

# ELEVEN Data-Set: A Labeled Set of Descriptions of Goods Captured from Brazilian Electronic Invoices

Vinícius Di Oliveira<sup>1,2</sup><sup>a</sup>, Li Weigang<sup>1</sup><sup>b</sup> and Geraldo Pereira Rocha Filho<sup>1</sup><sup>c</sup>

<sup>1</sup>*TransLab, University of Brasilia, Brasilia, Federal District, Brazil*

<sup>2</sup>*Secretary of Economy, Brasilia, Federal District, Brazil*

**Keywords:** BERT, Electronic Invoice, Labeled Data-set, Short Text, Supervised Learning, Text Classification.

**Abstract:** The task of classifying short text through machine learning (ML) models is promising and challenging for economic related sectors such as electronic invoice processing and auditing. Considering the scarcity of labeled short text data sets and the high cost of establishing new labeled short text databases for supervised learning, especially when they are manually established by experts, this research proposes ELEVEN (ELEctronic inVoicEs in portuguese laNguage) Data-Set in an open data format. This labeled short text database is composed of the product descriptions extracted from electronic invoices. These short Portuguese text descriptions are unstructured, but limited to 120 characters. First, we construct BERT and other models to demonstrate the short text classification using ELEVEN. Then, we show three successful cases, also using the data set we developed, to identify correct products codes according to the short text descriptions of goods captured from the electronic invoices and others. ELEVEN consists of 1.1 million merchandise descriptions recorded as labeled short-texts, annotated by specialist tax auditors, and detailed according to the Mercosur Common Nomenclature. For easy public use, ELEVEN is shared on GitHub by the link: <https://github.com/vinidiol/descmerc>.

## 1 INTRODUCTION

There is a vast amount of information on the Web, including images, videos, documents and a colossal volume of texts. Most of the text data available are unstructured, which makes it arduous and onerous to search, analyze and retrieve valuable information from this source. A labeling process, such as manual annotation could be an expensive solution, which in some cases, would be prohibitive as a time-costly task made by expensive labor (Pandolfo and Pulina, 2021). When it comes to short texts, the scenario is no different. Domain experts are needed for trust-able labeling and they are hard to find and hire (Du et al., 2019; Sugrim, 2020). It is possible to notice a shortage of short texts labeled in English, which is even greater in Portuguese.


The challenge to build machine learning models for classifying short texts is huge, especially when the task needs to analyse beyond Twitter and other comments on social networks. To construct relevant


and valuable knowledge it is very important to surpass sentiment analysis and move on to semantic classification. The use of machine learning algorithms is mandatory to retrieve structured information from unstructured texts (Ambika, 2020). Thereby, to thrive on the supervised learning field, an open and truthful labeled data-set could contribute significantly.


Even with the impressive progress in these fields, machines are still far from being able to have a complete semantic understanding of the human language (Maulud et al., 2021; Hitzler et al., 2020; Lake and Murphy, 2021), so the supervised machine learning algorithms are quite useful, if not indispensable, in achieving relevant results.

Research success, both academic and industry, grows as findings are shared, tested and debated. Open and free data sets play a very important role in this challenge, as they can keep parameters of equal comparison between different models. So different algorithms and/or different settings can be trained and evaluated based on the same information (Gasparetto et al., 2022; Pintas et al., 2021).

This work introduces the ELEVEN Data-Set (ELEctronic inVoicEs in the portuguese laNguage) for that purpose. It presents a set of 1,117,623 la-

<sup>a</sup> <https://orcid.org/0000-0002-1295-5221>

<sup>b</sup> <https://orcid.org/0000-0003-1826-1850>

<sup>c</sup> <https://orcid.org/0000-0001-6795-2768>

beled records of merchandise descriptions. The annotation process was made by tax auditors specialists following the Mercosur Common Nomenclature pattern. The descriptions were extracted from the Brazilian Electronic Invoice data base, it was made available by the Secretary of Economy of Brasília - Brazil.

