Chapter 78

Automatic classification of products marketed by public agency of Rio Grande do Norte through a committee of classifiers

10.56238/tfisdwv1-078

Elvis Rafael Ferreira Dias
Federal University of Rio Grande do Norte, Brazil
E-mail: elvisnaopresley@gmail.com

João Carlos Xavier Jr
Federal University of Rio Grande do Norte, Brazil
E-mail: jcxavier@imd.ufrn.br

ABSTRACT

The use of text mining techniques has increased considerably in recent years due to the large amount of text information being produced and stored by electronic systems and the need to make this data information for organizations. In this context, the Court of Auditors of the State of Rio Grande do Norte (Tribunal de Auditors do Rio Grande do Norte, TCE-RN) receives daily a large amount of electronic invoices containing data of product's purchases that need to be analyzed for the society's benefit. Still, thief documents allow free filling, often erroneous, of some data by the sellers who issue the invoices. This way, the documents do not come to follow a pattern and make it possible to carry out analysis in a practical and efficient way through common tools for obtaining and filtering data. Therefore, there is a need for automated processing in order to standardize the data, make them available quickly and enable their use as information for audit purposes. So, this work presents a solution based on text mining and machine learning techniques for the problem of identifying commercialized products in the state of Rio Grande do Norte from the description field of Electronic Invoices as a way to enable the classification of these products into unique products

Keywords: Text mining, Electronic invoices, Data processing, Machine learning

1 INTRODUCTION

The Electronic Invoice (NF-e) is a document of existence only digital, issued and stored electronically, with the purpose of documenting for tax purposes an operation of movement of goods or provision of services occurred between the parties. Its legalvalidity is guaranteed by the digital signature of the sender (guarantee of authorship and integrity) and the Authorization of use provided by the Tax Office, before the occurrence of the generating fact (Ministry of Finance [MF], 2020a).

As stated by the Department of Taxation of Rio Grande do Norte - SET-RN (2020), this document, established by Adjustment SI-NIEF 07/05 of March 30, 2005, aimed to implement a national model of electronic tax document aimed at replacingmodels 1 and 1A on paper. Among the benefits, it came to simplify the ancillary obligations of taxpayers and allow the monitoring in almost immediate time of commercial operations by the Tax Authorities.

Based on the mandatory issuance of the Notes in economic activities involving public agencies (MF, 2020d), SET-RN, responsible for the receipt and storage of state notes, signed an agreement with the Court of Auditors of the State of Rio Grande do Norte (TCE-RN) in 2019 to pass on NF-e related to the courts jurisdictional to the TCE-RN, allowing internal interest analysis.

Among the various information present in nf-e (MF, 2020b), the initial motivation of the TCE-RN leaves forthe analysis of prices performed on a daily basis, especially the calculation of the reference price for the various products marketed in the state. The reference price allows the various agencies to research prices, a task that, as stated in federal decree No. 7,892/2013 item IV of Article 5, is up to each body to carry out its periodic market research as a prior and indispensable procedure for a contractual claim (Brasil, 2013b). Furthermore, as established in Judgment 169/2013 of the Court of Auditors of the Union, it was clarified that the absence of research that adequately represents market prices, in addition to constituting an affront to the case-law of the Court of Auditors, may be in favour of contracting services or acquiring goods at prices higher than practices by the market, thus hurting the principle of economics, according to the understanding of TCU 1785/2013 – Plenary (Brasil, 2013b; Silva, 2014).

To allow price search for unique products, these need to be grouped together so that assim their prices can be compared. Through the barcode, technology responsible for product categorization, ideally all products would have a way of being identified only, however, this task is not possible in a way due to the recurrent practice of issuing these documents filled out negligently or erroneously (GS1, 2019). Thus, the task of identifying unique products falls on the interpretation of the text field that describes the products.

In view also of the large amount of purchases made by the state periodically and the use of this information be focused on *Business Intelligence panels*, the analysis made by the agency should be ideally automated and close to real time. Therefore, the process ofutomatizing product identification requires an intelligent solution that can learn different patterns based on the descriptions of the products themselves.

Therefore, due to the limitation of the interpretability of textual descriptions automatically, this paper proposes a methodology for handling and controlling these texts as a way to group commercialized products. Textual processing methods will need to be used to extract crucial information from the essential productsfor the classification slap performed by Classifier Committees.

The rest of this article is divided into 6 sections. Section II describes some important theoretical concepts while section III presents some works related to the theme of this artigo. In Section IV, the methodology used in the development of this work is presented, while an analysis of the results is presented in Section V. Finally, Section VI presents the final considerations and future work.

