Publicação: Fevereiro de 2019

Digital Object Identifier https://doi.org/10.33911/singular-etg.v1i1.19

Classifying Descriptions of Goods with Artificial Neural Networks

VINICIUS DI OLIVEIRA and
MARCELO LADEIRA.

Computer Science Department, University of Brasília-(UnB), P.O. Box 4466, 70910-900, Brasília–DF, Brazil
(e-mail:vinidiol@gmail.com, mladeira@unb.br)
Autor Correspondente: Vinicius Di Oliveira (e-mail: vinidiol@gmail.com).

ABSTRACT - The present study aims to evaluate the performance of an artificial neural network in the classification of merchandise descriptions indicated in electronic bills, legal document used to record all commercial transactions in Brazil. For this, a significant sample of the actual descriptions will be used as well as a overlook about the performance of the neural network with a KNN and a GBM algorithms forecasting the category of the merchandise each description refers. This paper brings a method for classifying descriptions of goods with Artificial Neural Networks. The descriptions are small non structured texts, maximum of 120 characters, relating to goods traded in commercial transactions.

KEYWORDS - Text mining, classifying, neural networks

I. INTRODUCTION

This paper proposes to study a way to classify goods indicated in electronic invoices issued in Brazil (NFe) in order to contribute to tax inspection by automating the analysis of these notes. The development of a computational method capable of identifying the tax classification to which a merchandise belongs in an electronic invoice (NFe) would bring a significant gain of efficiency to the tax administration. Artificial neural networks have been shown to be an effective tool in the classification of texts observed in the field of text mining. In the case studied, the good description is presented in the NFe in a string-type field limited to 200 characters, but free filled, so the ways of describing the same good vary substantially according to the NFe issuer.

The methodology adopted was CRISP-DM, as it is widely known and used in the academy as well as in the industry. A real database was provided by the tax administration of Distrito Federal/Brazil was made up of 15,000 NFe records, 5,000 from each classification group, regarding to invoices issued in 2017. At the end of the study, a satisfactory result in the classification (AUC = 0.95) was achieved, matching the methodology officially used by the state currently. The tax administrations of the Brazilian states have sought to improve their control processes in a substantial way, this study aims to contribute to the continuous search for improvement. In this sense, the development of an artificial intelligence capable of reading an invoice and indicates the correct tax treatment to be dispensed for that merchandise, this study aims to present the "starting point" in that reach.

II. STATE OF ART REVIEW

A. CRISP-DM

In the context of the literature studied, the CRISP-DM project structure proved to be the most robust in terms of use and easy understanding. The great permeability of CRISP-DM in market and academic environment makes its use more experienced and documented. Thus, for the purpose of this paper, we selected the CRISP-DM standard as more suitable for this data mining project.

According to Azevedo (2008) [1], CRISP-DM consists of a six-step cycle: 1) Understanding the business - understanding the objectives and requirements of the project from a business perspective, defining the problem and a preliminary plan to achieve the objectives; 2) Understanding data - collecting data and understanding it, identifying problems, seeking insight into relevant data and/or subsets to reveal hidden information; 3) Data preparation - activities to build the final data set; 4) Modeling - various modeling techniques are selected and applied and their parameters are calibrated; 5) Evaluation - the model obtained is evaluated in detail and revised to achieve the business objectives; 6) Deployment - usually not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge acquired must be organized and presented in a useful way for the user/client.

B. TEXT MINING

The text mining is a kind of knowledge discovery about texts via data processing, however, a different way of data mining in general due to the nature of unstructured textual data, Bezerra (2010) [2]. Allayari (2017) [3] says that this simple but not trivial activity is based on the analysis of frequency of terms in each document and the construction of tables with their attributes and values, thus enabling the use of supervised and unsupervised machine learning algorithms for processing this information.

According to Nishanth et al (2012) [4], text mining is an intensive process of knowledge in which a user interacts with a collection of documents using a set of analytical tools. Text mining algorithms operate on representations of attributes present in documents. Characters, words, terms, and concepts are the potential resources used to represent documents. Text mining often involves the prior conversion of unstructured content into structured content before applying the usual data mining techniques. The data in the databases may be lost due to data entry errors, system crashes at the time of data recovery, or various other reasons. Thus, text mining becomes an effective tool in classifying texts and detecting patterns, categorization, or errors.

C. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks are mathematical models inspired by the organic neural structure of intelligent living beings, which learn by experience, applied in computational techniques, Ninness (2013) [5]. With the rise of the importance of social networks and their impact on social and economic relations, the techniques of text mining have evolved systematically. In particular, as regards the mining of posts and comments, which are short and unstructured texts, artificial neural networks have stood out as a reference in mathematical modeling in the prediction of these data, Ali (2013) [6]. According to Guerrero et al (2017) [7], artificial neural networks (ANNs) can be used to classify the different concepts extracted in a semantic category.

III. UNDERSTANDING THE BUSINESS

A. SCENERY

All commercial transactions in the purchase and sale of goods in Brazil are obligatorily registered with electronic bills, whose data are stored in the state databases. By law [8], this issuance is done through a system controlled by the Treasury, all the data reported are available to the State. On the other hand, there is a representative aspect in these records that can distort them in relation to the real activities: The tax evasion and the errors of filling. Due to diverse cultural and socioeconomic factors, companies avoid taxes, therefore they could omit or alter information regarding their transactions in order to reduce in part or in a whole the tax burden.

B. DETERMINING THE APPLICATION'S OBJECTIVES

The verification of the tax due is done electronically in large masses of data. Such analyzes depend on reading the filled NFe fields. The field for product description, named "XPROD", is the most reliable because it is printed and the goods purchaser usually checks it for compliance reasons or guarantee terms. The XPROD field is a string type, up to 120 characters, with no padding pattern, so its reading and classification becomes unfeasible by traditional means of data crossing. In order to solve the problem of merchandise identification, it is used the field called NCM (Mercosul Common Nomenclature Product Code) as pattern, since it is a standardized 8 digits field (e.g. "28470000" - type: integer). However, it is also verified in the completion of these field, albeit rarer, an NCM that is totally different from the nature of the product described, such as in a observed case of "hydrogen peroxide", NCM "28470000", that was registered as NCM "01012100"which in the Mercosul pattern refers to "Live horses, asses and mules - Purebred breeding animals."

C. PURPOSE OF APPLICATION

- Given the above is defined as a proposal of this work to solve the following problem: "Is it possible to classify goods by text mining from an electronic bill through the Description field?"
- It is important to note in this study that there is a difference between this text mining and the others verified in relation to its prediction characteristic, not of understanding or identifying the subject matter to which the text is treated, but rather of an objective categorization through the description of the object;
- And there is still a need for high accuracy in the prediction.

D. APPLICATION SUCCESS CRITERIA

The criterion to be used to measure the success of the application will be the accuracy in the classification of the goods descriptions. The method currently used by the Department of Finance of Brasília, which uses the NCM field has an accuracy of 0.95 accuracy, so for the implementation of the proposed system this should overcome this assertiveness.

IV. UNDERSTANDING DATA

The database provided by the Transit Goods Inspection Bureau contains the record of a 15,000 (fifteen thousand) rows by 3 (three columns) worksheet. There are 5,000 records classified for each category. The visual representation of the worksheet is shown in Table 1.

- "XPROD" - Description of the product in NFe, type text, string of 120 characters. The texts has no standard of filling. Is a field of free registration by the issuer of the note, how to write or describe a product is free, there is no criticism or rules in the NFe issuance systems;
- "NCM" - The NCM code, 8 digits (integers numbers);
- "Item" (38, 39 or 40) - a reference to the tax legislation indicating the taxation form, which varies according to the item. "38" for cosmetics, "39" for cleaning material and "40" for foods.

Tabela 1. Sample Database Table

XPROD	NCM	Item
- REF.: (01254) - PENTE PROMO - SENS	96151100	38
0037540 ST DEO COL BLUE OCEAN	33030020	38
0040002 ST DEO COLONIA PINK DIAM	33049910	38
0041416 ST RIMEL EXTRA VOLUME	33042010	38
0041436 ST 8 SABONETES CORACAO	34013000	38
...	...	...

V. DATA PREPARATION

In this work the data preparation has the main focus to formatting the XPROD fields. As that is a text-type field, free and no padding pattern, it should be treated for use in modeling. The R platform was used in conjunction with the R Studio interface to carry out the work developed in this project. The database in a Excel spreadsheet was imported to R and a data base vector was created.

The NCM column has been deleted, the tax class columm (Item) remained, 38 for Cosmetics, 39 for Cleaning Materials and 40 for Foods. In the "Item"column the numbers 38, 39 and 40 have been replaced respectively by the corresponding names "cosmeticos", "matlimpeza"and "alimentos"as that will be used as the classification parameter.

