize formal texts. As discussed in Section 3, formal style often uses the passive voice. We extracted this feature from the parse trees returned by Connexor parser, and 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 could characterize informal texts. As discussed in Section 3, informal style often uses the active voice. 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 could characterize informal texts, as reported by Dempsey, McCarthy, and McNamara (2007); in addition, it is based on one of the main characteristics of informal style as discussed in Section 3. 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. **Average word length:** We hypothesized that this feature could characterize formal texts if the value is large (complex words), and characterize informal texts if the value is small (simple words). This hypothesis is based on the main characteristics of both styles, as discussed in Section 3.

11. **Type/Token Ratio (TTR):** This feature refers to how many distinct words are in a text, compared to the total number of words in the text. The TTR in formal texts is lower than in informal texts, as reported by Renkema (1984).

We used a parser to acquire some of the features. A part-of-speech tagger would have been adequate for most features, but some required the extra information provided by the parser (e.g., the active/passive voice and phrasal verbs).

Table 5 shows the number of passive voice, active voice and phrasal verbs that appeared and were extracted by Connexor parser for an informal text (Figure 2), and a formal text (from Figure 3). In addition, Table 6 shows a sample of the feature vectors of the informal text from Figure 2 and of the formal text from Figure 3. These features vectors are used in our classification model.

The class example	Passive Voice	Active Voice	Phrasal Verbs
Informal text (Figure 2)	0	13	4
Formal text (Figure 3)	2	11	0

TABLE 5 The number of passive voice, active voice and phrasal verbs for an informal text (from Figure 2) and for a formal text (from Figure 3).

Features	Features vectors for Figure 2 text	Features vectors for Figure 3 text
Formal words list	0.009	0.022
Informal words list	0.093	0.032
Formal pronouns	0.084	0.097
Informal pronouns	0.131	0.000
Contractions	0.039	0.000
Abbreviations	0.000	0.011
Passive voice	0.000	0.022
Active voice	0.121	0.118
Phrasal verbs	0.038	0.000
Average word length	4.028	5.559
Type/Token Ratio (TTR)	0.729	0.720
Class	Informal	Formal

TABLE 6 A sample of the feature vectors of informal text (Figure 2) and formal text (Figure 3).

6 Classification Algorithms

We used Weka (Hall et al., 2009) (Witten and Frank, 2005), which is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a particular dataset, or through a Java API. Weka has tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited to developing new machine learning schemes.

We chose the three machine learning algorithms because: Decision Tree (J48[9]) allows human interpretation of what is learned, Naïve Bayes (NB) works well with text, and Support Vector Machines[10] (SMO[11]) is known to achieve high performance. We trained the three classifiers on all the features described in Section 5, and used the default parameter settings from Weka. We also applied feature selection to examine the features and determine the weight of each one in our model, in order to identify the best feature for the model. We used the InfoGainAttributeEval[12] method from Weka, which shows the weight for each feature, and helps identify the weakest features that could be removed without affecting the performance of the classifier. More details about these experiments are provided in the next section.

7 Experiments and Evaluation

Our model was built with specific features extracted from the texts, based on the some of the main characteristics of both styles. We used 10-fold cross validation to evaluate the classifiers, and applied InfoGain feature selection. Applying the 10-fold cross validation method to the three algorithms showed different classification results. The comparison between these algorithms is based on numerous measures, including the Accuracy, Precision and Recall value, and the F-measure. We compared these values in order to determine which of the three algorithms was the best classifier.

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

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

[9] J48 implements decision trees algorithm C4.5 in Weka.

[10] We used the linear kernel. Other kernels were tested, but they were very slow and did not generate better performance than the linear kernel.

[11] Support vector machines algorithm is implemented in Weka by SMO.

[12] InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to each class.

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

For example, for the confusion matrix of the Decision Tree classifier in Table 7, based on the results produced by 10-fold cross validation for each data set, we can apply the formulae of Precision and Recall for each class in the Decision Tree as follows:

$$Recall\ (informal) = \frac{TP}{TP + FN}$$

$$Recall\ (formal) = \frac{TN}{TN + FP}$$

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

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

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.

Applying the different formulae gave the following results:

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

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

TABLE 7 The confusion matrix of the Decision Tree showing the distribution of the actual and the predicted classes of medical documents using 10-fold cross validation.

