Towards Intelligent Processing of Electronic Invoices: The General Framework and Case Study of Short Text Deep Learning in Brazil

Diego Santos Kieckbusch(✉), Geraldo Pereira Rocha Filho, Vinicius Di Oliveira, and Li Weigang

TransLab/CIC, University of Brasilia, Brasilia, DF 70910-900, Brazil
diego.kieckbusch@aluno.unb.br, {geraldof,weigang}@unb.br

Abstract. An electronic invoice (E-invoice) is a kind of document that records the transactions of goods or services and then stores and exchanges them electronically. E-invoice is an emerging practice and presents a valuable source of information for many areas. Dealing with these invoices is usually a very challenging task. Information reported is often incomplete or presents mistakes. Before any meaningful treatment of these invoices, it is necessary to evaluate the product represented in each file. This research puts forward a conceptual framework to explain how to apply machine learning technology to extract meaningful information from invoices at different levels of aggregation. Related work in the field is contextualized within a given framework. A study case based on real data from Electronic invoice (NF-e) and Electronic Consumer Invoice (NFC-e) documents in Brazil, related to B2B and retail transactions. We compared traditional term frequency models with the Convolutions sentence classification models. Our experiments show that even if invoice text descriptions are short and there are a lot of errors and typos, simple term frequency models can achieve high baseline results on product code assignment.

Keywords: CNN · Electronic invoice · Short-text classification

1 Introduction

The purpose of an invoice is to record the transactions of goods and services between buyers and sellers. Invoicing is very important in daily commercial and financial operations. It is also a rich source of information for financial analysis, fraud detection [10], value chain analysis, product tracking, and hazard alarms [3]. Even though local regulations may differ, the overall structure of these documents is similar in many countries. In Brazil, this process started in 2008, first Electronic invoice (NF-e), then Electronic Consumer Invoice (NFC-e), which is a nationwide B2B transaction reporting integrated system. Similar measures have also been taken in Italy [2] and China [27,30]. Invoices exist in various forms, from physical documents to semi-structured data, and each form has its challenges. Knowing how to deal with this type of document can bring many valuable applications. This scenario leads us to the problem of how to extract meaningful information from invoice documents.

© Springer Nature Switzerland AG 2023
M. Marchiori et al. (Eds.): WEBIST 2020/2021, LNBIP 469, pp. 74–92, 2023.
https://doi.org/10.1007/978-3-031-24197-0_5

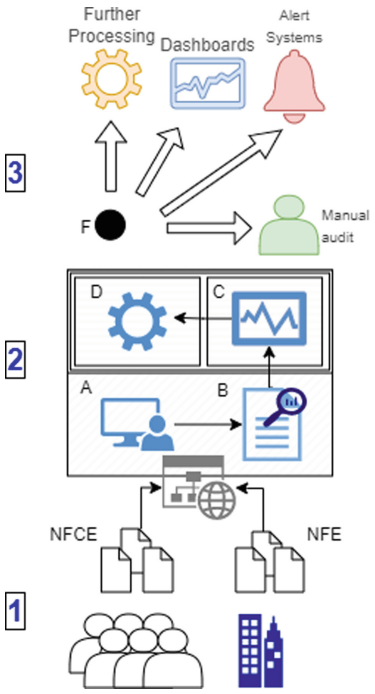

Fig. 1. Invoice processing in a nutshell.

Figure 1 presents an overview of invoice processing in three phases. At the bottom, Label 1, we have both retail and larger companies that issue invoices as part of their day-to-day activities. In Fig. 1, these invoices are represented by the NF-e and Electronic Consumer Invoice (NFC-e) documents, the Brazilian document for retail, and B2B invoices. These invoices are reported to a centralized system through web applications. Once reported, these invoices are processed to aid in a particular task. This process is depicted in Label 2, as an analyst selects relevant data to the core problem(A), data is then cleaned(B), explored(C), and used as the input to train a task-specific model, Label D. The trained model and analyzed data set (F) is then used as input for other applications and to aid manual auditing of other invoices, Label 3.

Due to a large amount of data, our first answer was to look at the problem through the lens of machine learning. Invoices are semi-structured documents with both tabular data and short text. Our previous work [11] focused on using the product description field to predict the universal code of each product. By further reading the existing literature on invoice processing, we step back and try to organize different stages, tasks, challenges, and techniques used in the process. We try to organize these topics in a conceptual framework to guide researchers and developers by creating a common thinking landscape to share knowledge and promote discussion.

This framework presents a layered structure for invoice processing. The three levels above represent different levels of abstraction: the more granular product transaction

level, in which invoices are broken down by the products listed, the invoice level, and the issuer level, in which invoices can be grouped to model a business or sector behavior. There is a need to create a structured database of invoice data at the base level. This type of task often involves extracting information from physical documents and user-oriented files such as scanned images of physical documents to a more computer-oriented representation.

We also expand our previous work on product-level invoice classification. In our last work, we presented a model, SCAN-NF, to classify products transactions based on the short-text product description contained in the transaction. We validated our model through a study case on two different Brazilian electronic invoice models: the NF-e and NFC-e models. These models report B2B and retail transactions, respectively. As mentioned by [6], short text processing has some special properties: 1) the contribution of the individual author is small; 2) grammar is generally informal and unstructured, and 3) the sending and receiving of information in real-time and mass; 4) large-scale data is the unbalanced distribution of interesting categories and presents the labeling bottleneck. Compared to other short texts, invoice description is very short, containing only a few words, which usually can not form a complete sentence. This exacerbates the problem of domain-specific vocabulary, abbreviations, and typos because the authors use their own logic.

Some related works on product-level invoice classification are mainly concentrated in China. Their solutions range from using hashing techniques to dealing with an unknown number of features [27,30], semantic expansion trough external knowledge bases [27], classification of paragraph embedding by k-nearest-neighbors [23] to different artificial neural network architectures [26,31]. Semantic expansion is prevalent not only on invoice classification but also on short-text classification [14,24]. These works are not suited for the Brazilian case either due to language differences or reliance on knowledge bases only available in English and Chinese [9]. In the literature, there are gaps in the models suitable for classifying languages other than Chinese.

