

# Examining Variations of Prominent Features in Genre Classification

Yunhyong Kim and Seamus Ross

Digital Curation Centre (DCC)

Humanities Advanced Technology and Information Institute (HATII)

University of Glasgow, UK

{y.kim,s.ross}@hatii.arts.gla.ac.uk

## Abstract

*This paper investigates the correlation between features of three types (visual, stylistic and topical types) and genre classes. The majority of previous studies in automated genre classification have created models based on an amalgamated representation of a document using a combination of features. In these models, the inseparable roles of different features make it difficult to determine a means of improving the classifier when it exhibits poor performance in detecting selected genres. In this paper we use classifiers independently modeled on three groups of features to examine six genre classes to show that the strongest features for making one classification is not necessarily the best features for carrying out another classification.*

## 1. Introduction

The research described in this paper examines genre classes of text documents and the role of different types of features in distinguishing these classes automatically. Automated genre classification (e.g. classification into scientific research articles, news report, or email), which identifies the function and structure of the document, supports metadata extraction ([12]) which is essential for efficient and effective information management in repositories. It also helps information seekers focus their search on specific genres, thereby enabling the retrieval of documents exhibiting different levels of detail, and, aids knowledge mining by performing a first-level classification of documents into documents of similar physical and conceptual structure.

The features which characterise a text often fall into well-defined groups. For example, some features capture the position of text blocks (visual layout), some describe indicative vocabulary (significant terms), and others attempt to identify the pragmatics or functional category of selected terms (style). In previous studies of automated genre classification (e.g. [4], [5], [11], [12],

[13], [19], [21]) these features have been combined to produce a single set of features to represent the documents which are to be classified. This approach optimises the overall performance of the classifier on the detection of the pre-defined classes, but makes it difficult to devise a means of improving the classifier when it displays poor performance in detecting selected genres. It also takes for granted that the predefined classes describe a comparable schema of a single classification task.

In this paper we give evidence that genre classification, as described in previous studies, may actually be a combination of several independent tasks. For example, the distinction between a Thesis and Scientific Paper is largely structural, while Meeting Minutes and Business Reports are mostly distinguished by topic and style. On the other hand, the distinction between a Table of Financial Statistics and a Financial Report lies mainly in the visual representation and style. Using the same features to model concurrently these different types of classification would be equivalent to estimating a single distribution for items which belong to distinct populations. If you examine previous literature (e.g. Table 5 in [12], Table 3 in [13]), classification errors range anywhere from seventeen percent to seventy-six percent ([12]), and six percent to eighty percent ([13]). Observing such big differences in error rate might indicate that a re-evaluation of the task, to determine if the task is actually a combination of many tasks disguised by the single term genre classification, would be productive.

Another prevailing notion in earlier analyses is that genre classification is orthogonal to topic or subject classification. This notion defines genre classification as a task independent from subject classification. While there may be a conceptual level at which this is true, within the probabilistic framework on which language processing is highly reliant, there is reason to believe that this is not generally the case. For example, consider the topic of *cohomology*, a well-known subject area in higher mathematics; this topic would not be expected to appear as frequently in the genre class Reportage as it would in

the genre class Research Article. This suggests that, at least on a practical probabilistic level, genre sometimes moves in close proximity to subject.

The present paper reports tests on two corpora of genre-labelled PDF documents conducted to examine the correlation between genre classes and three feature types, to demonstrate that the best feature types for detecting any one genre class are not necessarily the best for detecting other genre classes. The feature types we will examine are visual layout features, language modeling features and stylistic word frequency. Initially our corpus has been confined to one document format to narrow down the problem space. We have chosen PDF as this format because a tool for this format is likely to have immediate wide spread application given its popularity across library, archival, commercial and private sectors. The methods described here, however, do not use features dependent on elements available only in PDF documents. The process is dependent on the PDF only in so far as it depends on PDF tools to convert the documents into image and text.

It is not the intention of this paper to introduce a classifier optimised to perform genre classification (in contrast to [14]). Here we put forward evidence that establishing a correlation between feature types and genre classes may be a reasonable step forward in constructing a robust genre classification system.

## 2. Defining genre

Genre is a highly mutable context-dependent concept. Its mutability is apparent in its usage across the literature: Biber ([4]) characterised document genres using five dimensions (information, narration, elaboration, persuasion, abstraction), while others ([12], [5]) examined the categorisation of documents into common classes such as FAQ, Job Description, Editorial or Reportage. Genre classification have sometimes been defined as the analysis of particular aspects (narratives, fact versus opinion, intended level of audience, and, positivity or negativity of opinion) of text ([13], [11]), and even used to describe the detection of selected journals and brochures from one another using visual layout ([1]). Others ([19], [2]) have clustered documents into similar feature groups without delving into genre facets or classes, and some have championed a multi-genre schema for web page classification ([21]). Santini has reviewed different approaches to genre classification ([20]).

