# SCAN-NF: A CNN-based System for the Classification of Electronic Invoices through Short-text Product Description

Diego Santos Kieckbusch[a], Geraldo P. R. Filho[b], Vinicius Di Oliveira[c] and Li Weigang[d]

*Departamento de Ciência da Computação, Universidade de Brasilia, Brasilia, Brazil*

Keywords: Convolutional Neural Network, Invoice Classification, Short-text Classification, Tax Auditing, Few-word Classification.

Abstract: This research presents a Convolutional Neural Network (CNN) based system, named SCAN-NF, to classify Consumer Electronic Invoices (NFC-e) based on product description. Due to how individual issuers submit Consumer Electronic Invoices, processing these invoices is often a challenging task. Information reported is often incomplete or presents mistakes. Before any meaningful processing over these invoices, it is necessary to assess the product represented in each document. SCAN-NF is developed to identify correct products codes in electronic invoices based on short-text product descriptions. Real data from Brazilian NFC-e and NF-e documents related to B2B and retail transactions are used in experiments. Comparing base single model and proposed ensemble model approaches, the evaluation results using recall, precision, and accuracy show the satisfaction of the developed system.

## 1 INTRODUCTION

Invoices document the transaction of goods and services and other business activities. For companies, they are an important source of financial information and a fundamental basis for controlling tax funds. They are also the main source of information on taxation for regulators. Intelligent processing of invoices allows for applications in the context of financial analysis, fraud detection (He et al., 2020), value chain analysis, product tracking, and health hazard alarms (Chang et al., 2020). Since 2010, all Brazilian companies obligatorily report invoices to a central financial agency, such as the State Treasury Office (SEFAZ). Similar measures have also been taken in Italy (Bardelli et al., 2020) and China (Zhou et al., 2019)(Yue et al., 2020).

The Brazilian Electronic invoice is a standardized XML file. While fields are audited for fulfillment and type, there are breaches for exploits and errors. One fundamental vulnerability is on the reported product code, called NCM. NCM is a standardized nomenclature for products and services in Mercosur. It is used to define the correct taxation and if the product is eligible for tax exemption. One could miss-classify products to benefit from lower taxation. Brazil utilizes two types of electronic invoices: Electronic Invoice (NF-e), which records B2B transactions, and Consumer Electronic Invoices (NFC-e) that records retail transactions. Mandatory reports of the NFC-e begun 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 larger number of issuers and 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.

Invoice text data differs in grammar and vocabulary from regular language usage and can be seen as short-text. Short-text can be defined by the following characteristics (Enamoto et al., 2021): individual author contribution is small; grammar is generally informal and unstructured, sent and received in real-time and in large quantity; imbalanced distribution of classes of interest; large scale data presents a labelling bottleneck. Even when compared to other short texts, invoice description is very brief, containing only a handful of words, often not forming a complete sentence. This exacerbates the problem of

[a] https://orcid.org/0000-0002-9957-0059
[b] https://orcid.org/0000-0001-6795-2768
[c] https://orcid.org/0000-0002-1295-5221
[d] https://orcid.org/0000-0003-1826-1850



Kieckbusch, D., Filho, G., Di Oliveira, V. and Weigang, L.
SCAN-NF: A CNN-based System for the Classification of Electronic Invoices through Short-text Product Description.
DOI: 10.5220/0010715200003058
In *Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST 2021)*, pages 501-508
ISBN: 978-989-758-536-4; ISSN: 2184-3252

domain-specific vocabulary, abbreviations, and typos as authors use their individual logic.

Related works on invoice classification have focused on the Chinese case. These solutions have ranged from using hash trick for dealing with an unknown number of features (Zhou et al., 2019)(Yue et al., 2020), semantic expansion trough external knowledge bases (Yue et al., 2020), classification of paragraph embedding by k-nearest-neighbors (Tang et al., 2019) to different artificial neural network architectures (Yu et al., 2019)(Zhu et al., 2020). Semantic expansion is prevalent not only on invoice classification but also on short-text classification (Wang et al., 2017),(Naseem et al., 2020). 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 (Grida et al., 2019). There is a gap in the literature for models suited for the classification of Brazilian Electronic Invoices.