2 THEORETICAL REFERENCE

This section discusses the concepts most relevant to the development of this work, starting from the raw data, through textual processing, to finally arrive at the classification of products.

2.1 DEFINITION OF DATA

Among the fields that carachterize the products in an electronic Invoice, illustrated in Table 1, there are three fields related to identification: "NCM", "cEAN" and "xProd". The first is tied to taxation, grouping products similar to the same code (MF, 2020c), while the second (barcode) may not exist for certain products (GS1, 2020), being also free to fill and null in more than 70% of the notes analyzed. Finally, the field "xProd", which even inaccurate due to free typing, describes the products.

Table 1 : Fields containing characteristics of products marketed in an Electronic Invoice

Attribute	Description
NCM	Mercosur Common Name, a single eight-digit numerical code that groups goods traded in foreign trade operations in Mercosur.
cEAN	GTIN (*Global Trade Item Number*) of the product, old EAN code or barcode.
xProd	Textual description of the product or service.
uCom	Commercial Unit, informs the unit of measurement of the product, for example, liters and kilos.
qCom	Commercial Quantity, specifies how many products are being sold for each item marketed in the note.
vUnCom	Unit Value of Commercialization, indicates in reais the unit value of each product described in the note.
VProd	Total gross value of productor services.

Source: Ministry of Finance, 2020b.

2.2. NATURAL LANGUAGE PROCESSING (PLN)

It is an area of study responsible for manipulating human language, *natural language understanding* (NLU) and among its specialties, has PLN tools capable of extracting text information and enabling learning of these. Technical solutionsfor minimizing noise problems, for example, seek to remove useless textual terms for learning, while the transformation of words into numerical data, usable into algorithms, is done through word encoders.

Word encoders are responsible for transforming linguistic terms into numbers, enabling the use by algorithms. Among the types of representations for text, *stand out Count Vectorizer* and *Word Embeddings*; the first is oriented by the occurrence of words, being more used for classification, while the second is oriented by textual characteristics, capturing semantic sense and more used in the analysis of feeling and translation.

The Count Vectorizer, in addition to counting occurrences of words, can also analyze sentences by counting sequences *of n* words (n-gram), rather than isolated words, leading to a learning of terms together. The *Term Frequency-Inverse Document Frequency* (TF-IDF), defined by Formula 1, ensures that terms such as articles and prepositions, for example, that do not carry context information and often appear in texts, become irrelevant compared to the more frequent terms that carry more specific meaning to samples. *Tf(t,d) being* the frequency of the term *t* in document *d*; *n*, the total number of samples; *and df(t),* the document nude containing *t* (Buitinck et al. 2013).

$$tf\ idf(t,d) = tf(t,d).idf(t) \qquad (1)$$

Where

$$idf(t) = log\frac{1+n}{1+df(t)} + 1 \qquad (2)$$

2.3. MACHINE LEARNING

According to Mitchell (1997), "The ability to improve performance in the realisationof some task through experience", defines the class of computer programs that can learn and make up the machine learning area. These algorithms (also called models) can learn how to induce a function or hipothesis capable of solving a problem from data that represent instances of the problem to be solved (Faceli, Lorena, Gama & Carvalho, 2011).

In AM, the goal is to make algorithms generalize their learning, through a set of training data, in test data never seen before —a characteristic of learning (Goodfellow, Bengio & Courville, 2016). The classification problem can have solutions based on different forms of learning, being groupedinto: probabilistic methods, based on distance, search and optimization (Faceli, Lorena, Gama & Carvalho, 2011).

The learning of these algorithms can be measured through different performance metrics, such as accuracy, f1-score and ROC curve, by exemplo. In addition to accuracy, more intuitive, there are more robust ways to evaluate the performance of a model, taking into account *accuracy and accuracy (recall*). The metric called "f1-score" (or f-measure), Equation 3, measured through positive verdadeiros (TP), false positives (FP) and false negatives (FN) the relationship between the percentage of samples correctly classified as their class, by the percentage of the actual amount of samples that were correctly classified (Buitinck et al. 2013).

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad (3)$$

Knowing that there is no algorithm that is better for all classification problems, one possible strategy is to group sets of predictors whose individual decisions are combined or aggregated in some way to predict new examples. The main idea behind these methods, therefore, is slumed up in the fact that different machine learning techniques explore different search spaces and hypothesis evaluation functions, enabling them to work together to achieve possible better results (Faceli, Lorena, Gama & Carvalho, 2011).