In the XPROD field the accent, punctuation, numbers, double spaces and special characters were removed. All text has been converted to lowercase. The stop words were also withdrawn. The “Item” column was renamed to “Category” and the “XPROD” column to “Text”. The Table 2 shows how the database was after treatment. Such operations were made in R.

Tabela 2. Database after treatment

Text	Category
ref pente promo sens	cosmeticos
st deo col blue ocean	cosmeticos
st deo colonia pink diam	cosmeticos
st rimel extra volume	cosmeticos
st sabonetes coracao	cosmeticos
...	...

VI. MODELING

In order to better evaluate the performance of artificial neural networks, the k-Nearest Neighbor - KNN and Gradient Boosting Machine - GBM prediction models will be evaluated in a comparative way. Therefore, some manipulations at the base of the model are still necessary. The inclusion of the "Category"column in the data frame as a reference for the test and training classification. The index is the lines order maintained until then.

For a clear overview at the modeling possibilities, the Word Clouds were made for each of the three categories. This is a visual resource that plots the most recurring words in classes where their size print is proportional to their frequency. The result is shown in Figures 1, 2 and 3.

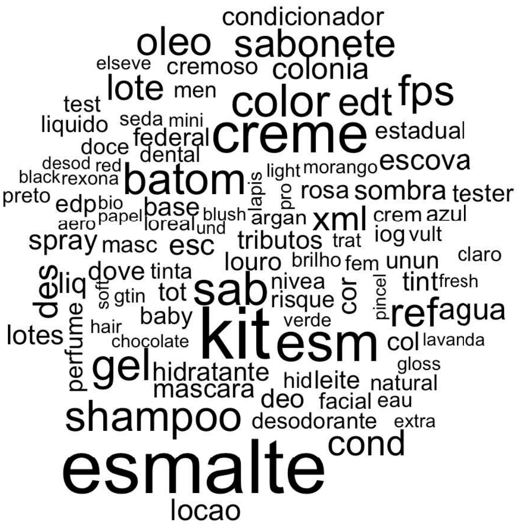

Figura 1. Word cloud for Cosmetics (*cosmeticos*).

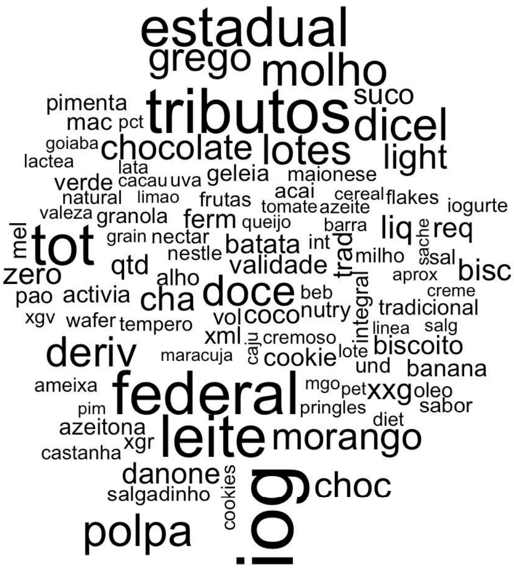

Figura 2. Word cloud for Foods (*alimentos*).

A. ALTERNATIVE MODELS - KNN AND GBM

For the construction of the training and validation bases a reference column is inserted with the correct classification for each description. The training base has a radon sample of 0.7 factor and the validation base has the remaining data, 0.3 fraction of the total initial base.

the KNN model was set in the native default form of the package used in R language (library:“class”). The GBM model was set with 1,000 trees and depth of 3. This arrange was founded by a test combination trial: 300, 500, 1000 and 1500 trees, crossing 3, 5 and 7 depth. The H2O package for R was loaded for that task (library: “h2o”). The Confusions Matrices resulted is shown in Tables 3 and 4.

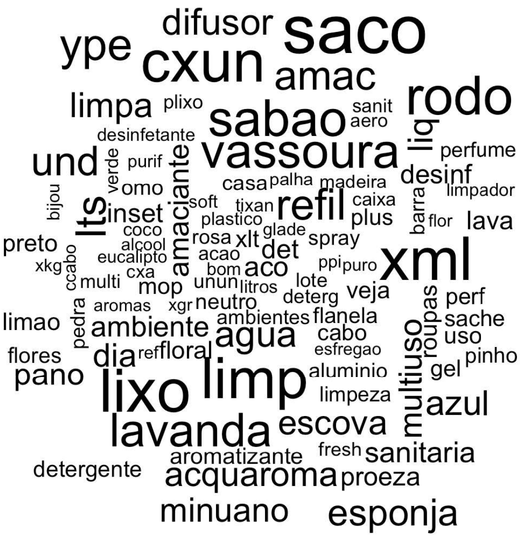

Figura 3. Word cloud for Cleaning Materials (*matlimpeza*).

