Using Text Mining to Categorize the Purpose of Public Spending for the Benefit of Transparency and Accountability

Mauricio Barros de Jesus
External Control Dept.
Goiás State Court of Accounts
Goiânia, Brasil
mbjesus@tce.go.gov.br

Gladston Luiz da Silva
Department of Statistics
University of Brasília (UnB)
Brasília, Brazil
gladston@unb.br

Marcelo Ladeira
Computer Science Dept.
University of Brasília (UnB)
Brasília, Brazil
mladeira@unb.br

Gustavo C. G. Van Erven
Computer Science Dept.
University of Brasília (UnB)
Brasília, Brazil
gvanerven@gmail.com

***Abstract*—Advertising is a fundamental and structuring principle of Public Administration, standing out as an instrument of supervision and social control. In Brazil, public managers are required by law to be accountable and transparent to public spending, and the Court of Accounts is responsible for overseeing and ensuring that the information provided is clear and complete. This paper presents a comparative study between the Support Vector Machine, Naïve Bayes, and Logistic Regression models to classify public spending according to the purpose of the expenditure, allowing to identify those that were omitted by managers. The constructed model detected 124 unpublished records totaling $ 3.1 million.**

***Index Terms*—Text mining, Accountability, Text Categorization, Support Vector Machine, Naïve Bayes, Dimensionality Reduction**

I. Introduction

Advertising is a fundamental and structuring principle of Public Administration, expressly provided in the Constitution of Brazil [1]. More than a duty of the public administrator, the publicity of administrative acts stands out as an instrument of oversight available to citizens, providing means of social control that strengthens democracy.

Transparency has a broader meaning than advertising in that it requires information to be precise, relevant, timely, and understandable. [2]. In this sense, access to information becomes not only the right of the citizen but a duty of the state [3].

As a result, managers are required to post on a specific Transparency Portal [1] web site all public spending to purchase advertising services. However, there are thousands of procurements records each month. Identifying, tracking, and enforcing compliance of each one of these records is a complex, costly, and often unfeasible task to perform manually, given the small number of people available for this task in the governing bodies, such as the State Court of Accounts.

Since the criterion used to decide which public expenditure will be published to society is its classification as to the purpose of the spending, the incorrect classification (intentional or unintentional) makes transparency and social control impracticable.

This paper aims to propose a predictive model, based on supervised learning and text mining, to classify public spending in (a) advertising services or (b) other services. Besides, we intend to identify omitted expenditures from the Transparency Portals and, based on the model response, propose a set of expenditure purposes that are related to advertising services spending.

The model creation is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) [4]. This model breaks down the six-phase data mining process into a top-down approach that begins with business understanding and extends to product deployment [5].

This paper is organized as follows. Section II presents the related work. Section III discusses the main aspects of the creation process and comparison of models. Sections IV and V approach the understanding of business and understanding of the data available for training and evaluation. Section VI shows all the preparation steps for the extraction and cleaning of data and which techniques were used. Section VII discusses the main aspects of the selected models. Section VIII completes the data mining work presenting the best model according to established metrics. Section IX proposes a model deployment in a production environment and presents the prediction results with non-categorized data. Section X provides final conclusions and suggestions for future work.

II. Literature Review

The text classification task aims to assign a category to a document from its content. For this, a set of previously classified documents is used to train a model and use it to categorize new documents. [6].

The use of text mining for classification is a recurring subject. Among the most used classification models are Naïve Bayes - NB [6]–[8], Support Vector Machine – SVM [6]–[10], and Logistic Regression [10]–[12].

[1] http://www.transparencia.go.gov.br

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00049

Carvalho et al. [13] developed a methodology for classifying public spending of the Federal Government of Brazil, resulting in a reference price database of purchases. K-means was used to group similar products and establish maximum and minimum values. Then the authors use the frequency of terms to classify each grouping according to the most relevant words. The study allowed us to establish the methodology for setting up a price bank to investigate public procurement of overpricing fraud.