Although the committee voting process may cause performance loss due to the presence of models withunsatisfactory apren said among the group, there are solutions based on committees with layers, for example, Stack or Cascade, in which overall performance is guaranteed by the best results. The Stack Generalization method, for example, has an architecture in camadas where there are levels of predictions, with different base classifiers. Zero-level classifiers receive the original data as input, and each classifier produces a prediction that will serve as input for the immediately successive layers. A singleclassifier at the highest level produces the final prediction, responsible for learning from the combination of the predictions immediately below (Faceli, Lorena, Gama & Carvalho, 2011).

Classifier committees have several possible architectures with provenbetter commitment compared to base classifiers. *Bagging*, for example, which trains models with modifications to the training set, is one of the most efficient ways to improve the accuracy of classifierssuch as cyst trees and neural networks. Conversely, Faceli, Lorena, Gama and Carvalho (2011) state that stack generalization still presents mystery about which conditions are best succeeded.

Stacking, also known as stack generalization, is a type d andcommittee that aims to minimize prediction error from classifiers in the higher layers that learn the type of error made from the classifiers immediately below. The rule of higher level classifiers, therefore, is to learn fromthe previous classifiers make mistakes, in which class they agree or disagree, and use their knowledge to make predictions (Faceli, Lorena, Gama & Carvalho, 2011).

Furthermore, as Faceli, Lorena, Gama and Carvalho (2011) explain, committeescan be applied with voting methods and serialization methods. Among several existing variations, uniform voting, for example, takes into account the opinion of all basic classifiers equally for the final classification. On the contrary,the s serialization methods present an improvement under the uniform method, because each classifier generates a probability that the example belongs to each class, rather than a label only. Serialization methods are useful for prediction among samples of unknown classes, because the model will not attempt to assign a class to it, instead it will return a low probability for all known classes, leading to a possible disposal of the example.

Thus, this work will conduct an empirical analysis on the performance of some base classifiers, as a way to choose the most appropriate components for the different committee architectures that will be analyzed through the serialization method.

3 RELATED JOBS

The challenge of classifying sentences using the descriptions in electronic invoices has been investigated by public agencies in Brazil, with some studies published on the Internet. Those reported here have a quantity of information limitedto that available via online search. The little variety and technical availability of the projects described in this section made it impossible to analyze the results and challenges when measuring whether they would serve the objective of the TCE-RN.

The public price bank divulgado by the ECA of Minas Gerais, for example, uses purchasing notes made by the state's public agencies to generate price analyses practiced. In this solution, instead of classifying descriptions into product classes, all database descriptions are grouped based on the similarity of these with user-informed keywords describing the product. The web system, open for public use, requires the user to decide among all the descriptions returned, which should enter or not in their price analysis, performing a process of classification in time of use. This approach has the advantage of working with error rate depending only on the user and there is no limitation in the number of products used, but the user experience of the system can be exhaustive in case of multiple products containing the same keyword (Court of Auditors of the State of Minas Gerais [TCE-MG], 2020).

The Court of Auditors of the State of Paraíba [TCE-PB] (2020a) also launched a web system for analysis of NF-es, entitled "Price Panels", related to purchases made by municipal and state administrations of food and fuel products (about twenty products in total). Based on AM algorithms, the team states that the system, ainda in the initial phase, mines millions of product descriptions through key classification and validation terms related to the different types of expenses with fuels and food, however, with the use of the system it was seen that products with different marketing formas are not differentiated from each other, for example, "Whole milk powder" has a single price (TCE-PB, 2020a).

Another system, made by the same team assigned to the TCE-PB, called "Hour Price", uses another type of Note, but countingsimilar information, Electronic Consumer Invoices (NFC-e). This solution is more focused on solving the problem of processing large amounts of data, seeking to make online note visualization available quickly and thus make it easier for the population to purchase local products; there are no tools for analyzing invoices for the purpose of *generating insights*. This approach, similar to that of the TCE-MG, has the same problem of listing all occurrences of the keywords informadas, requiring the user to classify the returned products in time of use (TCE-PB, 2020a).

In the Court of Auditors of the state of Rio Grande do Sul (TCE-RS), according to (Gandini, 2020), there was also an attempt to create a bank of prices to identifyoverprice in bidding processes in the State. As reported in its public repository, the approach to classifying descriptions used descriptive machine learning, seeking to obtain product descriptions grouped bythe algorithm. Even claiming to have arrived at good results and being used internally, however no details of the use are specified. The solution, very interesting, was tested and proved unfeasible for the use of TCE-RN due to the alta granularity of the required classes, its instability and because it has a heavy maintenance process (Gandini, 2020).

