MINERAÇÃO E PROCESSAMENTO DE TEXTOS NÃO ESTRUTURADOS PARA CATALOGAÇÃO DE DADOS DE CONFIABILIDADE

Luciana Velasco Medani

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Virgílio José Martins Ferreira Filho

Rio de Janeiro
Setembro de 2020

MINERAÇÃO E PROCESSAMENTO DE TEXTOS NÃO ESTRUTURADOS PARA CATALOGAÇÃO DE DADOS DE CONFIABILIDADE

Luciana Velasco Medani

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Orientador: Virgílio José Martins Ferreira Filho

Aprovada por: Prof. Virgílio José Martins Ferreira Filho
Prof. Ricardo Cordeiro de Farias, D.Sc.
Profª. Juliana Souza Baioco, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2020

Medani, Luciana Velasco

Mineração e processamento de textos não estruturados para catalogação de dados de confiabilidade / Luciana Velasco Medani. – Rio de Janeiro: UFRJ/COPPE, 2020.

XVI, 132 p.: il.; 29,7 cm.

Orientador: Virgílio José Martins Ferreira Filho

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2020.

Referências Bibliográficas: p. 122-132.

1. Text Mining 2. Classificação automática de textos

3. Confiabilidade. I. Ferreira Filho, Virgílio José Martins. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

"*Success is walking from failure to failure with no loss of enthusiasm*"

Winston Churchill

Dedico este trabalho a todos que se aventuram a buscar conhecimento e aprendizagem. Aos desbravadores da ciência e tecnologia deste país. Pesquisadores e professores de todos os níveis de ensino, que não recebem o devido mérito e prestígio que merecem frente a importância do trabalho que realizam.

AGRADECIMENTOS

Todos que acompanharam de maneira mais próxima o processo de aprendizado e amadurecimento desta pesquisa sabem o quanto o seu desenvolvimento foi lento e difícil. Assim, seria injusto se eu não dividisse os méritos deste trabalho com todos aqueles que contribuíram direta ou indiretamente no seu desenvolvimento e em sua conclusão. Tenho muita sorte de estar cercada de pessoas tão incríveis.

Agradeço à minha família e amigos, sempre presentes no meu coração.

Ao meu pai, *in memoriam*, por ter sido o primeiro incentivador da minha formação acadêmica. À minha mãe por provocar em mim um espírito questionador, característica tão pertinente no trabalho de um pesquisador. Ao meu irmão, a quem eu tenho imenso amor, por me desafiar a romper paradigmas e constantemente me recordar que a vida é muito mais do que eu penso ou imagino. Obrigada por sempre se orgulhar e comemorar minhas vitórias, inclusive as mais singelas.

Ao Rafael, por escolher multiplicar as felicidades e dividir as dificuldades da vida comigo. Por, a todo tempo, ser incansável em me fazer acreditar em mim. Obrigada por sempre me apoiar, incentivar, e, especialmente, pelas extensas contribuições no meu desenvolvimento pessoal e profissional.

À Polyana por ter me inspirado a seguir a carreira acadêmica e de pesquisa.

À Bruna, pela amizade verdadeira e por estar sempre comigo onde quer que estejamos.

Ao professor Virgílio por ser calmaria todas as vezes em que eu fui tempestade. Por ser um orientador acessível, sempre disposto a dialogar e ajudar os alunos. Por compreender as minhas necessidades e dificuldades pessoais e técnicas. Muito obrigada por sua orientação, incentivo, ensinamentos e zelo ao longo de todo este período.

Aos professores Ricardo e Juliana pela disponibilidade de tempo e solicitude de aceitar compor a banca avaliadora deste trabalho, por todas as contribuições valiosas a esta pesquisa.

Ao incrível grupo de confiabilidade do INEES II - Bryan, Rodrigo e Thonny, pela inestimável contribuição, orientação e companheirismo ao longo de todo o projeto.

Aos amigos que fiz na COPPE e UFRJ, que me inspiram a ser tão boa pesquisadora e aluna quanto eles.

Aos amigos e colegas do LORDE/SAGE pela extraordinária troca de experiências e aprendizado, mas principalmente, pelos risos. À Dani, Clarinha, Girão, Thonny e Gabi pela amizade, incentivo, paciência e valiosas contribuições à minha pesquisa (e por me ajudarem a me manter sã, principalmente neste período de quarentena). Ao Pedro pela

amizade, carinho e bom humor de sempre (obrigada pelas caronas também!). À Paulinha (minha dupla de trabalho preferida), Lúcia, Verônica e Marina que fizeram que o período de disciplinas fosse muito mais fácil e agradável. Aos colegas do grupo de carona Niterói/Fundão por tornarem mais leves os trajetos até a pós graduação.

A todos os funcionários do LORDE/SAGE por proporcionarem um ambiente de trabalho acolhedor e seguro.

Ao corpo docente, secretárias e funcionários do Programa de Engenharia de Produção, e demais programas da COPPE, pela excelente formação acadêmica.

Ao CNPq, pelo apoio financeiro e incentivo à pesquisa.

Não há palavras que descrevam o quanto aprendi com vocês e sou grata por me impulsionarem a ser uma pessoa-acadêmica-profissional melhor.

O mérito e o sucesso deste trabalho dedico a todos vocês. As falhas e imperfeições, são todas minhas.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MINERAÇÃO E PROCESSAMENTO DE TEXTOS NÃO ESTRUTURADOS PARA CATALOGAÇÃO DE DADOS DE CONFIABILIDADE

Luciana Velasco Medani

SETEMBRO/2020

Orientador: Virgílio José Martins Ferreira Filho

Programa: Engenharia de Engenharia de Produção

Informações valiosas para a compreensão de eventos de falha na indústria O&G podem ser obtidas e tratadas com métodos adequados para suportar a tomada de decisão. A gestão da manutenção de ativos industriais requer informações precisas quanto a falha de um ativo para modelar sua confiabilidade adequadamente, tais informações estão tradicionalmente dispersas em diferentes sistemas de informação não estruturados. Este trabalho propõe uma abordagem de mineração de textos em históricos de manutenção para facilitar o processo de coleta e catalogação de dados de confiabilidade. A metodologia é aplicada em um conjunto de dados de manutenção de turbinas, podendo ser replicada para outros equipamentos. Os modelos *Multinomial Naïve Bayes* (MNB) e *Complement Naïve Bayes* (CNB) são aplicados para a classificação automática dos registros em relação ao modo de falha do equipamento. A partir do resultados obtidos no estudo de caso, observa-se que o modelo CNB obteve melhor performance que o modelo MNB. A análise dos eventos identificados por meio do método proposto facilita a catalogação dos dados de confiabilidade, economizando tempo e contribuindo para melhoria do processo de tomada de decisão. Os resultados desta pesquisa destacam o potencial de melhoria na confiabilidade e gestão de ativos industriais que pode ser obtido através de um controle eficiente de registros históricos ao longo do tempo de operação.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

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

Luciana Velasco Medani

September/2020

Advisor: Virgílio José Martins Ferreira Filho

Department: Industrial Engineering

Valuable information for understanding failure events in the O&G industry can be obtained and treated with the appropriate methods to support the decision-making process. The industrial assets maintenance management requires accurate information about the failure of an asset to model its reliability accordingly, such information is usually dispersed into different unstructured information systems. This work proposes an approach to text mining in maintenance histories to facilitate the process of collection and cataloging of reliability data. The methodology is applied to a set of turbine maintenance data but can be easily escalated to other equipment. The Multinomial Naïve Bayes (MNB) and Complement Naïve Bayes (CNB) models are applied for equipment data automatic classification, based on their failure modes. From the results obtained in the case study, it is observed that the CNB model obtained better performance than the MNB model. The analysis of the events identified through the proposed method facilitates the cataloging of the reliability data, saving time, and contributing to improve the decision-making process. The results corroborate the reliability and asset management potential improvement that can be obtained through efficient control of historical records during an asset operational life cycle.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT – Associação Brasileira de Normas Técnicas

ANM – Árvore de Natal Molhada

AQR – Análise Quantitativa de Risco

BCP – Bombeio de Cavidade Progressiva

BHP – *Brake Horse Power*

BM – Bombeio Mecânico

BOW – *Bag Of Words*

BP – *Back-Propagation*

BSW – *Basic Sediments and Water*

CMMS – Computerized Maintenance Management Systems

DHSH – *Down Hole Safety Valve*

E&P – Exploração e Produção

FEPS – *Failure Elimination and Prevention Strategy*

FMEA – *Failure Mode and Effect Analysis*

FMECA – *Failure Mode Effects and Criticality Analysis*

FPSOs – *Floating Production Storage and Offloading*

FTA – *Failure Tree Analysis*

HSEQ – *Health, Safety, Environment and Quality*

ISO – *International Organization for Standardization*

LCC – *Life Cycle Cost*

LDA – *Latent Dirichlet allocation*

ML – *Machine Learning*

MCC – Manutenção Centrada em Confiabilidade

MTBF – *Mean Time Between Failure*

MTTR – *Mean Time to Repair*

NB – *Naïve Bayes*

NLP – *Natural Language Processing*

NLTK – Natural Language Toolkit

NPD – *Norwegian Petroleum Directorate*

NPSHR – *Net Positive Suction Head Required*

OCR – *Optical Character Recognition*

OREDA – *Offshore and Onshore Reliability Data*

OS – Ordem de Serviço

PDS – *Probability Design System*

PFD – Probabilidade de Falha na Demanda

PSA – *Petroleum Safety Authority*

pt-BR – português do Brasil

RAM – *Reliability, Availability and Maintainability*

RAMS – *Reliability, Availability, Maintainability and Safety*

RGO – *Razão Gás Óleo*

RI – Recuperação da Informação

RMS – *Reliability Management System*

RSLP – Removedor de Sufixo da Língua Portuguesa

SIL – *Safety Integrity Level*

SPE – *Society of Petroleum Engineers*

SVD – *Singular Value Decomposition*

SVM – *Support Vector Machine*

TF – *Term Frequency*

TF-IDF – *Term Frequency - Inverse Document Frequency*

TM – *Text Mining*

TBF – *Time Between Failure*

TTF – *Time To Failure*

TTR – *Time To Repair*

WOAD – *World Offshore Accident Database*

WO – *Work Orders*

1 INTRODUÇÃO

1.1 Contexto

A Indústria 4.0 e a evolução dos modelos de produção favoreceram que os processos se tornassem mais automatizados e informatizados, com maior adesão a máquinas de sistemas complexos (KARDEC; NASCIF, 2009). Essa transformação permitiu acesso a uma grande quantidade de dados, impactando diretamente na maneira de como são realizados os procedimentos de manutenção de equipamentos.

No que se refere a atividade de Exploração e Produção (E&P) de campos de petróleo em ambiente *offshore*, as decisões de como e quando realizar as manutenções são ainda mais complexas quando considerada toda a operação e logística envolvida devido ao distanciamento das unidades E&P da costa e da produção em águas cada vez mais profundas. O sistema é formado por diversos elementos essenciais para a atividade (árvores de natal, compressores, turbinas, bombas, válvulas, dutos, etc.) responsáveis por favorecer o deslocamento do fluido (óleo e/ou gás) do reservatório até uma unidade marítima de produção.

Ao longo do tempo, diversos fatores, como corrosão, danos mecânicos de componentes, mudanças nas condições de produção e do reservatório ou até mesmo causas aleatórias, podem fazer com que os equipamentos e outros ativos de uma unidade de produção não executem adequadamente suas funções, prejudicando a atividade de E&P. A ocorrência de eventos não planejados e falhas críticas de ativos podem ter sérias consequências ambientais, financeiras e de segurança para as atividades de E&P de petróleo. No entanto, a identificação da origem e ocorrência de falhas antes que os equipamentos estejam danificados pode ser prejudicada devido a inúmeros fatores, entre eles à complexidade de operar e realizar a manutenção deste sistema.

Assim, para uma gestão mais eficiente dos ativos industriais é realizada a coleta e catalogação de dados de falha e manutenção de equipamentos. Os dados catalogados podem ser utilizados para obter *insights* sobre a confiabilidade e o desempenho do equipamento, auxiliar na identificação de padrões de falha ou de comportamentos anômalos.

1.2 Motivação e justificativas

Embora importante e necessária, a catalogação de dados de falha é uma tarefa árdua que demanda grandes esforços da indústria em geral e, especificamente, a do petróleo devido ao grande volume de dados armazenados nos bancos de dados desse setor (BLANCO-M *et al.*, 2019).

Dados de confiabilidade fornecem informações quanto a probabilidade e natureza de falha dos componentes do sistema. São comumente associados à avaliação de desempenho de equipamentos, para verificar a frequência necessária das ações de manutenção e entender como sistemas falham regularmente. Fornecem parâmetros não apenas da operabilidade e de qualidade do sistema, mas também, quanto a quantificação de risco e segurança dos processos industriais.

BRYNJOLFSSON; MCAFEE (2012) indicam que companhias que baseiam suas decisões em dados, conseguem determinar objetivos e métricas mais adequadas, possibilitando melhores ganhos e resultados. Sob a perspectiva dos dados de manutenção e confiabilidade, identificar padrões, tendências ou anormalidades no comportamento equipamento a partir das informações históricas confiáveis representam grande valor para a indústria (GONÇALVES *et al.*, 2018).

O esforço utilizado para analisar relatórios de manutenção torna possível obter informações valiosas, desde taxas de falhas até a integridade dos ativos (SALO; MCMILLAN; CONNOR, 2019). Ao mesmo tempo, apresentam aos operadores diversas dificuldades analíticas devido a qualidade dos dados, o que pode influenciar diretamente nos ganhos de produção, antecipação de problemas e mitigação dos riscos.

ZHANG *et al.* (2020) relatam que históricos de ordens de serviço de manutenção podem ser empregadas para selecionar equipamentos mais confiáveis, minimizar interrupções, planejar e agendar atividades de manutenção mais econômicas e que, quando classificadas corretamente levam a conclusões precisas e rápidas que resultam em uma programação de manutenção eficiente.

No entanto, apoiar o processo de decisão em dados pouco representativos, de baixa qualidade e confiabilidade duvidosa podem ser mais prejudiciais que benéficos. MOBLEY (2002) indicou que cerca de 33% de todo o custo de manutenção são gastos devido à a manutenções desnecessárias ou impróprias, e que estes custos podem representar em torno de 15 a 60% dos custos das mercadorias produzidas, dependendo da indústria de aplicação.

Deste modo, esta pesquisa pretende auxiliar o processo de descoberta do conhecimento a partir dos registros históricos das atividades de manutenção, extraindo informações quanto à natureza da falha (isto é, o modo de falha) em textos de formato livre. Assim, o desenvolvimento de uma metodologia que auxilie o processo de catalogação de falhas representa um incremento na velocidade de catalogação e na qualidade/confiabilidade dos dados obtidos.

A metodologia também é valiosa pois possibilita extrair informações relevantes e obter conhecimento importante para melhorar a eficiência operacional de um ativo. Adicionalmente, viabiliza reduzir paradas críticas e determinar cronogramas de manutenção e inspeções ideais, trabalhando com estratégias para mitigar falhas e riscos. E, crucialmente, permite evitar desde a ocorrência condições de produção sub-ótimas até problemas mais graves que levam à interrupção da produção por determinado tempo ou definitivamente.

1.3 Contextualização do Problema

Ao longo da vida útil de um equipamento, este poderá passar por diversas intervenções para substituir, modificar ou conservar as funções de seus componentes de modo a manter ou restaurar sua operacionalidade.

No entanto, apesar da identificação de eventos de falha em registros históricos permitirem informações extremamente relevantes para a confiabilidade e disponibilidade em um sistema produtivo, o processo de extração e recuperação de informação em relatórios de manutenção é bastante complexo. Principalmente, por causa do formato não estruturado dos textos, do relato subjetivo de cada operador, a pouca ou nenhuma uniformidade nas descrições dos eventos, o baixo nível de completude, a qualidade dos dados e demais obstáculos da catalogação dos dados de confiabilidade que serão relatados no Capítulo 5.

Assim, o problema de pesquisa que esse trabalho pretende responder é como desenvolver uma metodologia eficiente que seja capaz de catalogar registros quanto ao modo de falha e fornecer maior confiabilidade na sua classificação, utilizando dados de manutenção de equipamentos da produção de petróleo em ambiente *offshore*.

1.4 Tema e Objetivos

Esta pesquisa é centralizada no tema de mineração de documentos textuais provenientes das atividades de manutenção para a catalogação de dados de confiabilidade, especificamente modos de falha de equipamentos da indústria do petróleo.

O objetivo geral desta pesquisa consiste na construção de um método para facilitar o processo de coleta e catalogação de dados qualitativos de confiabilidade, isto é, quanto a natureza da falha – modo de falha. A metodologia proposta objetiva utilizar técnicas para minerar textos não estruturados, escritos em português, obtidos a partir de dados históricos de manutenção de qualquer equipamento da indústria do petróleo.

Deste modo, o objetivo inicial deste trabalho é propor e implementar uma metodologia para minerar registros de manutenção com o auxílio de técnicas de Mineração de Textos (MT), *Natural Language Processing* (NLP) e *Machine Learning* (ML) para classificá-los automaticamente em modos de falha, seguindo os padrões da norma brasileira que trata da coleta e intercambio de dados de confiabilidade na indústria de óleo e gás.

Como objetivos específicos, deseja-se comparar os documentos obtidos em relação aos termos obtidos a partir de exemplos de descrições contidos na norma brasileira NBR ISO 14224 (ABNT, 2011). A comparação irá servir para avaliar se os termos mais representativos dos documentos de fato garantem que os modos de falha sejam corretamente identificados. Ademais deseja-se observar o quanto destes documentos textuais realmente são significativos para a classificação.

Por fim, deseja realizar a comparação de dois modelos de classificação de maneira a identificar o modelo mais adequado para a tarefa. Assim, é possível auxiliar o analista no processo de catalogação dos dados, acelerando o processo. Após a validação final de um especialista, será possível obter uma catalogação apropriada, colaborando na obtenção de dados com melhor qualidade e maior confiança.

1.5 Estrutura do Trabalho

Esta seção introduz cada um dos 9 capítulos desta dissertação: Introdução, Revisão Bibliográfica (Confiabilidade e *Text Mining*), Fundamentação Teórica, Definição do Problema, Metodologia, Estudo de Caso, Conclusão e Referências Bibliográficas.

Este Capítulo 1 apresentou uma breve introdução, contextualizando o tema a ser estudado, sua motivação e desafios, além de uma breve descrição do problema e os principais objetivos desta pesquisa. Em seguida, o Capítulo 2 e 3 se dedicam a apresentar uma breve revisão bibliográfica de trabalhos afins ou relacionados ao tema proposto. Já o Capítulo 4 consiste em introduzir a teoria de mineração de dados, apresentando os principais conceitos e fundamentos da mineração e classificação automática de textos que serão necessários para a compreensão do trabalho.