The text mining process deals with the problem of high dimensionality, caused by the number of terms in documents, which can hinder the learning process of models and degrade their performance [6], [14], [15]. To avoid this problem, dimensionality reduction techniques are applied through attribute selection and extraction [6], including Term Frequency Inverse Document Frequency (TF-IDF), Information Gain (IG), Latent Semantic Indexing (LSI), and Principal Component Analysis (PCA) [9]. The TF-IDF technique is the most used in the literature for its simplicity and good results [9], [16], [17].

Shah et al. [9] review the literature about extraction and selection techniques, presenting the advantages and disadvantages of each one. Among the techniques discussed are TF-IDF, Information Gain, Chi-square Statistic, Gini Index, Ambiguity Measure, LSI, and, PCA.

The study [12] compares performance among Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression classifiers for different n-gram combinations in a multiclass classification problem. The study points out that Logistic Regression outperformed the other classifiers, but with more significant variation of results for different n-gram combinations.

Rahman & Rahman [7] developed a comparative study between Naive Bayes, K-Nearest Neighbor, Decision Tree, and Support Vector Machines for text classification. The study proposes a hybrid model with Bayesian Network and Naïve Bayes Multinomial, with superior performance among all tested.

III. Methodology

The general and strategic objectives were established based on the definition of the research problem. The literature review was performed through searches in the Web of Science and Scopus databases, filtering articles published from 2015 related to machine learning used to surveillance, compliance and fraud detection. Search keywords were used: “text mining”, “information categorization”, “compliance” and “database assessment”. The articles were selected considering the relevance and quantity of citations and co-citations, selecting those authors who form the basis of knowledge about text mining.

The study was conducted from an inspection action by TCE-GO [2] and was supported by four business experts. The dataset was collected directly from the Budgetary and Financial System (SIOFI), the official data source for financial information in the State of Goiás [3]. Experts classified the data in two classes: (a) advertising services and (b) other services. Details of these procedures will be presented in future sessions.

Mining tasks followed the steps of the CRISP-DM reference model. Figure 1 shows the applied techniques, and they are based on the state-of-the-art survey. The entire dataset and Python code to Apache Spark used in this paper are available for download at https://github.com/mbjesus/spendingcategory.

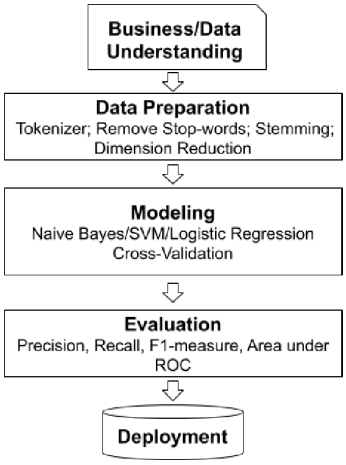

Fig. 1: Data mining process flow chart.

The six steps of the CRISP-DM model are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

1) Business Understanding: It is the stage in which the main aspects related to the business are raised and studied, and the acceptability criteria are established to build a project that adheres to the business needs.

2) Data Understanding: This stage aims at the initial understanding of the data, its identification, description of the format, databases, availability, and quality.

3) Data Preparation: Data is obtained, processed and made available in standardized formats for mining tasks. Data can be improved, reduced, aggregated, combined and, balanced. The final set is divided into test data and training data.

4) Modeling: In this step, several data mining techniques are applied to build the models. These models are tested and compared to get the one that best meets the business need.

5) Evaluation: The constructed model is tested, and its response is evaluated and compared to the acceptability criteria established in the business understanding phase, and the model can be accepted or not. You may need to return to previous phases for adjustments.

6) Deployment: In the Deployment phase, the model is presented to the business area for effective use, meeting the needs and objectives of the project.

IV. Business Understanding

The “Tribunal de Contas do Estado de Goiás (TCE-GO)” is a technical body with competence to assist the Legislative Power in the exercise of external control. Among the attributions of the Court is the protection of the State’s financial

[2]http://www2.tce.go.gov.br/ConsultaProcesso?proc=326492
[3]More information in: http://bit.do/fc6sM

resources, acting as guardians of the faithful enforcement of the rules by public managers.