While the definition of genre may not be easily pinned down, there is general agreement that *genre* is a concept used to categorise documents by structure and function. In fact, the structure of documents in the genre evolve to meet the functional requirements for its survival in the environment for which it was created, much the same as the structure of an organism evolves to optimise

its survival function in the natural environment (cf. [15]). The accepted layout, language, components and style of the document change dynamically to maximise its chances of fulfilling its role as

- a piece of communication reflecting the intention of the creator,
- a source of information for distribution to a user community,
- a part of a process such as publication, recruitment, or event,
- a type of data structure for representing information.

In this context, it seems intuitively clear that selected features will be dependent on one of five aspects: visual layout, style, topic, semantic patterns, and contextual elements which reflect the process for which the document was created and used (cf. [14]).

The proposed objective in this paper is to study these *feature types* in relation to genre classes to determine its effectiveness in the detection of *visual genres* (e.g. data structure type), *stylistic genres* (e.g. prescribed procedural style) and *topical genres* (e.g. business versus legal briefing paper) independently. To this end, we first examine white space analysis, stylistic term frequency and significant term analysis in relation to genre classification. Subsequently we will enrich this basic set to examine more sophisticated features. It seems important to keep a check on the number of parameters in the first analysis.

## 3. Data

A common problem in the study of automated genre classification is the lack of established experimental data. A limited classification of documents into genre is available in previously constructed datasets, but none of them span a large number of genres, nor do they employ a consistent schema. To alleviate the paucity of data, we have created two corpora which we describe in this section.

### 3.1. Corpora

There are two independent corpora which have been constructed in our research:

#### **RAGGED (RANDOMLY GENERATED GENRE DATA)**

This dataset consists of 570 PDF documents gathered from the Internet using random search words. For the retrieval of each item, the algorithm selects a random word from SCOWL (Spell Checker Oriented Word List - available from sourceforge.net), retrieves a list of PDFs containing the search word, and then saves a random

document from the first hundred documents returned. The data gathered by this method was labelled by one of the authors of this paper.

### KRYS I

This corpus consists of documents belonging to one of the seventy genres described in Table 1 of [16]. The corpus was constructed through a document retrieval exercise where university students were assigned genres from Table 1 of [14], and, for each genre, asked to retrieve from the Internet one hundred examples of that genre represented in PDF and written in English. They were not given any descriptions of the genres apart from the genre label. Instead, they were asked to describe their reasons for including the particular example in the set. For some genres, the students were unable to identify and acquire one hundred examples. The resulting corpus now includes 6478 items.

## 3.2 Experimental data

The experiments analysed in Section 6 have been conducted on two datasets, a subset of RAGGED (**Dataset I**) and a subset of KRYS I (**Dataset II**), consisting of all the documents in the corpora initially labelled as one of six genres comprising Academic Monograph (AM), Business Report (BR), Book of Fiction (BF), Minutes (M), Periodicals (P), and Thesis (T). The experimental Dataset I comprises 16 examples of AM, 16 examples of BR, 15 examples of BF, 19 examples of M, 19 examples of P, and 18 examples of T, while Dataset II comprises 99 examples of AM, 29 examples of BF, 100 examples of BR, 99 examples of M, 67 examples of P, 100 examples of T. Using two datasets which have been labelled by different class of labellers helps to gauge the consistency of the classifiers in modeling distinct classification standards.

The low proportion of Book of Fiction and Periodicals in Dataset II is due to the difficulty in finding publicly available examples of that genre. The number of documents in each genre are quite small in general; the sparsity of genre labelled data makes it difficult to collect large quantities of experimental data distributed across a variety of sources. Although we could have increased the data horizontally to include all the genre classes in KRYS I, this would only introduce more parameters and confusion without adding credibility to the analysis of the relationship between feature types and individual classes (which would still contain a small amount of data).

## 4. Classifiers

Eight classifiers are examined in this paper. They are each trained on one of three feature types and one of three statistical methods.

The three statistical methods employed are **Naïve Bayes (NB)** [16], **Support Vector Machine (SVM)** [26] and **Random Forest (RF)** [6]. These methods were chosen to represent a variety of approaches to classification which have proven to be successful in the past. Naïve Bayes is based on simple maximisation of likelihood of class occurrence given context. Support Vector Machine is based on constructing hyperplanes which separate the data into clusters of data in the same class. Random Forest is a decision tree method which is based on the construction of a committee of classification trees based on a random selection of features for each tree. There are other methods (e.g. Maximum Entropy) which we would have liked to examine as well, however, the lack of time and space has led us to postpone it to future research.