Here we note the following:

1. The higher the Recall, the better the classifier, because the Recall value shows how close the number of predicted instances is to the original number of instances in that class (i.e., how well the classifier predicts the instances).
2. A high value of Precision is preferred. This is explained by examining the higher Precision value of the formal class. Only seven instances from the total of 500 in the formal class were incorrectly classified (i.e., FN= 7).
3. Recall and Precision conflict with each other; while Recall is a good indicator of the classifier's performance in learning the informal class, the Precision is lower for this class.

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

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

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 results show that the classifier has almost the same predictive power (performance) for both the first and second classes.

Many evaluation measures can be used to measure the overall quality of a classifier. The Accuracy (success rate) shows the instances from both classes that are classified correctly by the following formula:

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

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

In our case the Accuracy was 0.985.

By applying the above analysis, we can measure the performance of the three algorithms at predicting the formal class and the informal class. The difference between the classifiers is apparent if we compare the values of the F-measure and Accuracy measures; higher values mean better performance. Comparing the overall values of the measures helped us choose the best classifier.

7.1 Classification Results of General-Domain Texts (Document Level)

The results of the text classification of general-domain texts at the document level for all three classifiers are shown in Table 8. They indicate that best classifier of the three algorithms is the Decision Tree. Based on a paired t-test (significance level 0.05), both the Decision Tree and the SVM accuracies are significantly better than that of the NB. Table 9 shows the detailed F-measure per class for the Decision Tree algorithm, while Figure 6 shows the visualization of the Decision Tree of text classification using 10-fold cross validation. Finally, we evaluated all the features through feature selection, using InfoGain attribute selection (InfoGainAttributeEval) from Weka. Table 10 shows each feature and its weight according to the InfoGain attribute selection, ranked in descending order from the strongest feature to the weakest. The most useful feature for the task was the informal pronouns feature, which is retained in our model. The weakest feature was the formal pronouns feature; this feature could be removed from our model.

Machine Learning Algorithm	F-measure (Weighted Avg.)	Accuracy
Decision Trees (J48)	0.985	0.985
Support Victor Machine (SMO)	0.983	0.983
Naïve Bayes (NB)	0.970	0.970

TABLE 8 Classification results for the SVM, Decision Trees and Naïve Bayes classifiers for general-domain texts at the document level.

Class	Precision	Recall	F-Measure
Informal	0.984	0.986	0.985
Formal	0.986	0.984	0.985
Weighted Avg.	0.985	0.985	0.985

TABLE 9 Detailed accuracy for both classes of Decision Tree for general-domain texts at the document level.

Attributes	Weight
Informal pronouns	0.9031
Average word length	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	0.0570
Formal pronouns	0.0183

TABLE 10 The features of our model and their InfoGain scores for general-domain texts at the document level.

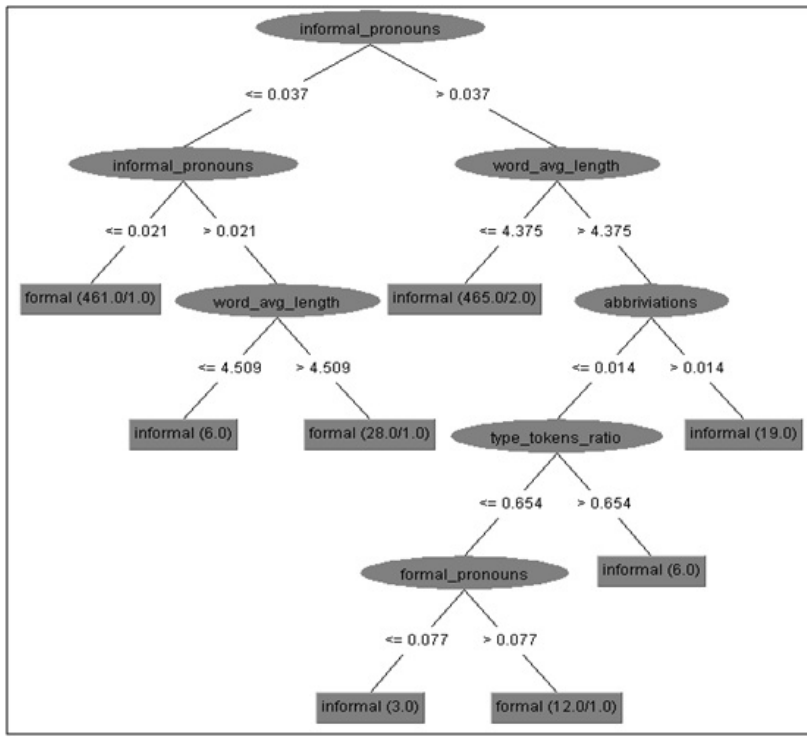

FIGURE 6 Decision Tree visualization for text classification using 10-fold cross validation.