In Brazil, the law stipulates that advertising expenses should be published monthly in Transparency Portals. However, some organizations had been omitting these expenses, breaking the law and the principles of transparency and publicity. Accordingly, TCE-GO set a deadline for organizations to publish spending on advertising services, discriminating the beneficiary of financial resources, the value, and purpose of public expenditure, under penalty of fine [4].

Subsequently, TCE-GO's technical department collected 143,000 public expenditure records from 2015 to 2018 from SIOFI to (a) identify which of these expenditures were related to the advertising service and (b) verify whether all identified payments were published on the Transparency Portal web site.

In SIOFI system, each expense record is categorized according to the purpose of the expense. The purpose of the expense is an aggregation of instances with the same characteristics as the expense goal. This field was used as the initial parameter to classify a public expenditure. However, there was no law or regulation to standardize the purposes of advertising spending.

In the absence of such standardization, TCE-GO specialists manually analyzed 436 SIOFI system spending purposes and identified eight related spendings on advertising services. With this list of purposes, experts identified 1,168 records for advertising spending, out of 143,000. These instances were categorized as "advertising service".

In SIOFI, the expense classification is done manually, so there is no way to guarantee that the registration is correct, either due to unintentional human error or some attempt of financial fraud. When expenditure is misclassified, it is not published on the Transparency Portal, which violates the law, compromises transparency and accountability. Besides, hundreds of purchases are made daily and controlling them all manually is an expensive task but can be automated with machine learning.

In SIOFI, there is a free-fill textual field, in which information about the description of the object contracted, the favored company or service provider, payment steps, and others are entered. As pointed out by the experts, reading the descriptive field allows identifying the purpose of spending.

Therefore, it was proposed to build a model for predicting spending on advertising services or other services based on text mining in the "Description" field in SIOFI System. The model response was compared to the categorization of SIOFI and discrepancies were forwarded to the Court's inspection area for appropriate action.

V. Data Understanding

Only the SIOFI database is required for the study. This database is the primary and official data source for financial release. The required fields for the mining process are the primary key (PK SIOFI), the expense purpose code (COD Fin), and the expense description (Description), which contains the text to be mined. Table I displays a preview of the dataset.

[4]http://www.tce.go.gov.br/ConsultaProcesso?proc=326492

TABLE I: Data set preview

Description	PK SIOFI	COD Finalidade
renov do contrat n..	2017590101..	33903935
prestaçã de servic de...	2016590102..	33903935
prestaçã de servic..	2016590102..	33903935
empenh complement...	2015115100..	33903935
despes de exercíci...	2018115100..	33909260

VI. Data Preparation

The dataset contains 143,000 instances of spending made between 2015 and 2018, divided between 436 purposes of the expenditure. Experts rated 1,168 instances as "advertising services". A total of 2,454 instances were randomly collected, stratified by purpose of the expenditure and classified by the experts as being or not spending on advertising services. Thus, the training and final test dataset contain 3,622 ranked instances (as shown in Table II).

TABLE II: Dataset configuration

Dataset Category	Number of Instances
Other Services	2,454
Advertising Services	1,168
Unknown Category	139,865
Total	143,487

The Python programming language with Apache Spark was used for data preparation, training, and model evaluation. We used the Scikit-Learn[5] package for data balancing, Spark SQL and Spark MLIB for creating training and evaluation sets for model training and results in the evaluation. Normalization techniques were applied to reduce dimensionality [18], turning the text into tokens to remove stop-words, performing stemming and weighting with TF-IDF [8].

Stop-words remove consists of eliminating words that have merely syntactic function in the text, such as articles, pronouns, conjunctions, prepositions. These words occur at a high frequency in the text, but do not add semantic relevance to data mining. The stemming technique reduces a word to its radical by removing suffixes and prefixes. The tokens were then transformed into a sparse integer matrix [19].

