








Classification of Untranscribed Handwritten Notarial Documents by Textual Contents

Juan José Flores¹ , Jose Ramón Prieto¹ , David Garrido² ,
Carlos Alonso³ , and Enrique Vidal¹ 

¹ PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
{[juafloar](mailto:juafloar@prhlt.upv.es), [joprifon](mailto:joprifon@prhlt.upv.es), [evidal](mailto:evidal@prhlt.upv.es)}@prhlt.upv.es

² HUM313 Research Group, Universidad de Cádiz, Cadiz, Spain

³ tranSkriptorium IA, Valencia, Spain

Abstract. Huge amounts of digital page images of important manuscripts are preserved in archives worldwide. The amounts are so large that it is generally unfeasible for archivists to adequately tag most of the documents with the required metadata so as to allow proper organization of the archives and effective exploration by scholars and the general public. The class or “typology” of a document is perhaps the most important tag to be included in the metadata. The technical problem is one of automatic classification of documents, each consisting of a set of untranscribed handwritten text images, by the textual contents of the images. The approach considered is based on “probabilistic indexing”, a relatively novel technology which allows to effectively represent the intrinsic word-level uncertainty exhibited by handwritten text images. We assess the performance of this approach on a large collection of complex notarial manuscripts from the *Spanish Archivo Histórico Provincial de Cádiz*, with promising results.

Keywords: Content-based image retrieval · Document classification · Historical manuscripts

1 Introduction

Content-based classification of manuscripts is an important task that is generally performed by expert archivists. Unfortunately, however, many manuscript collections are so vast that it is not possible to have the huge number of archive experts that would be needed to perform this task.

Current approaches for textual-content-based manuscript classification require the handwritten images to be first transcribed into text – but achieving sufficiently accurate transcripts are generally unfeasible for large sets of historical manuscripts. We propose a new approach to perform automatically this classification task which does not rely on any explicit image transcripts.

Hereafter, bundles or folders of manuscript images are called “image bundles” or just “bundles” or “books”. A bundle may contain several “files”, also called “acts” or just “image documents”. The task consists of classifying a given image document, that may range from a few to tens of handwritten text images, into a predefined set of classes or “types”. Classes are associated with the topic or (semantic) content conveyed by the text written in the images of the document.

This task is different from other related tasks which, are often called with similar names, such as “content-based image classification”, applied to single, natural scene (not text) images, and “image document classification”, where classification is based on visual appearance or page layout. See [12] for a more detailed discussion on these differences, as well as references to previous publications dealing with related problems, but mainly aimed at printed text.

Our task is comparable to the time-honoured and well known task of *content-based document classification*, were the data are plain text documents. Popular examples of this traditional task, are *Twenty News Groups*, *Reuters*, *WebKB*, etc. [1, 9, 11]. The task here considered (textual-content-based handwritten text image document classification), is similar, except for a severe difference: our data are sets of digital images of handwritten text rather than file of (electronic) plain text. The currently accepted wisdom to approach our task would be to split the process into two sequential stages. First, a handwritten text recognition (HTR) system is used to transcribe the images into text and, second, content-based document classification methods, such as those referred to above, can be applied to the resulting text documents.

This approach might work to some extent for simple manuscripts, where HTR can provide over 90% word recognition accuracy [18]. But it is not an option for large historical collections, where the best available HTR systems can only provide word recognition accuracies as low as 40–60% [4, 15, 18]. This is the case of the collection which motivates this work, which encompasses millions of handwritten notarial files from the Spanish Archivo Histórico Provincial de Cádiz. A small subset of these manuscripts was considered in the Carabela project [4] and the average word recognition accuracy achieved was below 65% [15], dropping to 46% or less when conditions are closer to real-world usage [4]. Clearly, for these kinds of manuscript collections, the aforementioned two-stage idea would not work and more holistic approaches are needed.