As the contributions of this paper beside the ELEVEN Data-set, we will construct Bidirectional Encoder Representations from Transformers (BERT) and other machine learning models to demonstrate the short text classification using ELEVEN. Then, we will show three successful cases in the literature, also using ELEVEN, to identify correct products codes according to the short text descriptions of goods captured from the electronic invoices (Kieckbusch et al., 2021) and others (Marinho et al., 2022; Schulte et al., 2022).

The article is organized as follows. After this introduction, section 2 describes the Brazilian electronic invoices and Mercosur Common Nomenclature (NCM) code. Section 3 studies the related work about the development of the electronic invoice and short text data-sets. ELEVEN data-set is presented in section 4. To show the possible application, section 5 shows BERT and other two model for text classification using the proposed data-set. Section 6 reports three successful cases using ELEVEN data-set. The last section gives the conclusions of the article.

## 2 THE BRAZILIAN ELECTRONIC INVOICE

All trade transactions of goods in Brazil are electronically recorded. The digital document that keeps each transaction's information is called *Nota Fiscal Eletrônica - NFe* - Electronic Invoice or Electronic Tax Bill as a free translation from Portuguese to English. This document is a XML file transmitted by the internet between the issuer (seller/remittent) and the tax administration system - in the case of Brasília, the Secretary of Economy (CONFAZ, 2013).

As shown in Figure 1, the issuer sends the NFe file to the tax administration system which validates, authorizes, and records the transaction. The transport of the goods is authorized and the delivery to the buyer is completed. After, the buyer can confirm the compliance requirements consulting the tax administration website. The communication between the companies and the tax administration system are made by a web service link (da Rocha et al., 2018).

The NFe document contains all the information to identify the seller (or sender), the buyer (or addressee), the tax information related, the transport

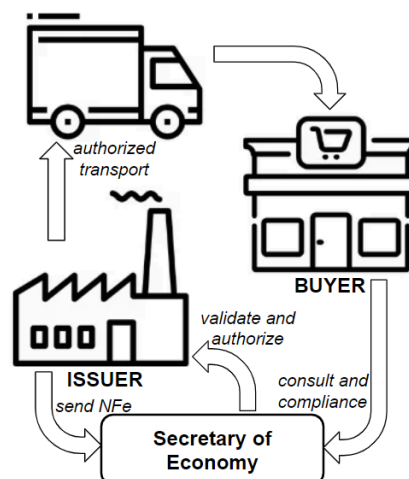


Figure 1: NFe generation and operation flow.

and freight information, and, finally our study object, the goods linked to that transaction, including descriptions, quantities, volumes, tax rate, tax value, tax codes, accounting codes and NCM code (Mercosur Common Nomenclature - NCM).

The NFe document is free to fill, i.e., there is no validation rule for the inputs. This feature is due to the declaratory nature of this document. The validation of the document's compliance is carried out later by the tax inspection. So, there is no guarantee that the input codes (e.g., tax, accounting, and NCM) are correct. Hence, there is no parameterized reference to compare the correctness of what has been declared. Filling errors or the intention to deceive the tax authorities can happen. Anyway, the description of the goods must correspond correctly with the goods to which that NFe refers, after all, the buyer will check the correctness of the description, as he needs to confirm what was actually purchased and delivered (de Aguiar and Gouveia, 2020).

Machine learning models that can read the goods' description field would be extremely useful, as they would be used to detect fraud and errors in filling out invoices, consequently contributing greatly to reducing tax losses currently verified (Sinayobye et al., 2018; Raghavan and El Gayar, 2019).

The tax administration of Brasília began recording and labeling descriptions to facilitate the automated process of data cross-referencing. The database produced resulted in the first version of ELEVEN.