We focus on the Brazilian electronic invoice model due to its maturity. Standardization of electronic invoices was initiated in 2008 in Brazil and has evolved since then. Currently, every business transaction must report a standardized electronic invoice to a centralized system. Brazil utilizes two types of electronic invoices: Electronic Invoice (NF-e), which records B2B transactions, and Consumer Electronic Invoices (NFC-e), which records retail transactions. Mandatory reports of the NFC-e only began in 2017, and audition processes performed on NF-e documents are not performed in NFC-e data. Manual auditing of these invoices is expensive and time-consuming, especially for NFC-e data, due to a more significant number of issuers and the low quality of reported data. Since tax auditing is a fundamental activity for the Treasury Office, autonomous or semi-autonomous tools for processing large invoice datasets are of great value [15].

While fields are audited for fulfillment and type, there are breaches for exploits and errors. One fundamental vulnerability is in the reported product code, called MERCOSUR Common Nomenclature (NCM), which is a standardized nomenclature for products and services in MERCOSUR. It defines the correct taxation and if the product is eligible for tax exemption. One could miss-classify products to benefit from lower taxation.

As the main contribution of this research, we present both a contextual framework for invoice processing and present a study case on product level classification of invoices based on Brazilian data. We expand the points presented in our last article [11], in which we proposed a system to aid fiscal auditors to recognize product transactions. We present experiments using character-level CNN and support vector machines. Character level representation may be used to tackle typos and abbreviations, such tokens would not be correctly represented when using pre-trained word embedding. Support Vector machines trained over TF-IDF representation act as an example of a term count model. Our case study focuses on invoices in Brazil because the relevant data can be obtained through cooperation with the Treasury Office. Although the case study in this article is aimed at Brazil's data, we have briefly outlined the resources in other languages that could help to process invoices.

This article is organized as follows. In Sect. 2, we give a context framework, which provides the prospect of e-invoice processing. In the third part, we introduce the related work on invoice and short text classification. Section 4 describes the architecture of the SCAN-NF system and classification model. In Sect. 5, we show a case study on real NF-e and NFC-e data. Results are presented in Sect. 6. We present closing remarks and future works in the final section.

2 Contextual Framework

In this section, we present a contextual framework to understand the landscape of invoice processing. The framework is organized in a layered structure, with each layer representing a sequential step in invoice processing. Figure 2 presents a visual representation of the proposed framework. At the base level, there is the data structuring layer.

Although electronic invoices have become more and more popular in recent years, in many cases, useful documents only exist in physical forms or user-oriented digital files, such as document pictures and PDFs. Before processing any meaningful information, we need to extract data from these documents and store it in some semi-structured mode. Related works have shown that computer vision solutions are useful for extracting useful information from physical documents directly [8,17,22,28,28]. These methods can greatly reduce the costs and workload for generating invoice data sets. This task is especially important in auditing, because it is necessary to cross the information reported in invoices with sales records in other systems.

The remaining steps in our framework relate to different levels of abstraction that can be applied to invoice modeling. These steps include product transaction processing, invoice processing and issuer processing. Each level serves as the stepping stone for the next. Product transaction is the first layer of processing, representing each individual product or service transaction represented on every invoice in the data. At this level, we are interested in extracting granular information such as product description, product price, due taxes as well as other task-oriented attributes. The main form of input at this level is the product description. Our work, both in this chapter as well as in our last article, is situated at this level, as we treat product description as a short-text classification problem to predict the correct product code for each transaction. This exemplifies

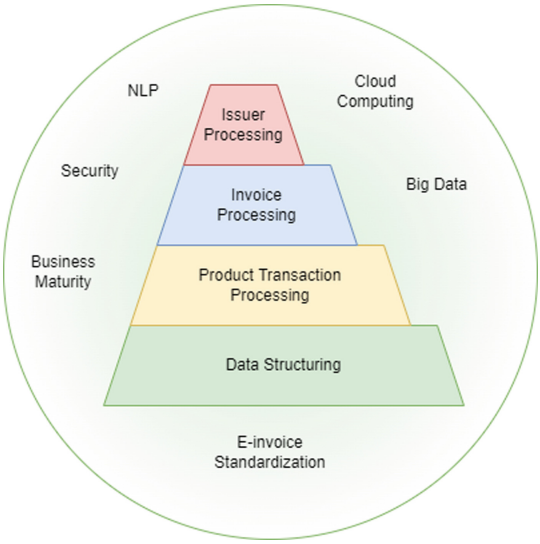

Fig. 2. E-invoice processing framework.

the main concern at this stage: we are interested in creating a good representation for each product transaction in order to produce the input for later tasks. It is much easier to analyze products transactions from a standardized product taxonomy than processing text descriptions [13].

At the invoice processing level, individual product transactions are aggregated and used to represent each invoice in conjunction to other meta-data. It is possible to track the relationship between multiple products in the same invoice. For example, Paalman [16] utilized two-step clustering to track fraudulent invoices. Auto-encoders have also been employed in fraud detection by measuring the distance between the reported text and the expected text produced by the model [20]. At this level, we can also model consumer behavior by utilizing association rules based on common product transactions. Another example is the usage of invoices issued by healthcare centers to extract association rules between commonly used medication [1].

At the higher level of abstraction, the behavior of parties involved in transactions is taken into account. One approach is to utilize previously known troubled issuers as a flag in processing invoices. An example of this kind of procedure is Chang's work [3], in which information about companies involved in violations is used to select and mark invoices to create an alarm system for safe edible oil. Another way to include issuer analysis in invoice processing is through graph analysis. It would be possible to model an oriented graph, each node representing an issuer with invoices being used to create the links between issuers. From this structure, it would be possible to look for communities, cycles, and other graph-oriented sub-structures and correlate them to real-world issues. At the time of this work, we have not been able to find works that model invoice processing utilizing graphs. We hope to address this tackle this problem in the future.

