MINING AND PROCESSING OF NON-STRUCTURED PORTUGUESE TEXTS FOR RELIABILITY DATA CATALOG

Luciana Velasco Medani[1]
Virgilio Jose Martins Ferreira Filho[1]

[1]*Universidade Federal do Rio de Janeiro, UFRJ*

ABSTRACT

The digitization movement has enabled the acquisition and treatment of a big data volume in a consistent, agile, and efficient manner. However, the data analysis is still underutilized in the maintenance area. This work proposes a novel supervised learning approach to classify work orders in failure class according to NBR ISO 14224 standard. In the proposed methodology, every maintenance record is preprocessed with standard Natural Language Processing (NLP) procedures. Term Frequency-Inverse Document Frequency (TF-IDF) technique is employed to vectorize the preprocessed data. Four machine learning algorithms are used to evaluate task performance. The classifiers employed were the Logistic Regression (LR), Support Vector Machine (SVM) with a linear kernel, and two Naïve classifiers, the Multinomial Naive Bayes (MNB) and the Complement Naive Bayes (CNB). The methodology was tested in a real-world case from the O&G industry to classify turbine maintenance data in failure mechanisms categories. The results obtained indicate that the MNB model presented the best performance to the proposed classification problem. This methodology application speeds up the Reliability of data cataloging associated with less needed efforts than conventional methods. The application also contributes to the decision-making process improvement and corroborates the asset management and maintenance planning through efficient control of historical records during an asset operational life cycle.

1. INTRODUCTION

Failures are common events that can occur in industrial assets such as equipment and Instrumentation devices. Failure occurrences may lead to accidents with serious environmental, financial, or safety consequences for Exploration and Production (E&P) activities in oil fields. To prevent unexpected equipment failures, downtime, and, most important, major accidents in Oil and Gas (O&G) industry, the industry is investing in failure studies to build Reliability and Maintenance (R&M) databases. A suitable R&M database allows companies to optimize processes, reduce losses and costs related to maintenance activities, unscheduled downtime, and unexpected shutdowns [1]. Therefore, improvements can be achieved on operability, quality, and risk management of E&P assets [2].

When a failure is detected, the asset undergoes interventions to preserve or restore its operability. The interventions are maintenance or repair procedures stored as maintenance reports called Work Orders (WO) for asset management purposes. Rich in raw data, but without valuable information for the companies, the work orders contain unstructured data presented as long and brief text descriptions of failure events and maintenance activities. The data is processed to extract the information, building the R&M databases with structured data. The R&M databases contain information such as failure occurrence probability, failure mode, and failure mechanism; that can be used to identify patterns, trends, or abnormalities in equipment behavior.

The R&M data collection and cataloging can be executed manually or automatically. In Oil and Gas industry, it is commonly manual labor performed in batches by an expert team. When performed by a human analyst, the cataloging task is arduous for a massive number of records, requiring a great deal of time and resources to evaluate each one [3]. Additionally, the manual analysis demands a specialized workforce; is

1 PhD Student, Industrial Engineer – UFRJ
2 PhD, Industrial Engineer / Professor – UFRJ

exhaustive for a large data volume, where only a small fraction of the data is cataloged; it can be too slow, biased, and prone to human errors [2].

Motivated by the digitalization of the O&G sector, this work proposed an Artificial Intelligence (AI) approach to automate the R&M cataloging processing regarding the failure mechanism. The proposed methodology is based on Natural Language Processing (NLP) and Machine Learning models to classify unstructured texts (written in Portuguese) related to failure data from maintenance reports. As a result, the methodology automatizes the process saving time and effort, providing a reliable and complete structured database.

2. DESCRIPTION

Operational and maintenance historical equipment data is stored on different systems in which it is possible to manage the Material flow and internal activities coordination to synchronize activities with equipment availability needs for production [4]. This data can provide implicit information obtained from the failure event descriptions. However, extract and recover failure information from maintenance reports such as WO or maintenance logs is quite complex, and is far away from trivial, both automatically and manually [5].

Maintenance reports are traditionally composed of a mixture of fields with a structured and non-structured format. The data is characterized by its poor quality, where only a fraction of the data contained in the records is relevant for the purpose [1]. The data poor-quality is the main challenge for either manual or automatic approaches, defying the obtention and validation of information that adds value to analysis and algorithms and, being characterized by:

- subjectiveness, where event descriptions may not contain explicit information about failure or components of interest;
- lack of standardization in the tools and methods used to report maintenance data;
- low level of the completeness of the events' descriptions, the level of details may also depend on the repair technician.
- presence of jargon and abbreviations to describe maintenance events, which difficult in the employment of traditional NLP resources due to the specificity of the vocabulary and context.

In most cases, this problem is also markable because a single data source is unable to provide enough information for asset reliability studies. To obtain factual information, it is necessary to cross data with other data sources, which increases the process complexity.

The data is also characterized by a deficit in the volume of data for each type of equipment, for instance, one new machine has around 50-100 maintenance actions performed per year. Therefore, the number of failure classes varies according to the type of equipment analyzed. Some kind of equipment may have more than 10 failure modes, for example. They are also the fact that some failures are more frequent than others, resulting in an unbalanced dataset with multiple classes.

Another challenge faced specifically by the O&G Brazilian industry is difficult to find NLP models for Portuguese vocabularies that work as well as those usually used for English applications.

In this paper, the text mining and machine learning process have the main purpose of classifying failure categories from each register for any asset considered. This effort allows specialists to learn more about asset problems and abnormal behaviors, identifying from failure rates to the integrity of assets [6].