## 3 RELATED WORK

The adoption of the NFe framework for recording and controlling all the merchandise transactions in Brazil

increased the tax revenue (Vieira et al., 2019), enhanced mechanisms to face tax evasion (de Aguiar and Gouveia, 2020) and, by companies' side, it improved the tax compliance systems and internal audit procedures (Codesso et al., 2020). In the Logistic field, as the NFe document has the information of the sender and the receiver, the data stored in electronic tax invoices were used as the main input for the allocation process for generating distribution routes for cargo vehicles in urban areas (Pipicano et al., 2021).

As this work presents a data-set of labeled short texts extracted from the Brazilian electronic invoices (NFe), the following subsections will show works that illustrate de NFe relevance in information technologies applications and point out some short text data sets.

### 3.1 NFe for Information Technology Ground

In the field of high-performance processing of large data-sets, (da Rocha et al., 2018) analyzes the performance of SQL queries on Hadoop comparing it with an RDBMS-based approach. The study focuses on a large set of NFe electronic invoice data.

The Brazilian electronic invoices have been used in several studies across other fields, especially in Artificial Intelligence. A work (dos Santos Neto et al., 2022) proposes that using Artificial Intelligence it is possible to find potential customers for a product. They present a methodology developed to identify pent demand by analyzing the NFe electronic invoices. Using the information collected in the electronic invoices, it was possible to quantitatively evaluate the existence of pent-up demand for some product in a specific region and then create decision support mechanisms. The experiment observed that 13,6% of products presented a strong indication of pent-up demand.

More about Artificial Intelligence applications, (Lucena et al., 2022) proposes an approach to inspect invoices and extract information about measures and units from goods descriptions. They used a neural network with the BiLSTM-CRF architecture, a combination of a long short-term memory (LSTM) and a conditional random field (CRF). This method validates product quantity information to verify whether any product was bought or sold by the enterprise without issuing an NFe electronic invoice.

Other study (Mendes Thame Denny et al., 2021) defined the creation of a tax credit clearinghouse and analyzes the applicability of blockchain and Distributed Ledger Technology - DLT to the Brazilian Electronic Invoice System. They set the adoption of

Hyperledger Composer Fabric as an encrypted framework that would be able to create a secure environment for the storage and analysis of information by using DLT, so it presents itself as a solution to address privacy and security concerns of the stakeholders.

### 3.2 Short Text Data-Sets

Other relevant data-sets are embodying short texts. The term "common data sets" was used by (Tang et al., 2022) when referring to three well-known data sets: Yago (Suchanek et al., 2008), Freebase (Bollacker et al., 2007), and Probase (Wu et al., 2012). They are called "common" because they represent general content and can be used by non-professional information and any researcher.

YAGO, A Large Ontology from Wikipedia and WordNet, presents with high coverage and precision a large ontology. It brings content from Wikipedia and WordNet in more than 1.7 million entities and 15 million facts. There is a taxonomic hierarchy as well as semantic relations between the entities. It maintains compatibility with RDFS while allows representing n-ary relations in a natural way (Suchanek et al., 2008).

Freebase, a Shared Database of Structured General Human Knowledge, is a graph-shaped database of structured general human knowledge. It is a store of large data objects such as text documents, images, sound files, and software. The primary method of access to Freebase is through its public HTTP-based API which contains tools for the collaborative design of simple types and properties. The data in Freebase consists of millions of topics and tens of millions of relationships between them (Bollacker et al., 2007).

Probase, a Probabilistic Taxonomy For Text Understanding, contains 2.7 million concepts harnessed automatically from a corpus of 1.68 billion web pages as well as it uses probabilities to model inconsistent, ambiguous, and uncertain information it contains (Wu et al., 2012). It was used as a source for short text understanding studies with semantic concepts (Ji et al., 2019; Shi et al., 2018) and used to empower an engine for products review analysis as a knowledge source (Luo et al., 2019).

There are other short text data-sets for specific themes. The Twitter data set could be set or sectioned by specific subjects, e.g. for COVID-19 research (Chen et al., 2020), natural disaster perception (Alam et al., 2018), and the foremost use for sentiment analysis (Zimbra et al., 2018). And the Amazon Fine Foods reviews (McAuley and Leskovec, 2013). Those were just examples, as the whole list would be too extensive and uncountable.

