Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus*

Alex Chengyu Fang and Jing Cao

Department of Chinese, Translation and Linguistics
City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong SAR
acfang@cityu.edu.hk, cjing3@student.cityu.edu.hk

Abstract. In this article, we report significant findings resulting from an investigation into the correlation between adjective density, calculated as the proportion of adjectives in word tokens, and degrees of text formality as part of an attempt to examine the potential application of adjectives in automatic text classification and identification. Correlations obtained from the training corpus will be compared with human ranking of the text categories concerned in the study and then adapted to unseen data in the test set. A linear regression analysis suggests a strong correlation between degrees of text formality and adjective density. With a weighted average *F*-measure of 0.606 achieved by a Naïve Bayes classifier, the research establishes adjectives as a powerful differentia of text categories amongst the open word classes, an important feature that has been generally ignored by past studies in automatic text categorization. The empirical findings suggest that the use of adjective density will lead to enhanced practical systems for automatic text classification.

Keywords: Adjective density, text formality, text classification, Linear Regression, Naïve Bayes.

1 Introduction

1.1 Text Classification and Past Studies

Text classification according to degrees of formality has been a long-standing issue that is both linguistically complex and computationally challenging. Biber (1988) examines 23 text categories according to six dimensions involving 67 chosen linguistic features. The first dimension (involved vs. informational) is related to text formality although the study reveals no clear-cut distinction between spoken and written texts through the use of the selected linguistic features. Sigley (1997) proposes to evaluate text formality by integrating 29 linguistic features into a formality index. Although the study concludes that such an index can evaluate a wide variety of text categories, there are still overlapping written and spoken categories that the formality index cannot deal with satisfactorily.

More recently, linguistic features specifically involving word classes have been investigated and employed for automatic text classification either through combining various word classes into a joint index to measure text formality or through using specific sets of a certain grammatical class to distinguish text categories. For example, Heylighen and Dewaele (1999 and 2002) propose a formula of formality based on word classes:

* The work described in this paper was supported partially by research grants (Nos 7002190, 7200120 and 7002387) from City University of Hong Kong. The authors would like to thank the reviewers for their valuable comments and suggestions.

$$F = \frac{\sum freq(\text{noun, adjective, preposition, article}) - \sum freq(\text{pronoun, verb, adverb, interjection}) + 100}{2} \quad (1)$$

Results show that such a formality score is quite effective in separating speech from writing and that imaginative writing can be separated from informational writing. Dempsey *et al.* (2007) investigate whether phrasal verbs can distinguish writing from speech, and formal from informal registers. The frequency counts of 397 most frequently used phrasal verbs were calculated and used to measure the formality of texts in three corpora. The results show that in most cases phrasal verbs can significantly distinguish writing from speech, and formal from informal registers. It is worth noticing that although a wide range of text categories are represented in the chosen corpora, the study only proposes a broad dichotomy classification, writing vs. speech and formal vs. informal registers. Rittman (2008) employs a set of trait adjectives, speaker-oriented adverbs, and trait adverbs to examine three chosen genres (i.e. academic, fiction, and news) in the British National Corpus (BNC). The results show that it is possible to use the particular sets of adjectives and adverbs to classify genres. In particular, speaker-oriented adverbs are found to be more effective than trait adjectives and adverbs.

1.2 Adjectives and Text Formality

This article reports an investigation into the use of adjectives to classify texts. In particular, the investigation focuses on the density of adjectives, defined as the proportion of adjectives amongst word tokens, as a characteristic of text formality that can be applied to effective text classification. For this purpose, a wide range of text types are investigated, covering not only transcribed speech but also written texts divided into six sub-categories (Table 1).