Finally, this problem was also addressed in a master's thesis as a case study of a distributed platformof m data inection for *Big Data*. The author applies predictive learning techniques, through a set of data to classify beverages in water, beer, distillate, soda, juice or miscellaneous. Nine attributes were used to trainAM models, among them " NCM", "qCOM", "vUnCom" and fields extracted from "xProd" such as Brand, Type, Packaging and Volume. The applied data engineering made it possible to have for each sample all the characteristics necessary to characterize a unique product, such as: "Brand: coca, Type: glue, Packaging: can, Volume: 350, Quantity: 1". However, the work was limited to a data set of 3,000 samples and to classify products not as "Coca-cola 350ml", but as a large refrigerant groupand (dos Santos, 2018).

The solution proposed here has the differential of seeking to achieve a high degree of accuracy by automatically classifying a wide variety and quantity of products, which can have their prices compared. Using only the field of descrição, investigate feasibility of AM tools to enable integrated use of the automated data flow, which partly from the receipt of notes via SET-RN, until reaching the interactive panels for business intelligence.

4 EXPERIMENTAL METHODOLOGY

In this section we describe the steps that constitute the methodology used to validate our hypothesis of classifying textual descriptions from applied textual processing. Starting for an experimental research, descriptions of someproducts were chosen as variables to analyze the problem, construct hypotheses and through their manipulations and different models of AM, evaluate the result with the hypothesis that models are able to classify the texts correctly (Koche, J. C., 2011).

4.1 PRODUCT PRE-CLASSIFICATION

The invoices available for this work are data that were previously manipulated from xml format to tabular format, and that refer to the period between Sep/2019 and Nov/2020. From this data, it was possible to create a subset of products with classes (labels) analyzed singularly from SQL queries made in the database.

The choice of the products that make up this subset took into account the large number of available instances (most marketed products) from key terms shared by descriptions. Thus, the assignment of classes was based on an analysis of the original and unique descriptions, drastically reducing the number of records to classify.

Dentre all characteristics that a product can have as a factor of change in price, were used those that are independent of brand, color, flavor or other perfumery for the same product. The pattern therefore sought consists of base product + possible feature + size / packaging, for example, "Refined Sugar 1kg", resulting in more than twenty thousand labeled samples and more than 400 unique products.

The price difference can often be negligible for products such as "1kg Refined Sugar" and "1kg Crystal Sugar", however, some products have been labeled in this more granular way to allow for a broader analysis. In other cases, for other products such as "Rice 1kg" and "1kg Beans" the specific types were maintained as belonging to a single class.

In this process, an accurate analytical capability is important to prevent the classifier from knowing only a few more popular versions when some products can be described in different ways. The product "Deodorizedair r" can also be called "Good Air" and "Air Odorizer ". Similarly, ethyl alcohol products may have their percentage of alcohol expressed based on °INPM or °GL to specify the same product.

Finally, incomplete descriptions, which puta large portion of the nf- data and, for example, "Gel Alcohol", were ignored. Even though in cases price change factors such as size and/or percentage of alcohol, allow inferences, these samples were left out aware of the limitation caused in the knowledge of the committees.

4.2 PRE-PROCESSING OF PRODUCT DESCRIPTIONS

Preprocessing product descriptions involves starting from a sentence in English, in free format, leaving it in a pattern that can be understood by the AM models. In this process, standard and specific textual cleaning techniques for this problem were applied together with pln, stemming, *steeming and* coding techniques.

4.2.1 Textual Cleansing

Knowing that the descriptions of the products have, due to their nature of free writing, infinite possibilities to be represented, the cleaning applied aims to reduce noise and ambiguities in general problematic terms and make the descriptions more genéricas and informative, maintaining the balance and integrity of the classes. Below are listed the cleaning steps that were applied to the original descriptions:

1. Removal of useless words, leading zeros, excess blanks, accents and symbols -with the exception of decimal numbers;
2. Removal of specific patterns, such as batch, address, validity , etc.;
3. Application of tiny in all letters;
4. Number separation next to the letter;
5. Standardization of units of measure and other measurement terms;
6. Number join with units of measure;
7. Combination of compound words describing products such as "Milk powder" for "Leitepo";
8. Removal of product dimensions and purchased quantities;
9. Removal of numbers from 5 digits, loose numbers without units, if you have a number with drive in the same description and numbers at the beginning of the descriptions;
10. Removal of loose letters if they are not units;
11. Removal of size one or empty descriptions;
12. Limitation ofs descriptions in 8 words.

Due to the continuous flow of data, the step of cleaning it needs to remain constantly updated and revised, even if there is no change in the classes of the products. New ways of describing them may arise, leading to confusion between classes, in the same way that descriptions can also change due to the emergence of new brands and/or packaging.