3. METHODOLOGY

This work presents a digital solution approach based on AI using the knowledge acquired through data already cataloged to speed up the process, classify more records with more reliable classifications, and reducing the number of human resources required for the task. The proposed methodology applies NLP and supervised learning techniques to historical equipment maintenance records, written in the Portuguese language, to automate the Reliability and Maintenance data cataloging process. For this objective, different Machine Learning methods are employed and have their performance compared to get task's the best model.

For this approach, each report is treated as a register of the dataset. The methodology scheme is presented in Figure 1.

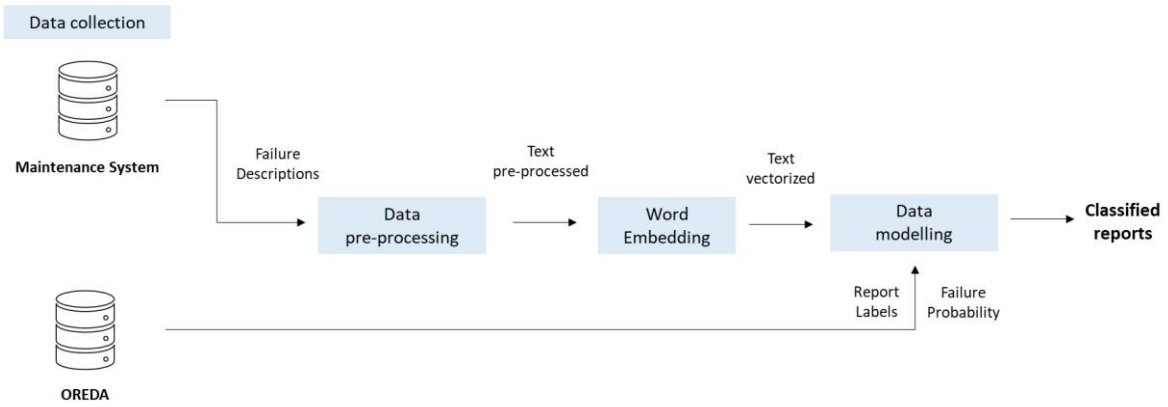


Fig.1 – Scheme of the maintenance record classification methodology

The first step comprises collecting the textual descriptions obtained during the equipment life cycle and its maintenance activities. Additionally, due to supervised learning, it is also necessary to collect the standard failure class labels that had been implicitly reported in each respective record. That is why the methodology obtains data from two databases. One, from an integrated maintenance management system that contains the description of historical maintenance records. Other, the standard reliability database for collect failure class labels of the same gather records, previously done manually by specialists.

Once the data is collected, it is necessary to clean and standardize raw texts. The main purpose of this step is to set up a meaningful set of words and terms that characterize the texts within each category. Every record is preprocessed with standard Natural Language Processing (NLP) procedures: remove regular expressions, standardizes text with case folding, filtering and lexical analysis, correct spelling errors, remove stop-words and stemmization. As a methodology proposes to text mining maintenance reports on the Portuguese Language it is necessary to use processing models that work specifically with this idiom.

After collecting and pre-processing the raw text from maintenance reports, the next step is to represent the registers as numerical vectors, known as word embedding. The third step is a key issue in the Text Mining and NLP process. The word embedding is performed because representing documents effectively using numerical vectors is fundamental for computer text processing and, for machine learning algorithms. Term Frequency-Inverse Document Frequency (TF-IDF) technique is employed to vectorize the preprocessed data.

After the word embedding step, the last step is to classify a maintenance record as a failure class label, according to the representative words of the documents. The data modeling step has two main sub-steps: calibration and adjustment evaluation of the models. Both steps are performed for the selected classification algorithms, in which the classifiers will be evaluated in each respective step simultaneously.

Four Machine Learning algorithms are employed to evaluate its performance and select the classification model that best solves the classification problem. The models applied in this methodology were: Support Vector Machine (SVM), Logistic Regression (LR), and two Naïve Bayes algorithms – Multinomial Naïve Bayes (MNB) and Complement Naïve Bayes (CNB).

The calibration procedure consists of testing different combinations of parameters specific to each learning model, to optimize its performance. The parameters considered for analysis for each model tested are shown in Table 1. In this stage, a Stratified K-Folds cross-validation was applied and an F-score metric to analyze the results' quality. The result considered is the mean value of the F-score metric obtained at all folds obtained by cross-validation. Besides that, class weight in Naïve Bayes models will be adjusted by the prior failure class probabilities.

The last step of the classification methodology is carried out for model evaluation. To evaluate the results, the calibrated models are run again with the cross-validation k-fold, with the same samples used in the calibration stage. The idea is to repeat the validation process, to check the robustness of the calibrated model, but at this step, more metrics will be obtained to verify the model's efficiency. Among the metrics of evaluation of the classification models, the adjustment step calculates the accuracy, precision, recall, and f-score (all weighted) for each round.

Tab.1 – Optimized parameters in model calibration

Classifier	**Parameters**	**Values**
Logistic Regression	C	0.0001, 0.001, 0.01, 1, 100
	Penalty	'l1', 'l2'
SVM	Kernel	linear
	C	10^{-3} até 10^{7}
	Gamma	10^{-5} até 10^{3}
MNB	Alpha	0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1
CNB	Alpha	0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1

4. DISCUSSION

In this case study, a data source from an O&G company is considered. This real case study is based on 377 maintenance reports from 7 years of gas turbine operation employed for power generation in a Brazilian offshore production unit. The dataset contains texts and their respective label, previously cataloged, as Table 2 shows. The records analyzed have five main failure mechanism categories, distributed as shown in Figure 2.