2.1 Larger Context

Invoice processing is also related to other concerns that are not directly related to extracting information from invoice documents. Due to a large amount of data, invoice-based systems require Big data architecture [5]. This may lead to solutions in distributed computing paradigm as storing and processing are more feasible in clusters than in single machines. The adoption of e-invoicing from the get-go is also a key factor, as it streamlines the data structuring layer, doing away with the need of using expensive image processing techniques to create digital representations of invoices. A maturity model for e-invoicing from the business perspective was provided by Cuylen [4].

3 Short Text Processing

In this section, we take a closer look at works related to invoice product transactions. We model product transaction processing as a short text classification problem, in which the main input is the short text snippet present in transaction descriptions. We present related work on traditional term-count-based methods and Neural Networks, as well as other product transaction processing models.

3.1 Traditional Methods

A Common representation technique in text classification is to create a term frequency vector to represent each document. Matrix factorization techniques can then be applied to engineer features in a smaller dimensional space. Due to the low word count in short text documents, there is lower co-occurrence of terms across the document-term matrix, which may hinder matrix factorization methods.

A possible solution to this problem is to directly address the brevity of short text by expanding on it. Document expansion utilizes the original text as the query to a secondary system. This system is then responsible to return similar documents to the query provided. The representation of the original text document is then calculated based on the collection of returned documents. This expansion can also be done term-wise by using lexical databases to extract terms with a strong semantic relationship to important terms in the documents. Early works attempted to address this problem by expanding available information through auxiliary databases [19,25]. Phan [18] proposed a framework for short text classification that used an external "universal dataset" to discover a set of hidden topics through Latent Semantic Analysis. Other work proposed to utilize web search engines as the query system [19].

For several reasons, document extension technology may not be suitable for invoice classification. Primarily there is an overhead mainly in processing and communication. The query of auxiliary documents increases the processing cost, which requires a good similarity function to identify related documents, and processing more documents than the initial data set also increases the cost. Communication with the auxiliary system may also bring bottleneck to the system. Finally, there is the additional cost of setting up and maintaining the auxiliary system in languages other than English. This is particularly important because these resources may not be easily available.

3.2 Neural Network Based Methods

Artificial neural networks (ANN) have become popular in many data-driven methods, because they allow better representation learning of problems with high dimensions (such as text and image classification). In Short-text classification, both Convolutions Neural networks (CNN) and Recurrent Neural Networks (RNN) have been used to create sentence embedding that could be classified. The general method follows two main steps: each item in the sentence is replaced by a vector with a fixed length, and input into the neural networks. These vectors can either be randomly initialized or trained independently to solve a self-supervised problem. These vectors generally incorporate underlying semantics of the corpus they were trained upon, demonstrated by the composition of vectors with similar meanings: the distance of the vector for the terms "King" and "man" is very similar to the distance between "Queen" and "woman".

The neural network will then perform sequential transformations of the input vectors representing a final output vector that will represent the whole input sentence. The classification itself is done on the final layer in which the learned vector is used to generate the classification label. The architecture proposed by Kim [12] serves as the basis for most CNN-based solutions. In CNN models, sliding windows of different sizes move through the input vectors learning to filter sub-structures throughout the training process. One common problem in short text is typos and abbreviations. Because of the training method of word vectors, typos and abbreviations are completely different from the original term. Zhang [29] utilized a 12-layer CNN to learn features from character embedding. On Character level CNN models, terms are created by forming sets of characters. This solves the problem of lack of vocabulary, misspellings and abbreviations, because words with similar structures will have similar embedding vectors. Wang [24] combined the word and character CNN with knowledge extension to classify short texts. The model used knowledge bases to return related concepts and included them in the text before the embedding layer. Knowledge bases included: YAGO, Probase, FreeBase, and DBpedia. A character-based CNN was used in parallel to the word concept CNN. Representations learned by both networks were concatenated before the final fully connected layer.

Naseem [14] proposed an expanded meta-embedding approach for sentiment analysis of short-text that combined features provided by word embedding, part of speech tagging, and sentiment lexicons. The resulting compound vector was fed to a Bidirectional long-short term memory (BiLSTM) with an attention network. The rationale behind the choice for an expanded meta-embedding is that language is a complex system, and each vector provides only a limited understanding of the language.

3.3 Invoice Classification

Invoice classification techniques have ranged from traditional count-based methods to neural-based architectures. In 2017, Chinese invoice data was made public for Chinese researchers, which motivated research in the area. This leads to the prevalence of works dealing with the Chinese invoice system.

Some works aimed to address the data sparsity problem by utilizing a hash trick for dimensionality reduction [30]. Yue [27] performed semantic expansion of features

through external knowledge bases before using the hash trick for dimensionality reduction. Tang [23] utilized paragraph embedding to create a reduced representation and then applied the K-NN classifier. Yu [26] utilized a parallel RNN-CNN architecture, with the resulting vectors being combined in a fully connected layer. Zhu [31] combined features selected through filtering with representation learned through the LSTM model.

Different from most western languages, in western languages, text is expressed through words with spaces as separators, while in Chinese, there is no separator and no clear word boundary. Words are constructed based on the context. Chinese invoice classification words leaned towards RNN-based architectures in a way to mitigate errors produced in the word segmentation step.

Chinese works aside, Paalman et al. [16] worked on the reduction of feature space through 2-step clustering. The first step was to reduce the number of terms through filtering and then cluster the distributed semantic vector provided by different pre-trained word embeddings. This method was compared to traditional representation schemes and matrix factorization techniques. In the experiments, simple term frequency and TF-IDF normalization performed better than the models of Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

3.4 Discussion of Related Work

Term count-based methods mainly address short-text processing through filtering and knowledge expansion. The problem with filtering is that there is information loss in a context where information is already poor. Semantic expansion is mainly done through knowledge bases. Communication with knowledge base becomes the bottleneck of the system, and because of the amount of invoice data, it is not suitable for invoice processing. Furthermore, knowledge bases may not be available in languages other than English and Chinese [9].