The investigation attempts to address the question how strongly adjective density correlates with text formality, a research question that past studies have not explicitly addressed. Such a correlation will be measured in both the training and test sets, and compared with the standard of human ranking. We shall report empirical results that suggest a strong and significant correlation between degrees of text formality and adjective density that can be used to mimic human ranking of the categories. It will be shown on empirical basis that adjective density successfully separates speech from writing and, within writing, academic prose from non-academic prose. Our study significantly extends past studies by further simplifying the set of characteristic features for text classification.

The article will be organized as follows: Section 2 will discuss the methodology and corpus resources adopted by the investigation. Section 3 describes both the manual ranking of the text categories concerned and the automatic ranking of the training set based on adjective density, followed by an evaluation. Section 4 will describe an attempt to adapt observations of the training set to unseen data in the test set. Section 5 describes the experiment using the Naïve Bayes classifier available in Weka, a general-purpose machine learning workbench (Holmes et al 1994), where adjective density is used as the sole feature for text classification. Finally, some conclusions will be drawn in Section 6.

2 Methodology and Corpus Resource

The investigation required a large corpus of texts that represents a range of text categories to be ranked manually according to degrees of formalities. The corpus needs to be grammatically tagged to enable the retrieval of adjectives, whose density, defined as the proportion in word tokens, will be computed and used to rank the same range of text categories. The two rankings will be subsequently analyzed for possible correlation. In the event of significant correlation between human ranking and automatic ranking according to adjective design, unseen data in the test set will be used to verify such correlation.

The British National Corpus (BNC) was used for the investigation for its size, its wide range of text categories and its part-of-speech annotation. This corpus has over 100 million word

tokens and represents a wide cross-section of British English from the later part of the 20th century. Eight text categories are identified:

1. Conversation
2. Other spoken
3. Academic prose
4. Fiction
5. Newspapers
6. Non-academic prose
7. Other published writing
8. Unpublished writing

A sub-corpus was created with samples randomly selected from the eight text categories in the BNC and a total of 3 million word tokens were selected for each category. See Table 1 for a summary of the composition of the sub-corpus.

Table 1: Total tokens in the subcorpus

Text Category		Token
Speech	Conversation	3,017,930
	Other spoken	3,019,043
Writing	Academic prose	3,124,550
	Fiction	3,026,196
	Newspapers	3,018,301
	Non-academic prose	3,083,486
	Other published writing	3,013,586
	Unpublished writing	3,001,746
Total		24,304,838

From this subcorpus, a training set was created that accounted for 80% of the total word tokens. The remaining 20% was kept as the test set for unseen data.

3 Manual and Automatic Ranking according to Adjective Density

3.1 Manual Ranking

Seven human subjects (six PhD students and one professor in linguistics) were invited to evaluate the formality of the eight text categories independently. They were asked to rank the text categories in the order of formality by specifying 1, 2, 3…etc, with 1 being the most informal and 8 the most formal. Inter-rater reliability was then tested by computing the intra-class correlation (ICC) coefficient. The value of the ICC coefficient is 0.857 with $p<0.001$, which is considered as outstanding inter-rater reliability (Landis and Koch, 1977). Next, the means of the human judgments were computed, according to which the eight different text categories were ranked. Table 2 summarizes the results with R_m indicating manual ranking.

Table 2: Manual ranking of the eight text categories

Text Category	R_m
Conversation	1
Other spoken	2
Unpublished writing	3
Fiction	4
Non-academic prose	5
Newspapers	6
Other published writing	7
Academic prose	8

As can be observed from Table 2, manual ranking of the eight text categories shows at least two features. First, spoken texts, including *Conversation* and *Other spoken*, are considered more informal than written texts. Second, among written categories, *academic prose* is regarded as the most formal, and expectedly separated from *non-academic prose*, which is ranked in the fourth place.

3.2 Adjective Density and Automatic Ranking

Adjective density is defined as the proportion of adjectives in all the word tokens for each category:

$$\text{Adjective density} = \frac{\text{Frequency of adjectives}}{\text{Frequency of word tokens}} \times 100 \quad (2)$$

All adjectives were retrieved from the training corpus and their proportion computed. Table 3 presents adjective density of the eight text categories in the training set in ascending order.