4.2.2 Encoding Descriptions

The encoding used (*Count Vectorizer*) was defined in order to count occurrences of words using *n-gram* range (2.4), that is, the model interprets sequences from two to four words in a row. For this interval value it was sought to generalize the amount of words cappeasees of identifying unique products in general, not taking into account products with description of a word. The choice of n-gram still directly influences the textual cleanings applied to the descriptions, and it should be sought to reduce repeat-rate sequesters from two words that do not characterize a single product, such as "1 liter".

4.3 CHOICE OF PRODUCTS

Considering the unfeasibility of using all products categorized in the database, it was decided to make a selection ofa limited amount of products, aiming to test the methodology presented here. Thus, it was oriented by selecting products with sufficient quantity and variability to apply in the AM models.

Also, due to the parameters *used in Count Vectorizer* (n-gram [2,4]), some products such as fruits and vegetables were excluded because they presented standard descriptions composed of only "Base Product" + "Common Feature". Because they do not present unique characteristics in the products, they end up having characteristics shared by other fruits and vegetables or unique descriptions.

Finally, among the more than four hundred products initially labeled, the ones that best fit the criteria already addressed were selected. Table 2 shows the 74 distinct products that will be part of the experiments.

Table 2: Product classes labeled and selected for experiments

1	Chocolate powder 400g	20	Tea 10g	39	LPG gas 13kg	58	Ream Paper A4 500fls
2	Sweetener 100ml	21	Cement 50kg	40	Condensed Milk 395g	59	Soap Powder 500g
3	Flavouring 360ml	22	Colorau 100g	41	Milk powder 200g	60	Soap in Bar 200g
4	Flavouring 400ml	23	Cup Desc. 150ml	42	Whole milk powder 200g	61	Plastic Bag 100lt
5	Rice 1kg	24	Cup Desc. 180ml	43	Whole Milk 1lt	62	Plastic Bag 150lt
6	Crystal Sugar 1kg	25	Cup Desc. 50ml	44	Clean glasses 500ml	63	Plastic Bag 200lt
7	Refined Sugar 1kg	26	Beef Rib	45	Polish Furniture 200ml	64	Plastic Bag 30lt
8	Milk Drink 1lt	27	Chicken Thigh/ Thigh	46	Sterile Surgical Glove	65	Plastic Bag 40lt
9	Bisc. Cream Cracker 400g	28	Sour Cream 200g	47	Graphite Pencil	66	Plastic Bag 50lt
10	Bisc. Maizena 400g	29	Disinfectant 1lt	48	Spaghetti Noodles 500g	67	Plastic Bag 60lt
11	Bisc. Maria 400g	30	Disinfectant 2lt	49	Margarine 500g	68	Refined Salt 1kg
12	Coffee powder 250g	31	Disinfectant 500ml	50	Dish Cloth	69	Brick 8 holes
13	Hydrated Lime 5kg	32	Disinfectant 5lt	51	Toilet paper with 4 rolls	70	Vinegar 500ml
14	Ballpoint pen	33	Detergent 500ml	52	Toilet paper with 64 rolls	71	Mineral Water 20lt
15	Text Mark Pen	34	Double-Sided Sponge	53	Paper Towel with 2 rolls	72	Bleach 1lt
16	Beef	35	Cornflour 500g	54	Chicken Breast	73	Ethyl alcohol 70% 1lt
17	Charque meat	36	Wheat Flour 1kg	55	Fruit Pulp 1kg	74	Soybean Oil 900ml
18	Sun Meat	37	Beans 1kg	56	Fruit Pulp 400g	-	-
19	Ground beef	38	Whole Chicken	57	Refrigerant 2lt	-	-

Source: Authors.

4.4 SEPARATION INTO TRAINING AND TEST SETS

Data separation into training and test sets followed the non-replication of data approach (without resampling), so that there was no intersection between the sets. Therefore, a random stratified separation of 75% of samples for training, 20% for tests and 5% for validation were proposed.

Finally, the validation set also aggregates noisy samples, remaining from manual labeling. Due to the real need for product analysis to involve the committeesdistinguishing known from unknown products, 5% of the total random non-categorized samples were applied to simulate the behavior of the classifiers in this distinction.

4.5 CLASSIFICATION MODELS

The methodology for understanding which models are appropriate and work best among the many available, was *based on cross-validation by* k-fold *cross validation*. For the trainings, the descriptions of the products were applied in their original distribution (with *repetitions), oversampling* and cross-validation *by stratified k-fold*.

The f1-score *metric*, in addition to being used in the process of choosing models, wasalso used to evaluate trained models with test and validation data. In the meantime, for the understanding of noise tests, due to noisy samples not being categorized, but still can contain known products, a contagem of the percentage of errors and correct answers was made — products with incomplete description predicted as a known product are considered correct, since because they do not have all the characteristics can not be refuted.