Tabela 3. Confusion Matrix - KNN Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	968	20	4	992
cosmeticos	49	825	32	906
matlimpeza	41	137	924	1.102
Total	1.058	982	960	Acc: 90,56%

B. ARTIFICIAL NEURAL NETWORKS MODEL

The training and validation bases of the artificial neural network model are the same used for the GBM model. This can shows a more effectively comparison of performance between the two models. A network with 3 layers of artificial neurons was dimensioned, being 100 in the first layer, 200 in the hidden layer and 100 in the output layer. The epochs were 20 (twenty). This set was determined by testing several variants of these configurations, where layers of 50, 100, 200 and 300 neurons were disposed in 2, 3 and 4 layers, as well as training cycles, or epochs, of 20, 40 and 60 .The set chosen was the one with the lowest error rate in predictions. The neural network confusion matrix was generated and its results are expressed in Table 5.

Based on these results, the superior performance of artificial neural network - ANN can be seen at looking to the KNN and GBM models in this specific case of text mining. At least, as regards the prediction accuracy of category described goods, the ANN model showed success rate higher than the other two models, although it is important to observe the tendency of greater error in the KNN for the "matlimpeza"(cleaning materials) category, GBM and the ANN models obtained a more uniform error between the categories as observed in their confusion matrices.

The R^2 statistical term of the ANN and GBM models can also be compared for a more accurate assessment. The GBM model obtained an R^2 of 0.9418, whereas the ANN model obtained R^2 = 0.9701. The parameter in the initial purpose of this study, set out in item III, D, as 0.95 success rate in predicting came close to being hit with only the table used in the ANN model, which presented a sample of 5,000 descriptions for each category of merchandise. With the construction of a new base model with a more representative sample, the currently success rate can be overcome.

Tabela 4. Confusion Matrix - GBM Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	878	103	13	994
cosmeticos	15	934	25	974
matlimpeza	10	120	867	997
Total	903	1.157	905	Acc: 90,35%

Tabela 5. Confusion Matrix - ANN Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	945	43	6	994
cosmeticos	16	939	19	974
matlimpeza	3	75	919	997
Total	964	1.057	944	Acc: 94,54%

VII. CONCLUSIONS

In view of the results presented, it was found that the classification of goods can be made by the descriptions given in the electronic bills through artificial neural networks. This study shows a satisfactory performance of the ANN algorithm for this specific text mining task. It is noteworthy that the text is not structured and has no fill rule, but it has small size, a maximum of 120 characters.

The classification of this type of text has its differential highlighted in relation to textual ratings commonly seen in the literature as well as in technical reports widely available on the Internet, is descriptive of the goods itself, not the meaning or significance of the analyzed text . When the most frequent need for cases of text mining shows the classification of the meaning of the text verified, in this issue the goal is the straight classification of merchandise which is being described in the text.

This type of categorization of goods can be expanded in future studies to more classes or even for all 99 classes listed in the Mercosul Common Nomenclature - NCM, which would help the FISCO identify filling errors, tax classification or even fraud.

Referências

[1] A. Azevedo, M. F. Santos, KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW, IADS European Conference Data Mining, 2008.

[2] E. Bezerra, R. Goldschmidt, Classification in Text Mining. journal Sistemas de Informação, n 5, 2010, pp. 42-62.

[3] M. Allahyari, P. Seyedamin, A. Mehdi, E. D. Trippe, J. B. Gutierrez, K. Kochut, A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. KDD Bigdas, Halifax, Canada, 2017.

[4] K. J. Nishanth, V. Ravi, N. Ankaiah, I. Bose, Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications, vol. 39, pp. 10583-10589, 2012.

[5] C. Ninness, R. Marilyn, C. Logan, L. David, J. T. Lacy, S. Halle, R. McAdams, S. Parker, D. Forney, Neural Network and Multivariate Analyses:

Pattern Recognition in Academic and Social Research, Behavior & Social Issues, vol. 22, 2013, pp. 49-63.

