machine learning & knowledge extraction

MDPI

Article

Automatic Electronic Invoice Classification Using Machine Learning Models

Chiara Bardelli [1], Alessandro Rondinelli [2], Ruggero Vecchio [2] and Silvia Figini [3,*]

[1] Department of Computational Mathematics and Decision Sciences, University of Pavia, 27100 Pavia, Italy; chiara.bardelli01@universitadipavia.it
[2] Datev.it S.p.a., 20090 Assago, Italy; alessandro.rondinelli@datev.it (A.R.); ruggero.vecchio@datev.it (R.V.)
[3] Department of Political and Social Sciences, University of Pavia, 27100 Pavia, Italy
* Correspondence: silvia.figini@unipv.it

Received: 1 October 2020; Accepted: 24 November 2020; Published: 30 November 2020

Abstract: Electronic invoicing has been mandatory for Italian companies since January 2019. All the invoices are structured in a predefined xml template which facilitates the extraction of the information. The main aim of this paper is to exploit the information contained in electronic invoices to build an intelligent system which can simplify accountants' work. More precisely, this contribution shows how it is possible to automate part of the accounting process: all the invoices of a company are classified into specific codes which represent the economic nature of the financial transactions. To accomplish this classification task, a multiclass classification algorithm is proposed to predict two different target variables, the account and the VAT codes, which are part of the general ledger entry. To apply this model to real datasets, a multi-step procedure is proposed: first, a matching algorithm is used for the reconstruction of the training set, then input data are elaborated and prepared for the training phase, and finally a classification algorithm is trained. Different classification algorithms are compared in terms of prediction accuracy, including ensemble models and neural networks. The models under comparison show optimal results in the prediction of the target variables, meaning that machine learning classifiers succeed in translating the complex rules of the accounting process into an automated model. A final study suggests that best performances can be achieved considering the hierarchical structure of the account codes, splitting the classification task into smaller sub-problems.

Keywords: multiclass classification; text mining; accounting control system

1. Introduction

In the digitization era, artificial intelligence systems can bring many opportunities in industrial and information areas to improve the efficiency and deliver more value to business. On the basis of the results obtained in the research conducted by Frey and Osborne [1], the accountancy profession is one of the job which will be highly automated and digitized in the near future. Even if, thanks to computer systems, the workload of accountants has been reduced [2], repetitive and monotonous routine tasks are still a relevant part of the accountants' work. For this purpose, artificial intelligence can be used to replace repetitive human activities with highly automated systems giving the possibility for the accountants to focus on more stimulating and motivating activities such as decision-making, problem solving, advising and business strategy development [3,4].

One of the tasks of accountants which can be easily automated is the creation of the general journal which includes each financial transaction made by a business company. Among all the records of the financial transactions, we find those related to supplier and customer invoices. When an invoice is received or issued by a company, the bookkeeper or accountant creates the related accounting journal entries which describes the economic nature of the transactions using a specific list of codes (i.e.,

the Chart of Accounts and the codes related to VAT rules). This task can be treated using machine learning models: input variables include information extracted from the invoices and the characteristics of the companies, while the codes used for the creation of the accounting journal entry can be considered as the target variables. Our methodological approach considers three different steps in the development of an automated system which could propose possible codes for the journal entry: collection of invoices and reconstruction of the training set, processing of the textual content of the invoices, and automatic classification of invoices to build the associated journal entry based on the training of a machine learning model. Figure 1 depicts the main steps used in this study, which are explained in detailed in Section 4.

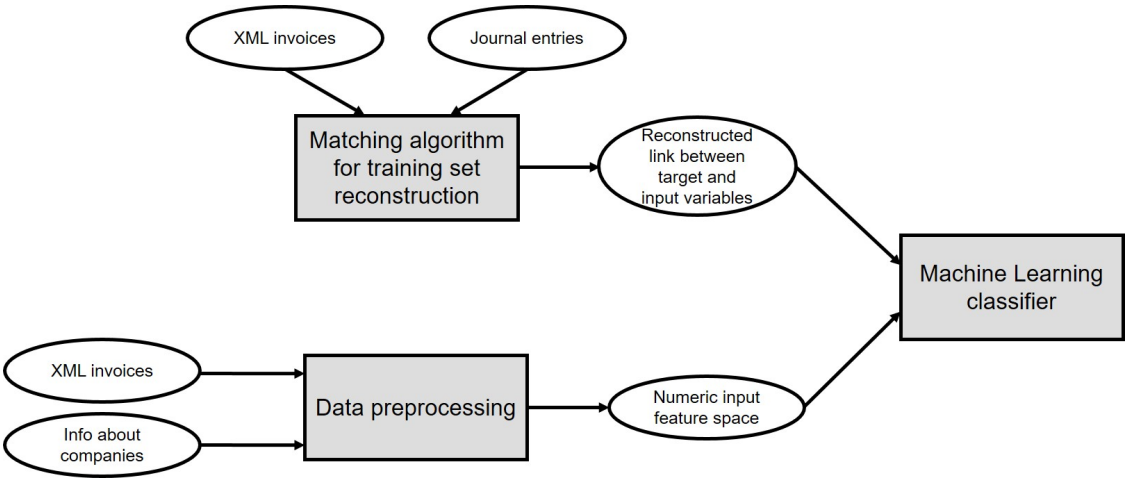

Figure 1. Methodology used in this study to process data and train a machine learning classifier. The rectangles represent the three methodological steps, while the circles contain input/output elements of the building blocks.