Apresentados os temas introdutórios, o Capítulo 5 descreve o problema de mineração de textos em português para catalogação de dados de confiabilidade. Já o Capítulo 6 é descrita a metodologia proposta para resolução deste problema. O Capítulo 7 consiste em apresentar o estudo de caso, assim como os resultados e discussões obtidas da experimentação do método desenvolvido. O Capítulo 8 elenca as conclusões consequentes da pesquisa, assim como as suas contribuições e sugestões para trabalhos futuros. Por fim, o Capítulo 9 apresenta as referências bibliográficas utilizadas como suporte para este trabalho.

2 CONFIABILIDADE

Este capítulo apresenta a revisão bibliográfica relacionada a confiabilidade, parte do tema desta pesquisa. Para melhor organização e compreensão, o capítulo foi dividido nos tópicos: estudos de confiabilidade na indústria do petróleo, dados de confiabilidade e bancos de dados de confiabilidade, apresentados a seguir.

2.1 Estudos de Confiabilidade na Indústria do Petróleo

Com a mecanização e automação dos processos na indústria em geral, e especificamente, a de óleo e gás, é possível observar um grande tendência das empresas do setor em empregar esforços para melhorar a confiabilidade de suas instalações e processos, evitando assim paradas não programadas.

As aplicações dos dados de confiabilidade em estudos de confiabilidade no setor de petróleo e gás são inúmeras. Dentre elas, as avaliações de riscos, a identificação de gargalos e equipamentos críticos do sistema, assim como ferramentas de melhoria para a disponibilidade do sistema e de sua capacidade produtiva, do planejamento e otimização das intervenções, tal como das estratégias de manutenção.

Em virtude da periculosidade presente nos processos da indústria do petróleo, a aplicação de estudos de confiabilidade na área de análise de risco é onde há maiores investimentos e esforços da indústria. Assim, o conceito de confiabilidade é comumente utilizado em estudos de segurança de processos, com aplicações de técnicas análises de risco qualitativas e quantitativas, como por exemplo: a análise de modo de falha e de criticidade de efeito (FMECA, do inglês *Failure Mode Effects and Criticality Analysis*), a análise de árvore de falhas (FTA, do inglês *Failure Tree Analysis*), o nível de integridade de segurança (SIL, do inglês *Safety Integrity Level*), a Análise Quantitativa de Risco (AQR) e outras.

Entretanto, outras aplicações de estudos de confiabilidade são observadas no âmbito da garantia da produção aplicada a gestão da manutenção de ativos da indústria do petróleo, com estudos de confiabilidade e de falhas, de diagnósticos e prognósticos de máquinas, Manutenção Centrada em Confiabilidade (MCC), RAMS (do inglês, *Reliability, Availability, Maintainability and Safety)* e outros.

Como nesta pesquisa deseja-se extrair informações de confiabilidade a partir das atividades de manutenção, serão apresentados a seguir alguns trabalhos de estudos de confiabilidade para melhorar o desempenho manutenção de equipamentos.

ARMITAGE (2003) propôs uma visão geral da aplicação do processo simplificado de MCC a equipamentos do sistema de perfuração de petróleo no intuito de melhor definir a necessidade de realizar a manutenção e a fim de preservar a função de equipamentos e sistemas de uma forma rentável. O procedimento proposto apresentou uma redução de 20% a 30% no trabalho de manutenção preventivo, identificando ou redefinindo os requisitos de manutenção de maneira mais eficaz. A aplicação da metodologia pode propiciar o desenvolvimento de um programa de manutenção sob medida, de acordo com situações e aplicações de equipamentos específicos.

BEVILACQUA *et al.* (2003) desenvolveram um estudo de falhas em 143 bombas centrífugas de uma refinaria, sob diferentes condições de operação, aplicando a técnica de árvore de regressão e classificação como uma solução não paramétrica para determinar os fatores críticos de operação e sua influência na confiabilidade destas bombas. Os resultados permitiram classificar as bombas em condições similares de falhas, demonstrando que o método foi útil na gestão da manutenção dos equipamentos, para melhorar sua eficiência e reduzir sua taxa de falhas.

PALMIERI *et al.* (2007) apresentaram um modelo de cálculo para estimar a probabilidade de falha de operação de uma instalação de produção de petróleo, para garantir uma operação sem acidentes, sem falhas perigosas, observando a política de segurança e meio ambiente estabelecida pela empresa. Assim, uma vez que a taxa de falha foi obtida, obteve-se o nível de confiabilidade do sistema. O estudo utilizou a técnica FTA para quantificar o sucesso ou falha e analisar separadamente e de maneira independente, os sistemas de segurança e operatividade. Por fim, os autores utilizaram o resultado das análises para identificar oportunidades de melhoria e realizar a reengenharia nos subsistemas e ter maior impacto no valor calculado da confiabilidade total.

WUTTKE; SELLITTO (2008) apresentaram uma metodologia para calcular a disponibilidade e localizar na curva da banheira uma válvula, pertencente à um processo petroquímico, ao longo de seu ciclo de vida. O estudo propôs modelar as distribuições dos tempos entre falhas e para reparo do sistema com a distribuição de Weibull para o cálculo das funções RAM (*Reliability, Availability, Maintainability*) e de taxa de falhas instantâneas. O trabalho considerou que a disponibilidade é afetada pela estratégia de manutenção adotada e que não pode ser generalizada para outras válvulas ou outras

plantas petroquímicas, mas que com o refinamento e robustecimento do método por repetição indutiva de casos pode-se realizar a generalização metodológica. Os autores apresentaram as limitações do estudo em razão do processo completo englobar mais equipamentos. O trabalho é de grande relevância devido a enorme quantidade de válvulas existentes em plantas produtivas e no fato que o comportamento da taxa de falha pode ser útil para a competitividade da empresa através da gestão da manutenção.

GUO *et al.* (2009) propuseram um modelo para avaliar a criticidade de equipamentos petroquímicos utilizando a avaliação abrangente de *Fuzzy* considerando os fatores de influência: perda de produção, efeito de segurança, efeito ambiental e custos de manutenção combinados a um algoritmo de *Back-Propagation* (BP) de redes neurais de três camadas. Além de estabelecer os critérios de avaliação e a função de pertinência do fator de influência, o estudo também realizou uma análise FMEA (do inglês, *Failure Mode and Effect Analysis*). Os autores utilizaram um estudo de caso em uma planta de etileno para avaliar a viabilidade do modelo, no qual os resultados dependem da função de associação e de um conjunto de fatores de peso. O método foi confiável e aplicável para avaliação de criticidade de equipamentos petroquímicos em Manutenção Centrada em Confiabilidade, mas para que a aplicação do modelo seja bem-sucedida e acurada em outros estudos, é necessário realizar o teste e as modificações nos fatores de peso para que o resultado da avaliação seja coincidente com a situação prática.

MAMMAN *et al.* (2009) estudaram como melhorar a confiabilidade de válvulas submarinas no âmbito da produção de óleo e gás, devido ao fato destas serem componentes críticos suscetíveis a falhas precoces e serem cada vez mais utilizadas em exploração em águas profundas. Foi utilizada o método FMECA para analisar as características de falha de válvulas dentro de uma Árvore de Natal Molhada (ANM), com o intuito de prever modos de falha críticos e as consequências resultantes e determinar as tarefas de manutenção apropriadas para os ativos. A partir da aplicação da FMECA foi possível identificar os modos de falha comum e dominante das válvulas, tendo em vista sua função primária e sua relação com a perda total e parcial da capacidade de controle de fluxo. Além disso, a metodologia proposta permitiu calcular financeiramente as consequências da falha, consideram as perdas de produção, custos de intervenção, reparação e custos ambientais. Os autores ainda utilizaram a Estratégia de Prevenção e Eliminação de Falhas (*Failure Elimination and Prevention Strategy*, FEPS) para evitar falhas no início da vida, melhorar a confiabilidade das válvulas submarinas e a disponibilidade geral da ANM.

DANTAS *et al.* (2010) apresentaram um estudo de confiabilidade aplicado a dados de tempo de vida de poços petrolíferos *onshore* da PETROBRAS, com o intuito de verificar a existência de relações entre o tempo de vida dos poços e características como: método de elevação, nível de produção, BSW (*Basic Sediments and Water*), razão gás óleo (RGO), unidade operacional de origem, entre outras. E, portanto, ajustar um modelo que indicasse o risco de falha dos poços. A modelagem probabilística dos dados foi realizada através do ajuste do modelo de regressão Weibull e as correlações entre o tempo de vida e suas características foi realizada através do teste da razão de verossimilhança. Como resultado, os autores puderam verificar que a distribuição de Weibull foi satisfatória para os dados avaliados e outras conclusões, como: (1) os poços produzindo com método de elevação BM (Bombeio Mecânico) são mais duráveis em relação aos poços BCP (Bombeio de Cavidade Progressiva); (2) os poços com alta produção de óleo tendem a apresentar menor tempo de falha em relação aos menos produtivos; (3) os poços com baixo e alto BSW mostram maior probabilidade de funcionamento do que os demais poços das outras unidades operacionais; (4) os poços mais antigos apresentam maior probabilidade de sobrevivência em relação aos poços mais jovens ao qual os autores julgam ser relacionado ao fato de um melhor conhecimento do poço; (5) que os altos valores de RGO provocam menor durabilidade dos poço; e (6) em geral, os poços com a bomba localizada em maior profundidade levam mais tempo até apresentar a primeira falha.

BRAGLIA *et al.* (2012) apresentaram uma abordagem baseada em estatísticas multivariadas, que permite classificar componentes mecânicos em termos de MTBF (*Mean Time Between Failure*) e identificar os parâmetros operacionais que influenciam sua confiabilidade. Especificamente, um conjunto de dados de falhas estruturadas foi classificado de forma significativa por meio de: análise de cluster; análise de variância multivariada; extração de características e análise discriminante preditiva. Isso tornou possível não apenas definir o MTBF dos componentes analisados, mas também identificar os parâmetros de trabalho que explicam a maior parte da variabilidade dos dados observados. A qualidade e a usabilidade industrial da técnica foram validadas por um estudo de caso aplicado a 126 bombas centrífugas instaladas em uma refinaria de petróleo.

GHAZVINIAN *et al.* (2012) trabalharam com a confiabilidade e incerteza na previsão de constantes dinâmicas-elásticas de rochas de reservatórios. Os autores adotaram como metodologia a regressão linear múltipla para apresentar uma equação empírica para previsão de velocidade de onda de cisalhamento (onda S), seguida de uma

análise probabilística utilizando simulação de Monte Carlo para avaliar a incerteza e a confiabilidade na predição de constantes dinâmicas-elásticas, com o módulo de Young e a razão de Poisson. Assim, a partir destas análises, consideraram a incerteza e a variabilidade das constantes dinâmicas-elásticas da rocha em relação às suas características como: porosidade, densidade, tensão horizontal mínima e pressão de sobrecarga, e determinaram os valores do módulo de Young e da razão de Poisson com certa probabilidade em uma amostra do reservatório. Ao que finalmente, avaliaram o impacto dos parâmetros log-data no valor das constantes dinâmicas-elásticas da rocha na amostra do reservatório.

WANG *et al.* (2012) realizaram a análise de confiabilidade da tubulação suporte da árvore de natal ao qual tem impacto direto no nível de segurança da árvore submarina. Para os autores, confiabilidade da resistência estrutural era alvo comum das publicações, apesar de outros fatores também afetarem a confiabilidade do sistema, como por exemplo: o nível de projeto, o erro humano e a validade do sistema de controle que provavelmente influenciariam a confiabilidade do suporte de tubulação. No estudo, foram consideradas a aleatoriedade de carga, materiais estruturais e geometria. A pesquisa foi realizada adotando como método de confiabilidade, o PDS (*Probability Design System*) do software ANSYS, que é capaz de avaliar a probabilidade de falha do sistema e a não-determinação dos parâmetros de saída, bem como a sensibilidade dos parâmetros de entrada. A análise de confiabilidade baseou-se na estatística matemática, análise de probabilidade e análise de elementos finitos, que tornavam o modelo de avaliação muito mais razoável, devido aos inúmeros fatores que afetam a incerteza do modelo estarem representados no cálculo.

CHO *et al.* (2013) estudaram sobre a melhora do serviço de confiabilidade para perfuração e avaliação de operações utilizando uma estratégia de MCC otimizado, com base no modelo desenvolvido pela empresa Baker Hughes para melhorar custo do ciclo de vida (LCC, do inglês *Life Cycle Cost*). O estudo considerava a exigência sofrida pelas empresas prestadoras de serviço para reduzir as taxas de falha de fundo de poço e estender a vida útil dos ativos de aluguel, devido à complexibilidade do sistema e o aumento dos custos das operações. O método da Baker Hughes definiu níveis de manutenção para ferramentas de perfuração e avaliação, a fim de padronizar o fluxo de trabalho e logística, enfatizando que a manutenção é conservadora, programada em uma tentativa de descobrir falhas antes que elas ocorram no campo. Os autores consideraram ser necessário observar

a degradação de todo o sistema deve para a avaliação. A distribuição optada pelos autores foi a distribuição de probabilidades de Weibull, um dos modelos mais utilizados em engenharia de confiabilidade. Ao final, os autores forneceram um modelo suporte descrito para aplicação e avaliação dos conceitos citados.

MENGUE; SELLITO (2013) realizaram um estudo para definição da estratégia de manutenção de bombas centrífugas na indústria do petróleo com base na manutenção centrada em confiabilidade. A metodologia foi aplicada a dados de tempos entre falhas e dos tempos para reparo do equipamento que foram modelados por distribuições de probabilidade, sendo possível calcular as funções RAM da bomba. Os ajustes de distribuição fatores RAM foi realizado pelo software computacional ProConf 2000, com base na distribuição de Weibull. Como resultado, os autores conseguiram identificar a fase da vida que o equipamento se encontrava e assim, definir a estratégia de manutenção de modo a eliminar os defeitos de projeto do equipamento, reforçar os itens que quebraram e remover as causas de origem das falhas.

NASERI; BARABADY (2015) propuseram avaliar a confiabilidade de um sistema de instalações de óleo e gás no Ártico com base em uma metodologia de árvore de falha utilizando análises gaussianas de *Fuzzy*. O intuito era utilizar a teoria dos conjuntos difusos para lidar com incertezas e sua propagação no emprego e agregação de opiniões de especialistas para previsão de confiabilidade de instalações em nível de sistema e componentes, em situações das quais os dados podem não estar disponíveis. A metodologia proposta fornecia uma base para decidir as medidas de adaptação ao clima gélido do Ártico que precisavam ser aplicadas se a confiabilidade do sistema estivesse abaixo do nível aceitável. O estudo se baseou também em dados de taxa de falha constante do equipamento obtidos no manual OREDA (*Offshore and Onshore Reliability Data*) e de distribuição de probabilidade de falha com perfil exponencial para estimar a confiabilidade do componente. Segundo os autores, o processo de julgamento de opiniões de especialistas é usado como uma ferramenta para modificar o tempo médio até a falha do equipamento, incluindo os impactos adversos das condições climáticas do Ártico no desempenho do equipamento. E que a estratégia pode ser modificada quando novos dados históricos estiverem disponíveis. Além disso, os autores também relatam que é necessário avaliar o número de especialistas selecionados e que estes devem ter uma compreensão adequada dos mecanismos de falha de vários componentes e dos efeitos das condições operacionais do Ártico nesses mecanismos para que se determine uma confiabilidade confiável dos resultados.

SANTOS; SELLITO (2016) realizaram uma abordagem quantitativa para estratégia de manutenção e de melhorias para aumento da disponibilidade em um posto de compressão de gases residuais do processo de destilação do petróleo, composto por dois compressores alternativos A e B em uma refinaria da indústria petrolífera. Foi realizado o cálculo da atual disponibilidade e a sugestão de ações para elevá-la com base nos dados modelados dos tempos até o reparo (*Time To Repair,* TTR) e até a falha (*Time Between Failure,* TBF) dos compressores individuais e do posto de compressão de gases como um todo. Além disso, as disponibilidades individuas de A e B, e do conjunto, foram calculadas em função dos valores médios dos modelos (MTBF e MTTR). Como resultado, foi possível notar que com a troca de matéria prima na refinaria (a planta passou a processar óleo mais pesado, originado do pré-sal, para o qual a instalação não foi projetada), o sistema passou a se comportar como se estivesse na fase de mortalidade infantil, indicando a existência de erros de projeto ou de definições inadequadas para o atual serviço. À estas divergências, os autores sugerem a estratégia de manutenção corretiva, ou seja, dado que uma falha ocorreu, não é feito apenas o reparo, mas tomadas medidas corretivas que sanem a deficiência. Assim, o artigo inclui uma lista de melhorias corretivas de projeto para aumentar a disponibilidade do posto e encerrar a fase de falhas prematuras. Os autores ainda sugerem estudos em indústrias cuja competição seja baseada em aspectos tecnológicos e de prestação de serviços e mais estudos em equipamento petrolíferos de grande porte e uso de simulação computacional para a avaliação técnica das soluções propostas.

SUN *et al.* (2016) analisaram a confiabilidade para manutenção estratégica de sistemas de *offloading* de FPSOs (*Floating Production Storage and Offloading*), em razão deste ser um módulo de chave, que sofre falhas ocasionalmente devido à grande imposição das intensas condições de trabalho. Para garantia do escoamento da produção é necessária uma manutenção eficiente, que muitas das vezes depende da precisão da previsão do período de manutenção. Os autores investigam a estratégia de manutenção pela análise de confiabilidade utilizando a técnica FMEA para identificar os principais eventos de falha e suas inter-relações. Os modos de falha do sistema são analisados com um modelo de árvore de falha dinâmica, considerando a conexão dinâmica entre modos de falha, sendo capaz de quantificar o período de manutenção para requisitos de confiabilidade em escalas diferentes. A importância crítica dos eventos de falha do sistema é calculada e permutada de cima a baixo, com resultados em conformidade com

as estatísticas WOAD (*World Offshore Accident Database*), que fornecem uma referência para a ordem de recondicionamento dos componentes no procedimento de manutenção. Os resultados apresentados no artigo demonstram os erros de projeto ou fabricação são o principal fator para acidentes em plataformas offshore. Ademais, os autores apontam que o projeto inadequado, pressão excessiva causada por *surge*, danos no flange e tensão causada por deslocamento, tem impacto direto nas causas de danos as mangueiras de *offloading* e consequentemente, grande impacto na segurança do sistema. Deste modo, o estudo sugere que as mangueiras são componentes vitais para o sistema de *offloading*, e que a redundância de confiabilidade deve ser considerada durante o procedimento de projeto e operação.