[6] M. P. J. Ali, and N. M. S. Surameery, A. R. M. Yunis, Gender Prediction of Journalists from Writing Style, ARO. The Scientific Journal of Koya University, vol. 1, 2013, pp. 22-28.

[7] J. I. Guerrero, C. Len, I. Monedero, F. Biscarri, J. Biscarri, Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection, Knowledge-Based Systems, vol. 71 2014, pp. 376-388.

[8] Sistema Integrado de Informações Econômicas e Fiscais- SINIEF. Ministry of Finance, Convênio s/n, National Council of Finance Policy CONFAZ, 1970.

VINICIUS DI OLIVEIRA Possui graduação em Engenharia Civil pela Universidade Federal de Goiás (1997). Especialista em Modelagem Estatística e Dinâmica Econométrica pela UnB, mestrando em Computação Aplicada na UnB e Atualmente é Auditor Fiscal da Receita do DF - Secretaria de Estado de Fazenda do Distrito Federal. Tem experiência na área de Fiscalização Tributária, atualmente com ênfase em Gerenciamento de Risco de Projetos.

MARCELO LADEIRA Possui graduação em Engenharia Mecânica pela Universidade de Brasília(1976), especialização em Curso Avançado de Operação de Sistema Hidrotérmico pela Universidade Federal do Rio de Janeiro(1981), mestrado em Análise de Sistemas e Aplicações pelo Instituto Nacional de Pesquisas Espaciais(1981), doutorado em Computação pela Universidade Federal do Rio Grande do Sul(2000) e pós-doutorado pela Universidade de Lisboa(2004). Atualmente é Professor Adjunto 2 da Universidade de Brasília, Membro de Comitê Consultivo do Ministério da Ciência, Tecnologia, Inovações e Comunicações, Revisor de periódico da Journal of Information and Data Management - JIDM e Membro de comitê assessor do Conselho Nacional de Desenvolvimento Científico e Tecnológico. Tem experiência na área de Ciência da Computação, com ênfase em Metodologia e Técnicas da Computação. Atuando principalmente nos seguintes temas:Inteligência Artificial, raciocínio probabilístico, diagramas de influências, redes bayesianas, Teoria da decisão e Representação de conhecimento incerto.

APPENDIX

R coding for the ANN and the GBM models.

```
set.seed(321)
library(readxl)
library(tm)
library(h2o)
h2o.init()
df <- read_excel("yourdatabase.xlsx")
docs <- Corpus(VectorSource(df\$Text))
docs <- tm_map(docs,
content_transformer(tolower))
docs <- tm_map(docs,
removeNumbers)
docs <- tm_map(docs,
removeWords, stopwords("pt"))
docs <- tm_map(docs,
removePunctuation)
docs <- tm_map(docs,
stripWhitespace)
dtm <- DocumentTermMatrix(docs)
mat.df <- as.data.frame(data.matrix(dtm),
stringsAsfactors = FALSE)
mat.df <- cbind(mat.df, df\$Category)
colnames(mat.df)[ncol(mat.df)]
<- "category"
mat.df.h <- as.h2o(mat.df)
data.split <- h2o.splitFrame
(data = mat.df.h, ratios = c(0.7, 0.2),
seed = 1234)
data.train <- data.split[[1]]
data.valid <- data.split[[2]]
data.test <- data.split[[3]]
myY <- "category"
myX <- setdiff(names(data.train),
c(myY, "ID"))
gbm.model <- h2o.gbm(myX, myY,
training_frame = data.train,
validation_frame = data.valid,
ntrees = 1000,
max_depth = 3,
model_id = "gbm_xprod_5mil")
conf.mat <- h2o.confusionMatrix
(gbm.model@model\$validation_metrics)
write.table
(conf.mat, file="confmat_gbm_5mil.csv",
sep=";")
r2.gbm.model.5mil <-
gbm.model@model\$validation_metrics@metrics\$r2
dl.model <- h2o.deeplearning(myX, myY,
training_frame = data.train,
hidden = c(100,200,100),
epochs = 20,
validation_frame = data.valid,
model_id = "dl_xprod_5mil")
conf.mat.dl5mil <- h2o.confusionMatrix
(dl.model@model\$validation_metrics)
write.table(conf.mat.dl5mil,
file="confmat_dl_5mil.csv", sep=";")
r2.dl.model.5mil <-
dl.model@model\$validation_metrics@metrics\$r2
```