Starting from January 2019, all Italian companies are required to use electronic invoices. They are composed of a structured and fixed xml template provided by "Agenzia delle Entrate", the Italian governmental agency which operates to ensure a high level of tax compliance. This allows solving the issue of extracting information from documents which, thanks to electronic invoicing, follow a regular structure shared by all the Italian companies. As a consequence, the acquisition of the data, contained in the invoice, is handled more easily representing a strong encouragement to focus the research on the automation of the classification of transactions for the creation of the general journal.

This paper introduces a machine learning model to predict the elements of a bookkeeping entry starting from the content of the electronic invoice. Each line of the invoice is represented by a numeric vector that combines textual description, other information related to the line, and the characteristics of the companies involved in the invoice. We test performances of our classifiers on two anonimyzed real-world accounting datasets which belong to two different Chartered Accounting firms. Data are collected and provided by Datev.it Spa, a company which develops software for professional accounting.

Our methodological approach considers: (i) the reconstruction of the training set; (ii) the definition of a complex structure of the account codes which are organized in a hierarchical taxonomy based on five different levels for more than 200 different labels leading to an imbalanced classification problem; and (iii) the strong heterogeneity in rules and methodologies used by different Chartered Accounting firms in the creation of the journal entries.

The paper is organized as follows. Section 2 reports the literature concerning the application of machine learning techniques to accounting systems. Section 3 defines the problem and the data available. Section 4 introduces the methods applied to reconstruct the training set and to solve our predictive problem. Section 5 shows the application and the most relevant results are displayed. Section 6 contains conclusions and further ideas of research.

2. Literature Review

In accounting systems, machine learning techniques have been recently applied for different purposes trying to improve the efficiency and the precision of monotonous and repetitive tasks.

When invoices are not required to fill a fixed template, one of the most laborious step of the bookkeeping process in terms of working hours is the extraction of structured information from the printed documents. Classical deterministic OCR (Optical Character Recognition) techniques are usually applied to this problem. Early approaches exploit the structure of documents to extract fields based on their position or detected forms [5,6]. However, to generalize extractions patterns with a flexible approach which can be applied to any template, some attempts have been made recently to combine deterministic OCR methods to machine learning models in presence of large quantities of historical data. In [7,8], deep neural networks are trained on a large dataset to extrapolate specific fields from the images of the invoices. In [9], the density of the black pixels and the density of image edges are used as input variables to train a classifier which predicts the class of the invoice. In both works, the need of a large historical dataset is highlighted as a possible drawback for these techniques.

Another aspect of interest which has been tackled with machine learning models is the anomaly detection problem in the accounting journal. With the increasing complexity of business processes and the growing amount of structured accounting data, identifying erroneous or fraudulent business transactions (and corresponding journal entries) represents a critical challenge for accountants and auditors. In [10,11], a novel unsupervised approach based on neural networks is proposed to detect anomalous journal entries and potentially fraudulent activities in large-scale accounting data. In [12], a similar architecture is applied also to a small subset of data to investigate if accurate results can be obtained also in case of more limited information. These methods, in contrast with the classical approach based on the experience of chartered accountants such as static red-flag tests, are able to detect novel schemes of fraud based on historical data. Therefore, the opportunity to develop an automated and high precision detection of accounting anomalies can save work time.

As the last area of application of machine learning models to accounting system, there are at least two works [13,14] which try to develop a classifier able to predict the details of an accounting journal entry; in particular, they focus on the prediction of the account codes. The performances of the data driven algorithms are compared to a deterministic approach based on a rule induction system built on the experience of accountants. In both works, the difference in accuracy between the machine learning classifier and the rule induction classifier cannot be considered statistically significant, concluding that there is potential for machine learning in this area, but it still does not outperform the existing deterministic implementation. A possible extension could be to combine the results of rule induction system to the predictions of the machine learning classifier. In this work, we extend the research in [13,14] to the prediction of the entire journal entry, considering as target variable not only the account codes but also the VAT codes.

3. Dataset Description

The accounting general journal is composed of a list of journal entries which collects financial transactions of a business firm and classifies them into specific codes. A consistent part of the journal entries derives from the recording of the customer and supplier invoices of a company. Thanks to the electronic invoicing introduced recently in Italy, the extraction of the relevant fields from the xml template can be easily handled.

Given all the information about a single line of an invoice and the characteristics of the companies which are involved in the invoice, the aim is to construct the journal entry related to a specific line of the invoice, thus predicting the account codes and the VAT codes. Figure 2 clarifies the relation between an invoice and its associated journal entries. Multiple lines of an invoice can be associated to the same journal entry with the same codes. For the sake of simplicity, in this study, we assume that lines inside an invoice are independent one from each other (ignoring the fact that lines which belong to the same invoice are all influenced by the context of the invoice itself).

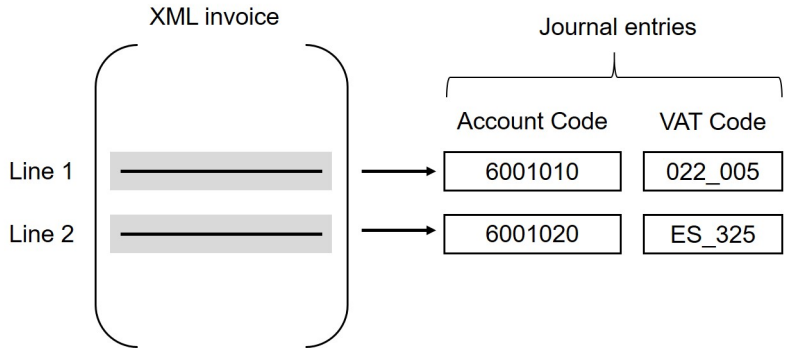

Figure 2. Classification approach to predict VAT codes and account codes.