## 4 THE ELEVEN DATA-SET

The ELEVEN construction started in 2017 with a team of tax auditors specialists in the inspection of that kind of operations. Every month the data-set is updated with new descriptions captured by the tax administration system and so on. The result of that work, the more than one million labeled descriptions data-set, was shared with the authors for academic study purposes.

The verification and calculation of the tax due is done electronically on large masses of data. Such analysis depend on reading the fields filled in the NFe. The field referring to the product description, named XPROD, is the most reliable as it is printed and sent attached with the merchandise, so the purchaser of the goods usually checks it for order compliance reasons or warranty issues, on the other hand, it is a text (field of type string), and upper limit of 120 characters, without any filling pattern, so its automatic reading and classification become unfeasible by traditional means of data crossing.

That said, it can be seen that the current main use of the database has its pros and cons. The direct comparison works effectively but is limited to perfect matching, character by character, between the checked fields. Therefore, to solve the problem of the perfect match must, an artificial intelligence capable of comparing text descriptions and tax codes, without the perfect match restriction, would improve the tax inspection actions. Finally, the ELEVEN Data-Set can be the necessary piece to build that artificial intelligence.

The ELEVEN Data-Set is composed by four mains columns in which the 1,117,623 elements are distributed. The columns are described below.

- **XPROD**: goods description in the NFe, max. length of 120 characters. The text is raw, exactly as it was inserted by the issuer;
- **NCM**: NCM code inserted in the NFe, 8 fix digits;
- **Rotulo**: specification of the goods identified by specialists according to the Common Mercosur Nomenclature - NCM, i.e., the label, text string;
- **Item**: A label code annotated by the experts, 2 digits. This can be read as the class label.

The class label codes are indicated in the *Item meaning* column on the Table 1.

The ELEVEN Data-Set is available on GitHub in a zipped csv file named *BaseDesc\_NCM.zip*. [<https://github.com/vinidiol/descmerc>]

The Common Mercosur Nomenclature - NCM is a regional nomenclature for the categorization of goods adopted by Brazil, Argentina, Paraguay, and Uruguay

Table 1: Label code annotated by the experts.

Code	Meaning
30	Cachaça and spirits
31	Wines, vermouths, ciders, other beverages
38	Perfumery, personal care and cosmetics
39	Cleaning materials
40	Food products
41	Building materials
42	Electrical Material

since 1995. It is used in all foreign trade operations of Mercosur countries. This nomenclature is an ordered structure that allows determining a single numerical code for a given commodity. This code, once known, starts to represent the commodity itself (Brazil, 2019).

By way of illustration, the sequence of NCM codes for “Milk” identification can be seen by the following:

- **40**: Milk;
- **4002**: Milk and cream concentrated or with added sugar or other sweeteners.;
- **400221**: Milk with no sugar and no sweeteners;
- **40210210**: Whole milk.

The Figure 2 illustrates the distribution of records among code classes, as well as shows the number of records per code class in descending order of occurrences.

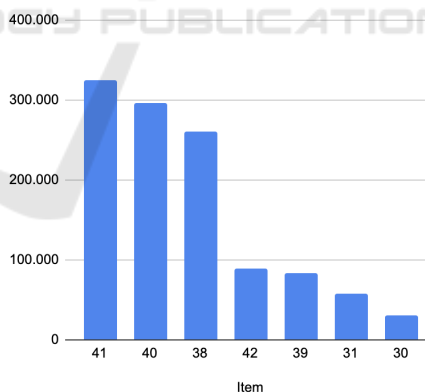


Figure 2: Illustration of records occurrences.