In previous works [4, 12], we have proposed an approach which strongly relies on the so-called *probabilistic indexing* (PrIx) technology, recently developed to deal with the intrinsic word-level *uncertainty* generally exhibited by handwritten text and, more so, by historical handwritten text images [3, 10, 14, 20, 21]. This technology was primarily developed to allow search and retrieval of textual information in large untranscribed manuscript collections [3, 4, 19].

In our proposal, PrIx provides the probability distribution of words which are likely written in the images, from which statistical expectations of *word* and *document frequencies* are estimated. These estimates are then used to compute well-known text features such as *Information Gain* and Tf-Idf [11], which are in turn considered inputs to a Multilayer Perceptron classifier [12].

In this paper, we consolidate this approach and, as mentioned above, apply it to a new collection of handwritten notarial documents from the Archivo Provincial de Cádiz. In contrast with [4, 12], where the underlying class structure was very limited (just three rather artificial classes), here the classes correspond to real typologies, such as *power of attorney*, *lease*, *will*, etc. Our results clearly show the capabilities of the proposed approach, which achieves classification accuracy as high as 90–97%, depending on the specific set of manuscripts considered.

2 Probabilistic Indexing of Handwritten Text Images

The Probabilistic Indexing (PrIx) framework was proposed to deal with the intrinsic word-level uncertainty generally exhibited by handwritten text in images and, in particular, images of historical manuscripts. It draws from ideas and concepts previously developed for keyword spotting, both in speech signals and text images. However, rather than caring for “key” words, any element in an image which is likely enough to be interpreted as a word is detected and stored, along with its *relevance probability* (RP) and its location in the image. These text elements are referred to as “*pseudo-word spots*”.

Following [14, 21], the image-region word RP is denoted as $P(R = 1 \mid X = x, V = v)$, but for the sake of conciseness, the random variable names will be omitted and, for $R = 1$, we will simply write R . As discussed in [22], this RP can be simply approximated as:

$$P(R \mid x, v) = \sum_{b \sqsubseteq x} P(R, b \mid x, v) \approx \max_{b \sqsubseteq x} P(v \mid x, b) \quad (1)$$

where b is a small, word-sized image sub-region or Bounding Box (BB), and with $b \sqsubseteq x$ we mean the set of all BBs contained in x . $P(v \mid x, b)$ is just the posterior probability needed to “recognize” the BB image (x, b) . Therefore, assuming the computational complexity entailed by (1) is algorithmically managed, any sufficiently accurate isolated word classifier can be used to obtain $P(R \mid x, v)$.

This word-level indexing approach has proved to be very robust, and it has been used to very successfully index several large iconic manuscript collections, such as the French CHANCERY collection [3], the BENTHAM PAPERS [19], and the Spanish CARABELA collection considered in this paper, among others.¹

3 Plain Text Document Classification

If a text document is given in some electronic form, its words can be trivially identified as discrete, unique elements, and then the whole field of *text analytics* [1, 11] is available to approach many document processing problems, including *document classification* (DC). Most DC methods assume a document representation model known as *vector model* or *bag of words* (BOW) [1, 6, 11]. In this

¹ See: <http://transcriptorium.eu/demots/KWSdemos>.

model, the order of words in the text is ignored, and a document is represented as a *feature vector* (also called “word embedding”) indexed by V . Let \mathcal{D} be a set of documents, $D \in \mathcal{D}$ a document, and $\vec{D} \in \mathbb{R}^N$ its BOW representation, where $N \stackrel{\text{def}}{=} |V|$. For each word $v \in V$, $D_v \in \mathbb{R}$ is the value of the v -th feature of \vec{D} .

Each document is assumed to belong to a unique class c out of a finite number of classes, C . The task is to predict the best class for any given document, D . Among many pattern recognition approaches suitable for this task, from those studied in [12] the Multi-Layer Perceptron (MLP) was the one most promising.

3.1 Feature Selection

Not all the words are equally helpful to predict the class of a document D . Thus, a classical first step in DC is to determine a “good” vocabulary, V_n , of reasonable size $n < N$. One of the best ways to determine V_n is to compute the *information gain* (IG) of each word in V and retain in V_n only the n words with highest IG.