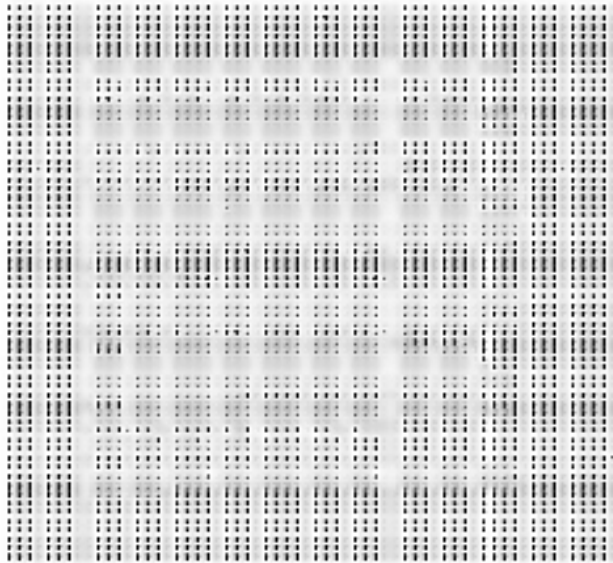
We have called the three feature types **image**, **style** and **Rainbow**. The features. image and style, have been modeled for comparison using all three statistical methods using the Weka machine learning toolkit ([21]). The features represented by Rainbow, on the other hand, have been modeled on Naïve Bayes and Support Vector Machine. The features of Rainbow are native to the Rainbow module of the BOW toolkit ([15]) and Random Forest (which was developed at a later date) was unfortunately not built into the module. We will refer to the eight classifiers by naming them with the feature type followed by the abbreviated name for the statistical method (e.g. **image NB** for image feature Naïve Bayes classifier). The parameters for feature selection and statistical methods have been optimised over a finite set of combinations tested for best overall accuracy on several samples taken from RAGGED. The final feature selection method is described below.

**Image features:** The first page of the document was converted into a low resolution grey-scale image and sectioned into a  $N \times N$  grid. Each region on the grid was examined for non-white pixels. All regions with non-white pixels were assigned a value of 1 and all the other regions were assigned a value of 0 to create a low resolution bit map. Several grid sizes were tested on samples taken from RAGGED, but we found  $N=62$  to produce the best results. This was also the coarsest level of granularity at which human subjects were able to distinguish particular documents as members of specific genre classes.

**Style features:** From an independent dataset consisting of documents retrieved from the Internet, the union of all words found to be *prolific* within each genre class was compiled into a list. A word is said to be *prolific in genre G* if it is found in many of the documents belonging to *G*. The dataset used in this process consists of 190 complete documents belonging to nineteen genres inclusive of the six genres being examined in this paper. The thirteen complementary genre classes include

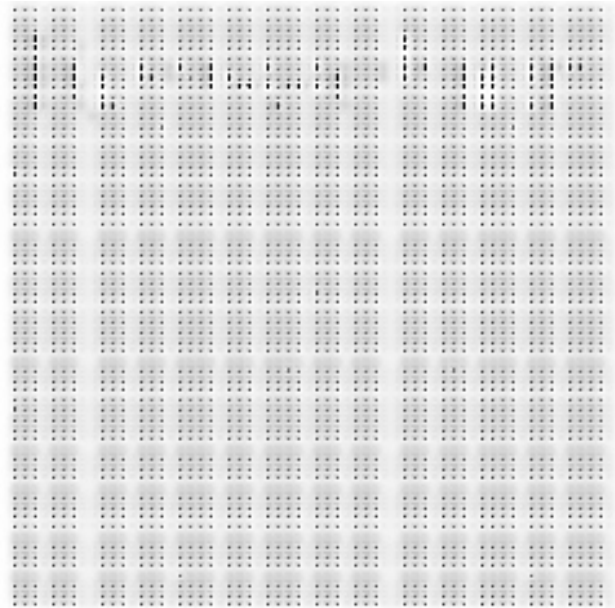
Abstract, Magazine Article, Scientific Research Article, Forms, Technical Manual, Technical Report, Email, Memo, Advertisement, Exam Worksheet, Slides, Speech Transcript, Poster. The test documents were represented by a vector constructed using the frequency of each word in the compiled list.

**Rainbow features:** This is a text classifier included in the BOW toolkit developed by McCallum ([15]). This toolkit indexes the alpha-numeric content of the text for an analysis of significant terms, to estimate the probability of each word against each class. We have adopted the default setting of using a stop-word list to capture significant topical words of documents. The rainbow classifier is popular with subject classification.



**Figure 1. Example of document image representation: Scientific Article.**

Our interest in image features reflects the recognition that documents of certain genres have more white space on the first page (e.g. title page of the book), are ruled by formatting conventions (e.g. first slide for a conference presentation), and are made visually elaborate to attract readership (e.g. the reversal of black and white on a magazine cover). Another benefit of examining documents using image processing methods is that the process does not depend on extracting text, can be language independent, and supports document analysis even when the content of the document is only accessible as an image. Examples of the image representation are given in Figure 1 and 2. These images should be interpreted as negatives of the original image: darker pixels which have been labelled 1 appear lighter than the white pixels which have been labelled 0.



**Figure 2. Example of document image representation: Periodicals (magazine cover).**

The image representation here is not the *shape* which is mentioned in Dillon and Vaughan ([10]). The notion of shape they introduce relates more to the visualisation of conceptual elements such as breadth of argument. The notion of *form* discussed by Toms and Campbell ([23]) is closer to the layout that we are trying to capture. However, while they concentrate on structural properties such as title, sections and bibliography, we are interested in the topology of the entire page. At the moment, only the first page of the document is used, later further pages can be incorporated.