**Data Protection and Tax Secrecy.** The data exposed does not violate the tax secrecy of companies issuing and/or receiving the invoices, as they only show the description of the goods listed in the electronic invoices, as can be verified by any consumer when purchasing a product, as well as no financial reference that can demonstrate pricing values or profit rates. The NCM codes are in the public domain and widely available. There will be mentions of product brands in the descriptions, but for the reasons explained above,

this indication remains to assure the tax secrecy of the brands. The authors had authorized access to the database by the Sub-Secretariat of Revenue of the State Department of Economy of Brasilia, respecting the premises indicated above regarding tax secrecy. All references to values, units, and quantities have been removed.

## 5 CLASSIFICATION CASE MODELING

In order to demonstrate the feasibility of using the Data-Set as a basis for a text classifier, a modeling essay was built and shared in this section. The chosen task was text classification, where the models will classify the goods' descriptions (inputs, "XPROD") conforming to the given label (outputs, "Item"). According to the nature and purpose of this Data-Set, the chosen models are supervised learning algorithms.

A sample of 15,000 labeled descriptions is also shared at the same GitHub page in another csv file named *AmostraDescMerc.csv*. This file is a random sample of three item categories: 38 - Perfumery, personal care, and cosmetics; 39 - Cleaning materials; and 40 - Food products. The file is composed of 5,000 records from each category. It is also available a R code for modelling the sample data-set, the results are shown in this section.

These are the tested classification prediction models: KNN, GBM, ANN, and BERT.

The sample data-set was splitted in three parts on the following proportions, 70%, 20% and 10% for training, validation and test, respectively. The Gradient Boosting Machine (GBM), Artificial Neural Network (ANN) and Pre-training of deep bidirectional transformers for language understanding (BERT) models settings are presented as follow.

- GBM: `ntrees = 1000`, `max_depth = 3`;
- ANN: 3 layers of 150, 300, 150 neurons, in 20 epochs.
- BERT: Number of Word = 5000, `max_length = 150`, batch size of 8, and learning rate of  $2 \times 10^{-5}$  in three epochs of training.

The complete code used in this study is shared on the GitHub folder indicated above. The predictions results of the three models are presented below. The KNN in Figure 3, the GBM in Figure 4, and the ANN in Figure 5. In those figures, the "Macro" value concerns the metrics of the model as a whole.

The neural network architecture was conceived by testing various layouts, some of them are shown in the

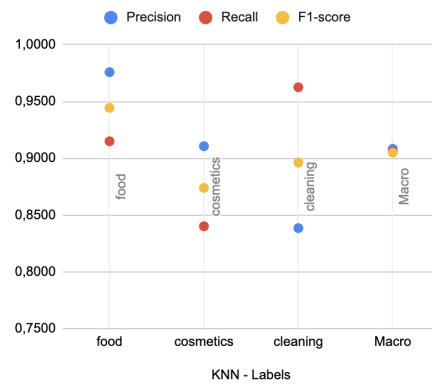


Figure 3: KNN model results.

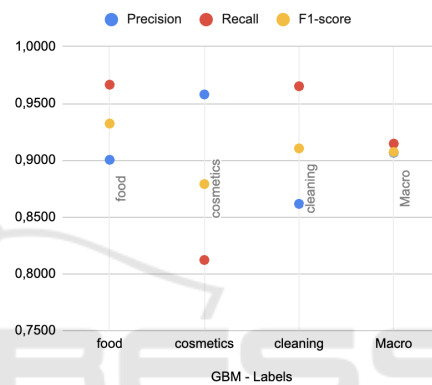


Figure 4: GBM model results.

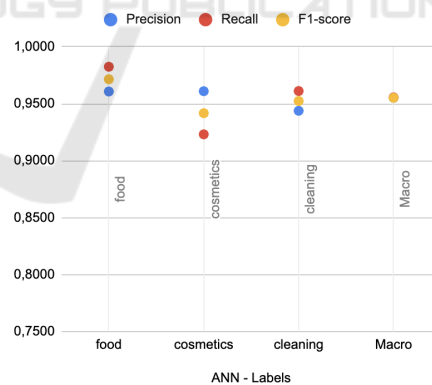


Figure 5: ANN model results.