Given that the separation into different data sets has a random factor, it may be that there is loss of information for training or creation of favorable scenario when removing from this samples contained in testing and validation. Therefore, all training, testing and valition stages were performed five times, on different data per execution, and grouped by the mean of the results, minimizing any existing trend.

The analysis of the classification models occurred in two stages, first the base classifiers were evaluated by cross-validation with re-sampling and then the classifier committees. In the first stage, two different configurations were explored for the k-NN algorithm with k=3 and manhattan and euclidean distance measurements; Decision Tree, optimized vertion of cart algorithm with separation criterion "gini" and "entropy"; SVM with kernel "rbf" and gamma "scale" and "auto"; and finally, Naive Bayes with smoothing 1 and 0.1. For the neural networks MLP ("adam") several nium architectures were exploredthat improved the value of the loss function, resulting in two hidden layers with (30.20) neurons (Pedregosa et al., 2011).

For the committees, the following methods were analyzed: *Random Forest* composed of 100 trees; *Bagging out-of-fold* with 10 MLP; *Stacking* using the four best base classifiers (AD, NB, SVM and MLP), leaving the fusion function of the predictions in charge of the Logistic Regression algorithm — it is important to note that only the best settings were considered in *stacking* form. The choice of committees was guided by models widely known in the literature, with the exception of committees by *Boosting methods*, due to initial tests with implementation of "sklearn" did not show satisfactoryresults.

Assuming that the classifier in a real environment will deal with all descriptions received, all models implemented are based on probabilistic prediction, similar to the committees by explained serialization. Thus, the determination whether an amostra belonging to a class when presented to a model is made if the model obtained a probability above 50% for any of the known classes, otherwise they are considered as non-belonging to the known products — it is worth mentioning that the probabilities mentioned here are not probabilities in the real sense (in AM, calibrated probabilities) and rather confidence factors of the model.

5 RESULTS FROM EXPERIMENTAL

The experimental results of this work will be shown according to the classification phases followed, starting with the analysis of the base classifiers, aiming to select the most appropriate models and ending in the different types of committees presented here.

5.1 ANALYSIS OF BASE CLASSIFIERS

As a way to perform a robust analysis on the performance of several supervised techniques of AM for data classification, cross-validation with *k folds* (k = 10) was used. In addition, 5 executions were performed for each technique, thus allowing the mean of thef1-score metric, used as a measure to evaluate their performance.

The results, in Table 3, allow us to know which techniques handled the data provided best. Note that among them, neural networks (MLP) performed best among all, followed closely by Naive Bayes (smoothing =0.1) and AD (with "gini"). The SVM technique ("auto") had a poor performance, being only ahead of *the* k-NN (Euclidean distance) which was the technique that presented the worst performance among the five (5) techniques.

Table 3: Mean and Standard Deviation of f1-score values for 5 executions of the base models

Classifier	F1-score	Standard deviation
Decision Tree	0,82	0,015
K-NN	0,65	0,019
MLP	**0,89**	**0,011**
Naive Bayes	0,88	0,012
SVM	0,73	0,017

Source: Authors.

5.2. ANALYSIS OF CLASSIFIER COMMITTEES

Based on the results shown in Table 3, the Neural Network (MLP) algorithm was chosen to compress the *Bagging architecture*, mainly because it presented the best performance in the training and testing stages, as well as because it presented little sensitivity to noise.

To make up the *Stacking architecture*, the four base classifiers with the best performance were chosen, leaving out the k-NN algorithm that presented the worst performance among all the algorithms analyzed. In addition to the committee components, the logistic regression algorithm with fusion method was chosen.

Table 4 shows the results obtained for the three (3) committees analyzed inthe experiments. Note that *Bagging performed* best (f1-average score) among all committees, with *Stacking and Random Forest* performing lower.

Table 4: Mean and Standard Deviation of f1-score values for 5 committee executions

Committees	F1-score	Standard deviation
Bagging	**0,88**	**0,004**
Stacking	0,87	0,004
Random Forest	0,87	0,004

Source: Authors.

In order to make a more robust analysis of the models reported above, it was decided to select samples that had not been seen by them, that is, they were not present in the training and testing stages. Thus, about 293 (+/- 5%) validation samples were used to evaluate the same three committees. Table 5 illustrates the results related to the performance of the models. Note Bagging again achieved the best performance among the three models, with *Stacking* second, and *Random Forest* in last place.

Table 5: Mean and Standard Deviation of f1-score values for 5 runs with validation data

Committees	F1-score	Standard deviation
Bagging	**0,89**	0,003
Stacking	0,87	0,008
Random Forest	0,84	0,002

Source: Authors.