7.2 Classification Results of General-Domain Texts (Sentence Level)

We built our model to classify sentences based on the same features that we used to classify longer texts (whole documents), because it achieved high accuracy in predicting the two classes. We also used the same method, 10-fold cross validation, to evaluate the classifiers.

The results of sentence classification of general-domain texts for all three classifiers are shown in Table 11. The best classifier of the three algorithms is the Decision Tree. Based on a paired t-test (significance level 0.05), both the Decision Tree and the SVM accuracies are significantly better than that of the NB. Table 12 shows the detailed F-measure per class of the Decision Tree algorithm. Finally, we examined all the features by performing feature selection using InfoGain attribute selection (InfoGainAttributeEval) from Weka. Table 13 shows each feature of our model and its weight according to the InfoGain attribute selection, ranked in descending order from the strongest feature to the weakest. The most useful feature for the task was the informal pronouns feature, which will be retained in our model. The weakest feature was the phrasal verbs feature; this feature could be removed from our model.

Machine Learning Algorithm	F-measure (Weighted Avg.)	Accuracy
Decision Trees (J48)	0.865	0.865
Support Victor Machine (SMO)	0.843	0.843
Naïve Bayes (NB)	0.784	0.786

TABLE 11 Classification results for the SVM, Decision Trees and Naïve Bayes classifiers for general-domain texts at the sentence level.

Class	Precision	Recall	F-Measure
Informal	0.863	0.867	0.865
Formal	0.867	0.863	0.865
Weighted Avg.	0.865	0.865	0.865

TABLE 12 Detailed accuracy for both classes of Decision Tree for general-domain texts at the sentence level.

7.3 Classification Results of Medical Texts

The results of text classification of medical texts for all three classifiers are shown in Table 14 (at the document level). The standard evaluation

Attributes	Weight
Informal pronouns	0.35444
Average word length	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 13 The features and their InfoGain scores for general-domain texts at the sentence level.

metric of F-Measure and the Accuracy were calculated. The results show that SVM achieved the highest performance, and was the best classifier for our model. Based on a paired t-test (significance level 0.05), both the Decision Tree and the SVM accuracies are significantly better than that of the NB.

Table 15 shows the detailed F-measure per class for the SVM algorithm. Finally, we evaluated all the features by performing feature selection using InfoGain attribute selection (InfoGainAttributeEval) from Weka. Table 16 shows each feature and its weight according to the InfoGain attribute selection, ranked in descending order from the strongest feature to the weakest. The most useful feature for medical texts was the average of words length feature, which will be retained in our model. The weakest feature was the abbreviations feature; this feature could be removed from our model.

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 14 Classification results of SVM, Decision Trees and Naïve Bayes classifiers for medical texts.

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 15 Detailed accuracy for both classes of SVM for medical texts.

Attributes	Weight
Average word length	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 16 The features of our model and their InfoGain scores for medical texts.

8 Discussion

Our experimental results show that it is possible to classify any kind of text (at the document or sentence level) according to formal or informal style[13]. We achieved reliable accuracies for all three classifiers, though the NB was less accurate than Decision Trees and SVM. This indicates that we selected high quality features for our model. This model can generate good results, whether it is applied on a single topic or different topics.

From Table 10, Table 13 and Table 16 we see that the abbreviations and phrasal verbs features have low InfoGain scores. This confirms our expectation that informal texts are not the only texts that use abbreviations and phrasal verbs; formal texts do this as well, despite the recommendations of the grammar books cited in Section 2.

The most useful features for predicting the formality level were the word length and the use of first and second person pronouns, according to the InfoGain values. This was the case for both document level

[13]Preliminary results, on medical documents only, were published in (Abu Sheikha and Inkpen, 2010a) and a shortened version of the classification at document level for general domain was published in (Abu Sheikha and Inkpen, 2010b).

and sentence level in general English texts, as well as in the medical text. The active voice was also important in the medical texts, probably having a stronger association with the informal documents than with the formal ones. Passive voice tends to appear often in scientific documents.

