Automatic Classification of Documents by Formality

Fadi ABU SHEIKHA
University of Ottawa, SITE
800 King Edward, Ottawa, ON, Canada
fabus102@uottawa.ca

Diana INKPEN
University of Ottawa, SITE
800 King Edward, Ottawa, ON, Canada
diana@site.uottawa.ca

Abstract

This paper addresses the task of classifying documents into formal or informal style. We studied the main characteristics of each style in order to choose features that allowed us to train classifiers that can distinguish between the two styles. We built our data set by collecting documents for both styles, from different sources. We tested several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines, to choose the classifier that leads to the best classification results. We performed attribute selection in order to determine the contribution of each feature to our model.

Keywords:

Text Classification, Formal Style, Informal Style

1. Introduction

The need for identifying and interpreting possible differences in linguistic style of texts, such as between formal and informal styles, has increased nowadays as more and more people are using the Internet as a main resource for their researches. There are different factors that affect formality, such as words and expressions, as well as syntactical features. Vocabulary choice is perhaps the biggest style marker. Generally speaking, longer words and Latin origin verbs are formal, while phrasal verbs and idioms are informal. There are also many formal/informal style equivalents that can be used in writing.

Formal style is used in most writing and business situations and in speaking with people with which we do not have close relationships. Some characteristics of this style are using long words and passive voice. While Informal style is used in casual conversation, for example, that often happens at home between family members. It is used in writing only when there is a personal or closed relationship, like between friends and family. Some characteristics of this style are using word contractions like "won't", abbreviations like "phone", and short words.

In this paper we show how to build a model that will help to automatically classifying any text document into formal or informal style. So, we tested several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines in order to choose the classifier that leads to the best classification results.

The rest of the paper is organized as follows: In the second section, we review some existing methods for text classification by style and by genre. The third section addresses the main differences between both styles. In the fourth section, we discuss how we collect our data set that will be used to train our model. The fifth section presents our approach for extracting the features to build our model. In the sixth section, we describe the classification algorithms that we used to train our model. The seventh section addresses the result and the evaluation methods for our model. In The eighth section we discuss the results that we obtained. Finally, the ninth section concludes the paper and discuses the future work.

2. Related Work

There is little research on automatic text classification according to formal and informal style, but there is some work on automatic text classification by genre. 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 work on formal/informal classification, and the work on genre classification.

Reference [1] proposed a method to determine the degree of formality for any text using a special formula. This formula is the F-score measurement which is based on the frequencies of different word classes (noun, verbs, adverbs, etc.) in the corpus. The texts with high F-score are considered formal, while the ones with low F-score are considered informal. In our work, we want to build a model based on main characteristics of the two styles, rather than based on the frequency of word classes.

Reference [2] proposed that phrasal verbs can be used as a text genre identifier. 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.

Reference [3] discussed the task of web page classification by genre, namely how to distinguish home pages from non-home pages as noise, and then classify those home pages as personal home page, corporate home page or organization home page. The corpus they used is rather small: 312 web pages. They tried the hard task of

978-1-4244-6899-7/10/$26.00 ©2010 IEEE

subgenre discrimination. The best accuracy they obtained is (71.4%) on personal home pages with a single classifier, manual feature selection, and without noisy pages.

3. Learning Formal and Informal Style

In this section, we explain the main characteristics for formal versus informal style. We also show a sample of ready-made list of words for both styles, which we collected from different sources; this will help to understand the difference between the two styles.

A. Characteristics of Formal versus Informal Style

We studied and summarized the main characteristics of formal style versus informal style from [4], [5], [6], [7], and [8][2] to:

- Be able to distinguish between both styles.
- Identify each style from texts.
- Build the features based on those characteristics.
- Predict a class for new text documents.

Here we explain the characteristics of each style and provide examples:

1) Main Characteristics of Informal Style Text:

- 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) and abbreviations (e.g., TV).
- It uses phrasal verbs (Anglo Saxon words) 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 is more used in everyday speech than in writing.
- 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) Main Characteristics of Formal Style Text:

- 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 or abbreviations.
- 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 is more commonly used in writing than in speech.
- It uses an objective style, using facts and references to support an argument.
- It does not use vague expressions and slang words.