The features in style are intended to capture frequency of words popular to all genres as well as words which are only prolific within some genres. A typical example of the weight of this feature is illustrated in the fact that forms or slides are likely to contain a fewer number of definite or indefinite articles than flowing text. Sample average frequencies of words commonly found in three of the genres discussed in the current study are presented in Table 1 (the average is taken over ten documents).

In the reported experiment we have taken words which were found in 75% of the files in each genre. We have also tried with a few other percentages but found this to show the best results. To compile the common words, we also tried a focused method of compiling words from the six genres under consideration only, and even words from a range of genres which exclude the genres of interest. The combined list was adopted in the end because higher accuracies were consistently observed in all three style-based classifiers when using this list in comparison to the other two lists.

The prolific words list that we have constructed is distinct from the word list of common English words constructed by Samatatos, Fakotakis and Kokkinakis ([22]). They collected words frequently used in the English language. We are collecting a list of words which have high *document count* in at least one of a range of genres, the words are expected to appear in many of the documents in one or more genres but may not necessarily have a high absolute frequency within the corpus nor any one document.

**Table 1. Average frequency of words per document across three genres.**

Word	Genre	Business Report	Thesis	Minutes
have		47	109	0
with		71	210	13
do		11	0	0
case		0	10	0
meeting		0	0	8
information		12	0	0

Only using word frequency, when other more sophisticated analyses of linguistic structure has produced promising results (e.g. Santini 2007) may seem absurd, but, there are advantages to further examining the frequency analysis before incorporating linguistic analysis. For example, the word frequency method can be applied immediately to a wider range of languages than can be a method reliant on heavy linguistic engineering.

The frequency model is basic to further semantic modeling. Here, we use a one-dimensional representation of each word dependent on its relative frequency in comparison to all the words in the word list. However, this can be refined so that relative frequency of each word is calculated within several different semantic groups of words taken from the list.

## 5. Experiments

### 5.1. Method

Eight classifiers (**image NB**, **image SVM**, **image RF**, **style NB**, **style SVM**, **style RF**, **Rainbow NB**, **Rainbow SVM**) have been tested on Dataset I and II for their performance in recognising six genre classes including Academic Monograph, Book of Fiction, Business Report, Minutes, Periodicals, and Thesis. The performance is examined using 10-fold cross validation results.

The performances of the eight classifiers are first evaluated to identify, for each feature type, the statistical methods that generate the best overall performances on

Dataset I and II (Section 6.1). Then, on each dataset, the best classifiers, one for each feature type, are compared in detail across the six genres (Section 6.2 and 6.3).

## 5.2. Evaluation

The results, apart from those reported in Section 6.3, have been evaluated with three conventional metrics for classification: accuracy, precision and recall. To make precise what we mean by these terms, let  $N$  be the total number of documents in the test data,  $N_c$  the number of documents in the class  $C$ ,  $TP(C)$  the number of documents correctly predicted to be a member of class  $C$ , and  $FP(C)$  the number of documents incorrectly predicted as belonging to class  $C$ . Accuracy  $A$  is defined to be  $A = \{\sum_c TP(C)\}/N$ , precision  $P(C)$  of class  $C$  to be  $P(C) = TP(C)/\{TP(C)+FP(C)\}$ , and, recall,  $R(C)$ , of class  $C$  to be  $R(C) = TP(C)/N_c$ . Although some debate surrounds the suitability of accuracy, precision and recall as a measurement of information retrieval tasks, for classification tasks, they are still deemed to be a reasonable indicator of classifier performance.

## 6. Results

### 6.1. Overall accuracy

The overall accuracies of classifiers built on each feature type across statistical methods is reported in Table 2 (best performances are indicated in bold-face).

The tests on the two datasets, consistently indicate Naïve Bayes as the best statistical method for image features. Although the overall accuracies of Naïve Bayes and Random Forest are comparable on the larger dataset, averaging (with a heavier weight on the larger set) the performances on the two datasets, suggested Naïve Bayes as a better performer for image. On both datasets, Support Vector Machine and Random Forest are both better than Naïve Bayes for style features. Although Support Vector Machine and Random Forest performs comparably on the smaller Dataset I, we have chosen Random Forest as the better choice for style, because the difference was shown to be prominent on Dataset II. We have chosen Naïve Bayes for Rainbow for comparison on Dataset I, and Support Vector Machine for Rainbow on Dataset II: in both cases the difference in performance was too large to indicate an overall better method for Rainbow.

In passing, we observe that, based on the overall accuracies of the classifiers on the two datasets, the classifiers based on image features are the least affected by training dataset size (average difference in accuracy 0.036) and the classifiers based on Rainbow are the most affected by dataset size (average difference in accuracy 0.328). Also the results indicate that Support Vector

machine and Random Forest seem more affected by dataset size than Naïve Bayes.