9 Conclusion and Future Work

In this paper, we presented an approach to classifying texts according to formal or informal style. We presented the main characteristics of both styles, and from these we derived the features of our model. The features we identified are success indicators of our work, because they helped us obtain quality classification results. In addition, the parallel lists of formal versus informal words and phrases that we compiled from different sources were very important in building our classification model.

The experiments and evaluation results showed that it is possible to classify any text or sentence according to formal or informal style. The evaluation results showed high accuracies in predicting the class. Classifying texts into formal and informal is very useful in different applications (e.g., evaluating texts of research papers to determine their degree of formality). Finally, our classification model is applicable to different types of texts, since it also achieved high accuracies on medical texts.

In future work, we will expand our lists of formal and informal word pairs from different domains, in order to increase the capability of our model to classify texts from more specific domains. We plan to extract more pairs of words by using a bootstrapping technique, starting with our current word pairs as seed words.We also intend to build additional models that can differentiate between different styles and genre variations. Growing the dataset is another objective. This can be done manually or automatically, as long as we include only texts that are clearly formal or informal, according to either their source or human judges. Adding more genres of texts to our dataset is also planned.

Acknowledgments

Our research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Ottawa.

References

Abu Sheikha, Fadi and Diana Inkpen. 2010a. Automatic Classification of Documents by Formality. In *Proceedings of the 2010 IEEE International*

Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE2010), pages 1–5. Beijing, China.

Abu Sheikha, Fadi and Diana Inkpen. 2010b. Learning to Classify Medical Documents According to Formal and Informal Style. In *Proceedings of the Workshop on Intelligent Methods for Protecting Privacy and Confidentiality in Data, AI 2010*, pages 27–34. Ottawa, ON, Canada.

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

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

Dempsey, K.B., P.M. McCarthy, and D.S. McNamara. 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*, pages 217–222. Menlo Park, California: The AAAI Press.

Denison, David. 1994. A Corpus of Late Modern English Prose. In M. R. Merja Kyto and S. Wright, eds., *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, pages 7–16. St Catharine's College Cambridge, Amsterdam, Rodopi. 25-27 March, 1993.

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

Garner, Bryan A. 2001. *A Dictionary of Modern Legal Usage*, page xxv. US: Oxford University Press, 2nd edn.

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

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

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

Hayakawa, S. I. (Editor). 1994. *Choose the Right Word: A Contemporary Guide to Selecting the Precise Word for Every Situation*. NY, USA: HarperCollins Publishers, 2nd edn. revised by Eugene Ehrlich.

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

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

Karlgren, Jussi. 2010. Textual Stylistic Variation: Choices, Genres and Individuals. In S. Argamon, K. Burns, and S. Dubnov, eds., *The Structure of Style*, pages 129–142. Springer Verlag.

Kennedy, Alistair and Michael Shepherd. 2005. Automatic Identification of Home Pages on the Web. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, pages 236–251. Dalhousie University, Halifax, Canada. 03-06 January.

Lim, Chul, Kong Lee, and Gil Kim. 2005. Automatic Genre Detection of Web Documents. In L. J. K. O. Y. Su K., Tsujii J., ed., *Natural Language Processing*. Berlin: Springer.

Manning, Christopher D. and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

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

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

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

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. *How to Prepare for the California State University Writing Proficiency Exams*, page 3. Victoria, Australia: Faculty of Health Sciences, La Trobe University. Chapter 5.

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

Pyysalo, Sampo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jarvinen Jouni, and Tapio Salakoski. 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, Andrea La Pietra, and Sarah Poon. 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*. UK: Cambridge University press, 2nd edn.

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

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

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

Tapanainen, P. and Timo Jarvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71. Washington D.C.: Association for Computational Linguistics.

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

Woods, Geraldine. 2010. *English Grammar For Dummies*, page 147. NJ, USA: Wiley Publishing, Inc. 2.

Yu-shan, Chang and Sung Yun-Hsuan. 2005. Applying Name Entity Recognition to Informal Text. CS224N/Ling237 Final Projects 2005, Stanford University, USA.

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