Automated admissibility of complaints about fraud and corruption

Thiago de Paula Thiago Meirelles Andre Victor Andre do Amaral

Rodrigo Moreira Luis Alberto Sales Rafael Basso

Petróleo Brasileiro SA
Rio de Janeiro, RJ, Brazil

{thiago.depaula, thiago.meirelles, aovictor, andre.carpinteiro, rodrigo.moreira, luis.sales, rafael.basso}@petrobras.com.br

Abstract

This study proposes a natural language processing solution for the automated analysis of corruption complaints. The solution uses techniques such as text preprocessing, feature extraction, and machine learning to classify the complaints into admissible and inadmissible categories. The proposed system was evaluated on a corpus of real corruption complaints from Brazil. The results showed that the solution achieved an area under the ROC curve of 83% in the classification task, which was very close to the performance of tested and validated approaches for general complaint analysis.

1 Introduction

Fraud and corruption are serious threats to the integrity and performance of any organization. According to the 2020 ACFE Report to the Nations, a global study on occupational fraud and abuse, the typical organization loses 5% of its revenues to fraud per year, the median loss caused by fraud cases was $125,000, and 21% of the cases involved losses of at least $1 million (of Certified Fraud Examiners, 2020). Moreover, fraud can also damage the reputation and trust of an organization, leading to further losses of customers, partners, and investors. Therefore, it is vital for organizations to have effective mechanisms to prevent, detect, and respond to fraud and corruption risks. One of these mechanisms is the ombudsman and complaint channels, which play a central role in the compliance systems of companies, as they are essential for receiving and handling fraud and corruption complaints. This step is crucial for an efficient investigation process and mitigation of the financial and reputational impacts on the operations of the companies. The process of analyzing and investigating a corruption complaint is typically divided into two phases. The first phase, also called admissibility phase, aims to identify the elements of the complaint, such as suppliers, contracts, employees, customers and other stakeholders, and assess whether the reported facts are feasible and consistent. This phase exists because many complaints are unsubstantiated and do not provide any facts or elements that justify an investigation. The second phase of the process is the investigation itself, where analysts collect data, delve into documents and gather testimonies to assess whether the reported suspicions are confirmed (Kranacher and Riley, 2019).

The entire process is costly, time-consuming, and involves significant human and material resources. In this context, the present study proposes, for the admissibility phase, the development of a solution based on natural language processing (NLP) techniques for the automated analysis of corruption complaints received through the complaints channel. The objective is to provide a system capable of evaluating and classifying the relevance of the complaints for supporting the identification of cases that will proceed to the second phase of the process, the detailed investigation.

The results of the study demonstrated that the solution achieved the area under the ROC curve (Bradley, 1997) of 83% in classifying the complaints into admissible and inadmissible categories. This result was very close to that obtained by (de Paiva and Pereira, 2021), who used a similar approach to extract information from complaints in general. However, the proposed model focused on complaints about fraud and corruption, which are more specific and required a specialized corpus and a fine-tuned model to handle them.

2 Related work

Machicao and Arosemena (2019) applied NLP techniques to textual reports for detection and classification of reports from the Peruvian Ombudsman Office. They used document classification algorithms to categorize the reports into a set of classes, with a special interest in extracting reports related to social conflicts. Their work is relevant for the analysis of human rights violations and social justice issues in Peru.

de Paiva and Pereira (2021) also used NLP techniques to analyze the text of the report and enrich it with related data for the generation of an automatic report classification model. They extracted information such as names, dates, locations, and topics from the reports and used them as features for a machine learning classifier. Their work is similar to the one proposed in this article, but they focused on a different domain and task.

3 Dataset

The dataset, subsequently named *ComplaintFraud* in this paper, consists of a collection of reports of complaints received by the Ombudsman Office from a major brazilian company. These complaints were registered through a specific reporting channel that allows employees and third parties to report incidents related mainly to corruption and fraud, among other themes. Each complaint is composed of a descriptive text that contains relevant information for investigating the incident, such as details of what happened, people involved, dates and places.

3.1 Creation and Preprocessing

Since the reports are stored in PDF files, an isolated Python (van Rossum, 1995) pipeline was constructed to decrypt, scan and extract the text from these files using the Py2PDF library (Fenniak et al., 2022).The extracted texts were then processed to identify and store relevant entities, mainly through the use of pre-defined regular expressions and a Named Entity Recognition (NER) model (Souza et al., 2020). The most relevant categories extracted by the NER were dates, employee and companies, while information such as description of the incident, IDs, and contracts were extracted using regular expressions.

Given that the admissibility analysis requires the validation of consistency and relevance of the presented information, several rule-based validation routines were developed to verify the accuracy and internal consistency of the collected information against the corporate databases.

As a result, a list of numerical and categorical variables was created to identify various aspects of the reported incidents. For instance, these variables contained how many individuals mentioned in the reports were employees of the company, if the mentioned contract indeed took place on the reported date, or if the purchase order was genuinely issued in the name of the referenced company, among other criteria.

It was our hope that, by cross-referencing the extracted information with the corporate databases, these validation routines would result in highly discriminative features that would help improve the performance of the classification model.

3.2 Descriptive Statistics

The previous process resulted in a dataset comprised of 2082 complaints collected between the years of 2018 and 2022. The dataset exhibited significant class imbalance between the "admissible" and "inadmissible" classes, with an approximate proportion of 68% to 32%, respectively.