Table 3: Adjective density of the training set

Text Category	Total Tokens	ADJ Tokens	ADJ Density
Conversation	2,368,324	82,599	3.49
Other spoken	2,382,061	111,126	4.67
Fiction	2,382,786	139,894	5.87
Newspapers	2,360,843	159,046	6.74
Unpublished writing	2,395,601	162,826	6.80
Other published writing	2,354,825	197,100	8.37
Non-academic prose	2,451,482	213,128	8.69
Academic prose	2,468,802	237,709	9.63

As noted in Table 3, *academic prose* has the highest density of adjectives, 9.63%, whereas *conversation* has the lowest density of 3.49%. Form the viewpoint of speech and writing, we notice that the written texts are grouped together towards the bottom of the scale and that the spoken texts are clustered together at the top of the scale. Moreover, within writing, *academic prose* has the highest adjective density, similarly separated from *non-academic prose* as in human ranking. Also similar to human ranking, *fiction* has the lowest density among the written texts, boarding the spoken texts on the scale. Initial results therefore suggest that spoken categories have a generally lower adjective density while written texts show an overall higher use of adjectives. More importantly, results shown in Table 3 suggest that informal categories, such as the spoken ones, have a lower adjective density and formal categories tend to have a higher adjective density.

3.3 Evaluating Manual and Automatic Rankings

This automatic ranking according to adjective density (R_{adj}) was then examined by comparing it with the manual ranking (R_m). The absolute difference of each paired rankings (D) was calculated and Table 4 presents the results.

Table 4: Manual ranking vs. automatic ranking in the training set

Text Category	R_m	R_{adj}	D
Conversation	1	1	0
Other spoken	2	2	0
Fiction	4	3	1
Newspapers	6	4	2
Unpublished writing	3	5	2
Other published writing	7	6	1
Non-academic prose	5	7	2
Academic prose	8	8	0

Based on D, the Spearman rank correlation coefficient r_s was calculated between the automatic ranking of adjective density and human ranking according to the formula:

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)} \tag{3}$$

As a result, the value of the Spearman rank correlation coefficient is 0.833, which is significant at the level of 0.02. In other words, there is strong evidence of agreement between the automatic ranking of adjective density and the manual ranking. This result also suggests a strong correlation between adjective density and text formality.

3.4 Linear Regression Analysis

To further examine the relation between adjective density and formality of text categories, linear regression was computed and analyzed. The regression equation was first graphed to determine if there is a possible linear relationship (see Figure 1).

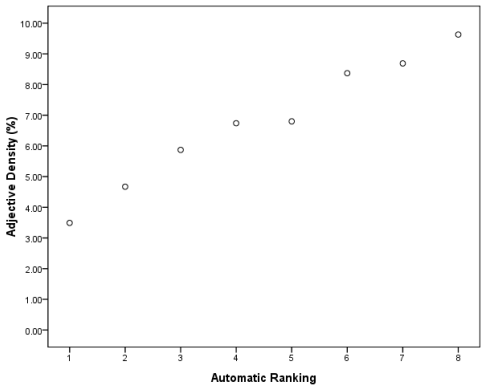

Figure 1: Graph of adjective density by automatic ranking

As can be seen in Figure 1, the points seem to follow a linear pattern with a positive slop. Next, the linear correlation coefficient (r) was computed. The value of r is 0.988 and again suggests a strong positive linear relationship between adjective density and degree of text formality. Accordingly, the coefficient of determination (r^2) is 0.977, indicating that about 97.7% of the variation in the density data can be explained by the degree of text formality. More importantly, since the value of r^2 is close to 1, the regression equation from the training dataset will be useful to make predictions of unseen datasets.