[2]http://webdelprofesor.ula.ve/humanidades/azapata/materias/english_4/formal_vs_informal_english.pdf

B. Formal versus Informal list of words

We collected informal/formal words, phrases, and expressions from different sources manually, also we extracted automatically more words from annotated text documents; such lists were very useful as two of the features in our model. In Table 1, we show an example of this list.

Table 1
An example of formal versus informal list of words

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

4. Data Set

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

A. Informal Texts

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

- Corpus of Late Modern English[3]: This corpus contains a set of annotated texts; most of these texts are informal texts (personal letters), Fig. 1 shows a sample of these informal texts.
- Enron Email Dataset/Corpus[4]: This corpus contains email texts; most of them are personal letters, therefore informal texts [9].
- Open American National Corpus (spoken texts): this corpus contains some categories that are informal texts such as spoken language texts[5].

> *" I'm never here, that's the pity of it, but I intend, when I write my War Office articles, to retire here solidly for the afternoons; otherwise I'm so terribly interrupted by visitors…"*

Figure 1. A sample of informal text file extracted from Corpus of Late Modern English (prose).

[3] http://ota.ahds.ac.uk/catalogue/index-id.html
[4] http://www-2.cs.cmu.edu/~enron/
[5] http://www.americannationalcorpus.org/OANC/index.html

B. Formal Texts

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

- Collection of news wire articles form the Reuters corpus[6]: This corpus contains a set of news texts; most of these texts are formal texts [9], Fig. 2 shows a sample of these formal texts.
- Open American National Corpus, written technical texts[7]: This corpus contains some categories that are formal texts such as written texts.

> " *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….*"

Figure 2. A sample of formal text file extracted from Reuters Corpus.

5. Features

We use several properties of the texts to encode texts as vectors of features. We built features that characterize formal and informal texts, based on the above analysis in the third section. We hypostasized that these features might be a good indicator to differentiate between both styles. 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 documents with the Connexor parser[8] [10], which helps to produce high-quality results for our model [11].

The features that we extracted are as follows:

- Formal words list: This feature is based on the formal list that we had mentioned in the third section. The value of this feature is based on its frequency in each text normalized by the length of the text for each document.
- Informal words list: This feature is based on the informal list, calculated based on its frequency in each text, normalized by the text's length.
- Formal pronouns: This feature characterizes formal texts. In the parse trees returned by the Connexor parser, we counted the frequency of impersonal pronouns, and we normalized by the length of the text for each document.
- Informal pronouns: This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has personal pronouns normalized by the length of the text for each document.
- Contractions: This feature characterizes informal texts. We counted the contractions words normalized by the text's length for each document.
- Abbreviations: This feature characterizes informal texts. We counted the abbreviations normalized by the length of the text for each document.
- Passive voice: This feature characterizes formal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has a passive voice normalized by the text's length for each document.
- Active voice: This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has an active voice normalized by the length of the text for each document.
- Phrasal verbs: This feature characterizes informal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has phrasal verbs normalized by the text's length for each document.
- Word length's average: This feature characterizes formal texts, if the value is large (complex words), and it characterizes informal texts if the value is small (simple words). We calculated the length's average for the words for each document.
- Type Tokens 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 [12].

We used a parser to obtain some of the features. For most of them, a part-of-speech (POS) tagger would have been enough, but for some features the extra information provided by the parser was needed, for example for active/passive voice and for phrasal verb.

6. Classification Algorithms

We used WEKA[9] [13], a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a certain dataset or called from Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

We chose three machine learning algorithms [13]: Decision Trees (J48) [10] because it allows human interpretation of what is learnt, Naïve Bayes (NB) because it is known to work well with text, and Support Vector Machines (SVM)[11] because it is known to achieve high performance. Table 2 shows the classification result for the three classifiers, by 10-fold cross-validation on our data set. Finally, we applied InfoGain attribute selection (InfoGainAttributeEval) from Weka to evaluate all features, in order to increase the performance of the classifiers.

[6] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[7] http://www.americannationalcorpus.org/OANC/index.html
[8] http://www.connexor.com
[9] http://www.cs.waikato.ac.nz/ml/weka/
[10] J48 implements Decision Trees algorithm C4.5 on Weka.
[11] Support Vector Machines algorithm is implemented on Weka by SMO.