This work presents a Convolutional Neural Network (CNN) system, named SCAN-NF, for tax auditors to identify suspicious invoices based on textual product descriptions. We utilize the sentence classification architecture proposed by Kim (Kim, 2014) as the basis for our model. Real data from Brazilian Electronic Invoices (NF-e) and Consumer Electronic Invoices (NFC-e) are used in our experiments. We compare single, and ensemble model approaches on recall, precision, and accuracy on both datasets. We then discuss performance and trade-offs between approaches. Comparing single model with ensemble model approaches, the evaluation metrics show the satisfaction of the developed system. Our ensemble approach achieved better precision on both datasets.

This article is organized as follows: In section 2, we present related work on the invoice and short text classification; in section 3, we describe the SCAN-NF system and the architecture of the classification model; in section 4, we present experimental setup on a study case on real NF-e and NFC-e data. Results of experiments are presented in section 5, and in the final section, we present closing remarks and future works.

# 2 RELATED WORK

In this section, we highlight other works related to short text and invoice classification. Taking invoices as an example of short text, short-text classification is a broader area, and some solutions may not suit invoice classification. In contrast, works aimed at invoice classification may not utilize short text processing techniques.

## 2.1 Traditional Methods

Traditional methods rely on bag-of-words representation and matrix factorization to create a representation for text processing. The low word count on short text documents leads to common co-occurrence of terms across the document-term matrix, which invalidate matrix factorization methods.

Early works attempted to address this problem by expanding available information through auxiliary databases. Document expansion seeks to substitute the representation of short text to represent a set of related documents. In query-based expansion, these documents are returned by using short text as the input on a search engine (Sahami and Heilman, 2006)(Yih and Meek, 2007). Phan (Phan et al., 2008) proposed a framework for short text classification that used an external ”universal dataset” to discover a set of hidden topics through Latent Semantic Analysis. The problem with document expansion is that it increases the computational cost to search and process a more significant amount of data. This new data also introduces noise to the model.

## 2.2 Neural based Methods

Neural-based methods represent short text as a sequence of vectors and utilize convolutional and recurrent neural networks to learn a suitable representation for classification.

The architecture proposed by Kim (Kim, 2014) serves as the basis for most CNN-based solutions. Zhang (Zhang and LeCun, 2016) utilized a 12-layer CNN to learn features from character embeddings. Character-based representation does not rely on pretrained word embeddings and could be used in any language. Wang (Wang et al., 2017) expanded the model proposed by Kim (Kim, 2014) by utilizing concept expansion and character level features. 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 (Naseem et al., 2020) proposed an expanded meta-embedding approach for sentiment analysis of short-text that combined features provided by word embeddings, part of speech tagging, and sentiment lexicons. The resulting compound vector was fed to a Bi-LSTM 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.

### 2.3 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 hash trick for dimensionality reduction (Zhou et al., 2019)(Yue et al., 2020). Yue (Yue et al., 2020) performed semantic expansion of features through external knowledge bases before using the hash trick for dimensionality reduction. Tang (Tang et al., 2019) utilized paragraph embedding to create a reduced representation and then applied K-NN classifier. Yu (Yu et al., 2019) utilized a parallel RNN-CNN architecture, with the resulting vectors being combined in a fully connected layer. Zhu (Zhu et al., 2020) combined features selected through filtering with representation learned through the LSTM model.

Unlike most western languages, in which text is expressed through words with white spaces as separators, text in Chinese is expressed without separators, with 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 (Paalman et al., 2019) 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 LDA and LSA.

### 2.4 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 bases becomes the bottleneck of the system and are unsuited for invoice processing due to the amount of invoice data. Furthermore, knowledge bases may not be available in languages other than English and Chinese (Grida et al., 2019).

The limitation of pre-trained word embeddings comes down to vocabulary coverage and word sense (Faruqui et al., 2016). 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 possess little to no context to disambiguate word sense.

Most invoice classification models did not utilize ANN. Yu (Yu et al., 2019) was the only one to combine both CNN and BiLSTM. However, CNN and BiLSTM were used in parallel over different fields of invoice data. Zhu (Zhu et al., 2020) combined an LSTM network with traditional methods using filtered features. While effective for the Chinese language, these architectures do not suit the Brazilian invoice model. We address these shortcomings by proposing a CNN-based model that does not rely on pre-trained word embedding and external knowledge bases.