4 Adapting to Unseen Datasets

The regression equation was determined from the training data to construct a model for possible adaptation to unseen datasets. The linear regression equation allows us to obtain two parameters: intercept (α) and gradient (β), given adjective density (Y) and automatic ranking (X) from the training set:

$$Y = \alpha + \beta X \tag{4}$$

Equation (4) is therefore seen as a model characterizing the correlation between adjective density and ranking along a continuum of text formality. Such a model can be used to predict the ranking and therefore the text category given an unseen text for which only adjective density is known. Effectively, an automatic classifier can be constructed that operates on adjective density alone. However, it is necessary to make sure that such a model will show a good level of consistency when tested with unseen data from the test set, that is, the high level of correlation can be replicated and observed on the test set. The following sections will first describe the construction of model based on the regression equation, and then the expected ranking of adjective density will be calculated, and finally the expected ranking will be evaluated by both manual and automatic rankings.

4.1 Linear Regression Analysis

Based on the data from the training set, the regression equation is determined and graphed in Figure 2.

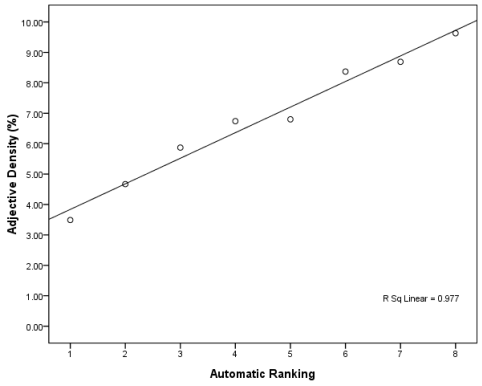

Figure 2: Graph of Y=2.998+0.841X.

Therefore, the acquired parameters are: α = 2.998 and β = 0.841. In this way, given adjective density of the unseen dataset (Y), Equation (4) can be converted into a model to predict the expected automatic ranking (X) of the unseen dataset:

$$X = \frac{Y - 2.998}{0.841} \tag{5}$$

4.2 Expected Automatic Ranking of Adjective Density in Test Set

The test set was employed as unseen data to test the consistency of performance by adjective density. A major objective is to see whether the test set will exhibit the same level of significant correlation between ranking and adjective density. Table 5 summarizes the test set.

Table 5: Basic stats of adjectives in text categories of the test set

Text Category	Total Tokens	ADJ Tokens	ADJ Density
Conversation	649,613	25,506	3.93
Other spoken	636,982	29,327	4.60
Fiction	643,410	37,508	5.83
Newspapers	606,149	41,441	6.84
Unpublished writing	657,458	46,807	7.12
Other published writing	658,762	52,466	7.96
Non-academic prose	632,003	53,816	8.52
Academic prose	655,748	60,468	9.22

As shown in Table 5, *ADJ Density* in the last column presents the *Y* of the test set. According to Equation (5), the expected automatic ranking of adjective density (ExR_{adj}) was computed and presented in Table 6.

Table 6: Expected automatic ranking of the test set

Text Category	ExR_{adj}
Conversation	1
Other spoken	2
Fiction	3
Unpublished writing	5
Newspapers	5
Other published writing	6
Non-academic prose	7
Academic prose	7

As can be seen in Table 6, there is again a clear dividing line between spoken texts and written texts, where two sub-divisions of spoken texts (i.e. *conversation* and *other spoken*) are grouped together towards the top of the scale. Among writing, *fiction* is the most informal while *academic prose* is the most formal one appearing at the bottom of the scale. It is also noticeable that *unpublished writing* vs. *newspapers* seem to have the same expected ranking, and that the same situation also involves *non-academic prose* vs. *academic prose*. The possible explanation is that the values of the expected rankings are in round numbers, when the actual values are 4.6 (*unpublished writing*), 4.9 (*newspapers*), 6.6 (*non-academic prose*) and 7.4 (*academic prose*) respectively. Therefore, although obtaining the same expected ranking number, those two paired sub-categories are placed in order.