Tab.2 – Dataset sample (some words were hidden for confidentiality reasons)

ID	**Failure Mechanism**	**Maintenance Text**
12	INSTRUMENT	**Note:** * 17.04.2009 10:20:27 [redacted] * Sanar problema de falsa indicação do [redacted] * * Localização: [redacted] * * Responsável: [redacted] **Order:** [redacted] Falha [redacted] com indicação falsa 17.04.2009 10:20:27 [redacted] Sanar problema de falsa indicação do [redacted] Tensão (volts): até 120 Vcc Condição: Energizado RESPONSÁVEL P/ PLANEJAMENTO: Operador/Supervisor ACOMPANHAMENTO PELO OPERADOR: PERIÓDICO APROVAÇÃO GERENCIAL: [redacted] **Operation:** [redacted] Elemento - Reserva [redacted] CORRIGIR INDIC FALSA_[redacted] CORRETIVA [redacted] INSTRUMENTTO COM DEFEITO. SERÁ NECESSÁRIO FAZER PEDIDO DE MATERIAL VIA OM. A CHAVE DE TEMPERATURA QUE SE ENCONTRA AO LADO DO [redacted] TAMBÉM ESTÁ COM DEFEITO. VOU APROVEITAR E FAZER PEDIDO PARA OS DOIS.

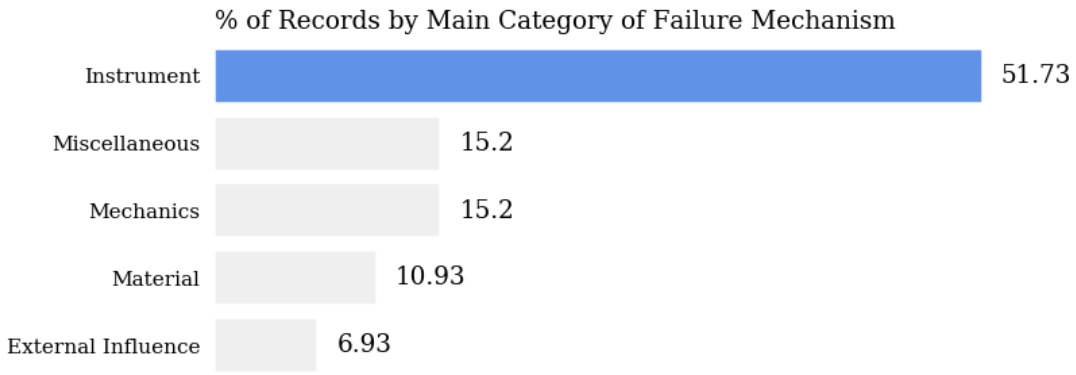
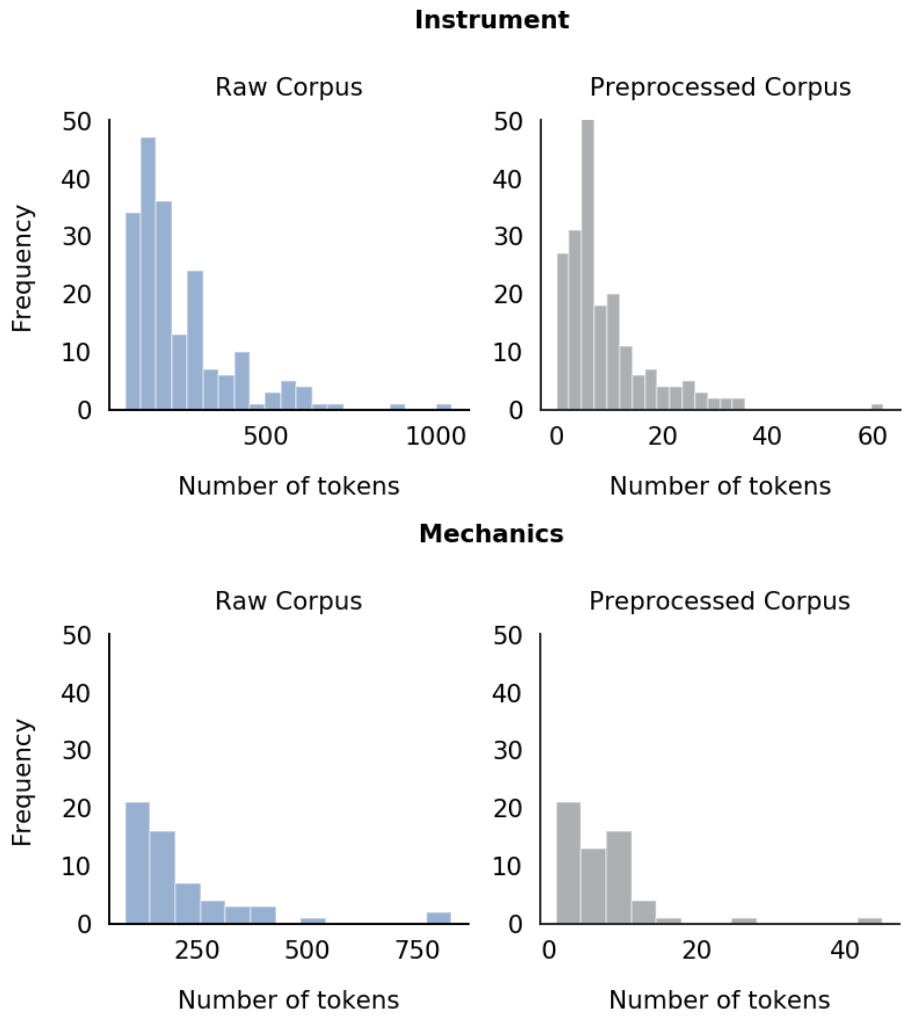


Fig.2 – Records main category of failure mechanism

Before evaluating each model tested is important to know the text size and the main terms occurring in the whole corpus.