Table 2. The choice criterion was the best accuracy, so the Table 2 shows the architectures in descending order of the achieved accuracy values. Therefore, the architecture chosen was one of three layers with 150, 300, and 150 neurons respectively, in 20 training epochs. The accuracy achieved is 95.52% and the AUC is 0.9998. The GBM model achieved an accuracy of 90.63% and the AUC of 0.9892.

The results presented indicates consistency with

the purpose of the ELEVEN Data-Set and with the results of these algorithms found in similar tasks of text classification (Kadhim, 2019; Kowsari et al., 2019; Thangaraj and Sivakami, 2018).

Relevant results have already been achieved with a CNN model using the ELEVEN Data-Set in another study by (Kieckbusch et al., 2021). The sampling classes chosen were different from those picked for this study, however the Accuracy achieved with the CNN model was 0.97. More details about this work are in the next section.

Table 2: ANN Architectures (20 epochs).

L1	L2	L3	L4	Acc.	AUC
150	300	150	-	0.9552	0.9998
100	200	100	-	0.9514	0.9998
40	40	-	-	0.9511	0.9999
30	60	60	30	0.9481	0.9999
30	60	30	-	0.9474	0.9998

The BERT model (Devlin et al., 2018), one of the most prominent state-of-art models (Acheampong et al., 2021; Bhavani and Santhosh Kumar, 2021; Minaee et al., 2021), achieved the best result with the data sample used for this experiment, an accuracy of 0.98. The Figure 6 compares the accuracy results from the models tested. The script used for the BERT model is also available on the aforementioned GitHub page.

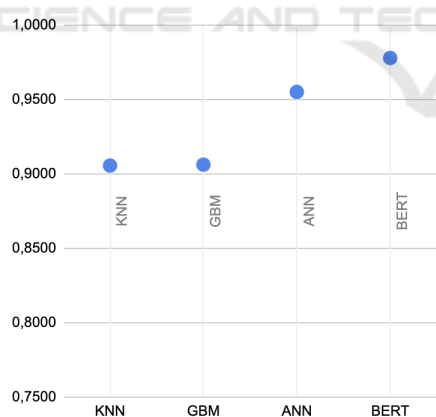


Figure 6: Tested Models Accuracy.

Although some advances achieved in how to interpret neural networks, they are still hard to explain and understand (Oh et al., 2019; Buhrmester et al., 2021). Explain why and how the model classified one description different from another one is not simple, especially when facing stakeholders' inquiries, explainability is a very important issue (Sokol and Flach, 2020). In some cases, there are legal requirements

that demand machine learning model interpretability (Bibal et al., 2021), i.e., the algorithm predictions (outputs) would be understood somehow by the users. That said, looking for the best implementation options, other models must be taken into account when their performance is close to the neural networks.

## 6 THE ELEVEN APPLICATION IN OTHER STUDIES

Recently, the labeled data-set from descriptions of goods, extracted from this electronic invoices, has been a source of data support for published studies. ELEVEN data-set has been applied in three cases in literature.

A Convolutional Neural Network (CNN) based system, named SCAN-NF (Kieckbusch et al., 2021), thrived to classify Electronic Invoices based on goods descriptions. SCAN-NF was built to identify correct products codes based on the short-text descriptions of goods captured from the electronic invoices. The SCAN-NF presents and compares two models. The first is a single CNN. The second is an ensemble model built from two binary classifiers which had achieved the best performance.

For Data visualization purpose, Marinho and others presented a method for visualizing electronic invoices to support tax inspectors to detect suspicious cases of tax frauds using point placement strategies (Marinho et al., 2022). Their experimental results with ELEVEN validated that proposed method according to the visualizations' quality by introducing a case study which simulate the discovery of suspicious invoices considering a subset of selected products.

A framework for clustering short-text data in the NF-es using an automatic encoder to cluster data was proposed by (Schulte et al., 2022) so called ELINAC. It makes the task of clustering similar data by the short-text descriptions and improve anomaly detection in numeric fields.