4.3 Evaluating Expected Automatic Ranking

Finally, the expected automatic ranking (ExR_{adj}) was evaluated through comparison with both manual and automatic rankings. The Spearman rank correlation coefficient (r_s) was calculated based on the difference of paired rankings (D) in three settings.

The first setting measured the correlation coefficient r_s between ExR_{adj} and manual ranking (R_m); the value of r_s is 0.869, which is above the significant level of 0.02, indicating a strong evidence of agreement between these two rankings. In the second setting, r_s was measured between ExR_{adj} and R_{adj}, the automatic ranking of the training set. The value of r_s in this setting is 0.976, which is significant at level of 0.01. In the third setting, r_s was measured between ExR_{adj} and the automatic ranking based on the adjective density in the test set (cf. Table 5). This time the value of r_s is 0.976, the same value as that in the second setting, which is significant at level of 0.01.

The results in all of the three settings show that the expected automatic ranking correlates significantly well with both manual and automatic rankings. In other words, the adaptation of the model to test dataset demonstrates a satisfactory level of performance. This finding suggests that the regression model can be adapted to unseen datasets with reliable performance.

5 Automatic Text Classification using Naïve Bayes Classifier

As previously shown, experiments on both training and test sets indicate that adjective density is significantly correlated to degrees of formality of different text categories and hence the prospect of using this measure to automatically classify texts. To verify how adjective density as a characteristic could contribute to text classification, experiments were carried out by using the Naïve Bayes classifier available in Weka, a machine learning system available from the University of Waikato, New Zealand (Homles *et al* 1994). Adjective density was calculated for all the individual texts in both the training and the test sets for each text category. Weka was used to perform an 80-20 split of all the instances (1,180 in total). NaiveBayes was selected to train a model based on adjective density as the sole feature from the training set (80% of all the instances), which was subsequently applied to the test set (20% of all the instances). Table 7 summarises the results for the eight text categories in terms of precision, recall and *F*-measure.

Table 7: Precision, recall and F-measure in eight BNC text categories

Text Category	Precision	Recall	*F*-Measure
Conversation	0.000	0.000	0.000
Other spoken	0.492	0.970	0.653
Academic prose	0.167	0.053	0.080
Fiction	0.000	0.000	0.000
Newspapers	0.000	0.000	0.000
Non-academic prose	0.246	0.538	0.337
Other published writing	0.000	0.000	0.000
Unpublished writing	0.250	0.214	0.231
Weighted Avg.	0.223	0.373	0.267

The classification results are summarized in the following matrix:

```
a  b  c  d  e  f  g  h   <-- classified as
1  0  0  0  8  4  5  1 |  a = ACPROSE
0  0  0  0  0  0 25  0 |  b = CONVRSN
0  0  0  0  0  1 12  2 |  c = FICTION
0  0  0  0 13  0  4 12 |  d = NEWS
2  0  0  0 14  1  3  6 |  e = NONAC
1  0  0  0  8  0  1  4 |  f = OTHERPUB
0  0  0  0  0  0 64  2 |  g = OTHERSP
2  0  0  0 14  1 16  9 |  h = UNPUB
```

It can be noticed that all the files of *conversation* have been classified into *other spoken*. There are two possible explanations. First, there is significant difference in instance number between the two categories: *conversation* has only 102 instances while *other spoken* has 378 instances. Second, *other spoken* also includes dialogues, which blurs the distinction of *conversation*. For the same reason, a majority of *fiction* has been classified as *other spoken*. In addition, *unpublished writing* and *other published writing* were observed as the less clearly defined categories. The training of the model and the subsequent prediction by the model were thus biased as a result of the unbalanced classes (Eibe and Bouckaert 2006).

The less clearly defined categories were excluded and a second try on the classifier was made. The results are presented in Table 8.