## 3 ARCHITECTURE OF SCAN-NF

In this section, we present an overview of the architecture of the SCAN-NF system and inner models, Figure 1. 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 SEFAZ 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 figure 1). Product description and corresponding NCM code for each product in each invoice are then extracted (label 2 in figure 1). Text is then cleaned from irregularities (label 3 in figure 1). A training dataset is constructed by balancing target classes samples and dropping duplicates (label 4 in figure 1). The training set is then fed to a CNN model that learns to classify product descriptions (label 5 in figure 1). Outputs at the training phase of the system are used to validate models before being put into production (label 6 in figure 1). 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 figure 1). Models trained in the training phase are then em-

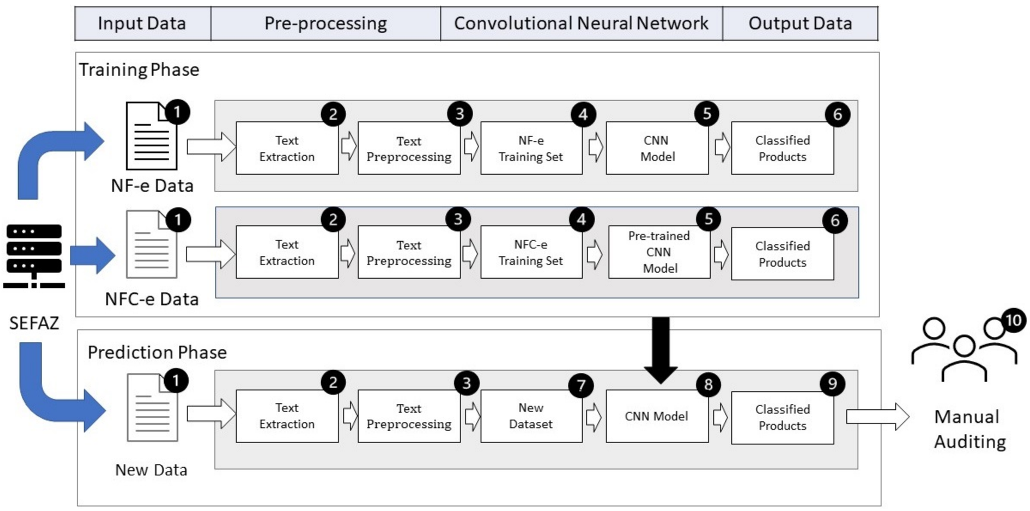


Figure 1: Architecture of SCAN-NF.

ployed for the task at hand (label 8 in figure 1). The final output of the model is the classified set of products inputs (label 9 in figure 1). This set of classified product transactions are then used in manual auditing by tax auditors (label 10 in figure 1).

The system is intended to aid tax auditors in the audition of 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 large 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 a larger amount of 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 (Kim, 2014) can be used as a single multi-label classification model. However, due to the high number of possible NCM codes and high amount of 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 the need of retraining other models.

Figure 2 presents the flowchart used in single models. The input layer takes the indexed representation of text. In the embedding layer, each word index is replaced by the word vector representation. Input is then reshaped to be fed to parallel channels of one-dimensional convolutions layers. Each convolution layer applies several filters of a given size to the encoded sentence. Max pooling is applied to the learned filters to extract the most useful features. Outputs of each channel are concatenated in a single vector flattened and fed to a Fully connected layer that will output the final classification. The categorical cross-entropy calculates loss, and the soft-max function acts as the activation function of the model.

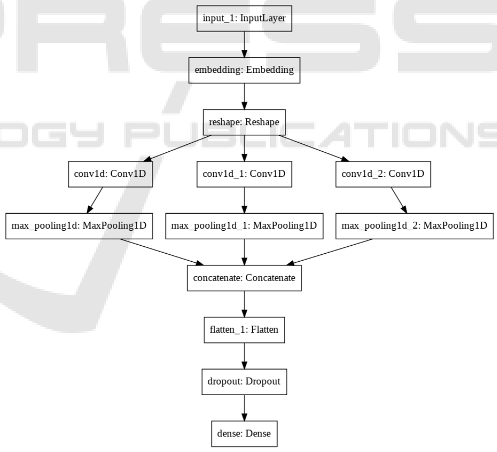


Figure 2: Flowchart of SCAN-NF CNN single model.