The limitation of pre-trained word embeddings comes down to vocabulary coverage and word sense [7]. These are significant to invoice classification. Words in invoices are often misspelled and abbreviated. Also, taxpayers often mix words of multiple languages depending on the kind of product being reported. Finally, invoices have little or no context to eliminate the ambiguity of word meaning.

Most invoice classification models did not utilize traditional ANN. The research of Yu and others [26] is the only one to combine both CNN and BiLSTM. However, CNN and BiLSTM were used in parallel over different fields. Zhu [31] combined a LSTM network with traditional methods using filtered features. While effective for the Chinese language, these architectures are not suitable for Brazilian invoice model. We propose a CNN-based model to solve these shortcomings, which does not depend on pre-trained word embedding and external knowledge base.

There is a gap between the general task of invoice text classification and similar tasks of Sentence and Short-Text classification in Natural Language Processing (NLP). Sentences are modeled as the components of a larger document. It is important to understand the context before and after the sentence, as well as the processing of the sentence itself. Even though we may draw parallels of the sentence role in a document being similar to that of an individual product transaction in an invoice, there is little meaning to

the product transaction order in an invoice. Invoices is just a simple report, and there is no potential intention to tell anything beyond the product transactions itself. The words on the invoices often doesn't even have complete sentences.

Another thing worth studying about invoice classification is that it's a very different use case from the traditional short text. The main object of study of short-text works addresses news snippets, review comments, and tweets. The task is generally either to identify a general very broad category, such as news topics or to identify sentiment-related attributes from the text. These tasks require a deeper understating of the text and need to address different challenges from Invoice Processing. In sentiment analysis, it is necessary to take into account not only sarcasm but negations, conjunctions, and adverbs as these change the meaning of the sentence.

We believe that although the invoice product descriptions is similar to other short text problems, the classification of invoice text is obviously different and solutions may be different. The NLP field has been moving towards language understanding through large self-supervised models such as Transformer models. The task of classifying an invoice is less dependent on understanding the meaning of the text fragments, but more dependent on finding key terms that allow us to assign the correct code. The problem then becomes the large shifting vocabulary used by issuers to describe their products.

4 Architecture of SCAN-NF

In this section, we present an overview of the architecture of the SCAN-NF system and inner models, Fig. 3. The system's goal is to assign the proper NCM product code to each product transaction based on the product description. The labeled transaction is then used as inputs for other analyses by Tax Auditors and Specialists. The system works in two phases: a training phase and a prediction phase. During the training phase, the system is fed audited data from the tax office server of the Department of Economy of the Federal District (SEFAZ) in Brasilia to train a supervised model. Two models are trained, one for the classification of NF-e Documents and another for NFC-e Documents. After training, these models are used on new data during the prediction phase.

The system works as follows: Data is extracted from the tax office server (Label 1 in Fig. 3). Product description and corresponding NCM code for each product in each invoice are then extracted (Label 2 in Fig. 3). Text is then cleaned from irregularities (Label 3 in Fig. 3). A training dataset is constructed by balancing target classes samples and dropping duplicates (Label 4 in Fig. 3). The training set is then fed to a CNN model that learns to classify product descriptions (Label 5 in Fig. 3). Outputs at the training phase of the system are used to validate models before being put into production (Label 6 in Fig. 3). During the Prediction Phase, trained models are utilized to classify new data. These datasets may be composed of invoices issued by a suspected party or a large, broad dataset used for exploratory analysis (Label 7 in Fig. 3). Models trained in the training phase are then employed for the task at hand (Label 8 in Fig. 3). The final output of the model is the classified set of products inputs (Label 9 in Fig. 3). This set of classified product transactions is then used in manual auditing by tax auditors (Label 10 in Fig. 3).

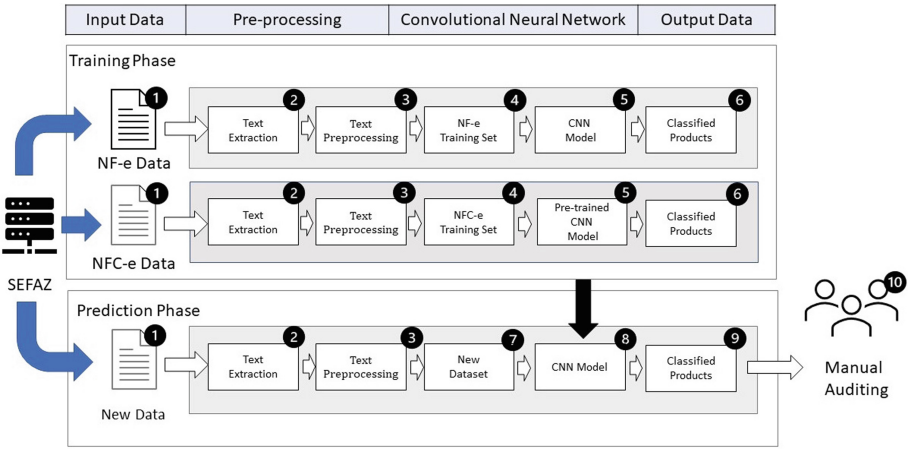

Fig. 3. Architecture of SCAN-NF. Extracted from [11].

The system is intended to aid tax auditors in auditioning invoices issued by already suspicious parties to pinpoint inconsistencies and irregularities. Currently, NFC-e documents are not audited due to the amount of data, a more significant number of issuers, and the nature of the data. Our solution helps auditors pinpoint inconsistencies in documents reported by an already suspicious party and allows for the automatic processing of more data. We hope that this solution will improve the productivity of tax auditors regarding NF-e processing and be the first step towards NFC-e processing.