Using the notation of [12], let t_v be the value of a boolean random variable that is *True* iff, for some random D , the word v appears in D . So, $P(t_v)$ is the probability that $\exists D \in \mathcal{D}$ such that v is used in D , and $P(\bar{t}_v) = 1 - P(t_v)$ is the probability that *no* document uses v . The IG of a word v is then defined as:

$$\begin{aligned} \text{IG}(v) = & - \sum_{c \in C} P(c) \log P(c) \\ & + P(t_v) \sum_{c \in C} P(c | t_v) \log p(c | t_v) \\ & + P(\bar{t}_v) \sum_{c \in C} P(c | \bar{t}_v) \log P(c | \bar{t}_v) \end{aligned} \quad (2)$$

where $P(c)$ is de prior probability of class c , $P(c | t_v)$ is the conditional probability that a document belongs to class c , given that it contains the word v , and $P(c | \bar{t}_v)$ is the conditional probability that a document belongs to class c , given that it does *not* contain v . Note that the first addend of Eq. (2) does not depend on v and can be ignored to rank all $v \in V$ in decreasing order of $\text{IG}(v)$.

To estimate the relevant probabilities in Eq. 2, let $f(t_v) \leq M \stackrel{\text{def}}{=} |\mathcal{D}|$ be the number of documents in \mathcal{D} which contain v and $f(\bar{t}_v) = M - f(t_v)$ the number of those which do *not* contain v . Let $M_c \leq M$ be the number of documents of class c , $f(c, t_v)$ the number of these documents which contain v and $f(c, \bar{t}_v) = M_c - f(c, t_v)$ the number of those that do *not* contain v . Then, the relevant probabilities used in Eq. (2) can be estimated as follows:

$$P(t_v) = \frac{f(t_v)}{M} \qquad P(\bar{t}_v) = \frac{M - f(t_v)}{M} \quad (3)$$

$$P(c | t_v) = \frac{f(c, t_v)}{f(t_v)} \qquad P(c | \bar{t}_v) = \frac{M_c - f(c, t_v)}{M - f(t_v)} \quad (4)$$

3.2 Feature Extraction

Using information gain, a vocabulary V_n of size $n \leq N$ can be defined by selecting the n words with highest IG. By attaching a (real-valued) feature to each $v \in V_n$, a document D can be represented by a n -dimensional feature vector $\vec{D} \in \mathbb{R}^n$.

The value D_v of each feature v is typically related with the frequency $f(v, D)$ of v in D . However, absolute word frequencies can dramatically vary with the size of the documents and normalized frequencies are generally preferred. Let $f(D) = \sum_{v \in V_n} f(v, D)$ be the total (or “running”) number of words in D . The normalized frequency of $v \in V_n$, often called *term frequency* and denoted $\text{Tf}(v, D)$, is the ratio $f(v, D) / f(D)$, which is a max-likelihood estimate of the conditional probability of word v , given a document D , $P(v|D)$.

While Tf adequately deals with document size variability, it has been argued that better DC accuracy can be achieved by further weighting each feature with a factor that reflects its “importance” to predict the class of a document. Of course, IG could be used for this purpose, but the so-called *inverse document frequency* (Idf) [2, 8, 17] is argued to be preferable. Idf is defined as $\log(M / f(t_v))$, which, according to Eq. (3), can be written as $-\log P(t_v)$.

Putting it all together, a document D is represented by a feature vector \vec{D} . The value of each feature, D_v , is computed as the TfIdf of D and v ; i.e., $\text{Tf}(v, D)$, weighted by $\text{Idf}(t)$:

$$\begin{aligned} D_v &= \text{Tf} \cdot \text{Idf}(v, D) &&= \text{Tf}(v, D) \cdot \text{Idf}(v) \\ &= P(v|D) \log \frac{1}{P(t_v)} &&= \frac{f(v, D)}{f(D)} \log \frac{M}{f(t_v)} \end{aligned} \quad (5)$$

4 Textual-Content-Based Classification of Sets of Images

The primary aim of PrIx is to allow fast and accurate search for textual information in large image collections. However, the information provided by PrIx can be useful for many other text analytics applications which need to rely on incomplete and/or imprecise textual contents of the images. In particular, PrIx results can be used to estimate all the text features discussed in Sect. 3, which are needed for image document classification.