CORVARO *et al.* (2017) trabalharam com uma nova abordagem para avaliar as estratégias de manutenção com base em métodos, ferramentas e técnicas de engenharia (tempo médio de falha, tempo ocioso do equipamento e valores de disponibilidade do sistema) para identificar e quantificar falhas de compressores alternativos. O trabalho teve como objetivo identificar e avaliar os efeitos dos fatores tipo RAM com relação ao comportamento dos estados definidos para cada componente individual e componentes do compressor alternativo. Ademais, utilizou a análise de fatores RAM para comparar com futuras estratégias de manutenção dos compressores alternativos. Como resultado do estudo, foi possível: (a) definir e classificar os equipamentos e subsistemas que contribuem para a indisponibilidade do sistema; (b) propor e avaliar potenciais de otimização custo-efetivas para garantir a disponibilidade; (c) permitiu ter um tempo janela para programação de atividades efetivas de manutenção e falhas e principais eventos que contribuem para comprometer o processo produtivo com base na análise de criticidade dos componentes; (d) propor o planejamento da manutenção preventiva a partir do estudo de manutenção centrada em confiabilidade. Ao final, os autores ainda reconhecem a necessidade de uma abordagem pela qual os *targets* quantitativos de RAM são definidos no estágio de projeto conceitual e usados ao longo do ciclo de vida da planta para controlar e revisar o desempenho da RAM.

A Tabela 1 apresenta um resumo dos trabalhos expostos anteriormente neste item.

Tabela 1 – Resumo dos trabalhos de estudos de confiabilidade na indústria do petróleo

Autor	Abordagem de confiabilidade utilizada	Aplicações
ARMITAGE (2003)	Manutenção Centrada em Confiabilidade (MCC)	Perfuração de petróleo
BEVILACQUA *et al.* (2003)	Algoritmo de árvore de regressão para classificar fatores críticos de operação e sua influência na confiabilidade	Bombas centrífugas
PALMIERI *et al.* (2007)	Análise de árvore de falhas (FTA - Failure Tree Analysis)	Instalação de produção de petróleo
WUTTKE; SELLITTO (2008)	Modelagem das distribuições de tempo entre falhas por distribuição de Weibull e cálculo dos fatores RAM	Válvula de processo petroquímico
GUO *et al.* (2009)	Avaliação da criticidade de equipamentos utilizando uma abordagem abrangente Fuzzy, algoritmo de Back-Propagation (BP) e FMEA (Failure Mode and Effect Analysis)	Planta de etileno
MAMMAN *et al.* (2009)	FMECA (Failure Mode Effects and Criticality Analysis) e Failure Elimination and Prevention Strategy (FEPS)	Válvulas de uma Árvore de Natal Molhada (ANM)
DANTAS *et al.* (2010)	Modelagem probabilística através do ajuste do modelo de regressão Weibull	Poços petrolíferos onshore
BRAGLIA *et al.* (2012)	Estatísticas multivariadas para classificar componentes mecânicos em termos de MTBF (Mean Time Between Failure) e identificar os parâmetros operacionais que influenciam sua confiabilidade	Bombas centrífugas de refinaria

Autor	Abordagem de confiabilidade utilizada	Aplicações
GHAZVINIAN *et al.* (2012)	Simulação de Monte Carlo para avaliar a incerteza e a confiabilidade na predição de constantes dinâmicas-elásticas, com o módulo de Young e a razão de Poisson	Rochas de Reservatório
WANG *et al.* (2012)	PDS (Probability Design System), análise de probabilidade e análise de elementos finitos	Tubulação suporte da árvore de natal
CHO *et al.* (2013)	Manutenção Centrada em Confiabilidade (MCC), distribuição de probabilidades de Weibull e Life Cycle Cost (LCC)	Ativos de poços
MENGUE; SELLITO (2013)	Estudo RAM e distribuições de probabilidades de Weibull	Bombas centrífugas
NASERI; BARABADY (2015)	Árvore de falha utilizando análises gaussianas de Fuzzy em dados do OREDA	Sistema de instalações de óleo e gás
SANTOS; SELLITO (2016)	Abordagem quantitativa RAM para estratégia de manutenção e de melhorias para aumento da disponibilidade	Posto de compressão de gases residuais do processo de destilação do petróleo
SUN *et al.* (2016)	FMEA (Failure Mode and Effect Analysis) em dados da World Offshore Accident Database (WOAD)	Sistemas de *offloading* de FPSOs
CORVARO *et al.* (2017)	Cálculo RAM	Compressores alternativos

2.2 Dados de confiabilidade

De modo geral, os estudos de confiabilidade necessitam que a coleta e análise de dados de falha seja previamente realizada. Embora algumas empresas já realizem a coleta e o armazenamento de dados de confiabilidade há algum tempo, a exigência desta prática para atender às normas e regulamentos é um movimento mundial recente (ABNT, 2011). Em razão disso, a falta de padronização, baixa qualidade e ausência de dados de falha ainda são existentes e podem variar significativamente entre as empresas, representando perdas com elevado custo para indústria, tanto em termos econômicos quanto ambientais (SANDTORV *et al.*, 1996).

A incerteza nas ocorrências de falha e suas consequências é uma questão fundamental nas análises de confiabilidade. Neste sentido, a coleta de dados confiabilidade, seja em relação aos tempos de falhas (TBF ou TTR, e outros) quanto sobre as análises da natureza da falha e sua de causa raiz, tem como objetivo obter maior grau de confiabilidade em componentes e dispositivos do equipamento (ZIO, 2009).

Para auxiliar o processo de coleta e intercâmbio de dados de confiabilidade e manutenção na indústria do óleo e gás, a Associação Brasileira de Normas Técnicas (ABNT) publicou a Norma Técnica Brasileira – NBR ISO 14224 (ABNT, 2011) idêntica a norma internacional da *International Organization for Standardization* – ISO 14224 publicada em 2006. Apesar de desatualizada em relação a nova edição da norma internacional ISO 14224:2016 (ISO, 2016), ambas as normas são bastante equivalentes.

No entanto, mesmo com os investimentos e esforços dedicados à coleta e análise de dados de falha e manutenção, grande parte dos registros ainda são realizados em desacordo com as normas. A baixa qualidade dos dados disponíveis para catalogação dificulta a utilização de técnicas avançadas nos estudos de confiabilidade e manutenção de equipamentos. Além disso, impossibilita desenvolver modelos de manutenção adequadamente calibrados e adaptados, aumentando o ceticismo sobre sua eficácia nas previsões de confiabilidade (ZIO, 2009).

A seguir serão apresentadas alguns desafios na aplicação das técnicas de confiabilidade na indústria do petróleo devido a inexistência ou a pouco acurada informação disponível e confiável, que foram relatadas em alguns dos trabalhos do citados item 2.1.

WUTTKE; SELLITTO (2008) descrevem que independentemente do software ou ferramenta utilizada para a gestão da manutenção de válvulas de um processo petroquímico, a existência de bancos de dados consistentes, com informações confiáveis é muito importante para o desenvolvimento de trabalhos em confiabilidade e modelagem de falhas.

BRAGLIA *et al.* (2012) relatam que se o conjunto de dados disponível for grande o suficiente para realizar uma análise estatística significativa, seria possível dividir os itens em grupos homogêneos e obter uma discriminação eficiente, tanto em termos de classificação de dados quanto de predição de falhas.

NASERI; BARABADY (2015) relatam que é necessário uma compreensão adequada dos mecanismos de falha de vários componentes e dos efeitos das condições operacionais nesses mecanismos para que se determine a resultados confiáveis e deste modo, que se possa avaliar a confiabilidade de um sistema de instalações de óleo e gás com base em uma metodologia de árvore de falha utilizando análises gaussianas de *Fuzzy*.

SUN *et al.* (2016) relatam que a insuficiência dos dados de falha e manutenção disponíveis para analisar a confiabilidade para manutenção estratégica de sistemas de *offloading* de FPSOs pode limitar a precisão da avaliação. Ademais, também mencionam que em pesquisas futuras, um banco de dados de falhas de FPSO específico seria necessário para fornecer uma base para a formulação de estratégias de manutenção.

CORVARO *et al.* (2017) reforçam a dificuldade de trabalhar em estudos de casos reais para avaliar as estratégias de manutenção para identificar e quantificar falhas de compressores alternativos, devido às dificuldades que a indústria possui em captar, armazenar e de compartilhar seus dados operacionais.

Por esta perspectiva, é necessário empreender esforços para coletar dados representativos que alimentarão os bancos de dados utilizados nos estudos de confiabilidade. Porém, ainda existe uma grande lacuna sobre trabalhos que tratem da recuperação e catalogação de dados de confiabilidade da área de óleo e gás segundo as normas técnicas ou bancos de dados de confiabilidade existentes.

BENDELL (1988) apresentaram uma visão geral da coleta, análise e aplicação de dados de confiabilidade nas indústrias de processo. Os autores fornecem motivos por que, o que e como coletar dados de confiabilidade, além de identificar áreas problemáticas na obtenção de dados como modos e causas comuns, fatores humanos e confiabilidade de software, dentre outros.

SANDTORV *et al.* (1996) apresentaram que em um estudo realizado com as empresas participantes do projeto OREDA que a padronização e qualidade dos dados do OREDA forneciam um enorme benefício para a realização destes estudos, mas que 82% enfrentavam problemas de indisponibilidade de dados para realizar seus estudos de confiabilidade.

O estudo de AZADEH *et al.* (2010) utilizaram o banco de dados de confiabilidade OREDA para melhorar o processo de manutenção. Os autores desenvolveram um mecanismo de diagnóstico de máquinas baseado em lógica *Fuzzy* por meio da extração de regras linguísticas e aquisição de conhecimento obtida a partir do OREDA handbook. O trabalho ainda associa o impacto interativo dos modos críticos de falha nos parâmetros operacionais hidráulicos e mecânicos da bomba, incluindo vazão, pressão de descarga, NPSHR (*Net Positive Suction Head Required*), BHP (*Brake Horse Power*), eficiência, vibração e temperatura.

HAMEED *et al.* (2011) se dedicaram em desenvolver um banco de dados específico para a compreensão do comportamento real de turbinas eólicas em ambiente *offshore*, desde o projeto até a operação e de acordo com padrões internacionais sobre RAMS (*Reliability*, *Availability*, *Maintainability and Safety*) e iniciativas de risco. Os autores apresentam diversos desafios relacionados a criação deste tipo de bancos de dados, os principais em razão do aprimoramento da qualidade dos dados e do gerenciamento do banco de dados. Outros obstáculos citados pelos autores foram associados a fatores como desenvolvimento e aprimoramento tecnológico, introdução de novos conceitos, qualificação de novas tecnologias e otimização de estratégias de Operação e Manutenção. O banco de dados proposto embora estruturado para a implementação dos conceitos RAMS de forma fácil e eficiente, também permitia aplicações que não se limitavam a apenas esta estrutura.

Avaliar as probabilidades, modos e mecanismos de falha requer um processo padrão para coletar, analisar e usar dados, levando em consideração onde esses dados serão armazenados após a catalogação, tal como a modernização dos equipamentos e das ferramentas computacionais existentes. Deste modo, o próximo item irá apresentar os principais bancos de dados de confiabilidade da indústria do petróleo.

2.3 Bancos de Dados de Confiabilidade na Indústria do Petróleo

Os benefícios de dispor de um banco de dados de confiabilidade são inúmeros. Desde fornecer informações para tomadas de decisões quanto, por exemplo, ao momento adequado para realizar intervenções em equipamentos, à avaliação de custo de ciclo de vida de sobressalentes, a melhorias em instalações operacionais, entre outros. Ademais, outros benefícios se relacionam a coleta e análises dos dados diretamente, como possível redução de falhas catastróficas e de impactos ambientais, análises de tendência de desempenho, aumento da disponibilidade de unidades de processo, dentre outros (ABNT, 2011).

De maneira geral, as bases de dados de confiabilidade atualmente existentes dispõem, principalmente, de estimativas de taxa de falha na operação (λ), Probabilidade de Falha na Demanda (PFD), tempo médio de reparo (MTTR) tal como seus intervalos de confiança, modos e mecanismos de falha, dentre outros (NTNU, 2019).

Uma pesquisa realizada pelo *ROSS Gemini Centre* pertencente à universidade norueguesa NTNU (2019) reuniu vinte e duas (22) de bases de dados de confiabilidade, que dispõe principalmente de estimativas de taxas de falhas. Dentre elas, há bases que tratam de equipamentos eletrônicos, de partes eletrônicas ou não, modos de falha e mecanismo de distribuição, de componentes nucleares e da indústria *offshore*.

Os bancos de dados de confiabilidade mais conhecidos e utilizados estão constantemente sendo atualizados e aperfeiçoados conforme o surgimento de novos métodos e técnicas de análise, tal como da modernização dos equipamentos e das ferramentas computacionais (AKHMEDJANOV, 2001).

No que se refere àqueles da indústria de óleo e gás destacam-se: OREDA, WellMaster e iQRA (SPARKE, 2015). O primeiro, abordando diversas fases da vida dos sistemas na indústria de óleo e gás, e os dois últimos, muito utilizados como base de estudos na fase de completação de poços.

Apesar de apresentar sucintamente os três principais bancos de confiabilidade utilizados na indústria do petróleo, este trabalho se baseia em catalogar dados segundo a NBR ISO 14224 (ABNT, 2011) de forma a possibilitar a intercambialidade de dados de falha especificamente com o banco de dados OREDA.

2.1.1. OREDA

O OREDA - *Offshore and Onshore Reliability Data* é um abrangente banco de dados de confiabilidade e manutenção de equipamentos de exploração e produção. É considerado um dos mais importantes banco de dados para a indústria de óleo e gás. Somente na base de dados científicos ScienceDirect há pelo menos 385 produções citando-o, com publicações distribuídas entre os anos 1996 e 2020 conforme apresentado na Figura 1 (SCIENCEDIRECT, 2020).

As informações fornecidas pelo OREDA cobrem tanto o ambiente *offshore* quanto *onshore*, englobando dados de confiabilidade coletados em equipamentos *Topside* (e *Subsea*) para operações no Mar do Norte, Golfo do México, Oeste de *Shetland*, Angola, Adriático, Cáspio, entre outros. Além disso, abrangem especificidades tais como: tipos de equipamentos e instalações, condições de operação e áreas geográficas no qual os dados são captados (OREDA, 2019).

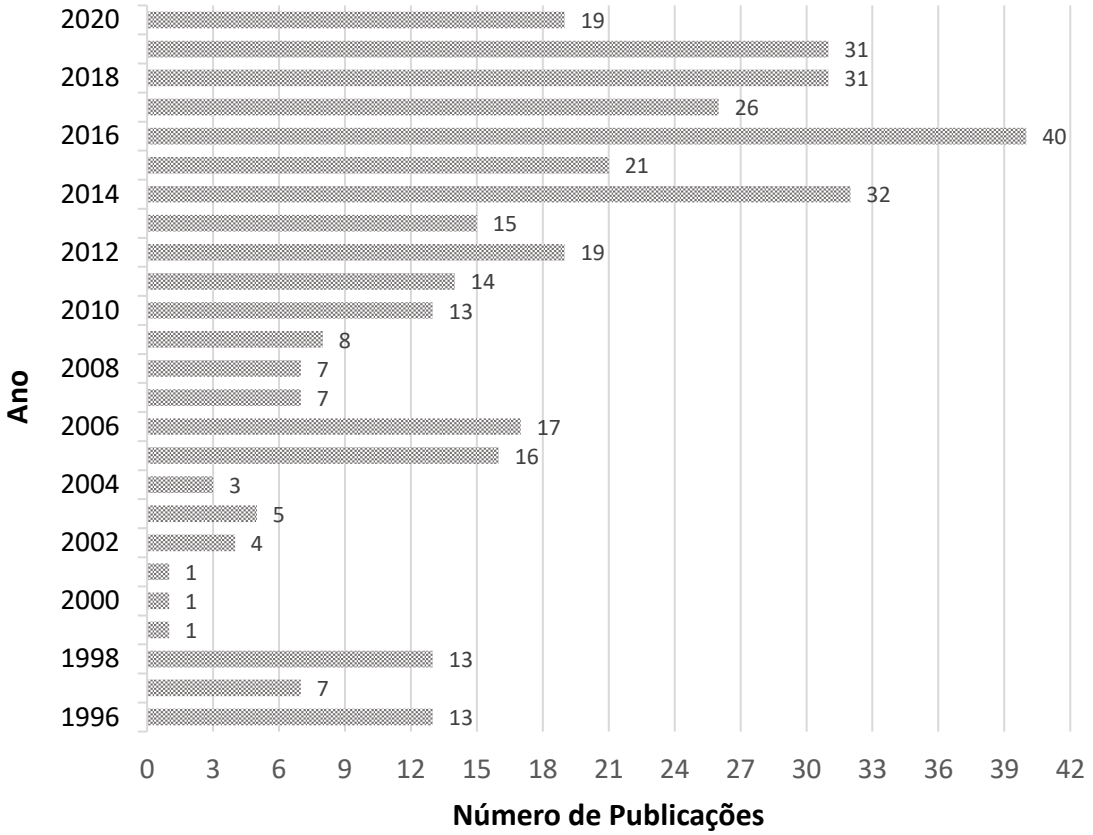


Figura 1 – Publicações que referenciaram o OREDA

Fonte: SCIENCEDIRECT (2020)

O projeto OREDA foi iniciado em 1981 por iniciativa da Diretoria Norueguesa de Petróleo (*NPD - Norwegian Petroleum Directorate*), desde 2004 conhecida como Autoridade de Segurança de Petróleo (*PSA - Petroleum Safety Authority*). O objetivo principal era coletar e avaliar dados de confiabilidade de equipamentos de segurança em condições operacionais. Em 1983, o OREDA passou a ser administrado por um grupo de companhias de óleo e gás. Assim, o objetivo inicial foi ampliado, passando a coletar dados de operações nas instalações de produção da indústria de petróleo *offshore* para melhorar dados básicos em estudos de confiabilidade de segurança (OREDA, 2019).

Da primeira até a quarta publicação, os manuais eram publicados em um único volume. A partir da quinta edição em 2009, o Handbook OREDA passou a ser publicado em dois volumes: o primeiro abrangendo os dados de confiabilidade de equipamentos de *topside* e o segundo, os dados de confiabilidade de equipamentos submarinos. Os dados de produção e exploração *onshore* passaram a ser incorporados nos manuais apenas na edição mais recente em 2015 (OREDA, 2019).

Estes manuais oferecem, de maneira geral, uma fonte de dados única com informações sobre taxas de falhas, distribuição de modo de falha e tempos de reparo para equipamentos usados na indústria de petróleo. Alguns exemplos das aplicações de estudo para o OREDA são listados a seguir:

- Análises de confiabilidade, disponibilidade e mantenabilidade (RAM);
- Planejamento e programação de manutenção, inspeção e testes;
- Análises de riscos e segurança de processo;
- Estudos de custo-benefício;
- Estudos regulares;
- Seleção de projetos de sistemas alternativos.

Os dados coletados e armazenados no OREDA compreendem, atualmente, "um banco de dados composto por: 278 instalações, 17.000 unidades de equipamentos com 39.000 falhas e 73.000 registros de manutenção". São registrados também de dados de classificação, especificação, manutenção e de operação; além de dados de identificação, de eventos com falha, de evento e recurso de manutenção e observações (EISINGER; CLAVÉ, 2018).

Os manuais também contemplam os seguintes dados de falhas: quanto ao número de falhas, valor mínimo de taxa de falhas a 90% do IC (Intervalo de Confiança), valor médio de taxa de falhas, valor máximo de taxa de falhas a 90% do IC, desvio padrão, razão entre número de falhas e tempo total em serviço e modos de falhas (OREDA COMPANIES, 2002).