Table 8: Performance of using ADJ density for text categorization

Text Category	Precision	Recall	*F*-Measure
Conversation	0.846	0.917	0.880
Academic prose	0.556	0.333	0.417
Fiction	0.524	0.733	0.611
Newspapers	0.667	0.467	0.549
Non-academic prose	0.448	0.591	0.510
Weighted Avg.	0.626	0.613	0.606

Almost every category shows a better performance. *Conversation* shows the best results, with an *F*-measure of 88% based on a recall of 91.7% and a precision rate of 84.6%. It is encouraging to see that the average precision, recall and *F*-measure have been all over 60%, which indicates that with distinctively defined categories, adjective density can be used as a powerful characteristic for automatic text classification. Indeed, the same experiments were carried out on the other three open classes, namely, nouns, verbs and adverbs. Results show that adjectives remain a powerful indicative of text categories with a weighted average *F*-measure of 0.606, close to nouns (weighted average *F*-measure=0.756) and followed by adverbs (0.511) and verbs (0.448).

6 Conclusion

In this paper, we described our investigation into the relation between adjective density and text formality for the purpose of automatic text classification. According to empirical results collected from the training set, adjective density exhibits a significant positive correlation with text formality. The Spearman rank correlation coefficient shows a significant degree of agreement between such an automatic ranking and manual ranking. In other words, formal text categories tend to have a higher adjective density than informal text categories. A linear regression analysis also confirms such a positive linear relationship, and more importantly, it helps to construct a linear regression model to describe the relation between adjective density and category prediction. When adapted to assess unseen data in the test set, the model produced a satisfactory performance by obtaining a high value of correlation with automatic ranking as well as manual ranking. The results suggest that adjective density can be reliably used to predict text categories. A predictive model was created using a Naïve Bayes classifier on adjective density, which achieved a weighted average *F*-measure of 0.606 across a set of five text categories, compared with 0.756 for nouns, 0.511 for adverbs and 0.448 for verbs. The experimental results establish adjective density as a powerful characteristic of text categories compared with the other open classes.

In conclusion, our investigation has shown on empirical basis that adjective density is significantly correlated to degrees of formality of different text categories. To be more specific, adjective density can effectively distinguish speech from writing, and within writing, academic prose from non-academic prose. Our study has advanced past research in the sense that we have extracted a single linguistic feature that can be used to distinguish text categories according to degrees of formality. By employing adjectives alone, our study indicates that it would be technically more feasible when applied in automatic text classification and genre detection. A significant finding of the research reported here is the established of adjectives as an effective characteristic of text categories amongst the open classes, an important feature that has been generally ignored in past studies. In hindsight of our current investigation, each text category is treated as a homogeneous group without considering the effect that file size has on adjective density. In a future study, we plan to investigate the relation between file size and adjective density with a view to develop an enhanced model parameterized not only with adjective density but also with the file size.

References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

BNC User Reference Guide. URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/index.html

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Dempsey, K. B., P. M. McCarthy, and D. S. McNamara. 2007. Using Phrasal Verbs as an Index to Distinguish Text Genres. *American Association for Artificial Intelligence* (www.aaai.org).

Eibe, F. and R.R. Bouckaert. 2006. Naïve Bayes for Text Classification with Unbalanced Classes. In *Proceedings of 10th European Conference on Principles and Practice and Knowledge Discovery in Databases*, Berlin, Germany, 503-510.

Heylighen, F. and J.-M. Dewaele. 1999. Formality of Language: definition, measurement and behavioral determinants. *Internal Report*, Center "Leo Apostel", Free University of Brussels.

Heylighen, F. and J.-M. Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Context in Context: Special issue Foundations of Science*, 7(3), 293–340.

Holmes, G., A. Donkin, and I.H. Witten. 1994. Weka: A machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia.

Landis, R. J. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

Rittman, R. 2008. *Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology*. VDM Verlag.

Sigley, R. 1997. Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics*, 2(2), 199-237.