4.1 Estimating Text Features from Image PrIx 's

Since R is a binary random variable, the RP $P(R | x, v)$ can be properly seen as the statistical expectation that v is written in x . As discussed in [12], the sum of RPs for all the pseudo-words indexed in an image region x is the statistical expectation of the number of words written in x . Following this estimation principle, all the text features discussed in Sect. 3, which are needed for image document classification can be easily estimated.

Let $n(x)$ be the total (or “running”) number of words written in an image region x and $n(X)$ the running words in an image document X encompassing

several pages (i.e., $f(D)$, see Sect. 3.2). Let $n(v, X)$ be the frequency of a specific (pseudo-)word v in a document X . And let $m(v, \mathcal{X})$ be the number of documents in a collection, \mathcal{X} , which contain the (pseudo-)word v . As explained in [12], the expected values of these counts are:

$$E[n(x)] = \sum_v P(R | x, v) \quad (6)$$

$$E[n(X)] = \sum_{x \subseteq X} \sum_v P(R | x, v) \quad (7)$$

$$E[n(v, X)] = \sum_{x \subseteq X} P(R | x, v) \quad (8)$$

$$E[m(v, \mathcal{X})] = \sum_{X \subseteq \mathcal{X}} \max_{x \in X} P(R | x, v) \quad (9)$$

4.2 Estimating Information Gain and Tf·Idf of Sets of Text Images

Using the statistical expectations of document and word frequencies of Eqs. (6–9), IG and TfIdf can be straightforwardly estimated for a collection of text images. According to the notation used previously, a document D in Sect. 3 becomes a set of text images or *image document*, X . Also, the set of all documents \mathcal{D} becomes the text image collection \mathcal{X} , and we will denote \mathcal{X}_c the subset of image documents of class c . Thus $M \stackrel{\text{def}}{=} |\mathcal{X}|$ is now the total number of image documents and $M_c \stackrel{\text{def}}{=} |\mathcal{X}_c|$ the number of them which belong to class c .

The document frequencies needed to compute the IG of a word, v are summarized in Eqs. (3–4). Now the number of image documents that contain the word v , $f(t_v) \equiv m(v, \mathcal{X})$, is directly estimated using Eq. (9), and the number of image documents of class c which contain v , $f(c, t_v)$, is also estimated as in Eq. (9) changing \mathcal{X} with \mathcal{X}_c .

On the other hand, the frequencies needed to compute the Tf·Idf document vector features are summarized in Eq. (5). In addition to $f(t_v) \equiv m(v, \mathcal{X})$, we need the total number of running words in a document D , $f(D)$, and the number of times the word v appears in D , $f(v, D)$. Clearly, $f(D) \equiv n(X)$ and $f(v, D) \equiv n(v, X)$, which can be directly estimated using Eq. (7) and (8), respectively.

4.3 Image Document Classification

Using the Tf·Idf vector representation \vec{X} of an image document $X \in \mathcal{X}$, optimal prediction of the class of X is achieved under the minimum-error risk statistical framework as:

$$c^*(X) = \operatorname{argmax}_{c \in \{1, \dots, C\}} P(c | \vec{X}) \quad (10)$$

The posterior $P(c | \vec{X})$ can be computed following several well-known approaches, some of which are discussed and tested in [12]. Following the results

reported in that paper, the Multi-Layer Perceptron (MLP) was adopted for the present work. The output of all the MLP architectures considered is a softmax layer with C units and training is performed by backpropagation using the cross-entropy loss. Under these conditions, it is straightforward that the outputs for an input \vec{X} approach $P(c | \vec{X})$, $1 \leq c \leq C$. Thus Eq. (10) directly applies.