In TF-IDF, short for Term Frequency Inverse Document Frequency, each i term receives a w_i weight for each d document, generating a document vector $dv = \{w_i..w_n\}$, with n being the total number of terms in the corpus. For each i term in d, the frequency tf_i, which is the sum of the occurrences of i in the d document, is calculated [15].Terms with high frequency in a few documents can distort the analysis [20]. To avoid this problem, we used the term inverse frequency, idf_i, calculated according to the Equation 1. The df_i variable represents the number of documents containing the term i and the N variable represents the total number of documents in the dataset. The value of w_i is given by the Equation 2.

$$idf_i = log(N/df_i) \quad (1)$$

[5]https://scikit-learn.org

TABLE III: Hyperparameters adjustment

NB	SVM		LR	
Smoothing	*reg Param*	*max Iter*	*reg Param*	*elastic NetParam*
1.0	0.1	5	0.01	0.5
0.5	1.0	15	0.1	0.0

TABLE IV: SVM Hyperparam

reg Param	max Iter	AUROC
0.1	15	0.9785
0.1	5	0.9780
0.1	10	0.9778
1.0	15	0.9775
1.0	10	0.9765
1.0	5	0.9741

$$w_i = tf_i \times idf_i \tag{2}$$

This paper considered the occurrence of n-gram tokens with n equal to 2. The final dataset was divided into a training and evaluation set, in the proportion of 70% and 30%, respectively.

When generating the training data, it may be unbalanced: there are more records in one data class than another, and this may cause overfit in the model as it will tend to classify the data into the majority class. To work around this problem, under-sampling balancing was performed on the test data set (70% of the data) by randomly removing records from the majority class.

VII. Modeling

A literature review was searched for best data mining techniques for text classification, and the models Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR) were chosen for a comparative study. For training, 1642 classified instances were used. 10-folds cross-validation was used for hyperparameter adjustments as described in Table III.

The AUROC (Area Under the Receiver Operating Characteristics) curve measure is used to select the best parameter fit as Table V, IV, VI. NB with Smoothing 1 was selected. SVM with regParam equal to 0.1 and maxIter equal to 15 showed the best result. LG with regParam equal to 0.01 and elasticNetParam equal to 0 was selected.

TABLE V: NB Hyperparam

Smoothing	AUROC
1.0	0.9710
0.5	0.9677
0.0	0.5665

TABLE VI: Hyperparam adjust for Logistic Regression

reg Param	elastic NetParam	AUROC
0.01	0.0	0.9734
0.1	0.5	0.9732
0.1	0.0	0.9713
0.01	0.5	0.9343

VIII. Evaluation

Each model was trained from the hyperparameters with 10-folds cross-validation. To compare the models in the validation step, the adopted assessment measures were based on the result of the confusion matrix generated by each model [21]. In the adopted confusion matrix, TP is true positives, FP false positives, false negatives FN and TF true positives. The following measures were used:

- Precision: measures model precision in terms of positive hits:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- Recall: measures the percentage of positives hit by the model weighted by the total positives that exist:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- F1-Measure: represents a harmonic mean between precision and Recall.

$$F1 - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

The dataset used to test and compare models consisted of 1,062 instances without applying balancing techniques. Results were similar for all models, with little advantage for SVM, and this model has been selected for the deployment step. Table VII shows the results.

TABLE VII: Comparative result in test dataset

Metric	Naïve Bayes	SVM	LR
Precision	0.9600	0.9663	0.9684
Recall	0.9683	0.9829	0.9712
F1-Measure	0.9641	0.9745	0.9698
AUROC	0.9744	0.9834	0.9779

IX. Deployment

The developed model was used in a TCE-GO's inspection. The model classified 143,000 records raised from SIOFI and found 124 instances incorrectly classified in the database system and omitted from the Transparency Portal, totaling $ 3.1 million [6].

Another contribution of the model was the identification of 13 expense purposes on advertising services, in addition to the eight spending purposes pointed out by the experts at the beginning of the project. This information was forwarded for review by the agency responsible for the Transparency Portal web site.

Figure 2 presents the proposed deployment model to integrate the machine learning model for real-time spending prediction.