There are different possibilities for the classification model used in the system. The sentence classification model proposed by Kim [12] can be used as a single multi-label classification model. However, due to the high number of possible NCM codes and high invoice data, we propose an ensemble model built from binary classifiers. Binary classifiers trained on individual classes can be pre-trained, stored, and then combined in multi-label classifiers on demand. This allows individual models to be updated and added without re-training other models.

Figure 4 presents architecture used in single models. The input layer takes the indexed word tokens. Each word index is replaced by a randomly initiated word vector representation in the embedding layer. The resulting vector is then reshaped to fit one-dimensional convolutions layers. Each convolution layer applies different sized filters to the encoded sentence. Max pooling is applied to the learned filters to extract the most useful features. Each convolution is applied in parallel, and they are then concatenated in a single vector, flattened, and fed to a Fully connected layer that will output the final classification. Soft-max was used as the activation function of the model, with the loss being determined by the categorical cross-entropy function.

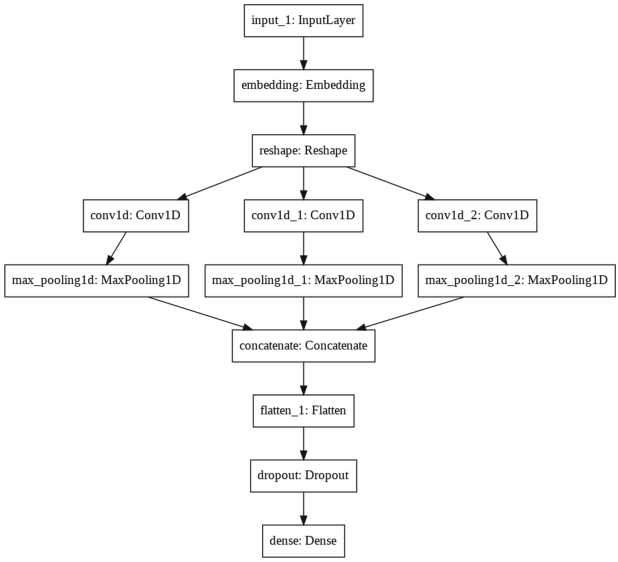

Fig. 4. Flowchart of single CNN word based model. Extracted from [11].

4.1 NF-e and NFC-e

The NF-e is the Brazilian national electronic fiscal document, created to substitute physical invoices, providing judicial validity to the transaction and real-time tracking for the tax office [21]. It contains detailed information about invoice identification, issuer identification, recipient identification, product, transportation, tax information, and total values. In our work, we utilize data present in product transactions, namely product description and NCM code. Data regarding issuer and recipient is kept hidden. NFC-e is a simplified version of the NFC-e used in retail services.

There are validations rules for the NCM field in the NF-e manual [21]. According to the experts engaged in tax audit and the schedule published in NF-e annual, although the verification procedures of NF-e documents have been implemented, NFC-e documents have not planned these verification procedures in the next few years. This leads to poor data quality.

5 Case Study of Brazilian E-Invoices

To validate our model, we conducted a case study based on real NFC-e and NF-e documents from SEFAZ. Data were separated into training and test sets, and different models were trained. Models were validated through cross-validation. Hyper-parameter optimization was conducted based on the average performance through all folders of cross-validations.

5.1 Dataset

In our experiments, we utilized data provided by the estate tax office of SEFAZ. Data provided included both NFC-e and NF-e documents. NF-e data consisted of invoices for cosmetics. NFC-e data consisted of a larger dataset of products from multiple sectors. We selected NCM codes present in the NF-e dataset and created a curated dataset with balanced classes. Due to disparity in market share, preserving product frequency would bias the models toward larger issuers and the most popular products. This could lead models to better classify invoices from large companies or learn their representation as to the norm. Our design decision was to drop duplicate product descriptions for each target class. While there is a significant vocabulary overlap between NF-e and NFC-e documents regarding NF-e data, NFC-e presents a much more vast vocabulary. Table 1 presents detailed information on the number of samples used in the experiment.

Table 1. Number of samples and datasets used in experiments. Extracted from [11].

	NF-E	NFC-E
Number of raw product samples	198882	99637515
Number of samples in balanced dataset	36234	49536
Number of balanced classes	18	18
Vocabulary Size	3646	15312
Shared Terms	2342	

5.2 Baseline Models

Besides the single and ensemble models presented in the SCAN-NF section, We utilize other classification models to create a baseline of comparison to our proposed model. We utilize SVM trained on the TF-IDF representation and Convolutional Neural Network trained on character representation.

SVM represents frequentist models and challenges the idea that traditional term count-based models fail at short text classification due to a sparse attribute matrix and low term count. We argue that while dimensionality reduction is particularly difficult due to low term co-occurrence, each product class will be defined by a handful of highly important terms. We expect these models to perform similarly to our CNN approach. Character-based Neural network is supposed to address typos and abbreviations.

5.3 Experiments

We conducted two experiments with different model sets. In the first one, we compare the single model and the ensemble model approaches. The single model is composed of a single CNN model trained on multi-label classification. The ensemble model is composed of a set of binary models. Each binary model is trained on a distinct class in a binary classification problem. The ensemble model takes the list of binary models

and is then fine-tuned as a multi-label classification problem. Callbacks are set to stop training based on validation error loss.

In the second experiment, we investigated models based on different representations. We trained a character-based convolutions neural network and an SVM classifier based on the TF-IDF representation of text. This experiment aims to address whether or not the points made by related work on the effectiveness of these representations hold for Invoice classification. We expect character-based representation to have a higher complexity than a word-based model due to the need to construct words from the ground up. We expect the TF-IDF-based SVM classifier to perform significantly worse than the CNN models based on related work on both invoice and short-text processing.

Data were separated into training and test sets. We utilized the validation accuracy score to set an early stop on the training of the CNN models. Hyper-parameter optimization was conducted based on the average performance through all folders of cross-validations.