Regarding the characteristics of the complaints, several descriptive aspects were analyzed. The average length of the complaint reports is 1904 words, varying depending on the complexity and level of detail of each reported incident. Additionally, temporal data was considered, such as the distribution of complaints over time, allowing the identification of trends or fluctuations in the occurrences.

Characteristic	**Value**
Mean complaint length (tokens)	1904 (3282)
Mean complaints per year	251 (148)
Mean proportion of admissible complaints per year	0.68 (0.14)
Proportion of complaints where the whistleblower was highly confident about the ocurrence of the fraud	0.78

Table 1: Some descriptive statistics of the complaints, with respective standard deviations

4 Methodology

The methodology presented in this study follows the approach suggested by (de Paiva and Pereira, 2021) and involves the creation and pre-processing steps to generate the *ComplaintFraud* dataset, as detailed in Section 3. Next, the feature selection and the training and evaluation of the complaint classification model are carried out, as explained in this and the following sections.

Figure 1 shows the methodology of the complaint classification model.

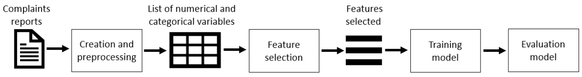

Figure 1: *Methology* of the proposed classification models

The complaint classification model proposed in this paper was trained and evaluated using the *ComplaintFraud* dataset, generated from the preprocessing of the texts of 2082 complaints. The experiments considered cross-validation with 5 folds and data split of 80% for training (*TrainSet*) and 20% for testing (*TestSet*) as used in (de Paiva and Pereira, 2021). The choice of classifiers was based on the preliminary evaluation of all the classification algorithms from the sklearn library (Pedregosa et al., 2011). The 4 best classifiers ordered by the ROC-AUC (Bradley, 1997) metric were chosen. The chosen classifiers were the XGBClassifier, Support vector classifier (SVC), MLPClassifier and LogisticRegression. Initially, 768 *features* were extracted from the text of the ComplaintFraud dataset. To reduce the dimensionality and select the most relevant features for classification, we applied the feature importance method based on decision trees from scikit-learn (Pedregosa et al., 2011) and arrived at a final set of 53 features. This dataset has the following characteristics:

- Features based on the TF/IDF (Hiemstra, 2000) vector that consider the most important words in the complaint texts. These features allows the model to assign greater weight to words that are characteristic of admissible complaints, as these words tend to be frequent in accepted complaints and infrequent in rejected ones;
- Features based on the verification of the existence of the entities (people, companies and contracts) in the corporate systems;
- Features extracted based on the dates of the complaints;
- Features resulting from calculations using the number of complaints in processing on the arrival date.

5 Results

We used the Area under the ROC curve (ROC-AUC) metric to evaluate the performance of different models (Bradley, 1997). The best model was the XGBClassifier, which achieved ROC-AUC score of 83% on *TestSet*. This classifier is based on decision trees and uses boosting techniques to improve performance. The table 2 shows the results of the other classifiers tested, which were inferior to the XGBClassifier. We also show precision and recall metrics for the "admissible" class.

Model	AUC	Recall	Precision
XGBClassifier	83%	75%	86%
SVC	77%	67%	85%
MLPClassifier	71%	94%	71%
Logistic	73%	66%	84%

Table 2: Results of the classifiers tested

The figure 2 shows the ROC-AUC plot for the models tested.

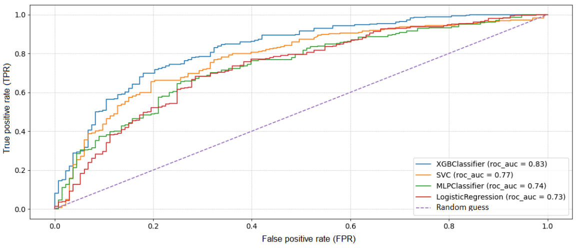

Figure 2: *ROC-AUC* of the proposed classification models

The proposed model, which used the corpus of complaints about fraud and corruption *ComplaintFraud*, achieved a similar performance to the reported 84% ROC-AUC of (de Paiva and Pereira, 2021). The model outperformed a random classifier in discriminating between admissible and inadmissible complaints. This capability can help reduce the time and cost of analyzing and investigating complaints about fraud and corruption.

6 Conclusion

The main contribution of this work is a model that can evaluate and classify complaints about fraud and corruption as admissible or inadmissible, based on pre-defined criteria. The solution applies natural language processing (NLP) and machine learning (ML) techniques to extract relevant information from the complaints and assign a confidence score to their classification. We compare different classifiers in this task and find that the XGBClassifier is the most effective.

The expectation for future works is to explore Large Language Models capabilities to extract finer semantic relationships between the entities cited in the complaints, enriching the discriminative power of the classification model. The initial tests were very promising.

References

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Eduardo de Paiva and Fernando Sola Pereira. 2021. Extraction and enrichment of features to improve complaint text classification performance. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 338–349. SBC.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and pypdf Contributors. 2022. The pypdf library.

Djoerd Hiemstra. 2000. A probabilistic justification for using tf× idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3:131–139.

Mary-Jo Kranacher and Richard Riley. 2019. *Forensic accounting and fraud examination*. John Wiley & Sons.

José C Machicao and Guillermo Miranda Arosemena. 2019. Peruvian ombudsman monthly social conflict reports analysis using knowledge management and artificial intelligence tools. In *2019 IEEE XXVI International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4.

Association of Certified Fraud Examiners. 2020. Report to the nations 2020 global study on occupational fraud and abuse.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

G. van Rossum. 1995. Python.