In our classification task, we construct the prediction rule given the training sample $\left\{(y_i^{(1)}, y_i^{(2)}, \mathbf{x}_i)\right\}_{i=1}^{n}$ where:

- $y_i^{(1)}, i = 1, \ldots, n$, are categorical observations which represent the account codes associated to the ith line of the invoice. The account codes belong to the Chart of Accounts which has a particular structure: codes are organized in a hierarchical structure and only the accounts, which are the leaves of the tree, are used as tag in the journal entry. In our problem, we consider only the accounts in the leaves of the hierarchical tree.
- $y_i^{(2)}, i = 1, \ldots, n$ represent the VAT codes to predict associated to the ith line of the invoice. This target variable is composed of two different sub-codes: one related to the tax rate applied to the line of the invoice, and the other one related to tax rule. In our problem, this two codes are considered as a unique variable to predict.
- $\mathbf{x}_i$ is the vector of predictors related to the content of the invoice and the characteristics of companies involved in the transactions.

The dataset considered in this study is the combination of two data sources:

- Customers and suppliers xml invoices of different companies
- Accounting journal entries related to the recording process of xml invoices (this data source contains account codes and VAT codes)

The match of these two different sources is possible only at document level: the general journal records can be directly associated only to their original invoice. On the other hand, it is not possible to recreate directly the link between a single line of an invoice and the related journal entry. This problem of data reconstruction can be addressed exploiting the information about the amounts which are included both in the xml invoices (detailed amounts) and in the accounting journal entry (aggregated amounts). Starting from this point, it is possible to translate the problem at hand into a combinatorial optimization task as the knapsack representation with equality constraints as explained in Section 4.1.

In our machine learning algorithm, we consider features related to the content of the invoice and characteristics of the companies. In particular, the information considered includes:

- textual description of the line of the invoice;
- codes associated to the line (e.g., tax rate); and
- information about activities performed by companies:
  - ATECO code (classification of economic Italian activities) provided by ISTAT (the Italian Statistics Agency);
  - ISA categories based on the level of fiscal reliability;
  - type of supplier, namely person, Italian firm, European company, or extra-European company;
  - type of accounting used by the company: ordinary or simplified; and
  - tax regime.

Machine learning classifiers are trained and tested on two anonymized real-world datasets of two different accounting firms which include invoices from January 2019 to March 2020. The first one contains about 32,172 electronic invoices which include more than 320,000 lines to classify. The second one is composed by 34,932 invoices and more than 200,000 lines. The number of distinct codes which the algorithm has to predict is shown in Figure 3. The most complex problem, in terms of accuracy, seems to be the one related to the prediction of the account codes for the received invoices since, in this scenario, the number of different categories to predict is very high, as reported in Figure 3a.

(**a**) (**b**)

Figure 3. Number of distinct codes used for sent and received invoices in the two datasets analyzed: (**a**) account codes; and (**b**) VAT codes.

4. Methodological Proposal

This section introduces methods applied to derive the training set, the pre-processing approaches adopted to prepare the training set and the predictive classification algorithm used to predict the account codes and of VAT codes of the journal entry.

4.1. Knapsack Problem

The reconstruction data problem can be represented as a multi knapsack problem with equality constraints [15] considering a single invoice at a time.

Let $i = 1, \ldots, N$ be the lines of the invoice and $j = 1, \ldots, M$ be the index that identifies a single entry of the accounting general journal. The combinatorial optimization problem can be formulated as follows:

$$\max \sum_{j=1}^{M} \sum_{i=1}^{N} p_i z_{ij} \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} c_i z_{ij} = b_j, \quad \forall j \tag{2}$$

$$\sum_{j=1}^{M} z_{ij} = 1, \quad \forall i \tag{3}$$

where c_i is the detailed amount related to the ith line of the invoice and b_j is the aggregated total of the jth journal entry. In our setting, the vector of weights p_i is equal to 1, for each i. The value of $z_{i,j}$ is 1 if the ith line of the invoice is associated to the jth journal entry and 0 otherwise.

The problem has been resolved through a heuristic which stops when the first feasible solution is found. First, the heuristic sorts the detailed amounts of invoice lines in a vector in decreasing order and the aggregated amounts in an increasing order vector. Then, the algorithm starts to match first

values in the first positions of the two vectors. Notice that lines of invoice with negative amounts have been excluded from the analysis (the algorithm can not converge in presence of negative values).

The algorithm has successfully matched 61% of the lines of our two initial datasets without any ambiguity. Eleven percent of the lines were matched using the first solution proposed by the heuristic. These two sources of data were used for the training phase of the classifier and the evaluation of its performances. The other 28% of lines were excluded from the analysis since the heuristic did not converge because of the high number of lines in the invoice or because of the presence of negative amounts in the invoice.

4.2. Data Pre-Processing

Most of the information extracted from the invoice, to construct the feature space of our problem, are categorical variables. Since many machine learning algorithms cannot operate on labeled data directly (they require all input variables and output variables to be numeric), we converted categorical features into numeric vectors through the one-hot encoding representation [16]. If K is the number of all different categories which the variable can assume, the variable is transformed into K binary vectors. Each binary vector represents a category of the original variable, and their elements are marked to 1 only if the original variable was equal to that specific category, leaving all the other elements equal to 0.