## 7 CONCLUSIONS

This research emphasizes the importance of providing open access to large labeled data sets for the development of machine learning algorithms in economic related applications. Especially because there are a few data sets of this type and they are expensive to build. Similarly, the challenge faced by the tax authorities in verifying the correctness of the electronic invoice information have also been proved. It is worth noting

that the description text is not structured and there are no filling rule, but it is very small, with a maximum of 120 characters.

The ELEVEN Data-Set was introduced as data resources and a tool to improve the solutions to both problems. It can be used as a benchmark to enhance machine learning models for short text classification, and to improve tax inspection behavior on fraud detection in electronic invoices. According to the descriptive nature of the labeled records themselves, the prediction tasks could focus on the semantic meaning of the analyzed text.

Given the presented results, it was shown that the product descriptions indicated in the electronic invoices could be input into the machine learning models used for goods classification. To demonstrate that fact, this research shows that the BERT model has satisfactory performance for this kind of text mining task. Three successful case studies in literature were also reported to show the applicability of ELEVEN Data-set for economic related applications.

**Future works.** Beyond text classifications by supervised learning, the ELEVEN Data-set could be used for web scraping tasks as the descriptions sets could be reference for searching rules. It would be valuable to use ELEVEN in the pre-training task of investigating deep learning model. The fitness of data sets of other Latin languages could also be verified and measured.

## ACKNOWLEDGEMENTS

Our sincere gratitude to the Secretary of Economy of Brasília, especially the auditor Ary Silva Júnior, who, in addition to leading the data-set construction project, kindly shared the data for this study.

## REFERENCES

- Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.
- Alam, F., Offi, F., and Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAI conference on web and social media*.
- Ambika, P. (2020). Machine learning and deep learning algorithms on the industrial internet of things (iiot). *Advances in computers*, 117(1):321–338.
- Bhavani, A. and Santhosh Kumar, B. (2021). A review of state art of text classification algorithms. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1484–1490.
- Bibal, A., Lognoul, M., De Streel, A., and Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2):149–169.
- Bollacker, K., Cook, R., and Tufts, P. (2007). Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pages 1962–1963.
- Brazil, R. F. d. (2019). Ncm - nomenclatura comum do mercosul. url: <https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/classificacao-fiscal-de-mercadorias/ncm>.
- Buhrmester, V., Münch, D., and Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989.
- Chen, E., Lerman, K., and Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273.
- Codesso, M., de Freitas, M. M., Wang, X., de Carvalho, A., and da Silva Filho, A. A. (2020). Continuous audit implementation at cia. hering in brazil. *Journal of Emerging Technologies in Accounting*, 17(2):103–118.
- CONFAZ, B. (2013). Sistema integrado de informações econômicas e fiscais SINIEF. Ministério da Fazenda/CONFAZ convênio s/n de 15 de dezembro de 1970.
- da Rocha, C. C., Bouffleur, M. P., da Silva Fornasier, L., Narciso, J. C., Charao, A. S., Maran, V., Lima, J. C. D., and de Oliveira Stein, B. (2018). Sql query performance on hadoop: An analysis focused on large databases of brazilian electronic invoices. In *ICEIS (1)*, pages 29–37.
- de Aguiar, G. N. and Gouveia, L. B. (2020). The benefits’ program of electronic invoice as a tool to tackle tax evasion. *International Journal of Advanced Engineering Research and Science*, 7(11):391–400.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- dos Santos Neto, A. B., Batista, M. d. C. M., and Ferreira, T. A. (2022). Support decision system based on invoices data mining to estimate commercial pent-up demands. *Expert Systems with Applications*, page 117204.
- Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.
- Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.
- Hitzler, P., Bianchi, F., Ebrahimi, M., and Sarker, M. K. (2020). Neural-symbolic integration and the semantic web. *Semantic Web*, 11(1):3–11.