Figure 3 present a simplified flowchart of the ensemble model. The ensemble model is built from binary classifiers, each trained on a singular target class. Each binary classifier is built on the flowchart presented in figure 2. In binary models, the loss is calculated by the binary cross-entropy, and the sigmoid function is used as the activation function for the last layer. To offset the imbalance between classes in binary models, we set class weights to a rate of 180 to 1. The ensemble model is built using previously trained binary models. The output of each model is concatenated and fed to a single fully connected layer that performs multi-layer classification. The categorical cross-entropy calculates loss for the ensemble model, and the softmax function gives the activation function of the model.

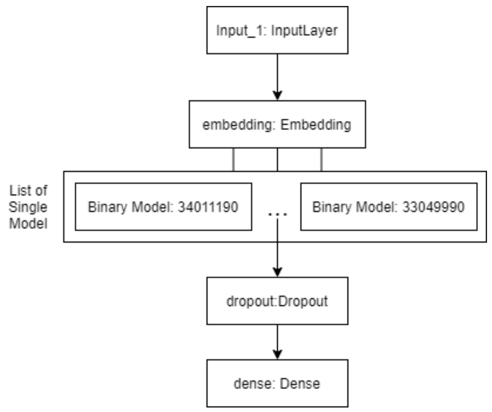


Figure 3: Flowchart of SCAN-NF CNN ensemble model.

### 3.1 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 (SEFAZ, 2015). 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 meant to be used in retail services.

There are validations rules for the NCM field in the NF-e manual (SEFAZ, 2015). According to a specialist working with tax audition and the schedule published in the NF-e manual, while validation procedures for NF-e documents are implemented, these validation procedures are not planned for NFC-e documents in the following years. This results in data of poor quality.

## 4 PERFORMANCE EVALUATION

We conducted a study case based on real NFC-e and NF-e documents from SE to validate our model. 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.

### 4.1 The Data

In our experiments, we utilized data provided by the estate tax office (SEFAZ). Data provided included both NFC-e and NF-e documents. NF-e data consisted of invoices of 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 the norm. Our design decision was to drop duplicate product descriptions for each target class. Table 1 presents detailed information on the number of samples used in the experiment. 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: Number of samples and datasets used in experiments.

| | 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 | |

### 4.2 Metrics

We evaluate models based on the following metrics: accuracy, precision, recall, and top k Accuracy. Metrics are calculated based on the occurrence of 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 how correct the set of retrieved samples is 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.

Table 2: Parameters used in optimization.

| Parameter | Values |
|---|---|
| Number of Filters on 1D Convolution #1 | {50,100,200,**300**, 400,500,600} |
| 1D Convolution Kernel Size #1 | {**3**,5,7,9} |
| Number of Filters on 1D Convolution #2 | {**50**,100,200,300, 400,500,600} |
| 1D Convolution Kernel Size #2 | {3,5,7,**9**} |
| Number of Filters on 1D Convolution #3 | {50,100,200,300, 400,500,**600**} |
| 1D Convolution Kernel Size #3 | {**3**,5,7,9} |
| Dropout rate | [0, **0.29**, 0.5] |
| Optimizer | {Adam, Adagrad, Adadelta, **Nadam**} |

## 5 RESULTS

In this section, we present the results of the experiments. The goal of the experiment is to compare single model and ensemble model approaches. The single model is composed of a single CNN model trained on multi-label classification. 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. Singular models and binary models were trained through 5 epochs, while the fine tune of the ensemble model is 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-tune of the ensemble model. Experiments were repeated 10 times.

### 5.1 NF-e Dataset

Figure 4 presents single and ensemble model performance on the NF-e dataset. We present results side by side. We can see that while model accuracy deviated sightly, differences in recall and precision were more evident. There is little spread for all metrics. The single model performed slightly better on both accuracy metrics. Both models presented accuracy above 0.85. The most notable difference between models comes from the trade-off between recall and precision. The single model performed better on recall at the cost of precision. The single model recall was 15% higher than the ensemble model at the cost of 5% precision.

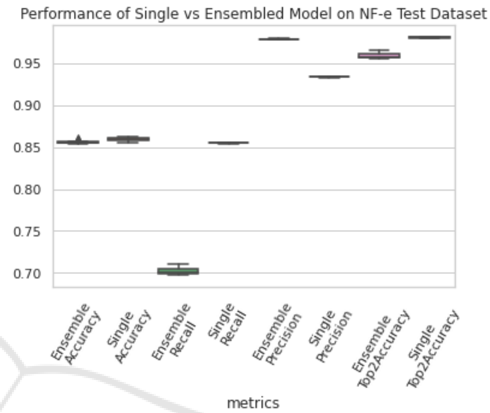


Figure 4: Performance of single and ensemble models on NF-e dataset.