Figure 3 shows the text size of each failure class before and after the preprocessing step. By comparing the text size in raw and preprocessed corpus from each class presented in Figure 3, it is possible to verify that the number of tokens in the text used in the classification model reduces significantly when compared to the original text. This corroborates with [1], who asserts that only a small fraction of the terms contained in the records are relevant for the purpose.



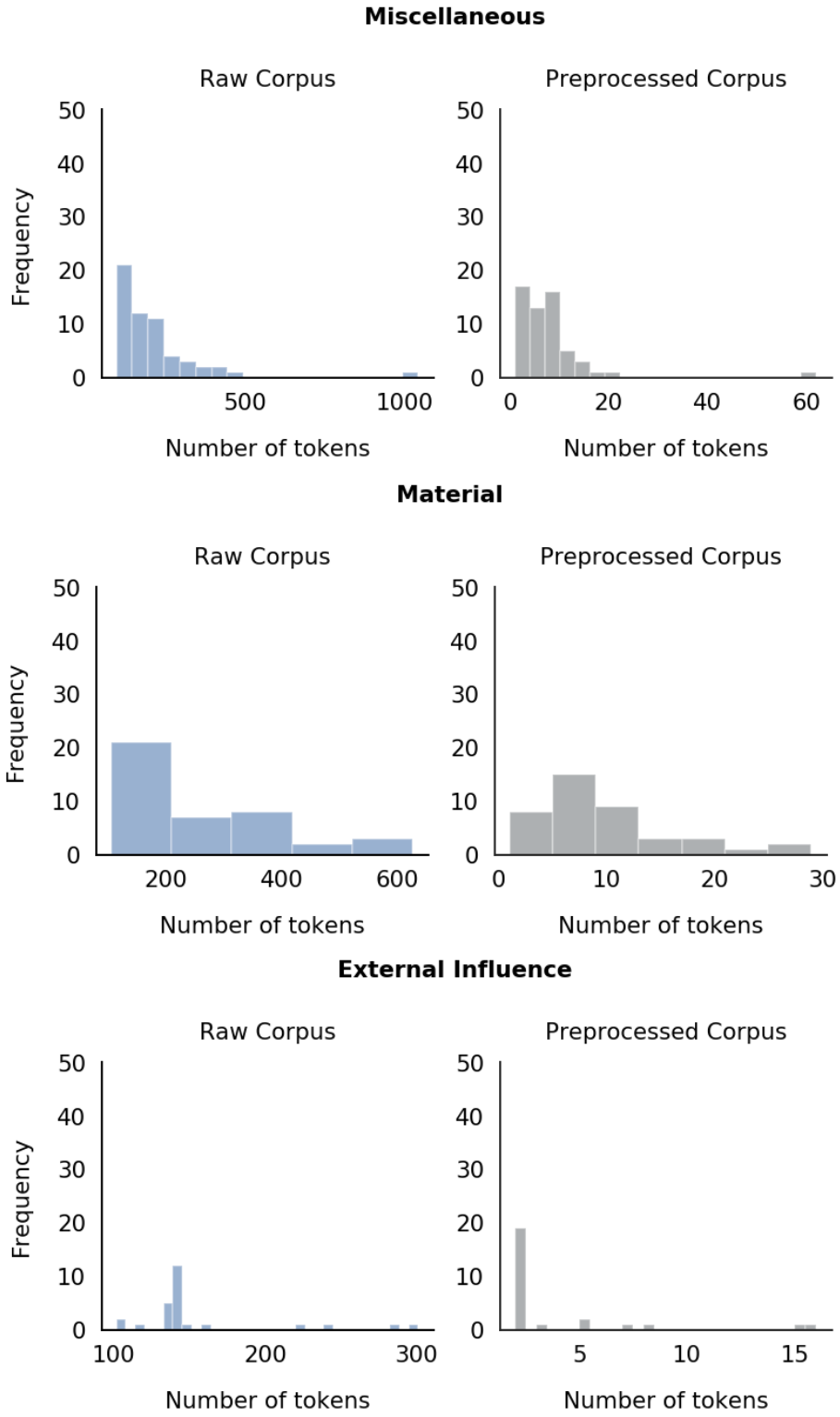


Fig.3 – Histogram of the number of tokens in each class of the collection

Figure 4 presents word clouds with the most frequent words for each class. For instance, for the failure mode of Instruments, the word "*falha*" ("fail") is the most frequent in this class. However, this is not an exclusive word for this class as others with high frequency such as "*Instrumento*" ("Instrument"), "*sensor*" ("sensor"), "*oscilação*" ("oscillation"), which characterize registers from Instrument failure mechanism. This shows the importance of using a TD-IDF instead of a simplistic TF (term-frequency) approach. Otherwise,

this could give more weight to words no so relevant resulting in misclassification. From other classes, this assumption is untrivial because they are more non-standard records than those from the Instrument class.