Three MLP configurations with different numbers of layers have been considered. In all the cases, every layer except the last one is followed by batch normalization and ReLU activation functions [7]. The basic configuration is a plain C -class perceptron where the input is totally connected to each of the C neurons of the output layer (hence no hidden layers are used). For the sake of simplifying the terminology, here we consider such a model as a “0-hidden-layers MLP” and refer to it as MLP-0. The next configuration, MLP-1, is a proper MLP including one hidden layer with 128 neurons. The hidden layer was expected to do some kind of intra-document clustering, hopefully improving the classification ability of the last layer. Finally, we have also considered a deeper model, MLP-2, with two hidden layers and 128 neurons in each layer. Adding more hidden layers did not provide further improvements.

5 Dataset and Experimental Settings

The dataset considered in this work is a small part of a huge manuscript collection of manuscripts held by the Spanish Archivo Histórico Provincial de Cádiz (AHPC). In this section, we provide details of the dataset and of the settings adopted for the experiments discussed in Sect. 6.

5.1 A Handwritten Notarial Document Dataset

The AHPC (Provincial Historical Archive of Cádiz) was established in 1931, with the main purpose of collecting and guarding notarial documentation that was more than one hundred years old. Its functions and objectives include the preservation of provincial documentary heritage and to offer a service to researchers that allows the use and consultation of these documentary sources.

The notarial manuscripts considered in the present work come from a very large collection of 16 849 bundles or “notarial protocol books”, with an average of 250 notarial acts or files and about 800 pages per book. Among these books, 50 were included in the collection compiled in the Carabela project [4].² From these 50 books, for the present work we selected two notarial protocol books, JMBD_4949 and JMBD_4950, dated 1723–1724, to be manually tagged with GT annotations. Figure 1 shows examples of page images of these two books.

The selected books were manually divided into sequential sections, each corresponding to a notarial act. A first section of about 50 pages, which contains a kind of table of contents of the book, was also identified but not used in the

² In <http://prhlt-carabela.prhlt.upv.es/carabela> the images of this collection and a PrIx-based search interface are available.



Fig. 1. Example of corpus pages from books JMDB_4949 and JMBD_4950.

present experiments. It is worth noting that each notarial act can contain from one to dozens of pages, and separating these acts is not straightforward. In future works, we plan to develop methods to also perform this task automatically, but for the present work we take the manual segmentation as given.

During the segmentation and labeling of the two notarial protocol books, the experts found a total of 558 notarial acts belonging to 38 different types or classes. However, for most classes, only very few acts were available. To allow the classification results to be sufficiently reliable, only those classes having at least *five* acts in each book were taken into account. This way, *five classes* were retained as sufficiently representative and 419 acts (i.e., documents) were finally selected: 220 in JMDB_4949 and 199 in JMBD_4950. So, in total, 139 acts (25%) were set aside, which amounted to 1321 page images (including the long tables of contents mentioned above), out of the 3186 pages of both books.

The five types (*classes*) we are finally left with are: *Power of Attorney* (P, from Spanish “Poder”), *Letter of Payment* (CP, “Carta de Pago”), *Debenture* (O, “Obligación”), *Lease* (A, “Arrendamiento”) and *Will* (T, “Testamento”). Details of this dataset are shown in Table 1. The machine learning task consists in training a model to classify each document into one of these $C = 5$ classes.

Table 1. Number of documents and page images for JMDB_4949 and JMBD_4950: per class, per document & class, and totals.