In addition, textual descriptions of the lines of invoices are also included in the input space since they contain helpful information for the creation of the journal entry. To process textual data, descriptions were previously cleaned through standard pre-processing steps [17]: punctuation, numbers, stop words and words of two characters were removed from the text. Finally, textual information was tokenized to create an array of words.

To transform textual data information into a suitable numeric feature space, we compared two different procedures:

- Bag of Words (BoW) approach [18] is a simple way to encode the array of words into a binary vector. The main drawback is that the length of the feature space grows linearly with the number of distinct words, leading to infeasible dimensions of the feature space. Different methods can be adopted for the dimensionality reduction [19]; in our case, we included in the vocabulary words with a frequency higher than 0.1% in the collection of documents.
- Word2Vec (W2V) algorithm [20] is a language modeling technique which maps similar sentences into similar numeric vectors of fixed size. Since Continuous Bag of Words (CBoW) model is faster with respect to Skipgram model and shows better performances in case of high sample size [20], we preferred to apply CBoW to our dataset fixing the dimension of the output vector to 100 and the window of words to consider for the prediction equal to 5. All words with total frequency lower than 10 were ignored.

These two procedures were applied on the data at hand before running the classification algorithms.

4.3. Classification Algorithm

The aim of this work is to understand if standard rules and logic of the accounting process can be learned from a machine learning classifier which is trained on the real data of an accounting firm. Two different models were trained separately to predict two different target variables:

- the account codes related to the economic nature of the transaction; and
- the VAT codes related to the tax rates coupled with tax regulations applied to the invoices.

The input variables chosen for the two target variables are slightly different. The textual data are part of feature space in both cases, but the set of variables which describe the characteristics of the companies changes depending on which target variable we aim to predict.

Various supervised machine learning techniques have been proposed in the literature for the automatic classification of text documents [21], such as Naïve Bayes [22], neural networks [23], Support

Vector Machine (SVM) [24], and decision tree, as well as ensemble methods [25]. However, considering the high number of distinct classes to predict (especially for the prediction of account codes in case of received invoice, as depicted in Figure 3a), we chose to train classifiers that can better handle this type of problem. For this reason, SVM was excluded from the empirical analysis, despite its well-known good performances obtained in classification task with textual data. Indeed, from a computational perspective, it is necessary to train a binary classifier for each class of the dataset, and thus tune the hyper-parameters of each classifier, leading to a very challenging optimization problem [26].

In this study, for each classification task, three different machine learning algorithms were compared: Random Forests (RF) [27], AdaBoost [28], and Multilayer Perceptron (MlP) [29]. The choice of these algorithms is motivated by the fact that they do not depend from any assumptions on the underlying distribution of input data, they all achieve good performances in case of large training set [21], and they implement a multi-class version.

5. Empirical Analysis

The different nature of sent and received invoices led us to split the data into two different sub-datasets and analyze them separately to obtain two sub-classification problems. To evaluate the performances of all the possible approaches described in Section 4, a 10-fold Monte Carlo cross validation method was used to estimate precision, recall, and f1-score on the validation sets.

Figures 4 and 5 show recall for the different combinations of algorithms applied, respectively, to sent and received invoices for the prediction of the account codes. As we expected, the algorithms trained on received invoices show, in general, an accuracy rate lower than sent invoices. This is mainly due to the fact that received invoices are more diversified in terms of content and number of possible account codes associated. As far as the pre-processing algorithm for textual data is concerned, Word2Vec obtains better results compared to Bag of Words both when it is used with Random Forests and AdaBoost. A better performance of the Word2Vec can be observed both in results of both the two Chartered Accounting firms in the case of received invoices (Figure 5). This can be explained by the limited vocabulary used in Bag of Words model: a selection of words was made following the parameters of Section 4.2 in order to reduce the dimensions of the feature space and keep computational costs under control. Fixing the method for preprocessing textual data, performances of RF, AdaBoost, and MIP are equivalent when applied to sent invoices; on the other hand, in the case of received invoices, the first algorithm shows better results in the case of both Word2Vec and Bag of Words. From these results, it seems that RF is able to better handle classification problems with a higher number of distinct categories to predict in our problem. Tables 1 and 2 report the mean precision and the mean f1-score of the cross validation with their associated standard deviation.

Figures 6 and 7 display recall indexes related to the prediction of the VAT codes. It is remarkable to note that the number of distinct VAT codes is smaller than the number of distinct account codes, especially in the case of received invoices, as shown in Figure 3. As a consequence, the precision of all classifiers for the VAT codes is on average higher with the respect to the results of prediction of the account codes as shown in Tables 3 and 4. In particular, as regards VAT codes, Word2Vec results the best approach to transform textual data in a numeric vector in the case of received invoices. In the case of sent invoices, there is no differences between the two textual approaches, probably due to the fact that prediction of the VAT codes is not highly influenced by the textual information contained in the invoice but other input variables of the invoice are sufficient to obtain high prediction accuracy. As regards the machine learning model, AdaBoost seems to perform slightly better for sent invoices but it is equivalent to RF in case of received invoices.

In general, we observe that results for the two different accounting firms (Datasets 1 and 2) are consistent with each other, concluding that, at least in these two cases, it is possible to choose a unique type of classifier and train it on historical data of a specific accounting firm. This can be considered an advantage from the business value point of view, since we have a tool adaptable to both the accounting firms, which learn new predictions from own historical accounting database.