In addition to the samples not seen used in the validation, it was decided to perform a deeper analysis, using just over 8,000 unlabeled samples, to comcant the validation set with noise from the 293 validation samples. Table 6 moyster scans the f1-score average among the 5 executions, with all committees reaching the same value.

Table 6: Mean and Standard Deviation of f1-score values for 5 runs with noise validation data

Committees	F1-score	Standard deviation
Bagging	0,97	0,000
Stacking	0.97	0,000
Random Forest	0.97	0,000

Source: Authors.

Knowing that among the unlabeled samples there may be samples with characteristics of the classes known by the models, they are expected to classify more than 293 samples, by probabilistic prediction. Therefore, among the samples predicted as conhecides mixed with noise, the correct quantity, product by product, was analyzed in each committee execution, to analyze whether the committees when predicting more samples caused a reduction in accuracy.

The results of the analysis, Table 7, *show random forest* with the best accuracy while predicting fewer samples among the three committees, while *Bagging* and *Stacking* present similar results even with the number of samples predicted by *Bagging* being higher. Thus, it is possible toaffirm that the models that return more samples consequently bring more erroneous samples.

Table 7: Average hit out of 5 executions with noise validation data

Committees	Accuracy	Standard deviation
Bagging	0,85	0,007
Stacking	0,86	0,009
Random Forest	**0,88**	**0,002**

Source: Authors.

Finally, after realizing that the committees encountered similar difficulties and facilities regarding the ability to recognize the products half noise, a final analysis was made regarding the learning of eachproduct. As illustrated in Table 8, some products such as "Beef", "Biscuit Cream Cracker" and "Chicken Breast" obtained only correct samples predicted by the three committees in all executions; while, on the contrary, products with"4-roll Toilet Paper", "Paper Towel 2 rolls" and "Disposable Cup 50ml" resulted in a low average hit.

Table 8: Mean accuracy, standard deviation and number of samples predicted among the 5 executions using the 3 committees under noise validation data

Goods	Accuracy	Standard deviation
Bisc. Cream Cracker 400g	1	0
Beef	1	0
Chicken Breast	1	0
Disposable Cup 50ml	0,59	0,036
Toilet paper with 4 rolls	0,51	0,161
Paper Towel with 2 rolls	0,48	0

Source: Authors.

When comparing the pattern of the samples that make up the products in Table 8, it is noticeable that the products that work perfectly have descriptions with a more constant pattern, with few variations in characteristic and absence of products unawareby the model with similarity in the descriptions. On the other hand, products such as toilet paper and paper towels have by default long descriptions, with many characteristics and share patterns with other products that the model does not know.

6 FINAL CONSIDERATIONS

Through the tests performed it was possible to draw conclusions about the methodology implemented for data modeling and its application in classifier committees. In data cleansing it was expected to reducec-onfuming noise for the models, in categorization using more and less granular products, while for the committees obtaining a more robust learning with respect to classifiers in unit.

The cleaning methodology proved to be fundamental to improve the accuracy of the models during their development, but it presents difficult previous tracking. Only from trial and error implementing the data in the models and investigating how they predicted the samples that it was possible to tracepattern responsible for confusion and act accordingly.

The choice of products proved satisfactory, because it allowed to prove that the models can seize very similar products and that some products concentrate the greatest amount of confusion problems. Given the data limitation, problematic products that have characteristics shared with products unknown by the model, in case there is no cleaning solution, should be temporarily removed from the models.

Through the experiments carried out it was possible to perceive the robustnessof the committees' security by presenting little variation in validation accuracy under different training data and showing greater learning capacity than singular models. According to the results presented even with the lower accuracy for the validation conjunto with noise, *the Bagging committee* proved to be the best option due to the f1-score presented and its greater capacity to predict new samples.

For the required scenario where classifiers do not know all products, but must distinguish probabilisically, half all notes, known products, there is no way to eradicate prediction errors with ambiguous samples due to the nature of *the text encoding technology Count Vectorizer*. However, from the noise tests it is concluded that the methodology presented can solve the problem of grouping of unique products from the textual field of description. Its use by the TCE-RN for price analysis is feasible, but oriented by the products that present better resultsthan for the possessed data.

Finally, in view of the difficulties encountered in this implementation, for the next steps of this work, it is possible to implement different machine learning techniques, such as *architectures based on Gradient Boosting*, model committees that favor weaker models and are popularly used for tabular classification problems (Chen & Guestrin , 2016). Also, there is room for experimentation with the methodology used to represent the descriptions in the numerical domain, in order to consider the context of the words described, such as *contextualized word embedding* (Reimers et al., 2019). Furthermore, statistical tests must be implemented to make the understanding of the different learnings and the choice of the best based models.