2.1.2. WELLMASTER

O projeto WellMaster foi iniciado por um grupo de operadores no Mar do Norte, quando a ExproSoft ainda era associada ao Instituto de Pesquisa SINTEF. O projeto foi pioneiro entre os trabalhos de confiabilidade de poços e tinha como objetivo armazenar e analisar dados históricos de equipamentos de completação de poços da indústria de óleo e gás. Em 2001, com a fundação da empresa norueguesa ExproSoft, a base de dados passou a ser comercializada para atender operações *offshore* e *onshore* globalmente. A ferramenta tem contribuído para os criação e atualização dos padrões NORSOK e ISO desde 2006, incluindo a última revisão da ISO 14224 (CHOI *et al.*, 2013; SPARKE, 2015; EXPROSOFT, 2018 e 2019).

Quando começou, na década de 80, o WellMaster era apenas um banco de dados de confiabilidade de válvulas de segurança de subsuperfície (*DHSH - Down Hole Safety Valve*). Com o passar dos anos, a informações contidas no banco de dados passou a abranger em totalidade a operação de completação de poços, cobrindo toda a operação do poço, árvores de natal e diversos outros equipamentos e dispositivos submarinos e de superfície (SPARKE, 2015).

O WellMaster *RMS* é capaz de coletar e estruturar dados em categorias que permitem realizar filtros em relatórios e análises. O *software* pode ser utilizado durante todo o ciclo de vida do poço, desde o projeto de poços; a seleção de equipamentos mais adequados a operação, até a avaliação de riscos, análises de integridade de poço e avaliações de vida remanescentes. Além disso, as informações de *input* inseridas, quando compartilhadas, são confidenciais (EXPROSOFT, 2019).

Até 2018, o WellMaster *RMS* (*Reliability Management System*) já havia registrado dados indústria de óleo e gás em 34 países, 30 operadoras, 352 ativos, 7.783 poços, 280 instalações, 67.336 componentes de equipamentos e 6.176 falhas (EXPROSOFT, 2018). Este mesmo relatório da EXPROSOFT (2018) também inclui o quantitativo total para cada poço presente no conjunto de dados disponibilizado pelo WellMaster RMS. A relação contendo a classificação dos poços é apresentada na Figura 2.

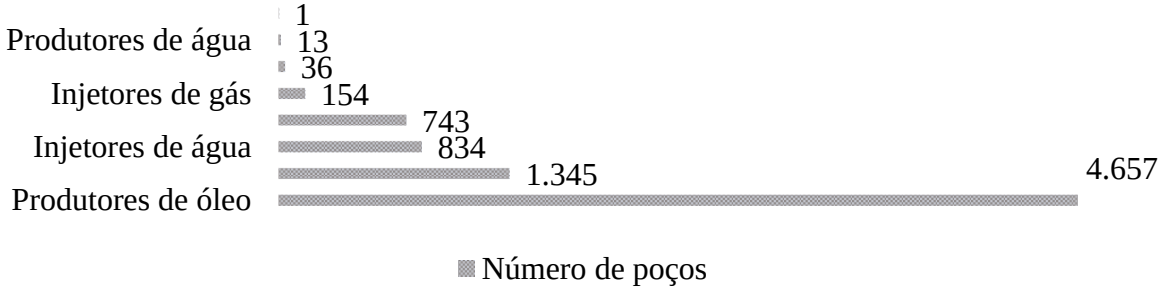


Figura 2 – Classificação dos poços disponibilizados no WellMaster RMS

Fonte: EXPROSOFT (2018)

Atualmente, a empresa comercializa além do gerenciador de bancos de dados em confiabilidade, o *software* Miriam RAM Studio para simulação e análises de fatores RAM (EXPROSOFT, 2019).

Os dados contemplados no WellMaster *RMS* são especificados pelas quatro categorias apresentadas (EXPROSOFT, 2019):

1. Dados do poço e dos ativos – Fornecem informações de alto nível sobre o operador, a localização do campo, os tipos de poço e condições operacionais.
2. Dados do equipamento – Os dados dos equipamentos dos poços são armazenados com atributos-chave no banco de dados, em um formato que permite a filtragem específica ao conduzir relatórios detalhados e comparativos entre campos e ativos.
3. Histórico do equipamento – Cada componente tem seus dados de experiência registrados, o que permite ao sistema calcular os principais indicadores de confiabilidade.
4. Dados de Falha – Os dados de falha registrados incluem detalhes categorizados para relatórios de alto nível e investigação de falha de poço único / equipamento.

2.1.3. iQRA

A base de dados iQRA foi desenvolvida pela Wood Group Intetech com o objetivo de ser uma ferramenta sofisticada com capacidade de análise de confiabilidade, ao qual foi inserida no *software* de gerenciamento de integridade de poços (WOOD GROUP INTETECH, 2019).

O banco de dados permite a avaliação da confiabilidade de maneira independente e imparcial, a partir do desempenho de todos os componentes presentes no poço. Assim, é possível realizar a comparação instantânea dos valores de confiabilidade da operação de cada empresa com os dados da iQRA, extraindo dados de confiabilidade crítica e de tempo médio de falha (MTTF) para avaliações quantitativas de risco e facilitar a tomada de decisão. Além disso, o conjunto de dados, datados desde a década de 70, cobre poços *onshore* e *offshore* em todo o mundo e também se baseia na experiência em dados de teste de componentes submarinos disponíveis no Wood Group Kenny (SPARKE, 2015; WOOD GROUP, 2019).

Segundo SPARKE (2015) este é um grande e diversificado banco de dados, capaz de fornecer dados desde o início da vida do poço. Ainda segundo o autor, há um grande potencial e benefícios para a indústria de confiabilidade, conforme há o aumento da participação dos operadores e a sua vinculação a estes projetos de *joint industry*. A

ampliação do espaço amostral dos dados permite a inclusão de dados de diferentes regiões, ambientes e experiências.

Este banco de dados permite comparar o desempenho de componentes individuais em seu próprio conjunto de dados em relação aos valores de confiabilidade média global e obter valores estatisticamente significativos. Além de possibilitar extrair dados de tempo médio para falha (MTTF) de modo a avaliar quantitativamente o risco e realizar decisões em momentos críticos do sistema; e, realizar consultas aos dados e gerar informações instantaneamente, por meio do recurso *online* via nuvem e a partir de qualquer lugar (WOOD GROUP, 2019).

3 TEXT MINING

Este capítulo apresenta alguns trabalhos da literatura que relatam o uso de métodos de mineração de textos, processamento de linguagem natural e outras recentes tecnologias de processamento e recuperação de informação no contexto da indústria do Petróleo e atividades de manutenção de equipamentos.

3.1 Mineração de textos na indústria do petróleo

Os trabalhos indicados a seguir foram obtidos a partir de uma busca na *One Petro* – uma base de dados de produção científica em aplicações da indústria do petróleo. O intuito da revisão é examinar trabalhos recentes que abordam o tema desta pesquisa especificamente neste setor industrial. A busca dos artigos foi realizada (em português e inglês) a partir dos tópicos "*text mining*", "*natural language*", "*maintenance*", "*work order*", "*maintenance logs*" e "*reliability data*".

Os termos "*text mining*" e "*natural language*" combinados com as características dos dados "*maintenance*", "*work order*", "*maintenance logs*" ou "*reliability data*" forneceram pouquíssimos resultados relacionados aos tópicos pesquisa. Assim, a busca foi realizada utilizando apenas os tópicos "*text mining*" e "*natural language*", que forneceram 44 resultados. Do número total de artigos encontrados todos foram pré-avaliados e apenas 10 serão detalhados a seguir.

WU *et al.* (2014) realizaram um estudo para demonstrar como extrair informações e descobrir conhecimento a partir de uma enorme quantidade de dados textuais não estruturados com valor de menor densidade utilizando a tecnologia Hadoop. O procedimento proposto foi aplicado em dados contendo mais de 10.000 resumos públicos da SPE (*Society of Petroleum Engineers*) que são relevantes para as práticas mundiais de exploração e produção de campos de petróleo pesado. Os dados foram coletados, pré-processados e armazenados no sistema de arquivos Hadoop. A eficácia e precisão do processo de descoberta de conhecimento foi validada com a comparação dos valores extraídos aos do sistema construído manualmente. Ao final os autores conseguiram que os processos de recuperação de informações e de mineração de conhecimento baseados na tecnologia de big data do Hadoop fossem capazes de recuperar valores numéricos e de texto não estruturados em dados brutos da indústria de exploração e produção, fornecendo tendências estatísticas semelhantes com o sistema manual.

SIDAHMED *et al.* (2015) apresentaram um estudo para melhorar a capacidade de monitoramento das operações de perfuração utilizando técnicas baseadas em dados e mineração de conteúdo não estruturado. A metodologia utilizou técnicas tradicionais de processamento de linguagem natural (NLP, do inglês *Natural Language Processing*) a relatórios de perfuração coletados diariamente de três poços para extrair padrões perspicazes dos logs de perfuração. Técnicas de aprendizado não supervisionados foram utilizadas para reduzir, simultaneamente, o tempo necessário para coletar e processar esses dados. E, o emprego de técnicas de extração de conceitos e frequência de padrões, forneceram a compreensão dos dados não estruturados relevantes para o processo, conseguindo rastrear e monitorar os sintomas relatados do comportamento observado para ajudar a identificar a causa raiz e a composição de fatores que levaram a ocorrência de um evento. Três importantes conclusões obtidas a partir deste trabalho são: (i) mesmo textos não estruturados podem fornecer novas informações sobre eventos que não são facilmente capturados com outros tipos de dados; (ii) informações ricas em forma de texto estão sendo subutilizadas pela indústria; e, (iii) o estudo não se beneficia dos recentes avanços na NLP e do aprendizado profundo, que podem escalar muito a recuperação de informações para ambientes industriais complexos.

ARUMUGAM *et al.* (2016) utilizaram diversas técnicas de mineração de texto como: recuperação de informação, extração, *clustering* e identificação de padrões; aliadas a gestão do conhecimento, para extrair eventos desfavoráveis (anomalias) nas atividades de perfuração relatadas e de modo a integrá-los às informações de subsuperfície. Os resultados podem ser integrados às técnicas geoespaciais para realizar a avaliação de riscos, aproveitando grandes conjuntos de dados de poços espacialmente distribuídos. A avaliação de risco pode ser executada para cada atividade da plataforma, faixa de profundidade definida pelo usuário, seção / tamanho do furo e marcadores de formação futuros. Assim, a abordagem proposta: (i) agiliza a avaliação de riscos no estágio de planejamento do poço e permite uma colaboração eficaz entre as equipes de perfuração, geológica e geofísica durante a execução; (ii) permite criar um inventário de risco eficaz que pode ser reutilizado para fins futuros de planejamento de poços; (iii) melhora as práticas de gerenciamento de riscos e gerenciamento de conhecimento.

ARUMUGAM; RAJAN; GUPTA (2017) realizaram a descoberta de conhecimento de textos em relatórios de perfuração para rastrear operações de perfuração contra formações geológicas e compensar poços em termos de perda de lama, vibrações e outros. O conteúdo é processado, e baseado em um conjunto de palavras e sua

frequência de ocorrência, os relatórios são classificados em temas utilizando a técnica de alocação de Dirichlet latente (*Latent Dirichlet Allocation* – LDA). O LDA possibilita que cada relatório analisado seja considerado como uma mistura de temas presentes no corpus textual, facilitando o processo de associação Ontológica entre os sintomas e suas causas correspondentes. Para os autores, a método utilizado auxilia que geólogos identifiquem os poços de compensação de maior desafio operacional, fornecendo uma análise de causa raiz mais rápida, tal como uma melhor compreensão da operação de perfuração e do planejamento do poço.

HOFFIMANN *et al.* (2018) propuseram uma metodologia para a classificação automática de sentenças escritas em relatórios de perfuração utilizando três rótulos: "evento", "sintoma" e "ação". Neste trabalho, desenvolvimentos recentes no processamento profundo de linguagem natural (NLP profundo) foram aplicados para classificar automaticamente sentenças em milhares de relatórios de perfuração de centenas de poços em um campo real. Segundo os autores, a ferramenta proposta pode ser usada *offline* por uma empresa de energia interessada em verificar relatórios antigos de perfuração para identificar padrões de operação, ou por uma agência governamental interessada em investigar as consequências de desastres ambientais. Sobre o trabalho é possível concluir que relatar e classificar completamente as atividades de perfuração são tarefas extremamente desafiadoras para uma força de trabalho humana limitada e que alguns dos principais desafios da recuperação da informação incluem: a alta frequência de símbolos técnicos; erros de digitação; abreviação de termos técnicos; e, a presença de frases incompletas nestes relatórios.

CASTIÑEIRA *et al.* (2018) desenvolveram um método para extrair automaticamente análises inteligentes e oportunidades a partir dos relatórios de perfuração e completação de poços. Inicialmente, utilizaram uma combinação de algoritmos de processamento de linguagem natural, mineração de dados e aprendizado de máquina para verificar a qualidade de um grande volume de dados de perfuração (incluindo o texto nos relatórios diários de perfuração), extrair informações cruciais e prever o tempo não produtivo e seu tipo. Em seguida, realizaram a integração dos conjuntos de dados de perfuração e completação a outras fontes de dados, como produção, geologia, reservatório, etc. de maneira a gerar um conjunto de métricas cruciais de gerenciamento de perfuração e gerenciamento de reservatório. O método resultou em uma redução significativa da tarefa de verificação da qualidade que exige muito trabalho para milhares de conjuntos de dados e na classificação imparcial dos eventos.

MAHASIVABHATTU *et al.* (2019) apresentaram o uso da Inteligência Artificial, e de técnicas como *Optical Character Recognition* (OCR) e NLP, no processamento de um desenho digitalizado, redesenhando-o automaticamente em uma plataforma digital. Segundo os autores a adaptação da abordagem proposta pode trazer uma vantagem considerável na busca pela digitalização.

Objetivando melhorar a eficiência do gerenciamento de dados da indústria do petróleo, ASFOOR *et al.* (2019) utilizaram técnicas de NLP e conceitos da Teoria *Fuzzy* para lidar com aspectos de aprimoramento da capacidade de pesquisa de arquivos e da redução do trabalho manual e a redundância de dados. A metodologia proposta consistia em extrair automaticamente palavras-chave de pesquisa em arquivos, marcando-os para rotulá-los com a respectiva categoria de negócios correta definidas previamente. Adicionalmente, a abordagem fornecia um grau (*Fuzzy*) de associação do arquivo a outras categorias, permitindo que os gerenciadores de dados encontrem arquivos semelhantes e duplicados em vários repositórios de arquivos da indústria de O&G. O algoritmo não usa metadados definidos manualmente pelos usuários e possuíam outras funcionalidades além das descritas anteriormente.

Enquanto MILANA *et al.* (2019) apresentaram uma metodologia que utiliza técnicas de *Machine Learning* (ML) e NLP para extrair padrões e correlações do texto, fornecendo à equipe de HSEQ (*Health, Safety, Environment, & Quality*) informações valiosas para a gestão da segurança e dos riscos em plantas de óleo e gás. O método foi proposto para desenvolver um sistema de segurança de pré-detecção que associa técnicas de aprendizado de máquinas ao processamento de linguagem para incorporar várias fontes de informação. A abordagem fornece meios de detectar, independentemente do idioma, semelhanças entre documentos escritos em línguas diferentes e agrupando-os de acordo para obter informações de HSEQ.

Já SALO; MCMILLAN; CONNOR (2019) trabalharam especificamente com a análise de ordens de serviço no contexto atual de digitalização. Os autores desenvolveram um software de mineração de texto que combina métodos clássicos de aprendizado de máquina – como cluster hierárquico, com o conhecimento especializado do operador obtido por meio de uma abordagem de aprendizado ativo. A métrica de distância foi adaptada da pesquisa teórica da informação para melhorar o desempenho do cluster. A ferramenta apresentou ótimos resultados, cerca de 90% do tempo de trabalho (de duas semanas úteis para um único dia), reduzindo de maneira significativa o esforço analítico para analisar dados reais de ordem de serviço. Os autores também avaliaram sobre a

incerteza dos resultados, fator chave para a implementação em contexto de tomada de decisão. O estudo corrobora para justificar o potencial das práticas de inteligência artificial para impulsionar a digitalização não apenas nas novas instalações, mas também naquelas mais antigas, onde, a grande quantidade de dados históricos não estruturados possui enorme valor, e são uma vantagem na compreensão dos eventos de falha.

NOSHI; SCHUBERT (2019) apresentaram um *survey* sobre as técnicas de mineração de textos, apresentando situações em que cada técnica pode ser benéfica e eficaz na recuperação de informações de bancos de dados textuais de várias fontes.

Além dos artigos citados na revisão realizada por NOSHI; SCHUBERT (2019), ANTONIAK *et al.* (2016), PRIYADARSHY *et al.* (2017), MA *et al.* (2018), TIAN *et al.* (2019), COLOMBO *et al.* (2019) e UCHEREK *et al.* (2020) são trabalhos que utilizaram as técnicas de mineração/processamento de textos a dados de poços de perfuração.

Um resumo dos trabalhos apresentados neste item é apresentado na Tabela 2. A partir dos resultados obtidos na *One Petro*, observa-se, que apesar da busca ter sido orientada ao tema da pesquisa, foram encontrados pouquíssimos trabalhos que tratem exclusivamente deste tema. É notório que os trabalhos utilizando técnicas de mineração de texto neste setor são bastante recentes, sendo majoritariamente empregadas à obtenção de informações nas atividades de perfuração. Assim, uma nova pesquisa bibliográfica com enfoque em processar dados de manutenção será apresentada na seção 3.2.

Tabela 2 – Resumo dos trabalhos de mineração de textosna indústria do petróleo

Autor	Abordagem de confiabilidade utilizada	Aplicações
SIDAHMED et al. (2015)	Mineração de dados não estruturados, NLP e aprendizado não supervisionado	Relatórios de perfuração
ARUMUGAM et al. (2016)	Técnicas de recuperação de informação, extração, clustering e identificação de padrões	Relatórios de perfuração
ARUMUGAM; RAJAN; GUPTA (2017)	Técnica de alocação de Dirichlet latente (Latent Dirichlet Allocation – LDA)	Relatórios de perfuração
HOFFIMANN et al. (2018)	Processamento profundo de linguagem natural (NLP profundo)	Relatórios de perfuração
CASTIÑEIRA et al. (2018)	Combinação de algoritmos de processamento de linguagem natural, mineração de dados e aprendizado de máquina	Relatórios de perfuração e completação de poços
MAHASIVABHATTU et al. (2019)	Optical Character Recognition (OCR) e NLP	Desenho técnico de plataformas de petróleo
ASFOOR et al. (2019)	NLP e Teoria Fuzzy	Repositórios de arquivos da indústria de O&G
MILANA et al. (2019)	Machine Learning (ML) e NLP	Informações de gestão da segurança e dos riscos em plantas de óleo e gás
SALO; MCMILLAN; CONNOR (2019)	Cluster hierárquico e conhecimento especializado do operador obtido por meio de uma abordagem de aprendizado ativo	Ordens de serviço
NOSHI; SCHUBERT (2019)	Revisão de trabalhos de mineração de textos na indústria do petróleo	N/A

3.2 Mineração de textos de manutenção

Esta nova revisão foi realizada nos repositórios de produções científicas *ScienceDirect* e *Google Scholar* utilizando os mesmos termos de busca empregados na pesquisa anterior, mas sem especificar a indústria de aplicação.