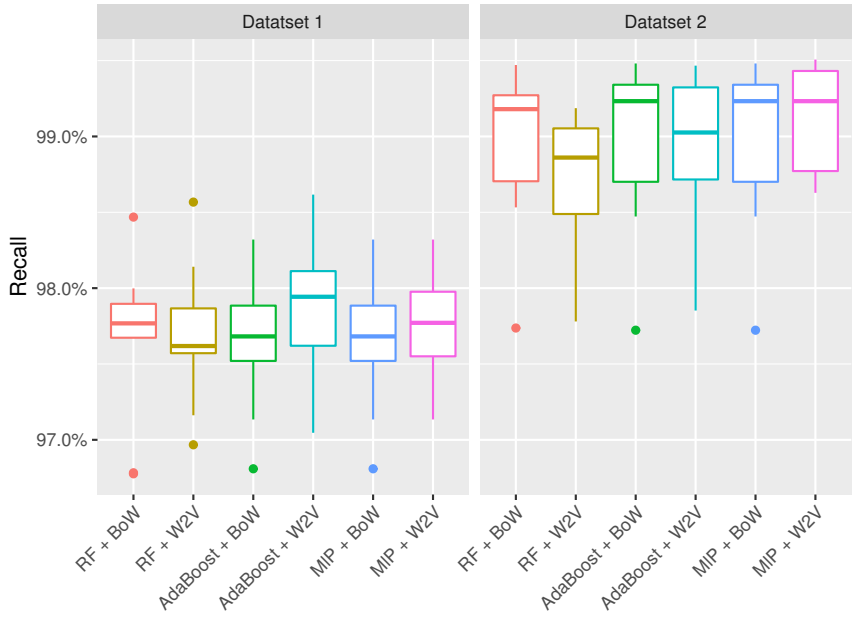

Figure 4. Cross validation values of the recall index for the prediction of the account codes (sent invoices).

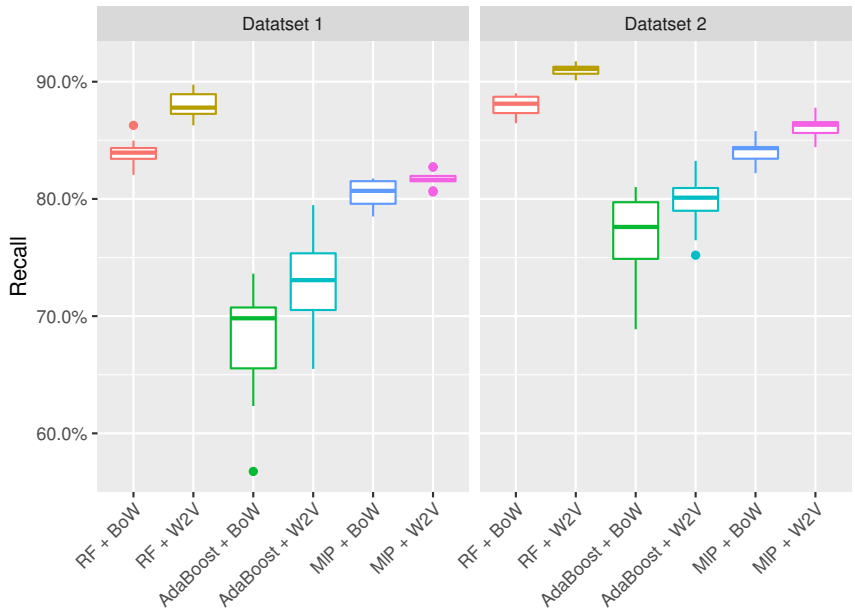

Figure 5. Cross validation values of the recall index for the prediction of the account codes (received invoices).

Table 1. Mean cross validation precision and f1-score of the classifiers which predict account codes (sent invoices). Standard deviation is reported in parentheses.

	Dataset 1		**Dataset 2**	
	Precision	F1-score	Precision	F1-score
RF + BoW	97.7% (±0.5%)	97.6% (±0.5%)	98.9% (±0.6%)	98.8% (±0.7%)
RF + W2V	97.8% (±0.4%)	97.6% (±0.5%)	98.7% (±0.5%)	98.6% (±0.6%)
AdaBoost + BoW	97.7% (±0.4%)	97.6% (±0.4%)	98.9% (±0.6%)	98.9% (±0.7%)
AdaBoost + W2V	98.0% (±0.4%)	97.8% (±0.5%)	98.9% (±0.5%)	98.8% (±0.6%)
MlP + BoW	97.7% (±0.4%)	97.6% (±0.4%)	98.9% (±0.6%)	98.9% (±0.7%)
MlP + W2V	97.9% (±0.3%)	97.6% (±0.4%)	99.0% (±0.5%)	98.9% (±0.7%)

Table 2. Mean cross validation precision and f1-score of the classifiers which predict account codes (received invoices). Standard deviation is reported in parentheses.