**Table 2. Overall accuracy of feature types across statistical methods**

Data & method	Dataset I (103 items)		
	NB	SVM	RF
image feature	<b>0.524</b>	0.35	0.417
style feature	0.505	0.573	<b>0.641</b>
Rainbow feature	<b>0.428</b>	0.25	N/A
Data & method	Dataset II (494 items)		
	NB	SVM	RF
image feature	<b>0.48</b>	0.395	0.48
style feature	0.63	0.724	<b>0.828</b>
Rainbow feature	0.618	<b>0.715</b>	N/A

## 6.2. Precision and recall

In this section we compare the precision and recall across genres of the classifiers for each feature type which have been shown to have the best overall accuracies in the previous section (on Dataset I, **image NB**, **style RF** and **Rainbow NB**; on Dataset II, **image NB**, **style RF**, **Rainbow SVM**).

The figures in Tables 3 and 4 show the precision and recall across the six genres of each classifier tested on Dataset I and II. The genres are indicated in the left most column of the tables, with the numbers of documents in each class noted in parenthesis. The statistical method being tested is indicated in the next column.

The results in Table 3 indicate that, on Dataset I, both precision and recall of **image NB** with respect to Periodicals are much higher than the other two classifiers. On the other hand, the results indicate that academic monographs and business reports are best recognised by **style RF**. Books of fiction seem to be best distinguished by **style RF** and **Rainbow NB**, but we also observe that the two classifiers seem to work in complementary positions (that is, where one has better recall the other has better precision). With the genre class Thesis, the complementary situation seems to be formed between **image NB** and **style RF**.

The performance on the genre class Minutes introduces some controversy: on the basis of precision, **Rainbow NB** shows a higher rate than the other two classifiers, but, on the basis of recall, **style RF** outperforms **Rainbow NB**. The comparison is further complicated by the observation that the average of precision and recall (given equal weight) suggests **image NB** as the best performer.

**Table 3. Genre classification across six classes on Dataset I, 10-fold cross validation.**

Genre (no. of items)	Method	Precision	Recall
Academic Monograph (16)	image NB	0.462	0.375
	style RF	<b>0.643</b>	<b>0.563</b>
	Rainbow NB	0.241	0.217
Book of Fiction (16)	image NB	0.4	0.125
	style RF	<b>0.813</b>	0.813
	Rainbow NB	0.763	<b>0.971</b>
Business Report (15)	image NB	0.273	0.2
	style RF	<b>0.667</b>	<b>0.4</b>
	Rainbow NB	0.453	0.173
Minutes (19)	image NB	0.667	0.526
	style RF	0.56	<b>0.737</b>
	Rainbow NB	<b>0.767</b>	0.272
Periodicals (19)	image NB	<b>0.773</b>	<b>0.895</b>
	style RF	0.565	0.684
	Rainbow NB	0.232	0.570
Thesis (19)	image NB	0.432	<b>0.889</b>
	style RF	<b>0.688</b>	0.611
	Rainbow NB	0.541	0.377

On the basis of average performance taken over precision and recall, the results in Table 4 presents **style RF** as the best overall performer. The precision of **style RF** is better than that of both of the other classifiers with respect to all genres except academic monographs and books of fiction, and recall is better with respect to all classes except Periodicals and Thesis.

The classifier **image NB** shows the best recall rate for detecting theses and displays a comparable recall rate for detecting periodicals.

Although the results of the experiments suggest style RF as the overall best performer on the two datasets, they do not identify genre classes for each classifier on which the classifier consistently outshines the other two classifiers. However, upon closer examination, the results do show that the binary partition of the genre classes, into classes with the three best performance and three worst performance, is preserved across the experiments on the two datasets: these partitions are (Minutes, Periodicals, Thesis) and (Academic Monograph, Book of Fiction, Business Report) for **image NB**, and (Book of Fiction, Minutes, Thesis) and (Academic Monograph, Business Report, Periodicals) for **style RF** and **Rainbow SVM**.

**Table 4. Genre classification across six genres on Dataset II, 10-fold cross validation.**

Genre (no. of items)	Method	Precision	Recall
<b>Academic Monograph</b> (99)	image NB	0.25	0.101
	style RF	0.718	<b>0.747</b>
	Rainbow NB	<b>0.74</b>	0.411
<b>Book of Fiction</b> (29)	image NB	0.111	0.069
	style RF	0.923	<b>0.828</b>
	Rainbow NB	<b>0.931</b>	0.807
<b>Business Report</b> (100)	image NB	0.385	0.05
	style RF	<b>0.825</b>	<b>0.825</b>
	Rainbow NB	0.797	0.609
<b>Minutes</b> (99)	image NB	0.604	0.818
	style RF	<b>0.913</b>	<b>0.949</b>
	Rainbow NB	0.91	0.874
<b>Periodicals</b> (67)	image NB	0.425	0.716
	style RF	<b>0.774</b>	0.716
	Rainbow NB	0.457	<b>0.794</b>
<b>Thesis</b> (100)	image NB	0.517	<b>0.91</b>
	style RF	<b>0.866</b>	0.84
	Rainbow NB	0.696	0.893

The general low level performance of the image features is partly due to the crude image representation. In the current model, the image features only capture the first page of the document, and each pixel value is strongly anchored to its position. This representation could be improved to combine representations of several pages of the document and to soften the positional information to embody the general shape or topology of the image. Likewise, for style, the size of the dataset and the variety of the documents in the datasets used for training and compiling word lists should be further examined for refinement.