A partir dos resultados obtidos a partir da palavras chaves definidas no item 3.1, é possível notar que mesmo realizando uma busca mais ampla em grandes repositórios há uma enorme lacuna de artigos que correlacionem métodos de mineração de textos e dados de manutenção para o estudo de falhas, seja no panorama da indústria do petróleo ou na indústria em geral. Talvez, a dificuldade de encontrar trabalhos com este tema deva-se ao desenvolvimento ainda recente das técnicas de mineração de textos ou em virtude da compreensão equivocada que dados históricos não estruturados possuem pouco valor frente aos esforços investidos na obtenção da informação. A seguir são apresentados os trabalhos encontrados nesta segunda pesquisa bibliográfica.

DEVANEY *et al.* (2005) propuseram uma arquitetura para extrair e categorizar componentes e subsistemas de equipamentos e suas falhas associadas analisando textos livres de logs de manutenção. A abordagem combinou técnicas de mineração de textos, processamento de linguagem natural, *Machine Learning* e ontologia de domínio. Embora os autores propusessem uma estrutura de análise baseada na construção de uma biblioteca de casos, nenhum estudo de caso com dados reais foi apresentado.

CHEN; NAYAK (2007) utilizaram técnicas de mineração de textos para extrair e classificar o modo de falha de cada registro. Os autores utilizaram dois tipos de métodos e ferramentas de *clustering*, o método de *clustering* de Ward com *SAS Text Miner* e o método de *clustering* por Histograma de Similaridade. Os resultados do experimento, no entanto, tiveram baixo desempenho devido ao tamanho do volume de dados utilizados e da falta de padronização das descrições dos eventos de falha. Outro fator dificultador apresentado pelos autores é que caso alguns termos e relacionamentos de cluster fossem conhecidos/disponíveis como conhecimento do domínio e configurados antes do experimento, o desempenho dos métodos utilizados poderia ser melhor. CHEN; NAYAK (2007) também ressaltam a dificuldade de avaliar o aprendizado não supervisionado, quando não há conhecimento prévio ou dados de treinamento disponíveis, principalmente devido à falta de padronização dos registros (os termos utilizados para descrever os eventos de falha são bem diferentes em cada registro). Os autores sugerem que em

trabalhos futuros sejam consideradas a criação de regras relevantes que trabalhem a confiança do agrupamento de registros. Adicionalmente, propõem que no futuro seja realizada a obtenção de um conjunto de dados de modos de falha, categorizados anteriormente por especialistas humanos, possibilitando a comparação das abordagens automatizadas com a manual, e entre si.

MUKHERJEE; CHAKRABORTY (2007) mineraram de dados de manutenção para extrair informações que são ou podem ser empregadas para enriquecer os modelos de confiabilidade. Os autores desenvolveram um processo para automatizar a construção de árvores de falha e melhorar a estimativa da confiabilidade, analisando os dados textuais de manutenção disponíveis em formato livre. O método proposto utiliza uma combinação de análise linguística e conhecimento de domínio para identificar a natureza da falha a partir de descrições textuais simples e breves a respeito das falhas do equipamento.

EDWARDS *et al.* (2008) analisaram logs de manutenção para verificar se recursos textuais dos dados de manutenção poderiam ser utilizados para classificar registros em atributos estruturados. O estudo objetivava categorizar os logs de manutenção em trabalhos de manutenção programada ou reparo não programado. A abordagem proposta foi empregada para analisar um estudo de caso contendo 12 anos de registros de manutenção de uma estação de bombeamento armazenados como texto livre em planilhas. Incialmente, o componente *Text Miner* do *SAS Enterprise Miner* foi empregado para realizar a conversão do texto livre em pesos de termo, executar a decomposição em pesos de termo empregando o *Singular Value Decomposition* (SVD) e criar os *clusters* de texto. Em uma etapa posterior, árvores de decisão foram treinadas usando o peso dos termos para fornecer algumas dicas sobre quais palavras melhor descreveriam cada *cluster*, e deste modo, tentar prever em qual dos *clusters* identificados um novo registro se encaixaria. Um segundo classificador foi treinado para prever o grupo de *cluster* visando a precisão sobre a interpretabilidade. Como os *clusters* não puderam determinar a classificação dos registros com precisão desejada, os dados foram rotulados manualmente como falha ou não falha, e assim, dois novos classificadores foram testados. De acordo com os modelos de classificação do *cluster*, uma árvore de decisão foi treinada usando o termo pesos como entrada, enquanto uma rede neural foi treinada usando os componentes SVD. As taxas de classificação incorreta obtidas foram de 15,0% (árvore de decisão) e 17,2% (rede neural com SVD). O estudo apresenta que com algumas informações de especialistas no assunto, os recursos de texto podem ser usados para classificar documentos em pequenos conjuntos de dados com um nível moderado de

precisão, e que mesmo em situações de dados de baixa qualidade, a mineração de textos é viável.

MARZEC *et al.* (2014) realizaram um estudo para verificar se os métodos de mineração de texto existentes e atualmente usados são suficientemente precisos para serem utilizados na classificação de dados não estruturados de manutenção e reparo. O objetivo do trabalho consistia em determinar se uma descrição específica estava relacionada à manutenção corretiva ou preventiva a partir de um estudo de caso com dados de garantia de um dos principais fabricantes de ônibus da Europa. Uma contribuição do estudo é a identificação da relação de dependência entre a precisão das técnicas de mineração de texto e fatores como a complexidade de um *target*, os algoritmos usados, a natureza, qualidade e quantidade dos dados.

ARIF-UZ-ZAMAN *et al.* (2017) propuseram uma nova abordagem para extração de dados de tempo de falha, aplicada a dois estudos de caso para aumentar a precisão da estimativa dos tempos históricos de falha e obter a confiabilidade de um ativo, determinando conjuntamente quando este falhou utilizando dados de manutenção contidos em ordens de serviço e de dados de tempo de inatividade. Os autores utilizaram dois algoritmos de classificação: *Naïve Bayes* (NB) e *Support Vector Machine* (SVM).

GONÇALVES *et al.* (2018) realizaram a análise bibliométrica de diversos estudos de mineração de textos na área de manutenção em diversas indústrias de aplicação. Os autores ressaltam a importância do tema e de suas aplicações para melhorar a atividade de manutenção, reduzir custos e auxílio a tomada de decisão. O trabalho apresenta que ganhos econômicos e operacionais podem ser obtidos a partir da extração de informações, correção de erros ortográficos nos documentos, extração de características, indicação de ocorrência de eventos e gestão do conhecimento.

GUNAY *et al.* (2019) e BORTOLINI; FORCADA (2019) propuseram minerar textos de manutenção de edifícios.

Enquanto GUNAY *et al.* (2019) apresentaram um método de mineração de texto para extrair informações sobre padrões de falhas a partir do pré-processamento das descrições de Ordens de Serviço (OS). A primeira etapa consistiu em converter as OS para realizar uma análise lexical quantitativa, e então, os dados foram agrupados em seções de interesse contendo ordens de serviço sobre falhas na construção de sistemas e componentes - em vez de atividades de manutenção e inspeção de rotina menos interessantes. Na última etapa, foram utilizadas regras de associação de mineração para identificar as tendências de coexistência entre os termos do cluster de interesse (por

exemplo, coexistência dos termos "radiador" e "vazamento"). A aplicabilidade da metodologia foi demonstrada usando dois conjuntos de dados. O primeiro de uma planta de aquecimento e refrigeração central com quatro caldeiras e cinco *chillers*; e o segundo, um cluster de 44 edifícios. Segundo os autores, os resultados forneciam informações detalhadas por eventos de falha de cada equipamento do sistema, modos de falha de seus componentes e suas frequências de ocorrência.

Já BORTOLINI; FORCADA (2019) propuseram uma abordagem de mineração sistemática de texto utilizando o algoritmo *MapReduce* para analisar e extrair informações das solicitações de manutenção dos usuários de edifícios, de modo a avaliar a condição da construção de sistemas. O estudo de caso consistiu em analisar 6.830 solicitações de manutenção derivadas de 46 edifícios localizados na Espanha, incluindo escritórios, edifícios acadêmicos e laboratórios. O objetivo do trabalho era obter informações sobre os problemas típicos de um edifício em uso, determinando se as características do edifício como uso (escritório, acadêmico e laboratórios), propriedade do edifício (público e privado), área bruta do solo (GFA), número de pisos e idade (ano da construção) e outros fatores eram determinantes nas solicitações de manutenção. Os resultados da pesquisa revelaram que as solicitações de manutenção mais comuns durante a operação e manutenção dos edifícios estão relacionadas a problemas em equipamentos elétricos e sistemas de climatização (aquecedor, ventilador ou ar condicionado). Embora o ano de construção não esteja relacionado às solicitações de manutenção dos usuários, o tipo de uso e a propriedade do edifício os influenciam. A implementação de estratégias de controle e preventivas com base nesses resultados pode aumentar a produtividade dos gerentes de instalações e o desempenho dos sistemas de construção.

Já a metodologia proposta por BLANCO-M *et al.* (2019) baseou-se em um conjunto de etapas para pré-processar e decompor o histórico de serviço e deste modo encontrar palavras e frases relevantes para identificar o período não saudável de um equipamento.

A Tabela 3 apresenta um resumo dos trabalhos expostos anteriormente neste item. A partir dos resultados encontrados verifica-se aplicações de uma variedade de técnicas e abordagens para a mineração de dados de falha a partir de relatórios de manutenção. Assim este trabalho apresenta uma proposta inovadora com um abordagem clássica para mineração de textos utilizando aprendizado supervisionado, ainda não estudada na literatura para minerar e catalogar dados qualitativos de confiabilidade em relação aos modos de falha provenientes de relatórios de manutenção escritos na língua portuguesa.

Tabela 3 – Resumo dos trabalhos de mineração de textos de manutenção

Autor	Abordagem de confiabilidade utilizada	Aplicações
DEVANEY et al. (2005)	NLP, Machine Learning e ontologia de domínio para extrair e categorizar componentes e subsistemas de equipamentos e suas falhas associadas	Textos livres de logs de manutenção
CHEN; NAYAK (2007)	Mineração de textos e clustering utilizando os métodos de Ward com SAS Text Miner e de clustering por Histograma de Similaridade	Textos livres de logs de manutenção
MUKHERJEE; CHAKRABORTY (2007)	Análise linguística e conhecimento de domínio para identificar a natureza da falha	Textos livres de logs de manutenção
EDWARDS et al. (2008)	Text Miner do SAS Enterprise Miner para minerar textos, Singular Value Decomposition (SVD) para executar a decomposição em pesos de termo e técnicas de clustering para classificar registros em atributos estruturados	Registros de manutenção de uma estação de bombeamento
MARZEC et al. (2014)	Modelo de classicação Árvore de decisão treinado usando pesos de termo e uma rede neural treinada em componentes SVD (Singular Value Decomposition)	Dados de manutenção corretiva ou preventiva a partir de dados de garantia deum fabricante de ônibus

Autor	Abordagem de confiabilidade utilizada	Aplicações
ARIF-UZ-ZAMAN et al. (2017)	Classificadores Naive Bayes e Support Vector Machine (SVM)	Dois estudos de caso, um com dados de manutenção de uma empresa de energia e outro em uma indústria açucareira
GONÇALVES et al. (2018)	Análise bibliométrica de estudos de mineração de textos na área de manutenção	N/A
GUNAY et al. (2019)	Análise lexical quantitativa e uma abordagem utilização regras de associação associação para identificar as tendências de coexistência entre os termos do cluster de interesse	Dados de manutenção de edifícios
BORTOLINI; FORCADA (2019)	Mineração de textos utilizando o algoritmo MapReduce	Dados de manutenção de edifícios
BLANCO-M et al. (2019)	Classificação de falhas utilizando algoritmo de árvore de decisão e floresta aleatória	Dados de manutenção de turbinas

4 FUNDAMENTAÇÃO TEÓRICA

Este capítulo contém uma sucinta fundamentação teórica sobre Mineração de dados, particularmente, da mineração textos. São apresentadas as principais etapas do processo de mineração para a classificação automática dos textos, fornecendo o embasamento teórico para a metodologia e arquitetura proposta nesta dissertação, dado que os experimentos realizados trabalham com dados provenientes de relatórios de atividades de manutenção de equipamentos da produção de petróleo em ambiente *offshore*. Tais aspectos serão mais detalhados também nos Capítulos 5 e 6.

4.1 Mineração de dados

O processo de Descoberta de Conhecimento em Base de Dados (do inglês, *Knowledge Discovery from Data* – KDD) ou Mineração de Dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) consiste em analisar um grande conjuntos de dados para descobrir correlações, tendências e padrões significativos a partir da utilização de técnicas estatísticas, matemáticas e de reconhecimento de padrões (ZANGL; OBERWINKLER, 2004).

A Mineração de Dados Textuais (FELDMAN; DAGAN, 1995) é uma área específica da Mineração de Dados trabalha com análises de textos. O processo de mineração de textos, ou Descoberta de Conhecimento em Bases Textuais (*Knowledge Discovery from Textual Databases* – KDT) é análogo ao processo de Descoberta de Conhecimento em Base de Dados (do inglês, *Knowledge Discovery from Data* – KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Em ambos, busca-se extrair informações úteis de uma base de dados através da identificação e exploração de padrões. No entanto, no KDT a base de dados é composta por corpus, isto é, coleções de documentos compostas por dados textuais não estruturados (FELDMAN; SANGER, 2007). Assim, a mineração de textos pode ser entendida como uma particularização da aplicação de técnicas de KDD a dados textuais, onde, anteriormente, são extraídas as representações estruturadas destes (FELDMAN; DAGAN, 1995).

Isto é, a Mineração de Textos se baseia no processo de conhecimento obtido da extração de informações de alta qualidade a partir de registros textuais escritos em linguagem natural, armazenados em formatos semiestruturados ou não estruturados. Detalhadamente, consiste em identificar tendências, regularidades ou padrões

significativos, não triviais, imperceptíveis ou ocultos, apresentando o conhecimento obtido de forma coerente e concisa. Todo o processo é multidisciplinar e pode combinar práticas de diversas técnicas interdisciplinares, como a extração e recuperação de informações, estatísticas, dados de mineração, aprendizado de máquina e linguística computacional (NOSHI; SCHUBERT 2019, HOTHO; ANDREAS; PAAß, 2005).

No contexto da indústria O&G, a mineração de dados, numéricos ou textuais, tem papel fundamental para favorecer melhorias nos processos e na otimização da produção de petróleo, ainda mais no cenário no qual grande parte da informação corporativa é registrada em linguagem natural (EBECKEN *et al.*, 2005).

ZANGL; OBERWINKLER (2004) relatam que a utilização de ferramentas de mineração de dados capazes de pré-processar dados brutos, verificar a sua qualidade e de extrair informações de uma grande quantidade de dados, pode, por exemplo, aprimorar e acelerar a produção de reservatórios.

4.2 Processo de Mineração de textos

O processo de mineração para a classificação automática de textos consiste resumidamente nas etapas: coleta dos documentos; definição da abordagem utilizada (análise semântica e/ou estatística); preparação dos dados (pré-processamento e modelo de representação de documentos); e, classificação dos documentos (aplicação de algoritmo de aprendizagem de máquina e análise dos resultados).

4.2.1. Coleta dos documentos

A primeira tarefa do processo de Mineração de Textos (MT) a ser executada é a aquisição dos dados textuais que serão processados. Nesta etapa é realizada a coleta de um conjunto de dados de interesse, realmente relevantes para compor a base de textos a ser trabalhada.

Dependendo do tamanho do corpus e da técnica que se deseja trabalhar na MT, o processo de descoberta de conhecimento pode ser incerto e extremamente difícil. Desta forma, o propósito dessa tarefa é criar uma coleção de documentos (corpus) de qualidade no qual o processo de MT será aplicado. O corpus representa uma conjunto de dados textuais escritos em linguagem natural, de formato não estruturado e *Fuzzy*, que usualmente é derivada de um documento real – relatórios, e-mails, artigos, notícias, entre outros (FELDMAN; SANGER, 2007, TAN, 1999).

Após o fim desta tarefa, é necessário definir a abordagem a ser utilizada no processo de mineração de textos.

4.2.2. Abordagem de análise dos textos

Usualmente, dados textuais podem ser analisados por meio de duas abordagens: semântica e estatística que podem ser adotadas exclusivamente ou de maneira combinada. A análise semântica é vantajosa quando o processamento linguístico é mais complexo e deseja-se melhorar a qualidade da Mineração de Textos, pois emprega fundamentos e técnicas de processamento de linguagem natural (em inglês, *Natural Language Processing* – NLP) de modo a avaliar a sequência dos termos no contexto dos textos, identificando a função de cada termo na frase de origem. Deste modo, para este tipo de análise, é fundamental o conhecimento morfológico, sintático, semântico, pragmático, do discurso e do mundo. Em contraponto, a análise estatística fundamenta-se em cálculos simples de frequência de ocorrência dos termos nos textos, isto é, a importância de um termo é dada pelo número (absoluto, relativo ou inverso) de vezes que este aparece no texto (EBECKEN *et al.*, 2005).

4.2.3. Pré-processamento de textos

Um único documento de texto pode conter uma enorme quantidade de palavras, frases e sentenças. O processo de mineração de textos, apesar de assemelhar-se ao processo de mineração de dados, trabalha com dados semi-estruturados ou não-estruturados.

Assim, uma etapa primordial para a Mineração de Textos (MT) se baseia em identificar um subconjunto de características que possam ser utilizadas para representar um documento, de tal modo seja possível, por exemplo, agrupar ou categorizar documentos conforme suas características.

Ao extrair informações de textos em linguagem natural, este deve passar por uma etapa de pré-processamento para normalizar, padronizar e selecionar as principais características (termos) que realmente agreguem valor ao entendimento do conteúdo de uma base de dados textual.

O pré-processamento consiste em um conjunto de tarefas de Processamento de Linguagem Natural (NLP) que permitem identificar informações relevantes e remover

palavras que são desnecessárias para o entendimento do texto, viabilizando a criação de modelos de representação mais simplificados.

A etapa é aplicada antes da transformação dos dados textuais, e é fundamental para reduzir os efeitos futuros no processamento de textos, principalmente no que se refere a MT de grandes coleções de documentos. Como por exemplo, o elevado custo computacional em razão da alta dimensionalidade dos dados, posto que cada característica representativa do documento pode ser vista como uma dimensão.