	Dataset 1		**Dataset 2**	
	Precision	F1-score	Precision	F1-score
RF + BoW	83.2% (±1.5%)	82.6% (±1.4%)	86.7% (±0.6%)	86.4% (±0.8%)
RF + Word2Vec	87.4% (±1.7%)	86.9% (±1.5%)	90.0% (±0.5%)	89.8% (±0.6%)
AdaBoost + BoW	68.6% (±5.0%)	67.7% (±4.7%)	76.2% (±4.5%)	77.0% (±3.6%)
AdaBoost + Word2Vec	73.6% (±4.3%)	72.8% (±4.3%)	79.6% (±3%)	79.5.2% (±2.8%)
MlP + BoW	80.8% (±1.4%)	81.2% (±1.5%)	83.2% (±1.6%)	83.9.1% (±1.4%)
MlP + W2V	79.5% (±1.2%)	80.0% (±1.0%)	83.1% (±1.0%)	84.0% (±1.1%)

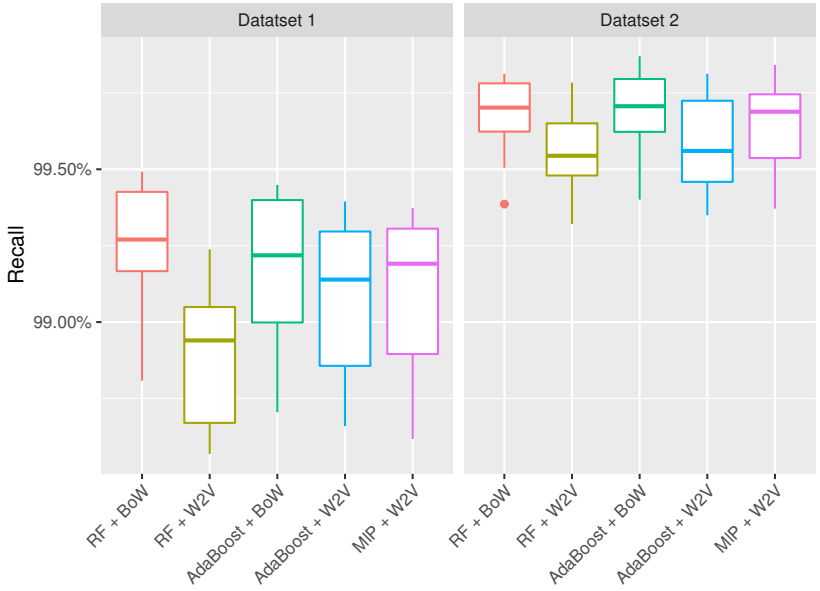

Figure 6. Cross validation values of the recall index for the prediction of the IVA codes (sent invoices).

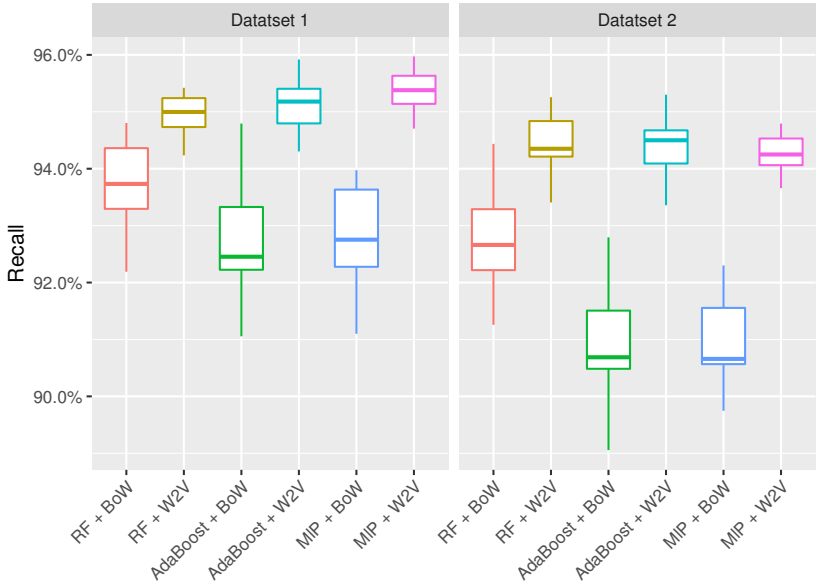

Figure 7. Cross validation values of the recall index for the prediction of the IVA codes (received invoices).

To improve performances of models for the prediction of the account codes for received invoices (which is the most problematic target variable considering the accuracy rates in Table 2 and the

recall index in Figure 5), we also investigated the hierarchical structure of the Chart of Accounts. RF algorithm was trained for each level of the hierarchy in order to show the improvement of the performances and to motivate a deeper study on hierarchical classification algorithms. Figure 8 shows the accuracy rates for the different levels of the target variable. For both datasets of the two accounting firms, the accuracy rates (in terms of recall) improve at lower levels. This is because the account codes associated to lower levels of the hierarchy are more generic and easier to predict. This aspect motivates the development of an algorithm which guides the classification output through the hierarchical structure of the account codes.

Table 3. Mean cross validation precision and f1-score of the classifiers which predict VAT codes (sent invoices). Standard deviation is reported in parentheses.

	Dataset 1		**Dataset 2**	
	Precision	F1-score	Precision	F1-score
RF + BoW	99.2% (±0.2%)	99.2% (±0.3%)	99.7% (±0.1%)	99.6% (±0.2%)
RF + Word2Vec	98.8% (±0.2%)	98.7% (±0.3%)	99.5% (±0.1%)	99.5% (±0.2%)
AdaBoost + Bow	99.2% (±0.3%)	99.1% (±0.4%)	99.7% (±0.1%)	99.6% (±0.2%)
AdaBoost + Word2Vec	99.1% (±0.3%)	99.0% (±0.4%)	99.6% (±0.2%)	99.5% (±0.2%)
MlP + BoW	99.3% (±0.3%)	99.2% (±0.4%)	99.7% (±0.2%)	99.6% (±0.3%)
MlP + W2V	99.1% (±0.3%)	99.0% (±0.4%)	99.6% (±0.1%)	99.6% (±0.2%)

Table 4. Mean cross validation precision and f1-score of the classifiers which predict VAT codes (received invoices). Standard deviation is reported in parentheses.