Fig.4 – Word cloud for each failure mechanism after the preprocessing step

The methodology presented in this paper compares four ML algorithms performances on the dataset. At the data modeling step, both minor stages (calibration and evaluation of the model) were performed with 5-folds cross-validation for all ML models and their variants (default and tuned). The Stratified K Fold was applied using 5-folds to guarantee the division of the sets in the cross-validation at least in the ratio 1: 4. Hence, in the cross-validation, the training set represents 80% (300 registers) of the data and the testing set 20% (75 registers). Thus, each fold of the training set contained approximately 39 records labeled as Instrument (INST),

12 Miscellaneous (MISC), 12 Mechanics (MECH), 8 Material (MAT), and 5 External Influence (EINF). In the test set, those categories have 39, 11, 11, 8, and 6 records, respectively.

Table 3 presents the calibration step results with F scores metrics for all ML algorithms proposed in the methodology. The model tuning was performed with the best parameter based on the mean weighted F1 score obtained with cross-validation for all models' parameter sets evaluated. From Table 3, it is observed that all classifiers obtained a higher performance after the parameter optimization in all cases, with Naïve classifiers with better performance than the linear ones.

Tab.1 – Model calibration results

Model	Parameter Type	Parameters	F Score Mean	F Score Std
CNB	**tunned**	**{'alpha': 0.2, 'norm': True}**	**0,72**	**0,04**
CNB	**default**	**{'alpha': 1, 'norm': False}**	**0,72**	**0,07**
MNB	**tunned**	**{'alpha': 0.1}**	**0,71**	**0,05**
SVM	tunned	{'C': 10.0, 'gamma': 0.1}	0,70	0,04
LR	tunned	{'C': 100, 'penalty': 'l2'}	0,68	0,05
LR	default	{'C': 1, 'penalty': 'l2'}	0,65	0,04
SVM	default	{'C': 1, 'gamma': 1}	0,61	0,04
MNB	default	{'alpha': 1}	0,55	0,03

The last procedure is model evaluation, where the testing samples were used to verify if the tuned model can correctly classify the failure modes labeled. Figure 5 presents the metrics Recall and F-score for the tuned classifiers, for each class from the testing set.

Figure 5 shows that the classes of the Instrument (INST) and Miscellaneous (MISC), even unbalanced among themselves, achieved high recall rates in LR and SVM models. While in MNB the high recall rates were Material (MAT) and Instrument (INST) classes. In CNB, the high recall rates were External Influence (EINF) and Mechanics (MECH).

From Figure 5 it is also possible to evaluate which model classifies better each failure mechanism. SVM showed a better performance for Instrument class (F-score = 75.6%) and Miscellaneous class (F-score = 84.8%). CNB and MNB had the same performance in the External Influence class (F-score = 60.0%). Mechanics classes classifications had better performance in CNB model (F-score = 86.4%). Finally, MNB model (F-score = 87.4%) better performed for Material classes.

Alone, the imbalanced data does not justify the high error rate of the classifiers concerning the Miscellaneous class. Since other classes even with a fewer number of records than the Instrument class had high recall rates. That way, it is valuable to consider the intrinsic difficulty in identifying Miscellaneous class, given the problem subjectiveness. Therefore, even applying balancing techniques of records such as synthetic data generation (SMOTE), few gains would be observed.