Classes	JMDB_4949						JMBD_4950					
	P	CP	O	A	T	Total	P	CP	O	A	T	Total
Number of documents	141	35	21	12	11	220	100	39	23	19	18	199
Average pages per doc.	3.6	4.5	4.2	4.6	6.0	4.0	3.7	4.8	5.4	5.2	10.0	4.8
Min-max pages per doc.	2-46	2-28	2-20	2-16	4-10	2-46	2-56	2-30	2-32	2-14	4-48	2-56
Total pages	514	158	90	56	66	884	370	188	124	100	179	961

5.2 Empirical Settings

PrIx vocabularies typically contain huge amounts of pseudo-word hypotheses. However, many of these hypotheses have low relevance probability and most of the low-probability pseudo-words are not real words. Therefore, as a first step, the huge PrIx vocabulary was pruned out avoiding entries with less than three characters, as well as pseudo-words v with too low estimated document frequency; namely, $E[m(v, \mathcal{X})] < 1.0$. This resulted in a vocabulary V of 559 012 pseudo-words for the two books considered in this work. Secondly, to retain the most relevant features as discussed in Sect. 3.1, (pseudo-)words were sorted by decreasing values of IG and the first n entries of the sorted list were selected to define a BOW vocabulary V_n . Exponentially increasing values of n from 8 up to 16 384 were considered in the experiments. Finally, a Tf-Idf n -dimensional vector was calculated for each document, $D \equiv X \in \mathcal{X}$. For experimental simplicity, $\text{Tf} \cdot \text{Idf}(v, D)$ was estimated just once all for all $v \in V$, using the normalized factor $f(D) \equiv E[n(X)]$ computed for all $v \in V$, rather than just $v \in V_n$.

For MLP classification, document vectors were normalized by subtracting the mean and dividing by the standard deviation, resulting in zero-mean and unit-variance input vectors. The parameters of each MLP architecture were initialized following [5] and trained according to the cross-entropy loss for 100 epochs using the SGD optimizer [16] with a learning rate of 0.01. This configuration has been used for all the experiments presented in Sect. 6.

The same leaving-one-out training and testing experiments were carried out for each book. In each experiment 10 runs were made with different initialization seeds, ending up with the average results for all runs. This amounts to 10 M leaving-one-out executions for each experiment, where M is the total number of documents in each book (see Table 1).

The source code and data used in the experiments presented in this paper are publicly available.³

6 Experiments and Results

The empirical work has focused on MLP classification of documents (handwritten notarial *acts*) of two books, JMBD_4949 and JMBD_4950. For each book separately, we classify its documents (groups of handwritten page images) into the five classes established in Sect. 5.1.

Classification error rates are presented in Fig. 2 for 12 increasing values of n , the number of (pseudo-)words selected with maximum Information Gain.

³ <https://github.com/PRHLT/docClasifbPRIA22>.

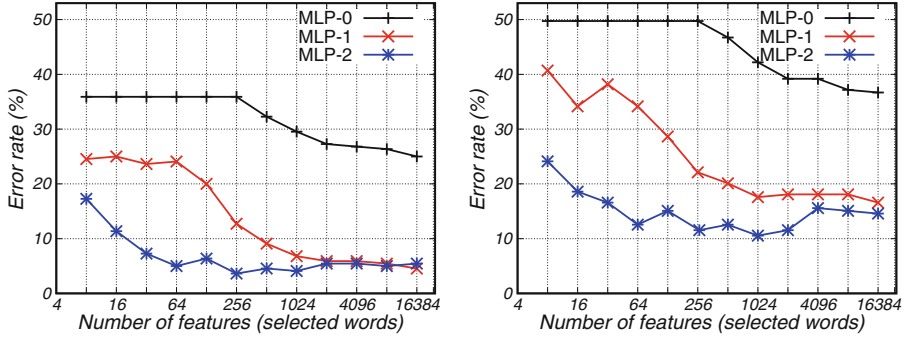


Fig. 2. Leaving-one-out classification error rate for three classifiers for JMBD_4949 (left), and JMBD_4950 (right). 95% confidence intervals (not shown for clarity) are all smaller than $\pm 6.0\%$ and smaller than $\pm 3.0\%$ for all the errors below than 6%.

Taking into account to the number of documents per each class (see Table 1), the error rates of a naive classifier based only on estimated prior probability per class would be 35.9% and 49.7% for JMBD_4949 and JMBD_4950, respectively.