### 6.3. Error analysis

In Section 6.2 we observed the style features as the best overall indicator for detecting genre, however, the situation may be more complicated than such a conclusion might portend. To understand fully the results in Section 6.2, a thorough error analysis is necessary.

**Table 5. Confusion matrix: Image NB on Dataset II.**

classified as ---->	AM	BF	BR	M	P	T
AM	10	4	4	20	25	36
BF	1	2	0	4	7	15
BR	17	9	5	16	32	21
M	6	0	1	81	0	11
P	5	1	3	8	48	2
T	1	2	0	5	1	91

**Table 6. Confusion matrix: Style RF on Dataset II.**

classified as --->	AM	BF	BR	M	P	T
AM	74	1	8	3	4	9
BF	0	24	0	0	2	0
BR	7	0	85	3	4	1
M	4	0	1	94	0	0
P	6	0	7	3	48	3
T	9	1	2	0	4	84

In Tables 5, 6, and 7, we have displayed the errors as six-by-six confusion matrices. The true genre labels of the documents are shown in the left most column and the genre labels assigned by the classifier is shown along the top row of the table. The genre class names have been denoted by their abbreviated names to save space. As reminder, **AM** stands for Academic Monograph, **BF** stands for Book of Fiction, **BR** stands for Business Report, **M** stands for Minutes, **P** stands for Periodicals, and **T** stands for Thesis.

**Table 7. Confusion matrix: Rainbow SVM on Dataset II.**

classified as --->	AM	BF	BR	M	P	T
AM	41	1	8	3	18	28
BF	0	23	1	0	4	1
BR	6	0	61	3	28	2
M	3	1	2	87	4	3
P	3	0	5	3	53	3
T	2	0	1	0	8	89

We have used two different measures of the confusion level displayed by the classifier: one based on belief ([8]) and another based on error impact. The belief  $B_C(C1:C2)$  of a classifier  $C$  that class  $C1$  is class  $C2$  is the number of documents in class  $C1$  labelled as being in  $C2$  divided by the number of documents in class  $C1$ . The error impact  $E_C(C1:C2)$  of the class  $C1$  in the documents

labelled by the classifier  $C$  as  $C2$  measures the percentage of errors arising from the predicted labels of documents in class  $C1$  within the errors arising from the classifier's decision to label documents as belonging to  $C2$ . More precisely, if  $C1 = C2$ ,  $E_C(C1:C2)$  is defined to be 0, and if  $C1 \neq C2$ ,  $E_C(C1:C2)$  is defined to be the number of documents of class  $C1$  which have been labelled as belonging to  $C2$  divided by the total number of documents *incorrectly* labelled as belonging to class  $C2$ . To compare values across classes, we have compensated for different numbers of document in each class by dividing  $B_C(C1:C2)$  and  $E_C(C1:C2)$  with the sum of  $B_C(C1:C2)$  over all  $C1$ , and  $E_C(C1:C2)$  over all  $C2$ , respectively. If the sum is zero then we simply define the belief and error impact to be zero. The same notation for belief and error impact has been retained to denote the normalised quantity.

We have introduced error impact in contrast to belief because belief is heavily influenced by the overall performance of the classifier itself. That is, having a high level of correct beliefs greatly reduces the incorrect beliefs of the classifier. In contrast, the greater or smaller number of academic monographs being labelled *correctly* as Academic Monograph does not have as predominant an influence over the relative distribution of different classes amongst the documents which have been *incorrectly* labelled Academic Monograph. We deemed error impact to be a better metric for accentuating the differences in confusion levels between classes within the performance of a single classifier.

Between two different classes  $C1$  and  $C2$ , the confusion level on the basis of belief,  $C_B(C1:C2)$ , is defined to be  $C_B(C1:C2) = B_C(C1:C2) + B_C(C2:C1)$ , and the confusion level on the basis of error impact,  $C_E(C1:C2)$ , is defined to be  $C_E(C1:C2) = E_C(C1:C2) + E_C(C2:C1)$ .

The contents of Table 8 indicate the feature type of the classifier exhibiting the lowest confusion level, between the pair of genre classes indicated on the two left most columns, based on the confusion metric noted on the top most row. Two feature types have been noted where the confusion levels were equal.