5.4 Metrics

We evaluate models based on the following metrics: accuracy, precision, recall, and top k Accuracy. Metrics are calculated based on True Positives, True Negatives, False Negatives, and False Positives.

Accuracy is given by the rate of correct predictions overall predictions: $(TP + TN)/(TP + TN + FP + FN)$. Top k Accuracy represents how often the correct answer will be in the top k outputs of the model. Accuracy is useful for getting an overall idea of model performance. In unbalanced datasets, recall and precision can paint a better picture of how the model behaves.

The recall represents the recovery rate of positive samples and is given by $TP/(TP + FN)$. Precision evaluates the correct set of retrieved samples and is given by $TP/(TP+FP)$. We utilize the F1-score, the harmonic mean of precision and recall, to get a balanced assessment of model performance on imbalanced classification.

In our experiments, we first set up a CNN architecture. We defined hyper-parameters through optimization using the hyper-opt library. Table 2 presents the parameters and values used in optimization, final parameters are highlighted in bold.

6 Results

In this section, we present the results of the experiments. We separate reports between the two experiments and datasets.

6.1 Single vs Ensemble CNN Approach

Figure 5 presents single and ensemble model performance on both the NF-e dataset e NFC-e datasets. We present results side by side. The key points of interest are that while both datasets model presented little deviation in the accuracy, there was a larger gap between precision and recall. In both datasets, the single model presented better

Table 2. Parameters used in CNN models optimization.

Parameter	Word CNN values	CHAR CNN values
Number of filters on 1D convolution #1	{0,100,200,300, 400,500,600}	{0,100,200,300, 400,500,**600**}
1D convolution kernel size #1	{**3**,5,7,9}	{**3**,5,7,9}
Number of filters on 1D convolution #2	{**50**,100,200,300, 400,500,600}	{50,100,200,300,400,500,**600**}
1D convolution kernel size #2	{3,5,7,**9**}	{3,5,7,**9**}
Number of filters on 1D convolution #3	{0,100,200,300, 400,500,**600**}	{0,100,200,300,400,500,**600**}
1D convolution kernel size #3	{**3**,5,7,9}	{3,5,**7**,9}
Dropout rate	[0, **0.29**, 0.5]	[0, **0.26**, 0.5]
Optimizer	{dam, Adagrad, Adadelta, **Nadam**}	{dam, **Adagrad**,Adadelta, Nadam}

recall at the cost of precision when compared to the ensemble model. We can see that while model accuracy deviated sightly, differences in recall and precision were more evident.

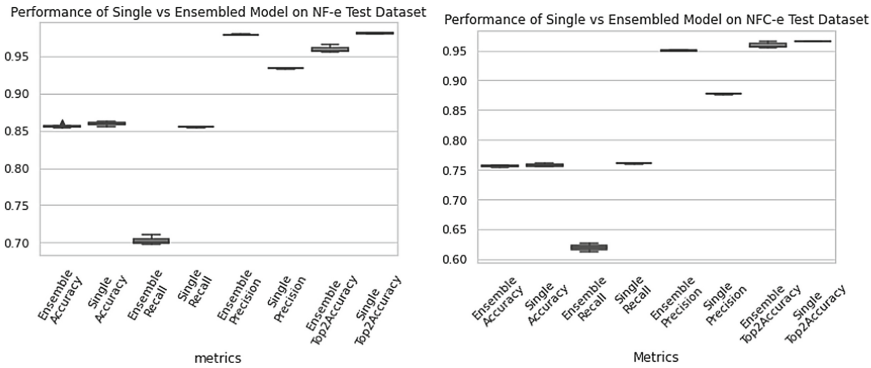

Fig. 5. Results of Experiment 1: single and ensemble models on NF-e and NFC-e datasets. Adapted from [11].

Singular models and binary models were trained through 5 epochs, while the fine tune of the ensemble model was done through 12 epochs. Each epoch took 4sec/10.000 samples to be performed. In practice, the ensemble model takes 20 times longer to be trained than the single model due to the training of binary models and fine-tuning of the ensemble model.

Individual class performance of the ensemble model is shown in Fig. 6. Due to the unbalanced nature of the problem, all classes presented high accuracy scores, scoring

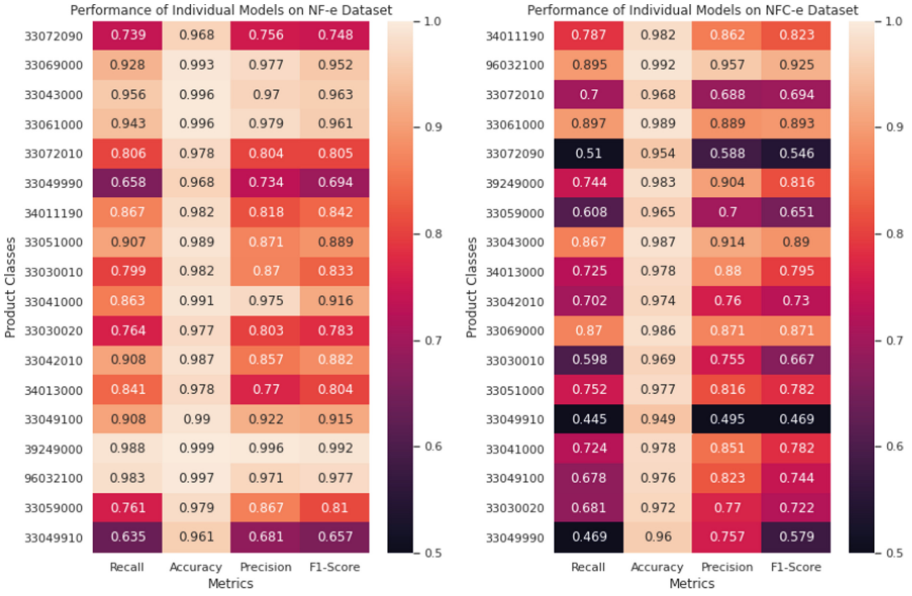

Fig. 6. Individual binary model performance on NF-e and NFC-e datasets. Extracted from [11].

higher than 96%. Of all models, 15 had an F1 score higher than 0.8, and 7 had an F1 score above 0.9. This signalizes that some classes are more challenging to predict than others, and some classification models are less trustworthy. Overall, there was a balance between recall and precision.