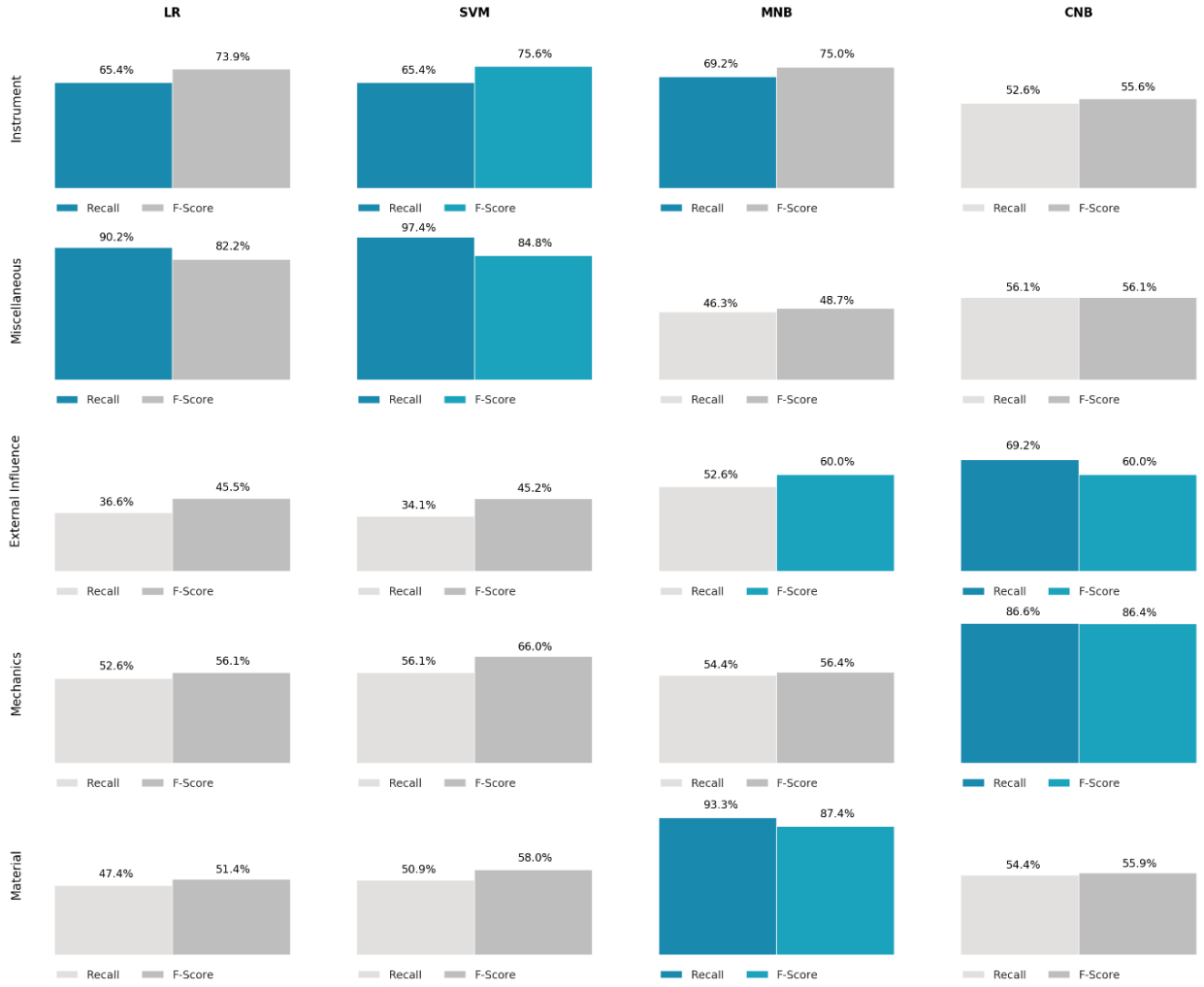


Fig.5 – Models' evaluation results for each class

Figure 6 presents the ordered overall results of all the metrics evaluated obtained with cross-validation, indicating how the chosen metrics vary for all classifiers. It is possible to identify the variability of these metrics results at the cross-validation procedure, which indicates a dependency on the samples selected for each folder. This behavior observed is expected, considering the small dataset. Also, comparing the results of all ML models (Figure 6), MNB was the best method considering all metrics evaluated, with less deviation than the CNB tuned and other models. Also, tuning the parameters of the models had increased all metrics, except for the CNB model.

From the classification results, tuning models of Logistic Regression, SVM-Linear, MNB, and CNB obtained balanced accuracy overall scores of 48%, 42%, 55%, 52%, respectively. Figure 7 illustrates the confusion matrix of the classifiers when applied to the dataset. The evaluating models' results are from all 375 records classification in the validation test sets (considering all rounds). As expected for unbalanced classification problems, there are more prediction errors in the less frequent class. It is worth mentioning that regardless of the records set, there is an inherent difficulty in separating the Miscellaneous of the other classes, as the characterization of the corpus indicates (being a complex task even for a human being). The Miscellaneous class had the worst rates because this class is characterized as a mix of unknowing failure mechanisms, others, or no cause found failure, being the most unstandardized records of the corpus.

In general, Naive classifiers obtained superior results when compared to linear ones. Given the high dimensionality of the input vectors of the models (838 dimensions) the lower performance of the linear models was expected, because as the number of explanatory variables increases and the more complex the surface of the problem becomes, the more significant is loss of performance in simpler models [7]. Based on the previous results, it is possible to state that, among all the models tested, the MNB tuned obtains a better overall performance among all models, being the one that best fits the proposed classification problem.