Os tópicos a seguir apresentarão as principais tarefas de pré-processamento de texto em linguagem natural utilizadas e encontradas na literatura.

- *Tokenization*

O primeiro passo do pré-processamento se refere, usualmente, à tarefa de quebrar ou dividir dados textuais em unidades menores, independentes e mais significativas (tokens). O processo conhecido como tokenização (em inglês, *tokenization*) é utilizado para separar um documento (ou corpus) de texto em sentenças e cada sentença em palavras. Usualmente, é executado para identificar a ocorrência de pontuações ou um espaço em branco – espaço, tabulação ou o início de uma nova linha, que costumam delimitar onde um token termina e outro começa.

No entanto, a especificação do que deve ser considerado como uma palavra é complexa, e o critério estabelecido para separação dos tokens não é necessariamente confiável. Alguns dos principais problemas encontrados no processo de tokenização são apontados por GASPERIN; LIMA (2001).

A vantagem, é que segmentar dados textuais segundo uma técnica de tokenização permite verificar as relações sintáticas e semânticas dos termos no texto. Os tokens gerados podem ser números, espaços, símbolos, pontuações, palavras ou termos compostos por mais de uma palavra. Nos casos em que o processo de tokenização separa o texto em uma única palavra (ex.: “manutenção”) este é denominado tokenização em unigrama, quando o token é formado por duas palavras (ex.: “manutenção corretiva”), bigrama, e assim por diante.

- *Case folding*

De modo a normalizar os textos e conferir maior agilidade na análise dos dados através do processo de indexação, o *Case Folding* se refere ao procedimento de converter todos os caracteres de um documento no mesmo tipo de letra – ou todas maiúsculas ou minúsculas.

A tarefa permite uniformizar termos de mesma grafia e capitalização distinta, de modo que após o *Case folding* eles possam ser reconhecidos como o mesmo termo. Por exemplo, os termos “Manutenção” e “manutenção”, serão considerados como tokens diferentes se não houver a normalização ou conversão dos caracteres para um mesmo formato.

- Remoção de Pontuação e Caracteres Especiais

A tarefa consiste em remover todas as pontuações e caracteres especiais do corpus (ex.: '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'), a fim de reduzir o número de tokens para representação do documento.

- Remoção de acentos

A tarefa se faz necessária em línguas que utilização acentuação – como o pt-BR, para normalizar os dados e evitar o problema de omissão de acentos em palavras do *corpus*, seja por erro de digitação ou por estarem escritos na forma informal da língua. Além dos acentos, são removidos caracteres com o cedilha, utilizado em palavras contendo c cedilha “ç”, como é o caso da palavra “manutenção”. Ao final do processo a palavra seria representada como “manutencao”.

- Remoção de *Stopwords*

A tarefa “remoção de *stopwords*” é uma das primeiras etapas no processo de preparação dos dados e visa desconsiderar palavras que tem pouco valor na representação de informação dos documentos ou não constitui conhecimento nos textos. Isto reduz significativamente a quantidade de termos e consequentemente, minimiza o custo computacional das etapas seguintes (MANNING; RAGHAVAN; SCHÜTZE, 2008).

As *Stopwords* consistem em um conjunto de palavras que serão descartadas pois não conteúdo semântico significante no contexto em que ela existe e são palavras consideradas não relevantes na análise de textos. Normalmente, essas palavras possuem ocorrência excessiva em uma coleção de documentos (ex.: preposições, pronomes, artigos e outras classes de palavras auxiliares) ou muito baixa, podendo ser removidas, pois não acrescentam à representatividade da coleção ou que sozinhas nada significam.

Uma estratégia para determinar que termos se referem a *stopwords*, é classificar os termos de acordo com a frequência da coleção (número total de vezes que cada termo aparece na coleção). Os componentes da lista de *stopwords* são removidos, usualmente, durante a identificação de tokens nos documentos e geralmente não são incluídas como termos indexados.

- Tratamento de abreviações, acrônimos e siglas

Para o tratamento de abreviações, acrônimos e siglas é necessário utilizar um dicionário de sinônimo/sigla para expandir estas expressões e realizar sua substituição. No entanto, conforme será citado na definição do problema, um dicionário tradicional de sinônimos ou siglas em português não garante a sua correta substituição em relação ao termo identificado.

- *Stemming*

Em morfologia linguística e recuperação de informação utiliza-se o *Stemming* para encontrar a forma primitiva da palavra. O processo é realizado considerando cada palavra isoladamente e tentando reduzi-la a sua provável palavra raiz, simplificando a representação dos termos envolvidos no documento ao transformar as variações de palavras de mesma raiz a um radical único.

Essencialmente, a tarefa de *Stemming* se baseia em reduzir palavras flexionadas ou derivadas a um radical comum, a partir da análise das características gramaticais dos elementos, como grau, número, gênero e desinência, eliminando as variações morfológicas como prefixos, sufixos, vogais temáticas e desinências. Entretanto, algoritmos de *Stemming* empregam linguística e portanto, a técnica é dependente do idioma que se deseja trabalhar.

Exemplificando: as palavras “quebra” (substantivo feminino), “quebrado” (adjetivo), “quebrou” (verbo no pretérito perfeito) e “quebrando” (verbo quebrar no gerúndio) tem significados em comum e podem ser reduzidas ao radical “quebr”.

Assim, ao executar a etapa de *Stemming*, as palavras de mesmo radical passaram a ser representadas por um único termo (o radical da palavra), reduzindo a dimensionalidade do problema e aumentando o desempenho do processo.

No entanto, o processo de *Stemming* não é livre de falhas. Dois erros comuns que podem dificultar ou impedir a não recuperação de documentos que seriam relevantes são: *over stemming* e *under stemming*. O primeiro acontece quando além do sufixo, é removido uma parte do radical. Enquanto, o último dá-se quando um sufixo não é removido, ou é apenas reduzido parcialmente. Desta maneira, é necessário configurar os parâmetros dos algoritmos que executam essa tarefa para evitar distorções. Por exemplo, as palavras “casa” e “casou”, podem ser entendidas como o mesmo radical mesmo quando possuem significados totalmente diferentes.

Dentre os algoritmos propostos para a língua portuguesa, há três que são bastante utilizados e citados na literatura: o Removedor de Sufixo da Língua Portuguesa (RSLP)

proposto por ORENGO; HUYCK (2001), o algoritmo STEMBR proposto por ALVARES; GARCIA; FERRAZ (2005) e a versão para português do algoritmo de PORTER (2006).

Comparações do desempenho desses três algoritmos podem ser encontradas em consulta a literatura, apontando que o algoritmo RSLP é o mais eficiente dos três, por cometer um número menor de erros de over stemming e de under stemming (ORENGO; HUYCK, 2001).

4.2.4. Modelo de representação de documentos

Minerar grandes coleções de documentos requer selecionar os termos mais representativos do *corpus* e realizar a estruturação dos documentos, de maneira a torná-los processáveis para modelagem preditiva e classificação automática dos textos. Deste modo, após o fim da etapa de pré-processamento, a terceira etapa consiste em aplicar um modelo de representação de documentos que será alicerce fundamental para o processamento de documentos de texto e, consequentemente, para a aplicação de algoritmos de aprendizado de máquina.

Esta etapa é uma daquelas que diferenciam o *workflow* de mineração de dados textuais (semi ou não-estruturados) em relação àquele aplicado à dados numéricos (estruturados). Para SALO; MCMILLAN; CONNOR (2019) um exemplo clássico relacionado à dados de equipamentos, é que no caso de modelos aplicados a dados contínuos provenientes de sistemas de monitoramento (ex.: dados de PI - do inglês, *Plant Information Management System*), é possível identificar um comportamento anômalo ou um modo operacional diferente realizando cálculos estatísticos básicos em relação a alguns parâmetros operacionais e desta maneira, definir um comportamento esperado do ativo em determinadas condições. Enquanto, a classificação de textos de manutenção é algo mais complexo, porque não é possível identificar diretamente qual é a "sentença média" ou do quão "distante" uma descrição de manutenção é de outras, sem que seja utilizado um modelo de representação de documentos para a obtenção de tais métricas.

Dentre os diversos modelos clássicos disponíveis na literatura (Booleano, Vetorial, Probabilístico, e outros) que são utilizados para a representação de documentos textuais na mineração de coleções de documentos, o modelo vetorial (SALTON; WONG; YANG, 1975) e suas variantes, são uma escolha bastante comum para realizar a tarefa.

A preferência em utilizar este modelo em relação aos demais deve-se ao fato que o modelo vetorial se destaca por ser um modelo algébrico simples, em que operações de vetores podem ser executadas muito rapidamente de modo a representar conteúdos textuais e importantes aspectos dos textos. Ademais, a existência de algoritmos eficientes capazes de realizar a seleção do modelo, a redução da dimensão e visualização de espaços de vetores, é um outro fator que pesa a favor desta escolha.

Este modelo permite o uso de algoritmos de aprendizado de máquina tradicionais, e pode ser utilizado em situações em que é necessário classificar documentos segundo um determinado critério. A utilização de um espaço vetorial para representação de documentos fornece ferramentas úteis para quantificar métricas que facilitem o processo de classificação automática dos textos.

O processo de vetorização, também conhecido como *Feature Extraction* ou *Feature Encoding*, realiza a codificação das palavras de um documento (ou *corpus*) como números inteiros ou valores de ponto flutuante (do inglês, *floating-point values*), e permite sua utilização como *inputs* nos algoritmos de aprendizado.

No modelo originalmente proposto por SALTON; WONG; YANG (1975), os documentos são representados como pontos (vetores de palavras) em um espaço Euclidiano de dimensão *i*, onde $D = \{d_1, d_2, ..., d_N\}$ corresponde ao conjunto de *N* documentos e $T = \{t_1, t_2, ..., t_i\}$ corresponde ao conjunto de *i* termos que compõem uma coleção de textos. Portanto, cada um dos *N* vetores dos documentos de uma coleção é composto por *i* dimensões. Ou seja, cada documento é um vetor em um espaço multidimensional e cada dimensão é um termo da coleção (FELDMAN; SANGER, 2007).

Este tipo de representação é também conhecido como BOW (do inglês, *Bag of Words*) e apesar de existirem vários métodos que tentam explorar também a estrutura sintática e a semântica do texto, a maioria das abordagens de mineração de texto se baseia na ideia de que um documento de texto pode ser representado por um conjunto de termos contidos nele, onde cada termo tem um peso associado para descrever sua significância. Desta forma, a BOW, isto é, união dos vetores das representações do documento em uma coleção pode ser representada por uma matriz documento-termo (TAN *et al.*, 2005). Uma representação desta matriz documento-termo (A) de uma coleção com N documentos e i termos é apresentada na Tabela 4. Cujo o termo da matriz $a_{dN,ti}$ representa um peso de um termo t_i em um documento d_N e c_{dN} representa a classe do documento *N*, isto é, os rótulos nominais de cada documento.

Tabela 4 – Representação de uma matriz documento-termo com *N* documentos e *i* termos

	$\mathbf{t_1}$	$\mathbf{t_2}$	**...**	$\mathbf{t_i}$	**Classe**
$\mathbf{d_1}$	$a_{d1, t1}$	$a_{d1, t2}$	...	$a_{d1, ti}$	c_{d1}
$\mathbf{d_2}$	$a_{d2, t1}$	$a_{d2, t2}$	...	$a_{d2, ti}$	c_{d2}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$\mathbf{d_N}$	$a_{dN, t1}$	$a_{dN, t2}$	...	$a_{dN, ti}$	c_{dN}

Assim, alguns aspectos negativos da representação vetorial, é que a matriz documento-termo é, geralmente, uma matriz esparsa de alta dimensionalidade. O problema decorre em razão do grande número de palavras diferentes contidas em uma coleção de documentos, em que a maioria das palavras ocorre apenas em uma pequena parte dos documentos. Outra complicação conhecida é a atribuição que quaisquer duas palavras são consideradas por definição não-relacionadas.

Uma das maneiras de lidar com os problemas de esparsidade e dimensionalidade é atribuir pesos para quantificar a relevância de cada termo em um documento e sua respectiva coleção. Este processo consiste em dar ênfase aos termos mais importantes para a representação, associando os pesos a coordenadas de vetor por meio de uma abordagem estatística, em que o peso representa a sua frequência ou uma função dela. Deste modo, as coordenadas dos vetores assumem valores numéricos conforme a relevância de um termo para o documento, de tal forma que valores maiores implicam em relevâncias maiores.

Dentre as medidas de atribuição de pesos para seleção de atributos disponíveis na literatura estão o sistema Binário, a frequência do termo (em inglês, *Term Frequency* – TF) e a frequência de termo - frequência inversa de documento (em inglês, *Term Frequency - Inverse Document Frequency* – TF-IDF).

O sistema binário consiste em atribuir valores 1 e 0 para apresentar se um termo existe em um documento ou não, respectivamente. A frequência do termo (TF) considera as ocorrências de um termo em um documento e utiliza esta métrica para atribuir os pesos, que são usualmente normalizados para valores no intervalo [0,1]. Enquanto, o TF-IDF consiste em avaliar a importância de uma palavra com base na frequência de sua ocorrência no documento e o inverso de sua ocorrência em todo o corpus, definindo pesos dos termos e suas características (POLETTINI, 2004).

4.2.5. Classificação automática de textos

Coleções de documentos textuais podem ser classificadas automaticamente em categorias pré-definidas utilizando técnicas de inteligência artificial via algoritmos de *Machine Learning* (ML) ou *Artificial Neural Network* (ANN).

Esta dissertação propõe empregar algoritmos de aprendizado supervisionada para automaticamente aprender a reconhecer padrões textuais e tomar decisões inteligentes, baseando se no conjunto de dados analisado. No entanto, o aprendizado de máquinas pode ser realizado de quatro formas: não-supervisionada, semi supervisionada, supervisionada e ativa, cuja escolha de abordagem adotada é realizada conforme o problema a ser estudado (HAN *et al.*, 2012).

O aprendizado não-supervisionado é empregado para resolver problemas cuja coleção textual não foi previamente rotulada. Assim, neste processo de aprendizado, os modelos são utilizados para agrupar os dados conforme as classes descobertas.

No aprendizado semi supervisionado, o conjunto de dados possui uma parte rotulada e outra não rotulada, e ambas as partes são utilizadas no aprendizado. Este tipo de aprendizado é comumente utilizado quando deseja-se utilizar os dados rotulados para separar as classes, e usar os exemplos não rotulados para refinar os limites entre as classes.

Enquanto no aprendizado supervisionado o classificador é construído baseando-se em um conjunto de documentos pré-classificados. Nesta tarefa, o conjunto de dados é dividido em dois: um conjunto de treinamento e outro de teste, em partições que não precisam, necessariamente, serem iguais. Os dados de treinamento rotulados são utilizados para construir ou treinar o classificador, encontrando os parâmetros ótimos que ajustam um determinado modelo. Então, emprega-se o conjunto de teste para testar a capacidade do classificador em prever os rótulos desconhecidos, até que seja encontrado um modelo aceitável.

Já no aprendizado ativo, o usuário desempenha um papel ativo no processo. O número de variáveis de entrada disponíveis no modelo é muito maior que o número de respostas disponíveis. Desta forma, o programa seleciona determinados registros para o usuário rotular capazes de otimizar a qualidade do modelo, permitindo que o usuário não tenha que rotular todos eles, mas exemplos específicos que tem uma grande importância para o aprendizado e elaboração do modelo.

Nos últimos anos, vários algoritmos de classificação foram desenvolvidos, testados e comparados para resolver problemas de categorização de textos. Dentre eles,

os classificadores de Árvore de Decisão, Regressão, classificadores Rocchio, K-Vizinhos mais próximos (do inglês, *K-Nearest Neighbor*), Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*), classificadores Naive Bayes e outros (KORDE; MAHENDER, 2012).

Dentre os algoritmos utilizados para classificação automática de textos, esta dissertação optou por desenvolver a metodologia proposta comparando o desempenho dos classificadores Naive Bayes. Esta escolha se baseou em razão da simplicidade do modelo e de resultados de desempenho frequentemente comparáveis com alguns métodos mais sofisticados, como árvore de decisão e classificadores de rede neural (AGGARWAL, 2014). Desta maneira, os classificadores Naive Bayes serão apresentados em detalhe a seguir.

4.2.5.1. Classificadores Naive Bayes

Classificadores Naive Bayes são amplamente empregados para resolver problemas de classificação de documentos, sendo baseado em um importante modelo probabilístico notavelmente bem-sucedido nesta tarefa, mesmo com a suposição imprecisa de independência (AGGARWAL, 2014). O modelo é baseado no aprendizado indutivo supervisionado com abordagem probabilística simples, que pode ser implementado muito eficientemente com uma complexidade linear (MCCALLUM; NIGAM, 1998). O classificador considera a probabilidade a priori de um documento pertencer à uma determinada categoria, sendo particularmente adequado em problemas cuja dimensionalidade das entradas é alta ou quando aplicado à um grande conjunto de dados (AGGARWAL, 2014).

O algoritmo é fundamentado no Teorema de Bayes – no qual considera a probabilidade condicional de determinadas palavras aparecerem em um documento de uma determinada categoria, combinado com a suposição ingênua (naive) de que os atributos dos documentos são condicionalmente independentes, ou seja, o classificador assume que existe independência entre as palavras de um texto (LEWIS; RINGUETTE, 1994, ZHANG, 2004).

Utilizando o teorema de Bayes é possível calcular a probabilidade a posteriori $P(c_i|t_k)$ conforme a Equação 1.

$$P(c_i|t_k) = \frac{P(c_i)\,P(t_k|c_i)}{P(t_k)} \qquad \text{Equação 1}$$

Onde $P(t_k|c_i)$ é a probabilidade condicional do atributo t_k ocorrer em um registro da classe c_i, $P(c_i)$ é a probabilidade a priori de um registro pertencer a classe c_i e $P(t_k)$ é a probabilidade de ocorrência de t_k em qualquer uma das m classes conforme apresenta a Equação 2.

$$P(t_k) = \sum_{i=1}^{m} P(t_k|c_i)\ .\ P(c_i) \qquad \text{Equação 2}$$

O classificador Naive Bayes estima a probabilidade a posteriori para cada uma das classes, escolhendo a classe com maior probabilidade segundo a Equação 3.

$$\hat{y} = arg\max_{i}\{P(c_i|t_k)\} \qquad \text{Equação 3}$$

Assim, quando o classificador assume que todos os atributos são independentes, o problema é simplificado, tornando possível que a probabilidade condicional $P(t_k|c_i)$ possa ser decomposta em um produto de probabilidades. Assim, a probabilidade de um registro d pertencer a classe c_i, $P(d|c_i)$, é calculada segundo Equação 4.

$$P(d|c_i) = P(t_1, t_2, \dots, t_k|c_i) = \prod_{j=1}^{k} P(t_j|c_i) \qquad \text{Equação 4}$$

A suposição de independência desconsidera as correlações que possam existir entre as palavras de um documento, simplificando bastante o aprendizado e permitindo que os parâmetros para cada atributo possa ser aprendido separadamente (ZHANG, 2004).