For JMBD_4949, best results are obtained using MLP-2, achieving a 3.6% error rate with a vocabulary of $n = 256$ words. For this model, accuracy remains good if the vocabulary size is increased, but it degrades significantly for lower values of n . The MLP-1 model cannot achieve the same accuracy as MLP-2 for $n = 256$ or less words, although it achieves good accuracy for larger n , reaching a 4.5% error rate for $n = 16384$ words. The plain perceptron classifier (MLP-0) fails to overcome the accuracy of the naive prior-based classifier for vocabulary sizes $n \leq 256$ words. Its lowest error rate is 25.0% for $n = 16384$.

Results for JMBD_4950 are also presented in Fig. 2 (right). Since it departs from a much higher prior-based naive classification error rate (49.7%), it is not surprise that all the accuracies for this book are significantly worse than those achieved for JMBD_4949. Best results are also obtained using MLP-2, with 10.6% error rate with a vocabulary of 1024 words. For this model, accuracy degrades either if $n < 256$ or $n > 2048$. The accuracy of MLP-1 is lower than that of MLP-2, the best result being a 16.5% error rate with $n = 16384$ words. Finally, for the plain perceptron (MLP-0), we see that it again does not achieve an accuracy good enough to be taken into account.

Model complexity, in terms of numbers of parameters to train, grows with the number of features, n as:

$$\text{MLP-0: } 5n + 5 \quad \text{MLP-1: } 128n + 773 \quad \text{MLP-2: } 128n + 17285$$

For all $n > 64$, the least complex model is MLP-0, followed by MLP-1 and MLP-2. For $n = 2048$, MLP-0, MLP-1 and MLP-2 have 10245, 262917 and 279429 parameters, respectively. Therefore, despite the complexity of the model, MLP-2 is the best choice for the task considered in this work.

Table 2 shows the average confusion matrix and the specific error rate per class, using the best model (MLP-2) with the best vocabulary (n) for each book.

Table 2. Confusions matrices using the MLP-2 classifier with 256 and 1024 words with largest IG for JMBD_4949 and JMBD_4950, respectively. Counts are (rounded) averages over 10 randomly initialized leaving-one-out runs.

	JMBD_4949							JMBD_4950						
	P	CP	O	A	T	Total	Err(%)	P	CP	O	A	T	Total	Err(%)
P	138	0	1	0	2	141	2.1	92	1	3	1	3	100	8.0
CP	1	34	0	0	0	35	2.9	2	35	2	0	0	39	10.3
O	2	0	18	1	0	21	14.3	2	2	18	1	0	23	21.7
A	0	1	0	11	0	12	8.3	2	1	0	16	0	19	15.8
T	0	0	0	0	11	11	0.0	1	0	0	0	17	18	5.5
All	141	35	19	12	13	220	3.6	99	39	23	18	20	199	10.6

7 Conclusion

We have presented and showcased an approach that is able to perform textual-content-based document classification directly on multi-page documents of untranscribed handwritten text images. Our method uses rather traditional techniques for plaintext document classification, estimating the required word frequencies from image probabilistic indexes. This overcomes the need to explicitly transcribe manuscripts, which is generally unfeasible for large collections.

The present work successfully extends previous studies, but its scope is still fairly limited: only 419 document samples of five classes. Nevertheless, the experimental results achieved so far clearly support the capabilities of the proposed approach to model the textual contents of text images and to accurately discriminate content-defined classes of image documents. In future studies we plan to further extend the present work by taking into consideration all the document samples and classes available. Using the two bundles JMDB4949 and JMBD_4950 together will allow us to roughly double the number of classes with enough documents per class to allow reasonable class modeling. Furthermore, in order to approach practical situations, experiments will also assess the model capability of rejecting test documents of classes not seen in training.