7. Results and Evaluation

As we mentioned in the sixth section, we trained three classifiers: Decision Tree, Naïve Bayes, and SVM.

The Experiments were run using a 10-fold cross validation test. Results are shown in Table 2 for all three classifiers. The standard evaluation metric of F-Measure, the weighted harmonic mean of precision and recall was calculated. The Results show that Decision Trees was the best classifier for our model that has achieved best performance. Table 3 shows the confusion matrix of the Decision Tree classifier which shows the distribution of the actual and the predicted classes on document level based on the results produced by 10-fold cross validation. In Table 4, we show the detailed F-measure per class of Decision Trees algorithm. Finally, we examined all the features by performing attribute selection using InfoGain attribute selection (InfoGainAttributeEval) from Weka. We tried to remove the weakest features to achieve better performance but we discovered that will decrease the accuracy for the three algorithms. So, we decided to keep all the features in our model, as all features are important to achieve good performance. Table 5, shows each attribute with its weight according to the InfoGain attribute selection, ranked in descending order from the strongest features to the weakest features. The most useful feature was the Informal pronouns.

Table 2. Classification results of Decision Trees, SVM, and Naïve Bayes classifiers

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

Table 3. The confusion matrix of the Decision Tree

		Predicted Class	
Actual Class		**Informal**	**Formal**
	Informal	TP = 497	FN = 7
	Formal	FP = 8	TN = 492

Table 4. Detailed accuracy for both classes of Decision Trees

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 5. Our model's features with their InfoGain scores

Attributes	**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

8. Discussion

Our experiments show that it is possible to classify any text according to formal and informal style. We achieved reliable accuracies for all three classifiers, especially on Decision Trees. This indicates that we selected high quality features to include in our model. This model can generate good results whether it is applied on a single topic or on different topics.

9. Conclusion and Future Work

In this paper we have discussed one approach to classify text documents according to formal and informal style. In doing so we presented the main characteristics of both styles. From these characteristics we derived the features of our model. The learning process was successful and the classifiers were able to predict the classes of new texts with high accuracy.

Our immediate future work will be on extracting more formal and informal lists which should increase the accuracy of the classifiers. We will also experiment with adding more features such as sentence length feature in order to obtain a classifier with close to 100% accuracy.

References

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

[2] K.B. Dempsey, P.M. McCarthy, and D.S. McNamara, "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 (pp. 217-222). Menlo Park, California: The AAAI Press, Feb. 2007.

[3] A. Kennedy and M. Shepherd, “Automatic Identification of Home Pages on the Web”, Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
[4] Deborah Dumaine and Elisabeth C. Healey, *Instant-Answer Guide To Business Writing: An A-Z Source For Today's Business Writer*, (pp. 153-156), 2003 ed., Writers Club Press, Lincoln, 2003.
[5] Fred Obrecht and Boak Ferris, *How to Prepare for the California State University Writing Proficiency Exams*, (pp. 173), 3rd ed., Barron’s Educational Series Inc., New York ,2005.
[6] Adrian Akmajian, , Richard A. Demers, Ann K. Farmer, and Robert M. Harnish, *Linguistics: an introduction to language and communication*, (pp. 287-291), 5th ed., MIT Press, Cambridge (MA), 2001.
[7] David Park, “Identifying & using formal & informal vocabulary”, IDP Education, the University of Cambridge and the British Council, The Post Publishing Public Co., Ltd, 2007.
[8] Argenis A. Zapata, “Inglés IV (B-2008)”, Universidad de Los Andes, Facultad de Humanidades y Educación, Escuela de Idiomas Modernos, 2008.
[9] Yu-shan Chang and Yun-Hsuan Sung, “Applying Name Entity Recognition to Informal Text”, Ling 237 Final Projects, 2005.
[10] P. Tapanainen and Järvinen Timo, “A nonprojective dependency parser”, In Proceedings of the 5th Conference on Applied Natural Language Processing, pages 64–71, Washington D.C. Association for Computational Linguistics, 1997.
[11] Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jarvinen Jouni, and Tapio Salakoski, “Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions”, International Journal of Medical Informatics, 75(6):430-442, June 2006.
[12] J. Renkema, “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, p. 109 – 119, 1984.
[13] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed., Morgan Kaufmann, San Francisco, 2005.