[6]http://www.tce.go.gov.br/ConsultaProcesso?proc=326492

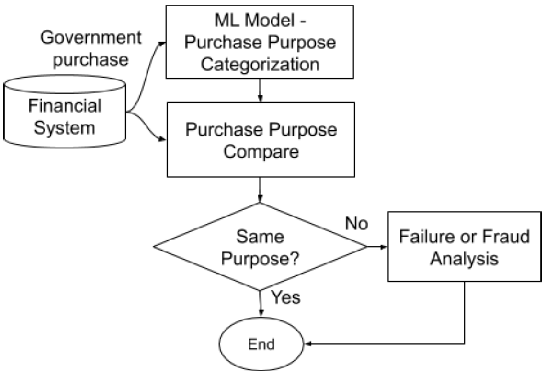

Fig. 2: Business model to audit public spending.

X. Conclusion

The text mining process for information categorization is widely used in academia, especially when it comes to sentiment analysis. This paper presents another approach when using a classification model for validating transactional databases.

The study proved to be very efficient, with high hit rates, which initially could indicate overfit. However, testing the model on unknown data found 1,376 positive cases with 184 false positives. Using the equation (3) corresponds to 94% precision. However, it is necessary to emphasize that, because it is a text mining process, the descriptive field mined must be enough and clear.

For future work, other dimensionality reduction processes can be used to improve model performance and reduce the possibility of overfittings, such as Principal Component Analyzes and Information Gain. In the text pre-processing steps, text correction techniques can be applied. Also, other supervised models may be used, such as decision trees or neural networks.

Acknowledgement

The authors would like to thank Tribunal de Contas do Estado de Goiás (TCE-GO) for providing the resources necessary to work in this research, as well as for allowing its publication.

References

[1] Brasil., "Constituição da República Federativa do Brasil," 1988.

[2] O. A. PLATT NETO, F. D. CRUZ, S. ROLIM ENSSLIN, and L. ENSSLIN, "Publicidade e transparência das contas públicas: Obrigatoriedade e abrangência desses princípios na administração pública brasileira," *Contabilidade Vista & Revista*, 2007.

[3] A. de Oliveira Reis, G. A. S. Sediyama, and E. L. de Castro, "Abordagens sobre a transparência em estudos de administração pública no Brasil," *Nucleus*, vol. 14, no. 2, pp. 35–46, 2017.

[4] P. Chapman, J. Clinton, R. Kerber, K. Thomas, T. Reinartz, C. Shearer, and W. Rudiger, "CRISP-DM 1.0. Step-by-step data mining guide," *CRISP-DM Consortium*, 2000.

[5] R. Balaniuk, "A mineração de dados como apoio ao controle externo," *Revista do TCU*, pp. 79–89, 2010.

[6] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowledge-Based Systems*, 2013.

[7] A. Rahman and U. Qamar, "A Bayesian classifiers based combination model for automatic text classification," in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2016, pp. 63–67.

[8] M. D. N. Arusada, N. A. S. Putri, and A. Alamsyah, "Training data optimization strategy for multiclass text classification," in *2017 5th International Conference on Information and Communication Technology, ICoIC7 2017*, 2017.

[9] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, 2016.

[10] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," 2018.

[11] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, 2016.

[12] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Baltic Journal of Modern Computing*, 2017.

[13] R. Carvalho, E. d. Paiva, H. d. Rocha, and G. Mendes, "Using Clustering and Text Mining to Create a Reference Price Database," *Learning and Nonlinear Models*, 2014.

[14] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems*, 2012.

[15] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, 2011.

[16] M. Abdel Fattah, "New term weighting schemes with combination of multiple classifiers for sentiment analysis," *Neurocomputing*, 2015.

[17] Z. Qu, X. Song, S. Zheng, X. Wang, X. Song, and Z. Li, "Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification," in *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*, 2018.

[18] F. F. dos Santos, M. A. Domingues, C. V. Sundermann, V. O. de Carvalho, M. F. Moura, and S. O. Rezende, "Latent association rule cluster based model to extract topics for classification and recommendation applications," *Expert Systems with Applications*, 2018.

[19] S. Joshi and B. Nigam, "Categorizing the document using multi class classification in data mining," in *Proceedings - 2011 International Conference on Computational Intelligence and Communication Systems, CICN 2011*, 2011.

[20] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, 1988.

[21] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, 2016.