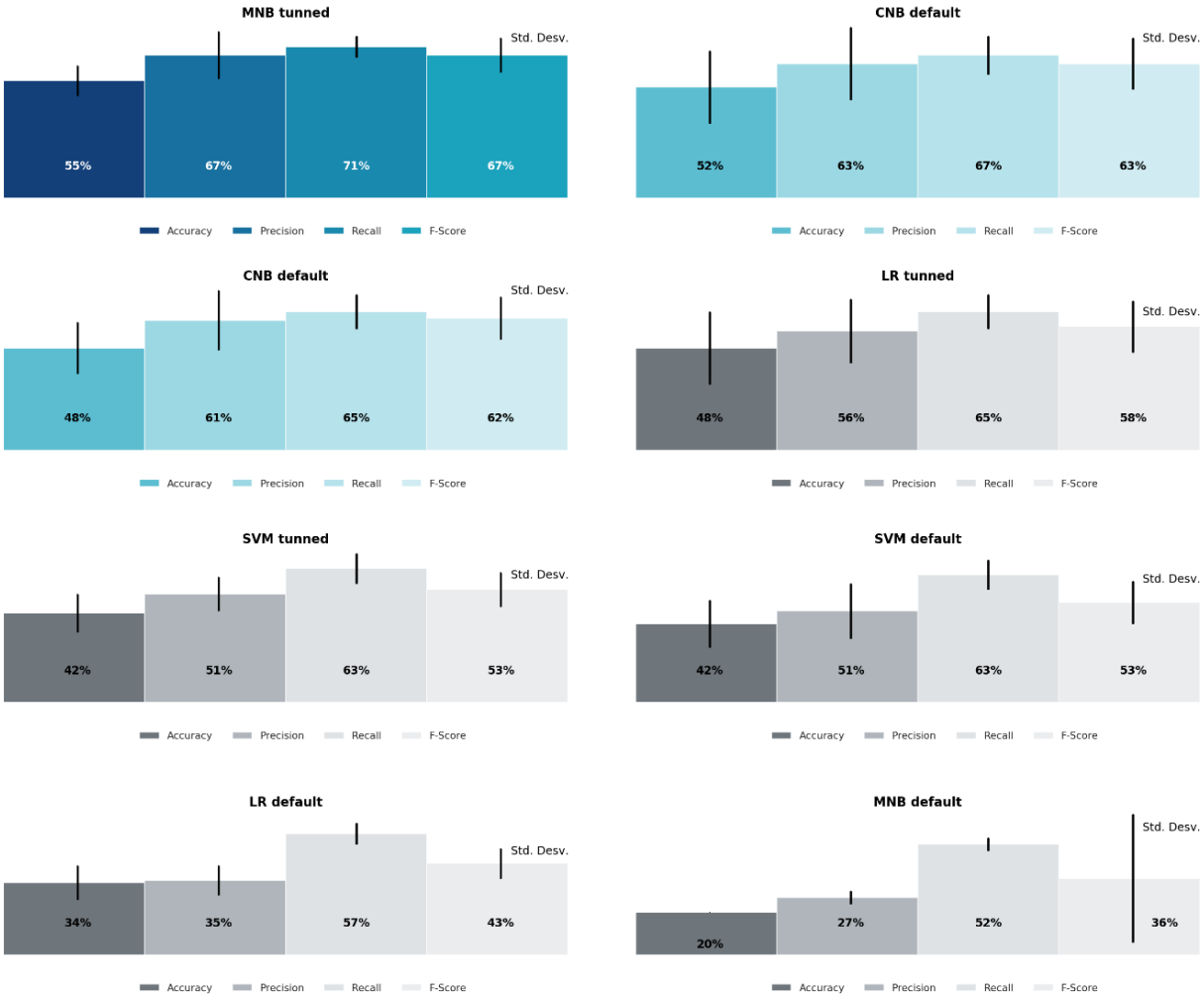
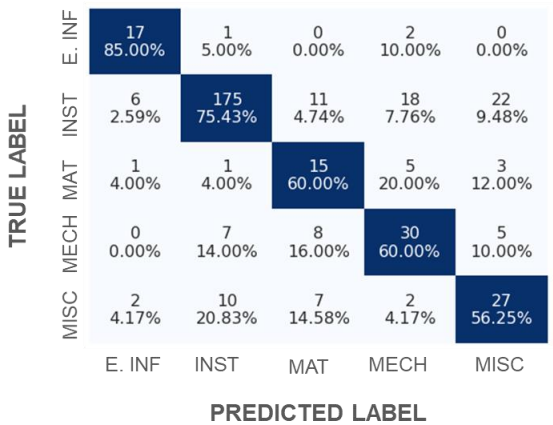
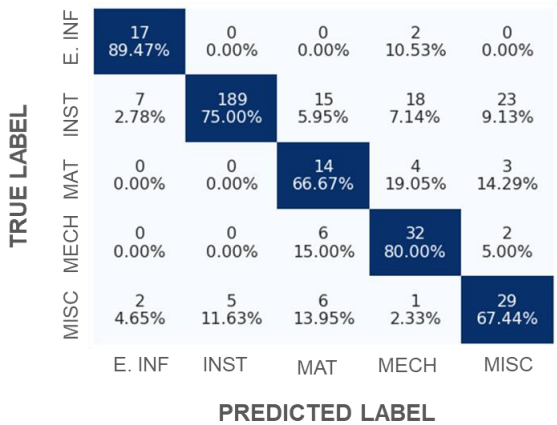


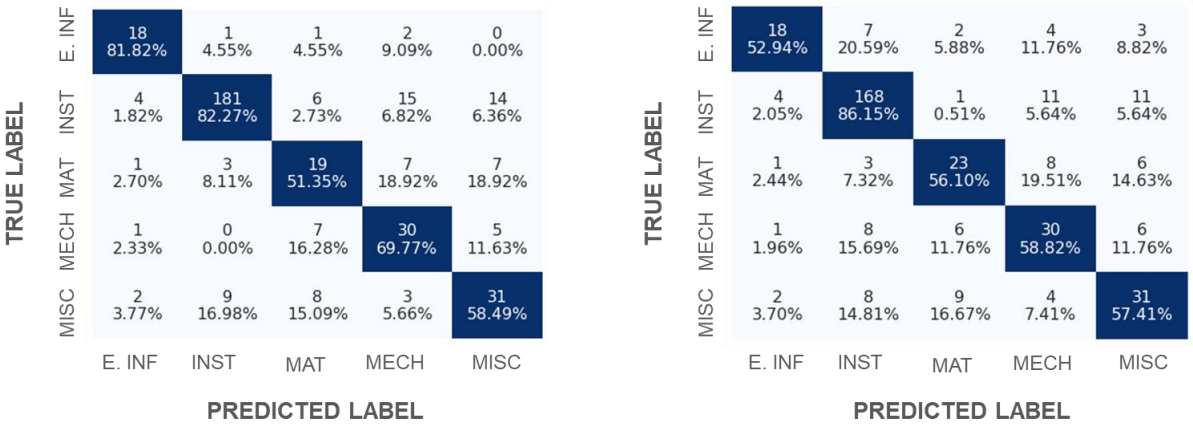
Fig.6 – Models' evaluation results



a) LR tuned



b) SVM tuned



c) MNB tuned d) CNB tuned

Fig.7 – Confusion Matrices

5. CONCLUSION

The Oil and Gas industry still struggles to embrace digital and data-based solutions into its decision-making process. From the literature overview, it was observed that textual documents containing valuable information may be underutilized by the industry. Especially, for supervised data-driven models in the Reliability and Maintenance data cataloging, which is not yet widely adopted. So, the development of a model that can make this analysis proves to be a great advance for the reliability data cataloging process, specifically for the O&G industry and in the Portuguese language.