	Dataset 1		**Dataset 2**	
	Precision	F1-score	Precision	F1-score
RF + BoW	93.4% (±0.8%)	93.7% (±1.0%)	92.2% (±0.9%)	91.8% (±1.1%)
RF + Word2Vec	94.8% (±0.5%)	94.6% (±0.5%)	94.3% (±0.6%)	94.2% (±0.6%)
AdaBoost + Bow	92.6% (±0.6%)	92.8% (±0.8%)	89.9% (±1.0%)	90.5% (±0.8%)
AdaBoost + Word2Vec	94.9% (±0.5%)	94.9% (±0.6%)	94.2% (±0.5%)	94.1% (±0.5%)
MlP + BoW	92.8% (±1.0%)	97.2% (±0.8%)	91.4% (±0.6%)	91.8% (±0.5%)
MlP + W2V	95.4% (±0.3%)	95.0% (±0.4%)	94.3% (±0.5%)	93.8% (±0.6%)

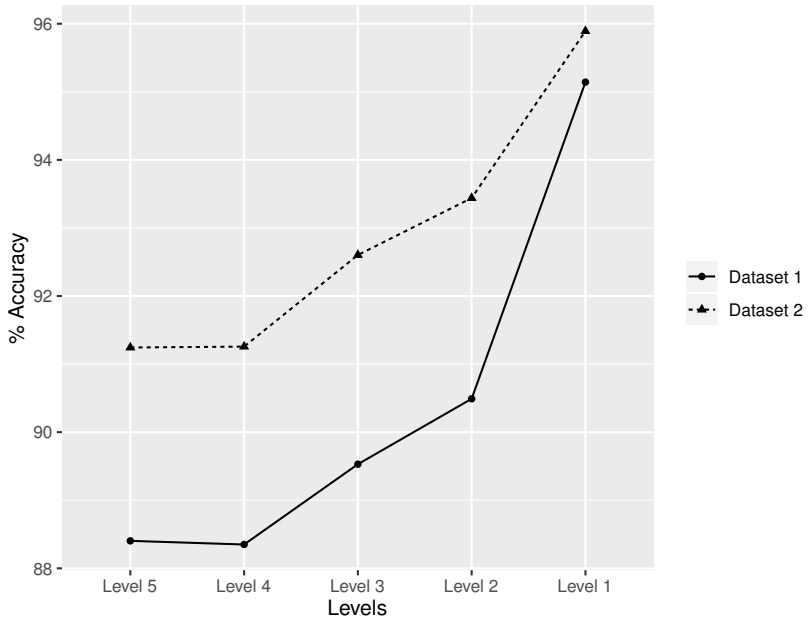

Figure 8. Accuracy computed at different levels of the account codes (target variable) for received invoices datasets. Model performances improve at lower levels since the target variable is more generic and the number of distinct categories is lower.

6. Conclusions

The part of the bookkeeping process which produces a journal entry is often time consuming. Classifying the transactions reported in invoices into specific codes can be translated into a machine learning classification task, whose predictions can facilitate the work of accountants.

This work introduces a methodological proposal to handle accounting data. First, the problem of the reconstruction of the training set is tackled, using a heuristic to solve the knapsack problem. Then, different classification approaches on two real datasets are tested; higher accuracy is achieved using the classification algorithm Random Forests combined with the pre-processing technique Word2Vec for the prediction of the account codes. Concerning VAT codes, the classifier AdaBoost combined with Word2Vec performs slightly better. The results show that account codes are the most problematic target variable, due to the high number of distinct labels which can be employed in the classification.

A possible workaround to solve this problem is to consider a hierarchical classification framework which reflects the structure of the Chart of Accounts: the target variable is structured in a predefined hierarchy which takes the form of a rooted tree. This is also motivated by the results obtained for the classification at different levels of our target variable. The application of hierarchical classification algorithm [30,31] can be considered a possible method to improve the results. Since the two target variables predicted in our problem, VAT codes and account codes, are dependent on each other, another possible approach to improve the accuracy of the account codes can be the application of a multi-label classification algorithm. In this case, we would not have two separately classifiers for the two target variable, but a unique classifier which exploits the information extracted from text as well as other categorical variables to predict VAT codes and account codes together.

The creation of the general ledger entries starting from an xml invoice seems to be a promising field in which machine learning models can be adopted to automate this repetitive and monotonous part of the bookkeeping process. These preliminary results support a deeper study on more advanced techniques which can solve some of the existent problems and improve the accuracy of the classifiers. Further ideas of research will consider the implementation of different classification models coupled with key performance indicators related to computational efficiency and predictive capability.

Author Contributions: This paper was written by C.B., including the contributions made in terms of methodological and computational approaches. A.R. and R.V. introduced the real case and business knowledge to solve the problem at hand. Supervision, S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Datev.it S.p.a. (research contract signed with the Department of Political and Social Sciences—University of Pavia).