In our opinion, probabilistic indexing opens new avenues for research in textual-content-based image document classification. In a current study we capitalize on the observation that fairly accurate classification can be achieved with relatively small vocabularies, down to 64 words in the task considered in this paper. In this direction, we are exploring the use of information gain and/or other by-products of the proposed MLP classifiers to derive a small set of words that semantically describes the contents of each image document. A preliminary work in this direction is described in [13]. It aims at automatic or semi-automatic creation of metadata which promises to be extremely useful for scholars and the general public searching for historical information in archived manuscripts.

Finally, in future works, we plan to explore other classification methods such as recurrent models that can take into account the sequential regularities exhibited by textual contents in successive page images of formal documents.

Acknowledgments. Work partially supported by the research grants: Ministerio de Ciencia Innovación y Universidades “DocTIUM” (RTI2018-095645-B-C22), Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121) and PID2020-116813RB-I00a funded by MCIN/AEI/ 10.13039/501100011033. The second author’s work was partially supported by the Universitat Politècnica de València under grant FPI-I/SP20190010.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer, Boston (2012). <https://doi.org/10.1007/978-1-4614-3223-4>
2. Aizawa, A.: An information-theoretic perspective of TF-IDF measures. *Inf. Proc. Manag.* **39**(1), 45–65 (2003)
3. Bluche, T., et al.: Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 311–316, November 2017
4. Vidal, E., et al.: The Carabela project and manuscript collection: large-scale probabilistic indexing and content-based classification. In: 16th ICFHR, September 2020
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **9**, 249–256 (2010)
6. Ikonomakis, M., Kotsiantis, S., Tampakas, V.: Text classification using machine learning techniques. *WSEAS Trans. Comput.* **4**(8), 966–974 (2005)
7. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift (2015)
8. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science (1996)
9. Khan, A., Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. *J. Adv. Inf. Technol.* **1**(1), 4–20 (2010)
10. Lang, E., Puigcerver, J., Toselli, A.H., Vidal, E.: Probabilistic indexing and search for information extraction on handwritten German parish records. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 44–49, August 2018
11. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
12. Prieto, J.R., Bosch, V., Vidal, E., Alonso, C., Orcero, M.C., Marquez, L.: Textual-content-based classification of bundles of untranscribed manuscript images. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3162–3169. IEEE (2021)
13. Prieto, J.R., Vidal, E., Sánchez, J.A., Alonso, C., Garrido, D.: Extracting descriptive words from untranscribed handwritten images. In: Proceedings of the 2022 Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA) (2022)
14. Puigcerver, J.: A Probabilistic Formulation of Keyword Spotting. Ph.D. thesis, Univ. Politècnica de València (2018)

15. Romero, V., Toselli, A.H., Vidal, E., Sánchez, J.A., Alonso, C., Marqués, L.: Modern vs diplomatic transcripts for historical handwritten text recognition. In: Cristani, M., Prati, A., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019. LNCS, vol. 11808, pp. 103–114. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30754-7_11
16. Ruder, S.: An overview of gradient descent optimization algorithms **14**, 2–3 (2017)
17. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manag.* **24**(5), 513/523 (1988)
18. Sánchez, J.A., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: A set of benchmarks for handwritten text recognition on historical documents. *Pattern Recogn.* **94**, 122–134 (2019)
19. Toselli, A., Romero, V., Vidal, E., Sánchez, J.: Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing. In: 15th IAPR International Conference on Document Analysis and Recognition (ICDAR) (2019)
20. Toselli, A.H., Vidal, E., Puigcerver, J., Noya-García, E.: Probabilistic multi-word spotting in handwritten text images. *Pattern Anal. Appl.* **22**(1), 23–32 (2018). <https://doi.org/10.1007/s10044-018-0742-z>
21. Toselli, A.H., Vidal, E., Romero, V., Frinken, V.: HMM word graph based keyword spotting in handwritten document images. *Inf. Sci.* **370–371**, 497–518 (2016)
22. Vidal, E., Toselli, A.H., Puigcerver, J.: A probabilistic framework for lexicon-based keyword spotting in handwritten text images. Technical report, UPV (2017)