- Ji, L., Wang, Y., Shi, B., Zhang, D., Wang, Z., and Yan, J. (2019). Microsoft concept graph: Mining semantic concepts for short text understanding. *Data Intelligence*, 1(3):238–270.
- Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292.
- Kieckbusch, D. S., Geraldo Filho, P., Di Oliveira, V., and Weigang, L. (2021). Scan-nf: A cnn-based system for the classification of electronic invoices through short-text product description.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.
- Lake, B. M. and Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological review*.
- Lucena, L. F., de Menezes e Silva Filho, T., do Rêgo, T. G., and Malheiros, Y. (2022). Automatic recognition of units of measurement in product descriptions from tax invoices using neural networks. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 156–165, Cham. Springer International Publishing.
- Luo, Z., Huang, S., and Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Information Processing & Management*, 56(3):408–423.
- Marinho, M., Oliveira, V., Neto, S., Weigang, L., and Borges, V. (2022). *Visual Analysis of Electronic Invoices to Identify Suspicious Cases of Tax Frauds*, chapter ICITS 2022 Lecture Notes in Networks and Systems, pages 185–195. Springer.
- Maulud, D. H., Zeebaree, S. R., Jacksi, K., Sadeeq, M. A. M., and Sharif, K. H. (2021). State of art for semantic analysis of natural language processing. *Qubahan Academic Journal*, 1(2):21–28.
- McAuley, J. and Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. *International World Wide Web Conference Committee IW3C2*.
- Mendes Thame Denny, D., Ferreira Paulo, R., and Crespo Queiroz Neves, F. (2021). Technological alternative to tax compensation (icms credits): Case study of the feasibility of using dlt in the brazilian electronic invoice system. *Braz. J. Pub. Pol'y*, 11:520.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning-based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.
- Oh, S. J., Schiele, B., and Fritz, M. (2019). Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144. Springer.
- Pandolfo, L. and Pulina, L. (2021). Arkivo dataset: A benchmark for ontology-based extraction tools.
- Pintas, J. T., Fernandes, L. A., and Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, 54(8):6149–6200.
- Pipicano, E. F. M., Arias-Rojas, W., dos Santos, E. M., and Fonseca, A. P. (2021). Roterisation allocation for urban freight transport, using data from electronic tax invoices: Case study federal district of brazil. *Journal of Tianjin University Science and Technology*, 54(11).
- Raghavan, P. and El Gayar, N. (2019). Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)*, pages 334–339. IEEE.
- Schulte, J., Giuntini, F., Nobre, R., Nascimento, K., Meneguette, R., Weigang, L., Gonçalves, V., and Filho, G. (2022). Elinac: Autoencoder approach for electronic invoices data clustering. *Applied Sciences*, 12:3008.
- Shi, Q., Wang, Y., Sun, J., and Fu, A. (2018). Short text understanding based on conceptual and semantic enrichment. In *International Conference on Advanced Data Mining and Applications*, pages 329–338. Springer.
- Sinayobye, J. O., Kiwanuka, F., and Kyanda, S. K. (2018). A state-of-the-art review of machine learning techniques for fraud detection research. In *2018 IEEE/ACM symposium on software engineering in africa (SEiA)*, pages 11–19. IEEE.
- Sokol, K. and Flach, P. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.
- Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.
- Sugrim, S. (2020). *Robust models and evaluation for systems security research*. PhD thesis, Rutgers The State University of New Jersey, School of Graduate Studies.
- Tang, Z., Dai, D., Chen, Z., and Chen, T. (2022). Short text classification combining keywords and knowledge. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 662–665.
- Thangaraj, M. and Sivakami, M. (2018). Text classification techniques: a literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 13:117.
- Vieira, P. A., Pimenta, D. P., Cruz, A. F. d., and Souza, E. M. S. d. (2019). Effects of the electronic invoice program on the increase of state collection. *Revista de Administração Pública*, 53:481–491.
- Wu, W., Li, H., Wang, H., and Zhu, K. Q. (2012). Probbase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492.
- Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. (2018). The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Manage. Inf. Syst.*, 9(2).