Essa hipótese é especialmente vantajosa quando o número de atributos é alto, mesmo em aplicações cotidianas em que a suposição de independência é inverídica (especialmente para problemas de domínio de texto). Em razão desta e outras propriedades, o algoritmo é surpreendentemente útil e preciso.

Talvez por isso, em comparação com métodos mais sofisticados, classificadores Naive Bayes podem ser extremamente rápidos e ter desempenhos impressionantes, mesmo para um pequeno conjunto de dados. No entanto, embora seja considerado um bom classificador, as estimativas de probabilidade não devem ser levadas em consideração (PEDREGOSA *et al.*, 2011).

A implementação de algoritmos classificadores Naive Bayes existentes variam de acordo com a função de densidade de probabilidade adotada para a estimativa da probabilidade condicional $P(t_j|c_i)$. Dentre eles, os principais: *Gaussian Naive Bayes* (GNB), *Multinomial Naive Bayes* (MNB), *Bernoulli Naive Bayes* (BNB).

Multinomial e *Bernoulli* são distribuições de probabilidade amplamente empregadas para resolver problemas de classificação de documentos, sendo o *Multinomial* um algoritmo muito adequado para representar a ocorrência de uma palavra em um único documento. Deste modo, este trabalho aplica o classificador Naive Bayes considerando uma distribuição multinomial.

A distribuição multinomial é uma generalização da distribuição binomial, sendo parametrizada por vetores $\theta_{y_i} = (\theta_{y_1}, \dots, \theta_{y_n})$, tal que θ_{y_i} é a probabilidade do evento i ocorrer, dado que a classe é c_i e n representa o número de atributos. Assim, $\theta_{c_i} = P(t_j|c_i)$ é estimado pela Equação 5.

$$\widehat{\theta_{c_i}} = \frac{N_{c_i} + \alpha}{N_c + \alpha n} \qquad \text{Equação 5}$$

Onde N_{c_i} representa o número de vezes que o atributo t_j aparece no conjunto de treinamento, N_c é o número de observações com classe c, n representa o número de atributos e α é uma constante que contabiliza os atributos que não estão presentes nas amostras de aprendizado e impede que haja probabilidade igual a 0. Assim alfa pode assumir valores contidos no intervalo [0,1]. Nos casos em que α é igual 1, ele é chamado de *Laplace smoothing* e, se α > 1, é chamado *Lidstone smoothing*. Se α = 0, não há correção.

4.2.5.2. Métricas de avaliação de modelos de classificação

O último estágio da classificação consiste em avaliar os classificadores experimentalmente. Para isso, é necessário que o conjunto de dados seja dividido em duas séries. A primeira para treinamento do modelo (série de treino), e a segunda para avaliar sua performance (série de teste). Após a etapa de ajuste do modelo com a série de treino, o modelo obtido é utilizado para prever classificações de uma série de teste. Assim, a avaliação dos resultados do modelo é realizada com a comparação da classificação prevista pelo teste e a real.

As métricas de avaliação são utilizadas para verificar o desempenho de um determinado modelo e avaliar a necessidade ou oportunidade de melhorar sua performance, isto é, sua capacidade de categorizar documentos corretamente. Outra finalidade da avaliação, é definir, dentre os diferentes algoritmos de aprendizado testados, qual é melhor para o caso estudado.

Diversas métricas podem ser empregadas para avaliar e comparar o desempenho dos classificadores, as mais usuais em problemas de classificação são: acurácia, erro, precisão, sensibilidade (*recall*), especificidade e medida F. A Tabela 5 apresenta as principais métricas utilizadas para avaliar os modelos de classificação binária (HAN *et al.*, 2012).

Tabela 5 – Medidas de Avaliação para Modelos de Classificação Binária

Medidas de Avaliação	**Formulação**
Acurácia	$\frac{VP + VN}{P + N}$
Erro	$\frac{FP + FN}{P + N}$
Sensibilidade ou *Recall*	$\frac{VP}{P} = \frac{VP}{VP + FN}$
Especificidade	$\frac{VN}{N}$
Precisão	$\frac{VP}{VP + FP}$
Medida F	$\frac{2 * Precisão * Recuperação}{Precisão + Recuperação}$

Fonte: HAN *et al.* (2012)

Tal que:

- VP (número de verdadeiros positivos) e VN (número de verdadeiros negativos) são a quantidade de valores que o modelo categorizou corretamente.
- FP (número de falsos positivos) e FN (número de falsos negativos) são a quantidade de valores categorizados erroneamente pelo modelo.
- P (total de positivos reais)
- N (total de negativos reais)

Acurácia e erro são métricas que indicam o desempenho geral do modelo sendo, respectivamente, os acertos e erros globais. Enquanto as medidas *recall* e especificidade representam o desempenho local do modelo em relação a cada classe.

O *recall* avalia, dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas. A precisão indica quanto, dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas. Já a medida F consiste em uma medida que considera a média harmônica das métricas precisão e *recall*.

Nos casos de classificações multiclasse as métricas utilizadas são as mesmas. No entanto, as métricas são calculadas para cada classe, como um problema de classificação binária após agrupar todas as outras classes (exceto uma) como uma classe que representa a "negação" da classe excluída no agrupamento. Por exemplo, sejam as classes: "vazamento", "leitura anormal do instrumento", "pequenos problemas em serviço". Para verificar o desempenho do modelo para a classe "vazamento", as outras classes restantes serão agrupadas na classe "negação" da classe "vazamento", e assim sucessivamente. Assim, para cada combinação haverá um valor da métrica. No final, o resultado das métricas para o problema multiclasse pode ser obtido por uma média (micro, macro ou ponderada) dessas combinações. As médias micro e macro tratam todas as classes igualmente, enquanto a ponderada considera um problema desbalanceado e assim a média é ponderada pela frequência da classe.

5 DEFINIÇÃO DO PROBLEMA

5.1 O problema de catalogação de dados de confiabilidade

A gestão de campos petrolíferos inteligentes/digitais (do inglês, *Smart/Digital Oilfields Management*) têm sido o principal foco da indústria de óleo e gás na última década (ARUMUGAM *et al.*, 2016). O processo de informatização e digitalização derivado deste movimento e daqueles provenientes últimas revoluções industriais, tornou possível armazenar uma imensa quantidade de dados de Exploração e Produção (E&P) em formato estruturado e, sobretudo, não estruturados (WU *et al.*, 2014; ARUMUGAM *et al.*, 2016; GONÇALVES *et al.*, 2018).

Apesar do movimento de digitalização ter proporcionado a aquisição de novos ativos com coleta automática de dados incorporada em seu design e operação, ainda existem instalações com ativos em que o nível de processamento de dados são anteriores ao processo de digitalização. Nesses casos, geralmente grande parte dos dados são coletados meramente para seguir o procedimento e seu valor é raramente explorado (SALO; MCMILLAN; CONNOR, 2019).

No entanto, informações extremamente valiosas para a tomada de decisão e melhoria dos processos podem ser obtidas a partir da análise destes dados e da informação obtida deles (DHAMODHARAVADHANI *et al.*, 2018; AL-ALWANI *et al.*, 2019).

Contudo, extrair ou recuperar informações úteis e padrões interessantes dos bancos de dados disponíveis requerem o manuseio e processamento de uma enorme massa de dados em um tempo de processamento razoável (ZIO, 2009).

Quando o conteúdo de cada texto não é conhecido e há enorme quantidade de registros, a ação de extrair informações úteis se torna quase impossível, especialmente quando não possuem um formato de dados estruturado, como na maioria dos casos (GONÇALVES *et al.*, 2018).

Estima-se que nos próximos anos o volume de dados de E&P provavelmente dobrará, aumentando o esforço necessário para extrair informações relevantes ou isolar informações específicas dos dados (ARUMUGAM *et al.*, 2016). Quando os dados crescem exponencialmente e estão disponíveis em muitas plataformas, a tarefa, tradicionalmente realizada de modo manual, é insuficiente e bastante restrita (SANDTORV *et al.*, 1996; GONÇALVES *et al.*, 2018).

Tais fatores têm conduzido a indústria a se afastar dos bancos de dados tradicionais e a adaptar o *Big Data*, recorrendo a métodos computacionais de tratamento e recuperação de informação semi ou automatizadas, cujo objetivo é analisar os registros de maneira consistente, ágil e eficiente (ZIO, 2009; ARUMUGAM *et al.*, 2016; GONÇALVES *et al.*, 2018).

Essas mudanças tornaram-se extremamente necessárias para processar os diferentes formatos de dados disponíveis e trouxeram gigantescos ganhos nas atividades de E&P de óleo e gás, permitindo análises melhores e mais rápidas, com recursos de tomada de decisão aprimorados (ZIO, 2009; ARUMUGAM *et al.*, 2016).

Dados de E&P são usualmente armazenados em bancos de dados onde um único documento ou registro pode incluir uma mistura de campos estruturados e outros componentes textuais semiestruturados ou não estruturados (NOSHI; SCHUBERT, 2019).

No entanto, a falta de padronização nas ferramentas e métodos usados para reportar dados e informações na indústria deste setor, tanto no aspecto documental (sistemas e políticas) quanto referente à sua utilização (conformidade com as normas), representa um desafio à validação e obtenção de informações que agreguem valor às análises e algoritmos.

No que se refere a dados de falha, a coleta e análise de dados de confiabilidade possui um grande potencial de ganho para a gestão da operação e manutenção de ativos, desde que sejam captadas/armazenadas informações de boa qualidade (SANDTORV *et al.*, 1996; CORVARO *et al.*, 2017). Para WUTTKE; SELLITTO (2008) independentemente da ferramenta ou método de gestão de manutenção escolhida, a melhoria na eficiência das operações de produção de petróleo requerem que os dados trabalhados sejam confiáveis.

5.1.1. Percepção, registros e análises de falhas

Ao longo do ciclo de operação de um equipamento é possível que haja a ocorrência de ao menos uma falha em seus componentes. Assim, uma vez detectada a falha, estes são passíveis de manutenção e os procedimentos requeridos são registrados. As intervenções são realizadas para que a função requerida possa ser restaurada e o equipamento tenha o desempenho adequado e esperado (SALGADO, 2008).

Muitos fatores podem influenciar a ocorrência de falha em um ativo, desde as condições a que ele está exposto até suas características e singularidades. Por exemplo, condições ambientais e de operação, indústria de aplicação, características do sistema, materiais de fabricação, componentes, fabricante, mecanismo de operação, dentre outros. Associado a isso, a complexidade dos sistemas e o comportamento não linear do equipamento ao longo do tempo possibilita a ocorrência de um número indeterminado de condições e diferentes eventos de falha em relação a cada um dos aspectos relatados. De modo geral, esses fatores dificultam que operadores prevejam e identifiquem a fonte de possíveis falhas no sistema de produção antes que esses tenham seu funcionamento interrompido ou já estejam muito danificados.

A detecção, diagnóstico e previsão de eventos são usualmente realizadas por uma equipe de especialistas que analisam conjuntamente as variações e combinações de sinais físicos pertencentes ao sistema de produção (percepção das falhas) associadas ao histórico das respectivas ações tomadas em virtude delas (PENNEL *et al.*, 2018).

Tipicamente, as análises de falhas se baseiam em três classes de problemas. O primeiro consiste em identificar que uma falha ocorreu, mesmo sem saber qual foi sua causa exata. O segundo, identificar a natureza da falha, isto é, as causas-raiz que resultaram na falha e seu modo, mecanismo e criticidade. O terceiro, baseia-se em realizar a previsão do momento em que o componente ou o sistema deixará de funcionar.

Entretanto, no contexto das atividades de produção de petróleo, o monitoramento de falhas é dinâmico e complexo. Isto porque o sistema é formado por um grande conjunto condições, parâmetros e variáveis de operação.

SINGH; KAZZAZ (2003), GUO *et al.* (2009), AHMAD; KAMARUDDIN (2012), AN *et al.* (2015) são alguns autores que trabalharam com dados de monitoramento em tempo real, manutenção baseada na condição de equipamentos e outros estudos desta natureza. Enquanto LIU *et al.* (2018) apresentaram uma revisão das técnicas de inteligência artificial para diagnóstico de falha. No entanto, apesar do empenho empregado nestes e outros estudos, observa-se que o desenvolvimento de modelos de monitoramento, detecção e diagnóstico de falhas em tempo real é uma tarefa complexa e que por vezes requer o conhecimento de falhas que aconteceram a priori.

Desta forma, esta pesquisa concentra seus esforços no processo de Recuperação da Informação (RI) histórica ou atuarial em dados textuais que descrevem os prováveis eventos de falhas e as ações de manutenção executadas. Em particular, na catalogação de

dados de confiabilidade em relação a natureza da falha, isto é, modos e mecanismos de falha.

5.1.2. Mineração de registros com conteúdo de formato textual livre

De maneira geral, a tarefa de catalogação modos e mecanismos de falha dispõe de ampla complexidade de solução, tanto em relação aos desafios tradicionais à mineração de textos e recuperação da informação, quanto àqueles específicos da tarefa de catalogação de dados de confiabilidade na indústria de óleo e gás, e da classificação de modos e mecanismos de falha.

Dados de confiabilidade são usualmente obtidos a partir de relatórios das atividades de manutenção de ativos, isto é, notas de falha, ordens de trabalho e descrições das operações realizadas. Os registros são tradicionalmente realizados em sistemas de gerenciamento de manutenção, que permitem o gerenciamento do fluxo de materiais e a coordenação das atividades internas para sincronizar as atividades com as necessidades de disponibilidade de equipamentos para produção (ZIO, 2009). Estes registros são tradicionalmente compostos por campos longos com as descrições dos eventos de falha e das atividades de manutenção, diversos campos textuais com descrições breves e outros campos estruturados.

Há dois métodos para a catalogação de falhas, o manual e o automático. Geralmente, a coleta e validação dessas informações do conjunto de dados é realizada em lotes e manualmente por especialistas, despendendo de uma grande quantidade de tempo e recursos humanos para avaliar cada registro (SANDTORV *et al.*, 1996). No entanto, é difícil encontrar mão de obra com as competências necessárias para realizar a coleta desse tipo de dados (SANDTORV *et al.*, 1996). Além disso, a exploração manual de dados históricos para o reconhecimento de falhas é uma tarefa repetitiva, exaustiva e inviável para um grande volume de dados, limitando a catalogação de apenas uma pequena parte dos dados disponíveis (CHEN; NAYAK, 2007, BLANCO-M *et al.*, 2019).

ZHANG *et al.* (2020) *et al.* (indicam que atribuir etiquetas categóricas a documentos de ordens de serviço de maneira manual é um processo propenso a erros humanos, o que exige o desenvolvimento de técnicas de mineração de texto e aprendizado de máquina capazes de classificar corretamente os relatórios humanos, levando a análises automáticas e precisas dos relatórios técnicos.

A abordagem automática baseia-se em utilizar as informações contidas nos registros para escolher sua categoria de falha a partir do aprendizado de máquina ou de outros recursos de inteligência artificial. Entretanto, apesar de automática a tarefa está longe de ser trivial.

O problema de classificação de registros de falha geralmente é desbalanceado, uma vez que cada modo de falha possui uma probabilidade de falha (isto é, há mais registros de uma classe que de outra), múltiplas classes (cada equipamento possui modos de falha específicos para cada componente). O número de classes pode variar bastante, no caso de turbinas a gás, por exemplo, a norma prevê dezenove modos de classe. Devido ao fato de grande parte dos eventos serem raros, o conjunto de dados disponível geralmente é pequeno mesmo avaliando um longo horizonte de tempo.

A catalogação também particularmente dificultada porque os registros são comumente armazenados em sistemas de informação que não são, necessariamente, desenvolvidos para a modelagem de confiabilidade de equipamentos, mas no qual é possível obter informações implícitas a respeito de um evento de falha (ARIF-UZ-ZAMAN *et al.*, 2017). Outro fator complicador é o fato de apenas uma fração dos dados contidos nos registros ser efetivamente relevante para a finalidade (GONÇALVES *et al.*, 2018).

A coleta e catalogação a partir das fontes de dados disponíveis também é especialmente complexa devido à qualidade e disponibilidade dos registros de falha e das atividades de manutenção, que podem variar significativamente entre as empresas e podem ser preenchidas de maneira incompleta, imprecisa ou fornecerem informações questionáveis (SANDTORV *et al.*, 1996).

Os registros na maioria dos casos são realizados em desacordo com os padrões para coleta de dados de confiabilidade tanto a nível nacional no que se refere à Norma Técnica Brasileira – NBR ISO 14224 (ABNT, 2011) ou internacional com a ISO 14224 (ISO, 2016) quanto em relação aos bancos de dados de confiabilidade usuais, prejudicando a catalogação dos dados (SANDTORV *et al.*, 1996).

Especificamente em relação os obstáculos da automatização da catalogação de dados de confiabilidade segundo os padrões brasileiros e em textos da língua portuguesa, é difícil encontrar técnicas de mineração de textos e modelos de processamento de linguagem natural que trabalhem tão bem com vocabulários em português quanto o inglês ou outros idiomas mais comuns aos estudos de mineração e processamento de textos.

A tarefa também é trabalhosa, pois a maioria das informações contidas nos registros é interpretada. Também, a pessoalidade empregada por cada operador no relato de eventos atrapalha encontrar padrões textuais que facilitem o processo de catalogação, reduzindo a acurácia dos dados disponíveis.

Cada operador pode usar palavras diferentes para descrever o mesmo evento. Apesar de existirem alguns bons dicionários de sinônimos (em inglês, thesaurus) disponíveis em português e ontologias lexicais em Português de fácil importação nos algoritmos com o PULO WordNET PT (WORDNET, 2020), nenhum ainda é específico aos sinônimos técnicos utilizados nas atividades de manutenção. Por exemplo, quando buscamos o sinônimo para o vocábulo "fluído" em um dicionário de sinônimos tradicional encontramos: "*emanado, descendente, filho, nascido, originário, oriundo, procedente, proveniente, derivado, originado, provindo, vindo, advindo, manado, exido*" (7GRAUS © 2011 - 2020, 2020). Enquanto para a manutenção o mesmo vocábulo pode ser sinônimo de "líquido", "gás", "lubrificante", "água", "óleo", "condensado" e outros. Desta maneira, a utilização de dicionários não técnicos pode custar tempo de processamento e o uso de recursos desnecessários para classificação dos registros.

Mais um problema relacionado à linguística, é a presença de sentenças desconexas ou sintaticamente mal formadas, que complica a busca e classificação dos modos e mecanismos de falha. Também é comum, que ao descrever um evento de falha, os operadores usem abreviações, acrônimos, siglas ou termos informais o que dificulta o processamento dos textos e a classificação da natureza da falha. O uso de pronomes para referenciar termos pode prejudicar similarmente a identificação dos termos referentes a cada modo e mecanismo de falha, pois são utilizados em sentenças que mencionam os termos de interesse de maneira implícita.

Outros obstáculos presentes no processo de catalogação destes tipos de dados são referentes aos erros ortográficos cometidos, seja por erro de digitação ou por estarem escritos na forma informal da língua.

