Learning to Classify Medical Documents According to Formal and Informal Style

Fadi Abu Sheikha and Diana Inkpen

School of Information Technology and Engineering,
University of Ottawa
{fabus102@uottawa.ca, diana@site.uottawa.ca}

Abstract. This paper discusses an important issue in computational linguistics: classifying sets of medical documents into formal or informal style. This 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. We used machine learning techniques in order to automatically classify documents into formal and informal style. First, we studied the main characteristics of each style in order to train a system that can distinguish between them. Then, we built our data set by collecting documents for both styles, from different sources. After that, we performed pre-processing tasks on the collected documents to extract features that represent the main characteristics of both styles. Finally, we test several classification algorithms, namely Decision Trees, Naïve Bayes, and Support Vector Machines, to choose the classifier that leads to the best classification results.

Keywords: Automatic Text Classification, Medical Documents, Formal Style, Informal Style.

1 Introduction

The need for identifying and interpreting possible differences in linguistic style of medical documents, 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 medical 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.

Automatic classification of medical documents into formal and informal might be important for patient safety, since informal documents are unlikely to be published by medical authorities; therefore, people should not trust informal documents found in Internet.

The rest of the paper is organized as follows: In Section two, we review some existing methods for text classification by style and by genre. Section three addresses the main differences between both styles. In Section four, we discuss how we collect our data set that will be used to train our model. Section five presents our approach for extracting the features to build our model. In Sections six, we describe the classification algorithms that we used to train our model. Section 7 addresses the result and the evaluation methods for our model. In Sections 8 we discuss the results that we obtained. Finally, Section 9 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. For instance, Heylighen and Dewaele (1999) 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.

Moreover, Dempsey, McCarthy & McNamara (2007) propose 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.

In addition, 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.

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 resources; this will help to understand the difference between the two styles.

3.1 Characteristics of Formal versus Informal Style

We studied and summarized the main characteristics of formal style versus informal style from Dumaine and Healey (2003), Obrecht and Ferris (2005), and Akmajian et al (2001) 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:

Main Characteristics of Informal Style Text

1. It uses a personal style, using the first and second person (I, you) and the active voice (e.g., *I have noticed that*...).
2. It uses short simple words and sentences.
3. It uses Contractions (e.g., *won't*) and abbreviations (e.g., *TV*).
4. It uses phrasal verbs (Anglo Saxon words) within the text (e.g., *find out*).
5. The words that express rapport and familiarity are often used in speech, such as *brother, buddy, and man*.
6. It is more used in everyday speech than in writing.
7. It uses a subjective style, expressing opinions and feelings (e.g., *pretty, I feel*).
8. 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*).

Main Characteristics of Formal Style

1. 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*….).
2. It uses complex words and sentences to express complex points.
3. It does not use contractions or abbreviations.
4. It uses appropriate and clear expressions, precise education, business, and technical vocabulary (Latin origin).
5. It uses polite words and formulas like (e.g., *Please, Thank you, Madam, Sir*)
6. It is more commonly used in writing than in speech.
7. It uses an objective style, using facts and references to support an argument.
8. It does not use vague expressions and slang words.

3.2 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.

Table1. Shows an example of this list

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 size of the data set that we collected is 1980 documents: 990 characterize informal text and 990 characterize formal text.

Informal Texts

We chose 990 texts that characterize the informal style (Yu-shan & Yun-Hsuan 2005) from Medical newsgroups collection, this corpus called 20 Newsgroups1 contains 20 topics, and each topic has 1000 texts. These texts characterize informal style. We use one of these topics which are medical texts. We excluded 10 documents which have less than two words.

Formal Texts

We chose randomly 990 texts that characterize the formal style from medical abstracts collection. This collection contains 23 cardiovascular diseases categories (Joachims, 1997).

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

5 Features

We built features that characterize formal and informal texts, based on the above analysis in section 3. 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[2], which helps to produce high-quality results for our model.

The features that we extracted are as follows:

1. **Formal words list**: This feature is based on the formal list that we had mentioned in section 3.2. The value of this feature is based on its frequency in each text normalized by the length of the text for each document.
2. **Informal words list:** This feature is based on the informal list. The value of this feature is based on its frequency in each text normalized by the length of the text for each document.
3. **Formal pronouns:** This feature characterizes formal texts. In the parse trees returned by the Connexor parser, we counted how many times the text has impersonal pronouns, and we normalized by the length of the text for each document.
4. **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.
5. **Contractions:** This feature characterizes informal texts. We counted the contractions words normalized by the length of the text for each document.
6. **Abbreviations:** This feature characterizes informal texts. We counted the abbreviations normalized by the length of the text for each document.
7. **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 length of the text for each document.
8. **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.
9. **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 length of the text for each document.
10. **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 average for the words for each document.
11. **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.

We used a parser to obtain some of the features. For most of them, a part-of-speech 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.

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

6 Classification Algorithms

We used WEKA3 (Witten & Frank 2005), 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 (Witten & Frank 2005): Decision Trees because it allows human interpretation of what is learnt, Naïve Bayes because it is known to work well with text, and Support Vector Machines (SVM) 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.

7 Results and Evaluation

As we mentioned in section 6, 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 SVM was the best classifier for our model that has achieved best performance. In Table 3, we show the detailed F-measure per class of SVM algorithm. Finally, we examined the all features by performing attribute selection using InfoGain attribute selection (InfoGainAttributeEval) from Weka. We tried to remove the weakest features 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 4, shows each attribute with its weight according to the InfoGainA attribute selection, ranked in descending order from the strongest features to the weakest features.

Table2. Classification results of SVM, Decision Trees, and Naïve Bayes classifiers.

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

[3] http://www.cs.waikato.ac.nz/ml/weka/

Table3. Detailed accuracy for both classes of SVM

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

Table4. Our model's features with its weights based on InfoGain Attribute selection

Attributes	Weight
word_avg_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	0.1598
formal_pronouns	0.0913
formal_list	0.0815
Phrasal_Verbs	0.0748
abbriviations	0.0168

8 Discussion

Our experiments show that it is possible to classify any Medical text according to formal and informal style. We achieved reliable accuracies for all three classifiers, especially on SVM. 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 medical 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 that close to 100% accuracy.

Acknowledgements

Our research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University of Ottawa. We thank Soumaya Chaffar and Alistair Kennedy for helpful comments on our draft.

References

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

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 (pp. 217-222). Menlo Park, California: The AAAI Press.

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

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

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

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

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

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