6.2 Experiment 2: Comparison of Models with Different Representations

Table 3 presents the mean accuracy and standard deviation of each based on ten runs of training and testing of each optimized model on both datasets. The character-based CNN performed very similarly to Word CNN, with a small trade-off between datasets. The character-based CNN performed better on the more unstructured retail invoices of the NFC-e dataset. The SVM model performed worse than the other models on both datasets. The Char took 13 epochs to train, significantly more than the single word CNN model.

Table 3. Accuracy metric of models on Experiment 2.

Accuracy	Word CNN	Char CNN	TF-IDF SVM
NF-e	0.869 ± 0,001	0.865 ± 0,001	0.776 ± 0,001
NFC-e	0.779 ± 0,001	0.784 ± 0,001	0.661 ± 0,001

6.3 Comparison of Approaches

From both experiments, it is clear that NFC-e product classification is a more complex problem than NF-e classification. Results also varied between different product classes. Regarding the comparison between single and ensemble approaches to word CNN models, we can see a trade-off between recall and precision, with the ensemble model presenting higher precision at the cost of the recall. This indicates that one approach may overcome the other based on the particular task. The single model will return a higher rate of classes of interest but may require more effort in the manual audit due to filtering out false positives. Models consistently achieved around 95% top2 accuracy on both datasets. This means that models can be used as recommendation systems to classify product descriptions.

There are also differences in the maintainability of approaches. The ensemble approach allows individual models to be updated without the need for all models to be updated. This also impacts the system's scalability, as additional classes can be added to the model without retraining the whole model at each addition.

Regarding text representation, word-based and character-based convolutional neural networks presented similar results. While the character-based model performed better on the NFC-e dataset, it is not clear if this resulted from handling typos and abbreviations. We could raise the question that the different results between word and character-based were the trade-off of handling typos at the cost of having to build word filters from the sum of character filters. In future work, this property of modeling typos can be better measured by introducing typos and abbreviations in a controlled dataset. Overall, CNN models managed to map product descriptions to the corresponding NCM code, while the SVM model struggled in both classes. This is in line with related work on short text processing findings.

7 Conclusion and Future Work

This work presented a general framework for invoice classification and expanded our previous work on invoice classification through a study case on Brazilian electronic invoices. Our experiments confirmed previous works on short-text classification, as the term-count model performed worse than text vector models. In our experimental datasets, both word and character-based CNN managed to map product descriptions to product code.

As a summary of this work, the main contributions include: 1) review the literature from the principle research and systems related to the studies of electronic invoices; 2) identify the characteristics and differences between short text processing and electronic invoice processing, especially using NCM code; 3) use machine learning to establish conceptual framework and SCAN-NF system for invoice classification; 4) experiments and analysis of NF-e and NFC-e data sets by SCAN-NF. Even though the invoice classification models are developed for E-invoicing in Brazil, it is easily extended for other countries by some reasonable adjusting.

We hope to improve the presented model by inserting it into real-world applications that can aid tax auditors, researchers, and public administrators in decision-making and day-to-day operations. Following our framework, the next step is to utilize the output

of the studied methods to engineer invoice-level attributes. One such attribute is the expected tax return for misreported invoices based on the expected tax return of individual product transactions.

On the computational side, we will focus on transfer learning. Transformers have emerged as the go-to method for transfer learning in NLP. We will focus on comparing the performance of models trained on the representation provided by pre-trained transformers and previously studied models and the need to fine-tune existing models.

Acknowledgements. This work has been partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under grant number 309545/2021-8. Thanks to Mr. Sergio Neto and other colleagues from the Department of Economy of the Federal District in Brasilia.

References

1. Agapito, G., Calabrese, B., Guzzi, P.H., Graziano, S., Cannataro, M.: Association rule mining from large datasets of clinical invoices document. In: Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, pp. 2232–2238 (2019). https://doi.org/10.1109/BIBM47256.2019.8982934
2. Bardelli, C., Rondinelli, A., Vecchio, R., Figini, S.: Automatic electronic invoice classification using machine learning models. Mach. Learn. Knowl. Extr. **2**(4), 617–629 (2020). https://doi.org/10.3390/make2040033, https://www.mdpi.com/2504-4990/2/4/33
3. Chang, W.T., Yeh, Y.P., Wu, H.Y., Lin, Y.F., Dinh, T.S., Lian, I.: An automated alarm system for food safety by using electronic invoices. PLoS ONE **15**(1), e0228035 (2020). https://doi.org/10.1371/journal.pone.0228035
4. Cuylen, A., Kosch, L., Breitner, M.H.: Development of a maturity model for electronic invoice processes. Electron. Mark. **26**(2), 115–127 (2015). https://doi.org/10.1007/s12525-015-0206-x
5. Da Rocha, C.C., et al.: SQL query performance on Hadoop: an analysis focused on large databases of Brazilian electronic invoices. In: ICEIS 2018 - Proceedings of the 20th International Conference on Enterprise Information Systems 1(ICEIS), pp. 29–37 (2018). https://doi.org/10.5220/0006690400290037
6. Enamoto, L., Weigang, L., Filho, G.P.R.: Generic framework for multilingual short text categorization using convolutional neural network. Multimedia Tools Appl. **80**(9), 13475–13490 (2021). https://doi.org/10.1007/s11042-020-10314-9
7. Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C.: Problems with evaluation of word embeddings using word similarity tasks, pp. 30–35 (2016). https://doi.org/10.18653/v1/w16-2506
8. Feng, Y., Jiang, P., Gu, Z., Dai, Y.: Study of recognition of electronic invoice image. In: 2021 IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC, vol. 5, pp. 1582–1586 (2021). https://doi.org/10.1109/ITNEC52019.2021.9586969
9. Grida, M., Soliman, H., Hassan, M.: Short text mining: state of the art and research opportunities. J. Comput. Sci. **15**(10), 1450–1460 (2019). https://doi.org/10.3844/jcssp.2019.1450.1460
10. He, Y., Wang, C., Li, N., Zeng, Z.: Attention and memory-augmented networks for dual-view sequential learning. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 125–134 (2020). https://doi.org/10.1145/3394486.3403055