Individual class performance of the ensemble model is shown in figure 5. Due to the unbalanced nature of the problem, all classes presented high accuracy scores, scoring higher than 96%. Overall, there was a balance between recall and precision. 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.

### 5.2 NFC-e Dataset

Performance on NFC-e is represented in Figure 6. We can see that the trade-off between recall and precision between models also occurs, with the ensemble model achieving lower recall and higher precision. There is a drop in accuracy for both models, while top2 accuracy remained the same. Individual binary model performance on NFC-e dataset is shown on Figure 7. Of 18 classifiers, 3 presented F1-Scores lower than 0.6, 12 presented F1-scores above 0.7, and only 1 achieved an F1-score higher than 0.9.

### 5.3 Comparison of Approaches

When comparing overall results between datasets, it becomes clear that NFC-e product classification

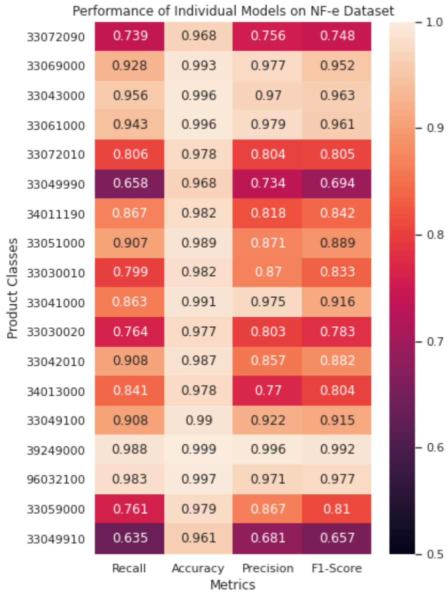


Figure 5: Individual Binary Model Performance on NF-e dataset.

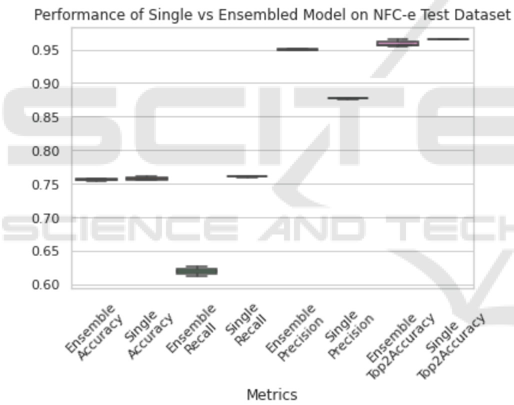


Figure 6: Performance of single and ensemble Models on NFC-e dataset.

is a more complex problem than NF-e classification. In both datasets, the worst-performing and best-performing classes were the same. This indicates that identifying certain products is more complex than others. Performance on NF-e data was higher than NFC-e data. This is in accord with the NFC-e document characteristics.

We can see that there is a trade-off between recall and precision, with the ensemble model presenting higher precision at the cost of the recall. The simple model approach is more suited for exploratory data analysis due to higher recall, while the ensemble model approach is more suited to the audition of suspicious issuers due to higher precision. There are also differences in the maintainability of approaches.

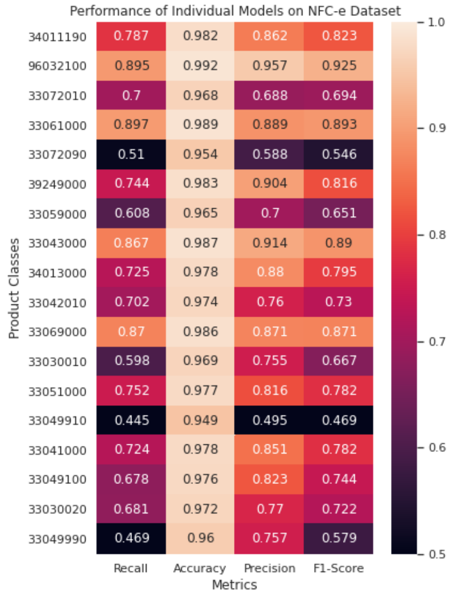


Figure 7: Performance of binary models on NFC-e dataset.

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.