Both metrics agree that style displays the lowest level of confusion in differentiating the pairs **Book of Fiction** and **Minutes**, **Academic Monograph** and **Periodicals**, **Book of Fiction** and **Minutes**, **Business Report** and **Periodicals** and **Minutes** and **Thesis**, and image displays the lowest level for **Periodicals** and **Thesis** (see Table 8). However, we would ideally like to minimise both error impact and out-of-class belief. For each pair of classes in Table 8, if we combine the features which have been calculated to have the lowest level of confusion on the basis of belief and error impact, the results seem to support our intuition. For example, style and image would be estimated as the best features to

distinguish most pairs which include Periodicals which conforms to instinct, since periodicals deal with a wide range of topics, and we do not expect Rainbow features which emphasise topical distinction to fare well. In particular, visually elaborate periodicals and structurally formal theses are unsurprisingly best distinguished by image features.

**Table 8. Feature types with lowest pairwise confusion level on two confusion metrics.**

Genre pair		Metric	$C_B$	$C_E$
AM	BF		style, Rainbow	Rainbow
AM	BR		style	style
AM	M		style	Rainbow
AM	P		style	style
AM	T		style	image
BF	BR		style, Rainbow	style
BF	M		style	style
BF	P		style	image
BF	T		style, Rainbow	Rainbow
BR	M		style, Rainbow	Rainbow
BR	P		style	style
BR	T		style, Rainbow	style
M	P		style	image
M	T		style	style
P	T		image	image

The same consideration of confusion levels indicates that topical features do little to distinguish genre classes likely to have similar topic areas such as Academic Monograph and Thesis. And it supports the expectation that distinctions between Academic Monographs and Book of Fiction, and Book of Fiction and Thesis would be highly topical.

Although the features indicated in Table 8 are the features exhibiting the lowest confusion levels with respect to belief and error impact, in some cases, the difference is very slight (e.g.  $C_E$  for Academic Monographs and Book of Fiction). In interpreting the information in Table 8, it seems reasonable to take the confusion level differences into consideration. We have merged the contents of Table 8 with these differences and presented the result in Table 9 for a convenient overview. For example, the pair of classes Book of Fiction and

Periodicals has been examined to be best distinguished by style and image, but the figures in Table 9 seem to suggest that the weight is more prominently on image.

**Table 9. The difference between maximum and minimum belief and error impact confusion.**

Genre pair		Metric	$C_B$	$C_E$
AM	BF		0.25 style, Rainbow	0.01 Rainbow
AM	BR		0.5 style	0.06 style
AM	M		0.2 style	0.51 Rainbow
AM	P		0.2 style	0.08 style
AM	T		0.08 style	0.42 image
BF	BR		0.39 style, Rainbow	0.22 style
BF	M		0.08 style	0.54 style
BF	P		0.13 style	0.66 image
BF	T		0.32 style, Rainbow	0.73 Rainbow
BR	M		0.14 style, Rainbow	0.18 Rainbow
BR	P		0.37 style	0.09 style
BR	T		0.07 style, Rainbow	0.28 style
M	P		0.68 style	0.24 image
M	T		0.08 style	0.64 style
P	T		0.07 image	0.39 image

Likewise, Academic Monographs and Minutes seem best distinguished by style and Rainbow with a higher weight placed on Rainbow.

## 7. Conclusions

The results in this paper provide evidence that genre classification is a task composed of several

classification tasks, where each task is defined by the set of classes which are distinguished by features of similar types.

As a method optimised to be effective over a wide range of classes, automated genre classification relies most heavily on linguistic style. However, the method fails to be at its best when dealing with classes such as Periodicals which encompass many different styles and topics. These classes are often more readily distinguished by visual style (e.g binary classifications of Periodicals and Thesis, using the image features show an overall accuracy range of 0.91-0.94, depending on the statistical machine, with precision and recall above 0.9 in all cases, while classifications using the style features show an accuracy range of 0.85-0.916).

The conventional approach to handling visually distinctive and stylistically distinctive classes together is to combine the visual features with the style features (either equally weighted or by estimating optimal weights) to get the best of both worlds, but this often degrades the recall or precision of the classifier modeled on one of the feature groups which may have been high with respect to selected classes (e.g. the recall of **image NB** with respect to Thesis is degraded from 0.91 to a recall range of 0.72-0.82 when a classifier modeled on combined features is used). We propose to tackle this problem by approaching it from two directions: first, we would like to build a domain-dependent classifier built on feature strengths by

1. Taking a small slice of the data in the domain to estimate distinguishing features of each pair of classes via analyses similar to that presented in this document.
2. Constructing several classifiers by taking, for each classifier, a different weighted combination of labellers, based on the feature analysis of class pairs.
3. Augmenting the set of classifiers by adding randomly weighted classifiers to prevent the classifier from over-fitting the initial slice of data.
4. Taking a poll of votes cast by all the classifiers to obtain the final label.

and, second, in parallel, we would like to build rigorous definitions of genre classes in terms of low level visual, stylistic, semantic, and contextual metrics (rather than high level functional concepts), which can be established by further studies of the type presented in this paper.

We have only discussed the genre classification of document genres. However, the same methods can be applied to web page genres. Web page genre classification has immediate application. by enabling genre based searches, such as the retrieval of scientific

articles and product or book reviews. Further, the minimal reliance of the approach on syntactic structure inspires a possible application in the classification of selected non-document objects such as databases.