REFERENCES

Brasil (2013a). *Acódão 1785/2013, de 10 de julho de 2013.* Tribunal de Contas da União, Brasília, DF. Recuperado de https://pesquisa.apps.tcu.gov.br/#/documento/acordao-completo/*/KEY%253AACORDAO-COMPLETO-1279889/DTRELEVANCIA%2520desc/0/sinonimos%253Dfalse

Brasil (2013b). *Decreto 7.892, de 23 de janeiro de 2013. Regulamenta o Sistema de Registro de Preços previsto no art. 15 da Lei nº 8.666, de 21 de junho de 1993.* Diário oficial da República Federativa do Brasil. Poder Executivo, Brasília, DF.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel ... & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop:* Languages for Data Mining and Machine Learning, p. 108–122.

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD*, São Francisco, CA. Recuperado de https://arxiv.org/pdf/1603.02754.pdf

Koche, J. C. (2011). Fundamentos de metodologia científica. Petrópolis: Vozes. Recuperado de: http://www.adm.ufrpe.br/sites/ww4.deinfo.ufrpe.br/files/Fundamentos_de_Metodologia_Cienti%CC%81fica.pdf

Ministério da Fazenda (2020a). *Conceito, uso e obrigatoriedade da nf-e (26 questões*). Recuperado de https://www.nfe.fazenda.gov.br/portal/perguntasFrequentes.aspx?tipoConteudo=E4+tmY+ODf4=

Ministério da Fazenda (2020b). *Manual de orientação do contribuinte - versão 6.00*. Recuperado de https://www.nfe.fazenda.gov.br/portal/listaConteudo. aspx?tipoConteudo=33ol5hhSYZk=

Ministério da Fazenda (2020c). *Ncm*. Recuperado de https://receita.economia.gov.br/orientacao/aduaneira/classificacao-fiscal-de-mercadorias/ncm

Ministério da Fazenda (2020d). *Protocolo icms 42, de 3 de julho de 2009*. Recuperado de https://www.confaz.fazenda.gov.br/legislacao/protocolos/2009/pt042_09

Tribunal de Contas do Estado da Paraíba (2020a). *Painéis preços*. Recuperado de https://sagres.tce.pb.gov.br/paineis-precos/

Tribunal de Contas do Estado da Paraíba (2020b). *Preço da hora*. Recuperado de https://precodahora.pb.gov.br/

Tribunal de Contas do Estado de Minas Gerais (2020). *Banco de preços tcemg*. Recuperado de https://bancodepreco.tce.mg.gov.br/

Secretaria de Tributação do Rio Grande do Norte (2020). *Nota fiscal eletrônica*. Recuperado de http://www.set.rn.gov.br/contentProducao/Aplicacao/SET_ v2/nfe/gerados/inicio.asp

dos Santos, D. S. (2018). *Uma plataforma distribuída de mineração de dados para big data:* um estudo de caso aplicado à secretaria de tributação do Rio Grande do Norte. Dissertação (Mestrado em Engenharia de Software). Universidade Federal do Rio Grande do Norte, Natal, Brasil.

Faceli, K., Lorena, A. C., Gama, J. & de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial:* Uma Abordagem de Aprendizado de Máquina. Barueri, SP: LTC

Gandini, A. (2020). *Banco de preços*. Recuperado de https://github.com/alexgand/banco-de-precos

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learnin*g. Cambridge, MA: MIT Press. Recuperado de http://www.deeplearningbook.org

GS1 (2019). *Código EAN 13: entenda o que é, para que serve e como usar*. Recuperado de https://blog.gs1br.org/codigo-ean-13-entenda-o-que-e-para-que-serve-e-como-usar/

GS1 (2020). *Gtin - número global do item comercial*. Recuperado de https://www.gs1br.org/codigos-e-padroes/padroes-de-identificacao/gtin

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: p. 2825–2830.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I. (2019) *Classification and Clustering of Arguments with Contextualized Word Embeddings*. Recuperado de: https://arxiv.org/pdf/1906.09821.pdf

Silva, D. S. (2014). *Manual de Orientação*: pesquisa de preços. Brasília, DF: Seção de Reprografia e Encadernação - Coordenadoria de Serviços Gerais. Recuperado de https://www.stj.jus.br/static_files/STJ/Licita%C3%A7%C3%B5es%20e%20contas%20p%C3%BAblicas/Manual%20de%20pesquisa%20de%20pre%C3%A7o/manual_de_orientacao_de_pesquisa_de_precos.pdf