Models consistently achieved around 95% top2 accuracy on both datasets. This means that models can be used as recommendation systems for the classification of product descriptions. This is particularly valuable for NFC-e documents, in which no homologation is currently done, and data is more varied. Recommendations can aid taxpayers in narrowing down the NCM code given a general text description of the product. In turn, this could lead to better reported NFC-e data. Overall, models managed to map product descriptions to the corresponding NCM code.

## 6 CONCLUSION AND FUTURE WORK

In this work, we showed SCAN-NF, an invoice classification system based on product description for tax auditing. We presented related work on short-text and invoice classification and a set of desired properties for invoice classification. We then presented SCAN-NF, a solution for the modeled problem, and the architecture of the CNN model to power the solution. We presented two possible configurations for the CNN models: a single model based on established sentence classification architecture and our proposed ensemble model. Both CNN configurations were validated on datasets of NFC-e and NF-e documents. Our ensemble approach presented higher precision on both datasets. Overall we managed to present an invoice classification system that can aid tax auditors in auditing a larger number of invoices and aid taxpayers in providing the correct classification of products.

In future work, we will focus on transfer learning. We hope that the parameters obtained from pre-training using better represented NF-e documents can improve performance on the training of NFC-e data. This would be of great value as manual auditing of individual invoices is quite expensive. Our main focus will be using Natural Language Processing (NLP) techniques such as pre-trained word embeddings and transformers into our concerning research.

## REFERENCES

Bardelli, C., Rondinelli, A., Vecchio, R., and Figini, S. (2020). Automatic electronic invoice classification using machine learning models. *Machine Learning and Knowledge Extraction*, 2(4):617–629.

Chang, W.-T., Yeh, Y.-P., Wu, H.-Y., Lin, Y.-F., Dinh, T. S., and Lian, I. (2020). An automated alarm system for food safety by using electronic invoices. *PLoS ONE*, 15(1).

Enamoto, L., Weigang, L., and Filho, G. P. R. (2021). Generic framework for multilingual short text categorization using convolutional neural network. *Multimedia Tools and Applications*, 80.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. pages 30–35.

Grida, M., Soliman, H., and Hassan, M. (2019). Short text mining: State of the art and research opportunities. *Journal of Computer Science*, 15(10):1450–1460.

He, Y., Wang, C., Li, N., and Zeng, Z. (2020). Attention and Memory-Augmented Networks for Dual-View Sequential Learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 125–134.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (2011):1746–1751.

Naseem, U., Razzak, I., Musial, K., and Imran, M. (2020). Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.

Paalman, J., Mullick, S., Zervanou, K., and Zhang, Y. (2019). Term based semantic clusters for very short text classification. In *International Conference Recent Advances in Natural Language Processing, RANLP*, volume 2019-Septe, pages 878–887.

Phan, X. H., Nguyen, L. M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, (January):91–99.

Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. *Proceedings of the 15th International Conference on World Wide Web*, pages 377–386.

SEFAZ (2015). *Manual de Orientação do Contribuinte - Padrões Técnicos de Comunicação*. ENCAT.

Tang, X., Zhu, Y., Hu, X., and Li, P. (2019). An integrated classification model for massive short texts with few words. In *ACM International Conference Proceeding Series*, pages 14–20.

Wang, J., Wang, Z., Zhang, D., and Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. *IJCAI International Joint Conference on Artificial Intelligence*, pages 2915–2921.

Yih, W. T. and Meek, C. (2007). Improving similarity measures for short segments of text. *Proceedings of the National Conference on Artificial Intelligence*, 2:1489–1494.

Yu, J., Qiao, Y., Shu, N., Sun, K., Zhou, S., and Yang, J. (2019). 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*, pages 64–71.

Yue, Y., Zhang, Y., Hu, X., and Li, P. (2020). Extremely Short Chinese Text Classification Method Based on Bidirectional Semantic Extension. In *Journal of Physics: Conference Series*, volume 1437.

Zhang, X. and LeCun, Y. (2016). Text understanding from scratch.

Zhou, M., Hu, X., Zhu, Y., and Li, P. (2019). 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*, pages 861–865.

Zhu, Y., Li, Y., Yue, Y., Qiang, J., and Yuan, Y. (2020). A Hybrid Classification Method via Character Embedding in Chinese Short Text with Few Words. *IEEE Access*, 8:92120–92128.