Acknowledgments: We thank Datev.it S.p.a. for the data provided and for the PhD scholarship.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Frey, C.B.; Osborne, M.A. The future of employment: How susceptible are jobs to computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280. [CrossRef]
2. Tekbas, I.; Nonwoven, K. The Profession of the digital age: Accounting Engineering. In *IFAC Proceedings Volumes, Project: The Theory of Accounting, Enginnering*; Elsevier: Amsterdam, The Netherlands, 2018.
3. Gulin, D.; Hladika, M.; Valenta, I. Digitalization and the Challenges for the Accounting Profession. In Proceedings of the 2019 ENTRENOVA Conference, Rovinj, Croatia, 12–14 September 2019.
4. ICAEW. Artificial Intelligence and the Future of Accountancy; Technical Report. Available online: https://www.icaew.com/technical/technology/artificial-intelligence/artificial-intelligence-the-future-of-accountancy (accessed on 29 November 2020).
5. Tang, Y.Y.; Suen, C.Y.; De Yan, C.; Cheriet, M. Financial document processing based on staff line and description language. *IEEE Trans. Syst. Man Cybern.* **1995**, *25*, 738–754. [CrossRef]
6. Cesarini, F.; Gori, M.; Marinai, S.; Soda, G. INFORMys: A flexible invoice-like form-reader system. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 730–745. [CrossRef]

7. Holt, X.; Chisholm, A. Extracting structured data from invoices. In Proceedings of the Australasian Language Technology Association Workshop 2018, Dunedin, New Zealand, 10–12 December 2018; pp. 53–59.
8. Wang, Y.; Gui, G.; Zhao, N.; Yin, Y.; Huang, H.; Li, Y.; Wang, J.; Yang, J.; Zhang, H. Deep learning for optical character recognition and its application to VAT invoice recognition. In Proceedings of the International Conference in Communications, Signal Processing, and Systems, Dalian, China, 14–16 July 2018; Springer: Berlin, Germany, 2018; pp. 87–95.
9. Palm, R.B.; Winther, O.; Laws, F. Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: New York, NY, USA 2017; Volume 1, pp. 406–413.
10. Schreyer, M.; Sattarov, T.; Borth, D.; Dengel, A.; Reimer, B. Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv* **2017**, arXiv:1709.05254.
11. Zupan, M.; Letinic, S.; Budimir, V. Accounting Journal Reconstruction with Variational Autoencoders and Long Short-term Memory Architecture. 2020. Available online: http://ceur-ws.org/Vol-2646/05-paper.pdf (accessed on 29 November 2020).
12. Schultz, M.; Tropmann-Frick, M. Autoencoder Neural Networks versus External Auditors: Detecting Unusual Journal Entries in Financial Statement Audits. In Proceedings of the 53rd Hawaii International Conference on System Sciences, Maui, HI, USA, 7–10 January 2020.
13. Bengtsson, H.; Jansson, J. Using Classification Algorithms for Smart Suggestions in Accounting Systems. Master's Thesis, Chalmers University of Technology, Gothenburg, Sweden, 2015.
14. Bergdorf, J. Machine Learning and Rule Induction in Invoice Processing: Comparing Machine Learning Methods in Their Ability to Assign Account Codes in the Bookkeeping Process. 2018. Available online: http://www.diva-portal.se/smash/get/diva2:1254853/FULLTEXT01.pdf (accessed on 29 November 2020).
15. Kozanidis, G.; Melachrinoudis, E.; Solomon, M.M. The linear multiple choice knapsack problem with equity constraints. *Int. J. Oper. Res.* **2005**, *1*, 52–73. [CrossRef]
16. Pyle, D. *Data Preparation for Data Mining*; Morgan Kaufmann: Burlington, MA, USA, 1999.
17. Khan, A.; Baharudin, B.; Lee, L.H.; Khan, K. A review of machine learning algorithms for text-documents classification. *J. Adv. Inf. Technol.* **2010**, *1*, 4–20.
18. Joachims, T. *Learning to Classify Text Using Support Vector Machines*; Springer Science & Business Media: Berlin, Germany, 2002; Volume 668.
19. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR)* **2002**, *34*, 1–47. [CrossRef]
20. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
21. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
22. Kim, S.B.; Han, K.S.; Rim, H.C.; Myaeng, S.H. Some effective techniques for naive bayes text classification. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1457–1466.
23. Wang, Z.; He, Y.; Jiang, M. A comparison among three neural networks for text classification. In Proceedings of the 2006 8th International Conference on Signal Processing, Beijing, China, 16–20 November 2006; IEEE: New York, NY, USA, 2006; Volume 3.
24. Wang, Z.Q.; Sun, X.; Zhang, D.X.; Li, X. An optimal SVM-based text classification algorithm. In Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 13–16 August 2006; IEEE: New York, NY, USA, 2006; pp. 1378–1381.
25. Kanakaraj, M.; Guddeti, R.M.R. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, CA, USA, 7–9 February 2015; IEEE: New York, NY, USA, 2015; pp. 169–170.
26. Colas, F.; Brazdil, P. Comparison of SVM and some older classification algorithms in text classification tasks. In Proceedings of the IFIP International Conference on Artificial Intelligence in Theory and Practice, Santiago, Chile, 21–24 August 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 169–178.
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-class adaboost. *Stat. Its Interface* **2009**, *2*, 349–360. [CrossRef]

29. Delashmit, W.H.; Manry, M.T. Recent developments in multilayer perceptron neural networks. In Proceedings of the Seventh Annual Memphis Area Engineering and Science Conference, MAESC, Memphis, TN, USA, 11 May 2005.
30. Silla, C.N.; Freitas, A.A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **2011**, *22*, 31–72. [CrossRef]
31. Cesa-Bianchi, N.; Gentile, C.; Zaniboni, L. Incremental algorithms for hierarchical classification. *J. Mach. Learn. Res.* **2006**, *7*, 31–54.