This paper presents an NLP and machine learning approach with the objective of mining reliability data in the form of Work Orders to extract failure information and classify maintenance records automatically in failure mechanism classes. This work is innovative, challenging and extremely relevant. In the context of text mining written in Portuguese, even though it is one of the most widely spoken languages in the world, it has a scarcity of works that address technical texts written in this idiom, with most applications in social media texts. In addition, in the O&G field, this paper relevance is even greater, since the Brazilian pre-salt is an important exploratory frontier, and its documentation is often in Portuguese. From the maintenance field, by applying the proposed methodology, would be possible to increase the speed of the reliability data cataloging process and reduces the subjectiveness of the analysis.

This methodology presented successful results in a real case study of maintenance text data from gas turbines in Petroleum E&P activities, representing improved textual data usefulness by applying NLP techniques and supervised learning for failure classification. The results confirm the potential improvement of reliability and asset management, that can be obtained through efficient control of historical records during an asset operational life cycle and the significant positive impact that the proper use of these data can bring to the industry, especially for older installations.

From the results obtained in the data modeling step, this approach showed up an accurate solution for classifying a big volume of data, saving time and human resources. Also, the methodology contributes to catalog possible abnormal behaviors or failure patterns on time. Another fact observed in this study was how the low quality and lack of data standardization on the registers directly implies the performance of all methods evaluated. Classification results for each failure mechanism tend to improve as records have more quality and are more standardized. Comparing the results of the four machine learning algorithms chosen for text classification it became clear that such an approach is promising and can be extended to different equipment in any kind of industry when maintenance reports are available. Contrary to the [8] states that CNB has superior performance in failure modes classification problems, in the case study presented for failure mechanism classes, MNB models had better performance than compared to all other models tested. The presented methodology also showed that it is possible to use the same maintenance reports to classify different R&M qualitative data.

By employing this methodology, it is possible to capture valuable information from free-text maintenance reports in oil production units. Additionally, reduces the analysis subjectiveness, takes advantage

of the knowledge acquired from previously cataloged records for classifying a new register, and demands less effort when compared to conventional methods. Simultaneously, a larger amount of data can be cataloged, increasing the process's confidence, which improves the decision-making process. Thus, possible problems in assets can be better studied and maintenance planning could be potentially optimized by implementing the proposed solution. Therefore, the safety work environment, unplanned downtime metrics, failure risk, and maintenance costs could be enhanced.

There is still room to grow and improve this methodology, especially on critical steps that have a significant impact on the result. It is possible to compare and evaluate others ML models as K-Nearest Neighbors, Decision Tree, Random Forest for the classification task. Other possibilities are the use of failure example descriptions content in NBR ISO 14224 for training the machine learning algorithms with different weights than reports description. Future research will be focused on other methods more efficient to text vectorization with a semantical approach as Word2Vec or in a specific O&G domain as PetroVec [9] may also be evaluated and employed in a new version of this methodology. Simultaneously, the author intends to cross historical equipment monitoring information (PI data) to automatically validate some classes categorized.

6. ACKNOWLEDGMENTS

The author acknowledges the support of Programa de Recursos Humanos da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – PRH-ANP during this research.

7. REFERENCES:

[1] GONÇALVES, Virgínia Siqueira; GONÇALVES JÚNIOR, Elias Rocha; CARVALHO, Álvaro de Azeredo Araújo de. Bibliometric Study in Text Mining and Maintenance. International Journal of Science and Research (IJSR), v. 7, n. 11, pp. 1796–1801, 2018. https://www.ijsr.net/archive/v7i11/ART20193184.pdf.

[2] BLANCO-M., Alejandro; MARTI-PUIG, Pere; GIBERT, Karina; CUSIDÓ, Jordi; SOLÉ-CASALS, Jordi. A Text-Mining Approach to Assess the Failure Condition of Wind Turbines Using Maintenance Service History. Energies, v. 12, nº 10, pp.1–20, 2019. DOI: https://doi.org/10.3390/en12101982.

[3] SANDTORV, Helge A.; HOKSTAD, Per; THOMPSON, David W. Practical experiences with a data collection project: the OREDA project. Reliability Engineering and System Safety, v. 51, n° 2, pp. 159–167, 1996. DOI: https://doi.org/10.1016/0951-8320(95)00113-1.

[4] ZIO, Enrico. Reliability engineering: Old problems and new challenges. Reliability Engineering and System Safety, v. 94, nº 2, pp. 125–141, 2009. ISSN 0951-8320. DOI: https://doi.org/10.1016/j.ress.2008.06.002.

[5] ARIF-UZ-ZAMAN, K.; CHOLETTE, M. E.; MA, L.; KARIM, A. Extracting failure time data from industrial maintenance records using text mining. Advanced Engineering Informatics, 33, pp. 388–396, 2017.

[6] SALO, Erik; MCMILLAN, David; CONNOR, Richard. Work orders - Value from structureless text in the era of digitisation. Society of Petroleum Engineers - SPE Offshore Europe Conference and Exhibition, Aberdeen, UK, pp. 3-6. 2019. DOI: https://doi.org/10.2118/195788-MS.

[7] KIRASICH, Kaitlin; SMITH, Trace; SADLER, Bivin. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review, v. 1, n° 3, 2018. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/9.

[8] MEDANI, L. V.; HALL, B. M.; JARDIM, Thonny S.; DA SILVA; Rodrigo B.; BAIOCO; Juliana S.; DE FARIAS; RICARDO C.; FERREIRA FILHO, Virgilio. J. M. Artificial intelligence to obtain reliable failure data from maintenance reports. Technical Papers - Rio Oil and Gas Expo and Conference, v. 20, pp. 420–421, Rio de Janeiro, Brazil, 2020. DOI: https://doi.org/10.48072/2525-7579.rog.2020.420.