11. Kieckbusch, D.S., Filho, G.P.R., Oliveira, V.D., Weigang, L.: SCAN-NF: a CNN-based system for the classification of electronic invoices through short-text product description. In: Mayo, F.J.D., Marchiori, M., Filipe, J. (eds.) Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST 2021, 26–28 October 2021, pp. 501–508. SCITEPRESS (2021). https://doi.org/10.5220/0010715200003058
12. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2011), pp. 1746–1751 (2014). https://doi.org/10.3115/v1/d14-1181
13. Marinho, M.C., Di Oliveira, V., Neto, S.A.P.B., Weigang, L., Borges, V.R.P.: Visual analysis of electronic invoices to identify suspicious cases of tax frauds. In: Rocha, Á., Ferrás, C., Méndez Porras, A., Jimenez Delgado, E. (eds.) ICITS 2022. LNNS, vol. 414, pp. 185–195. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96293-7_18
14. Naseem, U., Razzak, I., Musial, K., Imran, M.: Transformer based deep intelligent contextual embedding for Twitter sentiment analysis. Future Gen. Comput. Syst. **113**, 58–69 (2020). https://doi.org/10.1016/j.future.2020.06.050
15. Oliveira, V.D., Chaim, R.M., Weigang, L., Neto, S.A.P.B., Filho, G.P.R.: Towards a smart identification of tax default risk with machine learning. In: Mayo, F.J.D., Marchiori, M., Filipe, J. (eds.) Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST 2021, 26–28 October 2021, pp. 422–429. SCITEPRESS (2021). https://doi.org/10.5220/0010712200003058
16. Paalman, J., Mullick, S., Zervanou, K., Zhang, Y.: Term based semantic clusters for very short text classification. In: International Conference Recent Advances in Natural Language Processing, RANLP, vol. 2019, pp. 878–887 (2019). https://doi.org/10.26615/978-954-452-056-4_102
17. Palm, R.B., Laws, F., Winther, O.: Attend, copy, parse end-to-end information extraction from documents. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, pp. 329–336 (2019). https://doi.org/10.1109/ICDAR.2019.00060, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079851980&doi=10.1109%2FICDAR.2019.00060&partnerID=40&md5=29b092a6c8a3c0caf86779867d63d202
18. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: 2008 Proceeding of the 17th International Conference on World Wide Web, WWW 2008, pp. 91–99 (2008). https://doi.org/10.1145/1367497.1367510
19. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th International Conference on World Wide Web, pp. 377–386 (2006). https://doi.org/10.1145/1135777.1135834
20. Schulte, J., et al.: ELINAC: autoencoder approach for electronic invoices data clustering. Appl. Sci. **12**, 3008 (2022). https://doi.org/10.3390/app12063008
21. SEFAZ: Manual de Orientação do Contribuinte - Padrões Técnicos de Comunicação. ENCAT (2015)
22. Tang, P., et al.: Anomaly detection in electronic invoice systems based on machine learning. Inf. Sci. **535**, 172–186 (2020). https://doi.org/10.1016/j.ins.2020.03.089
23. Tang, X., Zhu, Y., Hu, X., Li, P.: An integrated classification model for massive short texts with few words. In: ACM International Conference Proceeding Series, pp. 14–20 (2019). https://doi.org/10.1145/3366715.3366734
24. Wang, J., Wang, Z., Zhang, D., Yan, J.: Combining knowledge with deep convolutional neural networks for short text classification. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 2915–2921 (2017). https://doi.org/10.24963/ijcai.2017/406
25. Yih, W.T., Meek, C.: Improving similarity measures for short segments of text. In: Proceedings of the National Conference on Artificial Intelligence, vol. 2, pp. 1489–1494 (2007)

26. Yu, J., Qiao, Y., Shu, N., Sun, K., Zhou, S., Yang, J.: Neural network based transaction classification system for chinese transaction behavior analysis. In: Proceedings - 2019 IEEE International Congress on Big Data, BigData Congress 2019 - Part of the 2019 IEEE World Congress on Services, pp. 64–71 (2019). https://doi.org/10.1109/BigDataCongress.2019.00021
27. Yue, Y., Zhang, Y., Hu, X., Li, P.: Extremely short Chinese text classification method based on bidirectional semantic extension. In: Journal of Physics: Conference Series. vol. 1437 (2020). https://doi.org/10.1088/1742-6596/1437/1/012026
28. Zhang, H., Dong, B., Feng, B., Yang, F., Xu, B.: Classification of financial tickets using weakly supervised fine-grained networks. IEEE Access **8**, 129469–129477 (2020). https://doi.org/10.1109/ACCESS.2020.3007528, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089215581&doi=10.1109%2FACCESS.2020.3007528&partnerID=40&md5=9fffb4e8a98ac64be2fa28de21f4e632
29. Zhang, X., LeCun, Y.: Text understanding from scratch (2016). http://arxiv.org/abs/1502.01710
30. Zhou, M., Hu, X., Zhu, Y., Li, P.: A novel classification method for short texts with few words. In: Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019, pp. 861–865 (2019). https://doi.org/10.1109/ITNEC.2019.8729520
31. Zhu, Y., Li, Y., Yue, Y., Qiang, J., Yuan, Y.: A hybrid classification method via character embedding in Chinese short text with few words. IEEE Access **8**, 92120–92128 (2020). https://doi.org/10.1109/ACCESS.2020.2994450