## 8. Acknowledgments

DELOS: Network of Excellence on Digital Libraries (G038-507618)<sup>1</sup> funded under the European Commission's IST 6th Framework Programme provides key framework and support for this research, as does the UK's Digital Curation Centre (DCC). The DCC<sup>2</sup> is supported by a grant from the Joint Information Systems Committee (JISC)<sup>3</sup> and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)<sup>4</sup> (GR/T07374/01). We would like to thank colleagues, Adam Rusbridge and Andrew McHugh, at HATII, University of Glasgow<sup>5</sup> who facilitated document retrieval and classification of the KRYSS I corpus with web support, and, Lea and Vera Beringer, who helped to validate the data used in the automated experiments of the research.

## 10. References

- [1] Bagdanov, A. and Worring, M. (2001) Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR2001)*, 79-90.
- [2] Barbu, E., Heroux, P., Adam, S., and Turpin, E. (2005) Clustering document images using a bag of symbols representation. In *Proceedings International Conference on Document Analysis and Recognition*, 1216-1220.
- [3] Bekkerman, R., McCallum, A., and Huang, G. (2004) *Automatic categorization of email into folders: benchmark experiments on enron and sri corpora*. Technical Report IR-418, Centre for Intelligent Information Retrieval, UMASS. <http://www.cs.umass.edu/~mccallum/papers/foldering-tr05.pdf>
- [4] Biber, D. (1995) *Dimensions of Register Variation: a Cross-Linguistic Comparison*. Cambridge University Press.
- [5] Boese, E. S. (2005) *Stereotyping the web: genre classification of web documents*. Master's thesis, Colorado State University.
- [6] Breiman, L. (2001) Random forests. *Machine Learning*, 45:5-32.
- [7] Burges, C. J. C. (1998) A Tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, Vol 2, 121-167.
- [9] Chen, L., and Tang, H. L. (2004) Improved computation of beliefs based on confusion matrix for combining multiple classifiers. *Electronic Letters*, Vol 4, No 4, 238- 239.
- [10] Dillon, A. and Vaughan, M. (1997) It's the journey and the destination: Shape and the emergent property of genre in evaluating digital documents. *New Review of Multimedia and Hypermedia*. Vol. 3, pp. 91-106.
- [11] Finn, A., and Kushmerick, N. (2006) Learning to classify documents according to genre. *Journal of American Society for Information Science and Technology*, 57(11), 1506-1518.
- [12] Karlgren, J., and Cutting, D. (1994) Recognizing text genres with simple metric using discriminant analysis. In *Proceedings 15th Conf. Comp. Ling.*, Vol 2, 1071-1075.
- [13] Kessler, G., Nunberg, B., and Schuetze, H. (1997) Automatic detection of text genre. In *Proceedings 35th Ann. Meeting ACL*, 32-38.
- [14] Kim, Y., and Ross, S. (2006) Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, *Proceedings European Conference on advanced technology and research in Digital Libraries (ECDL)*, Lecture Notes in Computer Science, Springer Verlag, Vol 4172, 63-74.
- [15] Kim, Y., and Ross, S. (2007) Detecting family resemblance: Automated genre classification. *Data Science Journal*, ISSN:1683-1470, Vol 6, , S172-S183.
- [16] Kim, Y. and Ross, S. (2007) Feature Type Analysis in Automated Genre Classification. <http://eprints.erpanet.org/128>.
- [17] McCallum, A. (1996) *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*. <http://www.cs.cmu.edu/~mccallum/bow>
- [18] Minsky, M. (1961). "Steps toward Artificial Intelligence." *Proceedings of the IRE* 49(1), 8-30.
- [19] Rauber, A. and Müller-Kögler, A. (2001) Integrating automatic genre analysis into digital libraries. In *Proceedings ACM/IEEE Joint Conf. Digital Libraries*, Roanoke, VA, 1-10, <http://doi.acm.org/10.1145/379437.379439>
- [20] Santini, M. (2004) State-of-the-art on Automatic Genre Identification, Technical Report ITRI-04-03, ITRI, University of Brighton, UK.
- [21] Santini, M. (2007) Automatic Identification of Genre in Web Pages, PhD Thesis, University of Brighton, UK.
- [22] Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000), Text genre detection using common word frequencies. In *Proceedings of 18<sup>th</sup> International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, Germany, 808-814.
- [23] Toms, E. and Campbell, D. (1999) Genre as interface metaphor: Exploiting form and function in gdigital environments. In *Proceedings 32<sup>nd</sup> Annual Hawaii International Conference on System Sciences (HICSS-32)*, ISBN:0-7695-0001-3, 2008-2015.
- [24] Witten, H. I., and E. Frank. (2005) *Data mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.

---

1 <http://www.delos.info>

2 <http://www.dcc.ac.uk>

3 <http://www.jisc.ac.uk>

4 <http://www.epsrc.ac.uk>

5 <http://www.hatii.arts.gla.ac.uk>