Ademais, identificar corretamente a que modo e mecanismo de falha esse registro se refere, e principalmente, identificar informações implícitas contidas em uma descrição, não é uma tarefa simples. Uma especificidade, é que um texto pode referenciar um ou vários termos de interesse e estes podem se referir a mais de um modo ou mecanismo de falha (múltiplos ou concorrentes) confundindo a classificação.

Por fim, a catalogação de dados de confiabilidade também possui algumas limitações. Em alguns casos, uma única fonte de dados pode ser incapaz de fornecer

informações confiáveis aos estudos de confiabilidade de ativos e o processamento dos dados. Além disso, o grande número de categorias possíveis em relação ao modo e mecanismo de falha e o pequeno número de registros disponíveis previamente catalogados usados no aprendizado de máquinas faz com que o problema tenha mais um grau de complexidade. Assim, a tarefa torna-se mais trabalhosa e as falhas, bem como seus respectivos modos e mecanismos, dificilmente identificáveis.

Portanto, a partir destas considerações, visa-se responder a seguinte pergunta: Como extrair dados de confiabilidade quanto ao modo de falha a partir de dados textuais não estruturados, e melhorar a qualidade dos bancos de dados de confiabilidade? E mais especificamente, como desenvolver uma metodologia eficiente que considere os fatores complicadores apresentados anteriormente, que seja capaz de catalogar esse tipo de dados de confiabilidade aplicado a equipamentos da produção de petróleo em ambiente *offshore*.

6 METODOLOGIA

Este capítulo apresenta as especificações técnicas e os procedimentos metodológicos para o desenvolvimento da arquitetura proposta.

6.1 Especificações Técnicas

A metodologia proposta nesta pesquisa foi implementada na linguagem Python (ROSSUM, 1995), em sua versão 3.7.3 e no ambiente de desenvolvimento Spyder, em sua versão 3.3.6.

Dentre as diversas linguagens de programação disponíveis atualmente, alguns dos fatores que justificam a escolha da programação em Python frente as demais, é que esta é uma linguagem orientada ao objeto e de bastante simplicidade de implementação, sendo bastante amigável e fácil de aprender. Ademais, Python é uma linguagem *open source* que conta com uma comunidade de desenvolvimento bastante ativa, o que permitem frequentes atualizações de seus recursos (PYTHON SOFTWARE FOUNDATION, 2020).

Outra considerável vantagem é o vasto número de módulos e bibliotecas bem documentadas, que facilitam todo o processo de pré-processamento, mineração dos dados e aprendizado de máquinas, sendo frequentemente utilizado pelos cientistas de dados. Na arquitetura proposta nesta pesquisa, foram utilizadas as bibliotecas Python disponíveis para processamento de linguagem natural - NLTK de BIRD; LOPER; KLEIN (2009), e aprendizado de máquina - *Scikit Learning* de PEDREGOSA *et al.* (2011).

Na etapa de pré-processamento do texto, a lematização das palavras foi realizada aplicando-se o algoritmo RSLP disponível na biblioteca NLTK (BIRD; LOPER; KLEIN, 2009). A correção ortográfica foi realizada utilizando a biblioteca Python *SpellChecker*, que foi desenvolvida segundo o trabalho de NORVIG (2016). Além disso, os algoritmos *CountVectorizer* e *TfidfTransforme*r da biblioteca *Scikit Learning* foram utilizados na etapa de vetorização das palavras.

Os experimentos realizados neste trabalho foram realizados em um computador de sistema operacional Microsoft Windows (versão Windows 10 e 64 bits) com processador Intel® Core™ i7-6700HQ 2.60 GHz e 8.00 GB de memória RAM.

6.2 Procedimento Metodológico

Conforme apontado por SALO; MCMILLAN; CONNOR (2019), apesar de documentos de texto livre não serem mais tão utilizados quanto agora com o avanço da digitalização, muitas horas de trabalho ainda são perdidas em todo o setor para a criação e análise manual desses tipos de dados.

Assim, esta pesquisa desafia a prática atual da análise manual de dados históricos de manutenção (Ordens de Serviço – OS), apresentando uma metodologia adequada ao contexto atual de eficiência e digitalização.

A metodologia proposta oferece facilitar e agilizar o processo de catalogação de dados de confiabilidade, empregando uma abordagem mais econômica e automatizada, evitando um esforço tedioso e demorado por parte dos especialistas. Para isso, o método proposto utiliza técnicas de mineração de textos nos registros históricos de manutenção de equipamentos para posteriormente realizar sua catalogação automática quanto ao modo de falha.

A metodologia consiste em quatro etapas principais esquematizadas conforme a Figura 3. Assim, a primeira etapa da metodologia consiste em coletar e verificar os registros contendo as informações de falha e manutenção. Enquanto na segunda etapa é realizado todo o pré-processamento dos textos, a terceira etapa realiza vetorização do conjunto de palavras que representam os documentos textuais no espaço para então realizar a sua utilização na quarta e etapa final – o aprendizado de máquina, para a classificação automática dos registros. Cada uma das etapas da metodologia será descrita com mais detalhes nos subtópicos a seguir.

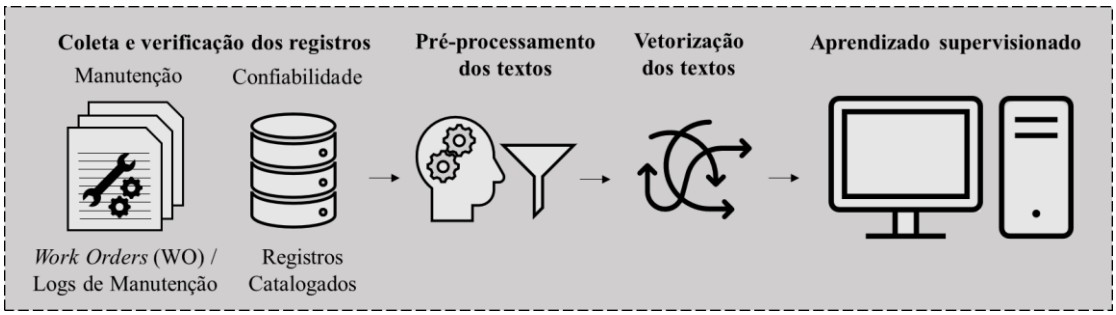


Figura 3 – Esquematização da metodologia de classificação de registros de manutenção

6.1.1. Coleta, verificação e caracterização dos registros

Por se tratar de uma proposta de metodologia automática de textos utilizando o aprendizado supervisionado, a primeira etapa consiste em coletar os registros de manutenção que foram previamente catalogados em relação aos modos de falha do equipamento, contendo as informações de falha e de manutenção para cada tipo de equipamento.

Deste modo, são utilizados dois bancos de dados. O primeiro para obtenção dos registros históricos de manutenção de ativos da E&P de petróleo (documentos textuais). Já o segundo, para obter o par "ID do documento - rótulo de classificação", que foram previamente e manualmente catalogados por especialistas da área e armazenados em um banco de dados de confiabilidade.

Nesta pesquisa, registros de manutenção são considerados quaisquer documentos criados de acordo com os procedimentos e gestão da manutenção de ativos com o intuito de notificar a percepção de uma falha (nota), a solicitação de sua correção (ordem), assim como a própria descrição da operação de manutenção realizada (operações). Na literatura, estes documentos são usualmente denominados logs de manutenção ou ordens de serviço (em inglês, *Work Orders* -WO).

Segundo apontado por SALO; MCMILLAN; CONNOR (2019), esses documentos servem, por exemplo, para dar início a uma tarefa de manutenção a ser executada em um ativo, descrever a tarefa a ser realizada ou anotar informações a respeito da integridade do ativo, com referência a falhas observadas para ações de manutenção corretivas ou planejadas seja em momento próximo ou futuro. Além disso, WO podem apresentar-se sob diferentes formatos, de acordo com o sistema de gestão empresarial adotado para o gerenciamento da manutenção (por exemplo, os sistemas SAP e Maximo), e por esta razão, outro modo comum para se referir a estes registros é utilizando o nome da plataforma de origem, por exemplo, "dados SAP".

Estes registros geralmente contém vários tipos de informações, de formato estruturado ou não estruturado. Dentre elas, a data e hora dos registros, os códigos de identificação de instalações, ativos e componentes, campo para identificação dos técnicos, e dois campos para preenchimento com texto livre. Conforme exemplificado por SALO; MCMILLAN; CONNOR (2019), o textos livres de menor tamanho (usualmente conhecidos como textos breves) se referem ao "título" da tarefa, isto é, a sua

sumarização. Enquanto os textos longos apresentam a parte que contém uma descrição mais longa da tarefa.

Para fins desta pesquisa, como ambos os tipos de texto fornecem informações valiosas sobre os eventos de falhas e as atividades de manutenção, o texto utilizado no treinamento será proveniente da concatenação das descrições breves e longas após a etapa de tratamento dos textos. A concatenação dos textos longos e breves de notas, ordens e operações em relação a um mesmo registro permite lidar com dados faltantes (uma vez que um registro pode não conter um texto longo (nota, ordem ou operação) mas possuir um texto breve, e vice-e-vesa, além de considerar mais textos para aumentar a confiabilidade dos dados e facilitar a recuperação da informação desejada.

Deste modo, a coleção será composta com as partes dos registros de manutenção que se referem aos textos (longos e breves) de formato livre e não estruturado, escritos em linguagem natural, acompanhados devidamente de seu número de identificação (ID). A coleção de documentos será disposta em uma tabela de histórico de dados conforme a o exemplificado na Tabela 6.

Tabela 6 – Exemplo da tabela de histórico de dados

ID	**Texto Breve**	**Texto Longo**
…	….	Nota: ……..... ……... Ordem: …….. ……... Operações: …… ……...

Os registros de manutenção de equipamentos podem ser criados devido a percepção de um defeito ou problema, devido a rotina de inspeção ou planejamento ou revisão de paradas. Usualmente, um registro é criado para cada atividade de manutenção em cada ativo individual. Assim, um registro de manutenção pode ou não se referir a ocorrência de uma falha. Ademais, por se tratar de documentos preenchidos manualmente por cada técnico, pode haver mais de um registro que se refira a uma mesma falha ou com dados inconsistentes.

Assim, um tarefa importante a ser realizada após a coleta dos registros nos bancos de dados de manutenção é a sua pré-avaliação para verificar e remover possíveis inconsistências (duplicidade de registros, registros que não se referem a falha ou que

possuem o *status* cancelado, por exemplo) para evitar que sejam avaliados registros duvidosos, que reduziriam a confiança dos resultados do classificador. Esta parte é realizada utilizando a ferramenta para visualização de dados BR-CHRONOS (MONTEIRO *et al.*, 2020), que permite verificar os registros ao longo do tempo, cruzando com outras bases de dados de operação e manutenção de equipamentos para identificar registros que se referem a mesma falha (dentro do mesmo período) ou que não são referentes a falha alguma.

Após a verificação dos registros, é necessário classificar manualmente os registros em relação ao modo de falha previsto nas normas de catalogação de dados de confiabilidade. Por se tratar de uma metodologia que objetiva minerar textos de manutenção de equipamentos da indústria do petróleo, escritos na língua portuguesa (pt-BR, português do Brasil), deseja-se realizar a catalogação de dados especificamente segundo a norma ABNT NBR ISO 14224 (ABNT, 2011). Esta é escrita neste mesmo idioma e trata da coleta e intercâmbio de dados de confiabilidade na indústria O&G. Por esta razão, é necessário utilizar modelos de processamento que trabalhem com o idioma especificado. Ressalta-se que apesar disso, a arquitetura proposta não inviabiliza a intercambialidade dos dados catalogados de acordo com a norma internacional ISO 142224 (ISO, 2016) ou o banco de dados OREDA, por exemplo.

Nesta fase deve-se considerar que cada equipamento avaliado tem modos de falha próprios segundo a norma de catalogação de dados de confiabilidade adotada. E que, portanto, registros de equipamentos diferentes não são comparáveis. Deste modo, as informações de falhas coletadas nos bancos de confiabilidade devem estar de acordo com cada registro e tipo de equipamento analisado.

Conforme comentado no item 2.3, os bancos de dados de confiabilidade possuem diversas informações relevantes quanto a falha. No entanto, para a abordagem proposta nesta pesquisa são utilizados apenas os dados de probabilidade de falha dos componentes do equipamento (obtidos no banco de dados OREDA) e seu modo de falha, isto é, o rótulo de falha referente a cada registro avaliado.

Por fim, serão adicionadas à tabela de histórico de dados as informações referentes a catalogação do modo de falha e a probabilidade de falha dos componentes, realizada anteriormente por especialistas, que se referem a mesma identidade de registros coletados. Assim, a nova tabela de histórico de dados é apresentada na Tabela 7.

Tabela 7 – Exemplo da tabela final de histórico de dados

ID	Texto Breve	Texto Longo	Modo de Falha (MF)	Probabilidade MF
…	….	…… …… …… …… …… ……	……..	%

6.1.2. Pré-processamento dos textos

Após a fase de coleta e verificação dos registros, ambos os textos (longos e breves) da Tabela 7 passam por uma etapa de pré-processamento para sua limpeza e padronização textual. O objetivo é obter um conjunto de palavras que caracterizem os textos em relação a cada determinada categoria, de tal modo que o conjunto obtido permita a classificação dos textos.

Nesta etapa, alguns dos processos comuns a quase todas as etapas de pré-processamento de dados textuais (ver item 4.2.3) são executados. A lista das etapas gerais do pré-processamento é apresentada abaixo:

- *Tokenization*: divisão dos dados textuais em componentes menores e mais significativos (tokens). Neste estudo, optou-se por realizar o processo de tokenização em unigramas, isto é considerando o termo como uma única palavra. Por exemplo: “manutenção corretiva” é considerada como dois tokens “manutenção” e “corretiva” na tokenização unigrama. Enquanto na tokenização bigrama, seria um único token considerando as duas palavras, “manutenção corretiva”.
- *Case folding:* padronização do formato de escrita dos textos. Transformação dos tokens para que todas as palavras sejam formatadas em letra minúscula (*lower case*). Por exemplo, os termos “manutenção”, “Manutenção” e “MANUTENÇÃO” seriam representados todos da mesma forma como “manutenção”.
- Remoção de tokens irrelevantes ou desnecessários: remoção de *stopwords*, caracteres especiais e numéricos.
- Correção de erros ortográficos: correção de erros devido à digitação ou por estarem escritos na forma informal da língua (por exemplo: erros de

concordância). Para a correção dos erros ortográficos foi utilizada a biblioteca Python *SpellChecker* (ver no item 6.1).

- *Stemming*: extração dos radicais das palavras utilizando o algoritmo RSLP (idioma: português) disponível na biblioteca NLTK (ver no item 6.1).

Ademais, faz parte da etapa de pré-processamento lidar com componentes específicos do problema a ser resolvido. Deste modo, é também realizada a identificação, remoção ou substituição de expressões regulares nos textos que não fornecem informação direta ou relevante sobre a falha e por isso dificultam a classificação.

Por exemplo, a parte textual relativa aos procedimentos administrativos da atividade são dados irrelevantes para o algoritmo de aprendizado, sendo necessário remover os ruídos. Um exemplo deste caso pode ser o nome/número de identificação do funcionário que realizou a manutenção/descrição do evento e o momento (data e hora) que a ação foi realizada. Assim, a informação é removida dos textos.

Ao fim da etapa de pré-processamento os textos breves e longos de cada respectivo documento são concatenados para serem considerados como um único texto.

6.1.3. Vetorização dos textos

Nesta metodologia optou-se por utilizar o modelo de representação de documentos vetorial, em que documento de texto pode ser representado por um BOW (conjunto de termos contidos no documento).

Em razão do grande número de palavras diferentes contidas em uma coleção de textos, onde a maioria das palavras ocorre apenas em uma pequena parte dos documentos, esta pesquisa emprega a abordagem estatística TF-IDF para seleção dos atributos.

Entende-se por TF-IDF a combinação do peso local que um termo possui em um documento (medida TF) e seu peso global em relação a coleção (medida IDF). A medida TF é definida como o número de vezes que um termo t ocorre em um documento d, enquanto o IDF para o termo t é definido como $\log(N/N_t)$, onde N é o número total de documentos e N_t é o número total de documentos contendo t.

Assim, conforme a singularidade do termo entre os documentos aumenta (ou seja, o termo é pouco frequente na coleção) e sua ocorrência em um documento específico aumenta, estes termos receberão pesos altos, pois são capazes de diferenciar um documento de outro.

Devido aos problemas associados com o tamanho (comprimento) do documento, é necessário normalizar o TF-IDF. POLETTINI (2004) apresenta alguns dos problemas relacionados a atribuição de pesos. Por exemplo, em casos em que a normalização não é realizada, um termo pode ter peso maior em um determinado vetor-documento simplesmente porque o documento correspondente é muito grande. POLETTINI (2004) justifica porque a normalização via cosseno é vantajosa pois esta lida em uma única etapa com os problemas ocasionados pelas frequências de termo mais altas e número de termos. HUANG (2008) apresentam com mais detalhes a teoria por trás da normalização via similaridade de cosseno. Desta forma, o TF-IDF normalizado via cosseno será calculado conforme apresentado na Equação 6 (LOPES, 2004).

$$Nt_k d_j = \frac{TF^* IDF(t_k d_j)}{\sqrt{\sum_{s=1}^{|T|}(TF^* IDF(t_s d_j))(TF^* IDF(t_s d_j))}} \qquad \text{Equação 6}$$

Onde t_k é o termo e d_j, o documento em consideração; o $TF^* IDF(t_s d_j)$ é a medida da coordenada de t_s em d_j, e $|T|$ é o número total de termos no espaço de termos.

Para esta etapa da metodologia utilizou-se os recursos das funções *CountVectorizer* e *TfidfTransformer* da biblioteca *Scikit Learning* para o processo de vetorização das palavras.

6.1.4. Classificação dos registros de manutenção

Após a conclusão da vetorização dos textos pré-processados, a última etapa é classificar um registro de manutenção como modo de falha do equipamento, de acordo as palavras representativas dos documentos.

O problema é resolvido como um modelo preditivo de classificação, em que as entradas são a representação vetorial das palavras contidas nos documentos e uma variável de saída que representa o modo de falha identificado no registro.

Conforme já apresentado, os registros de manutenção são classificados em modos de falha segundo os padrões das normas existentes para catalogação de dados de confiabilidade e de acordo com a rotulação dos especialistas. Desta forma, objetivo desta etapa é detectar, de forma supervisionada, as múltiplas classes de modos de falha (segundo a NBR ISO 14224) presentes na coleção, conforme o conteúdo de cada registro.

O fluxograma da Figura 4 apresenta em detalhes todos os passos da metodologia proposta nesta dissertação. A partir desta mesma figura observa-se que a etapa de

aprendizado possui duas sub-etapas principais, a calibração dos parâmetros e a avaliação do ajuste do modelo. Ambas as etapas são realizadas para os algoritmos de classificação selecionados, em que os classificadores serão avaliados em cada respectiva etapa simultaneamente.

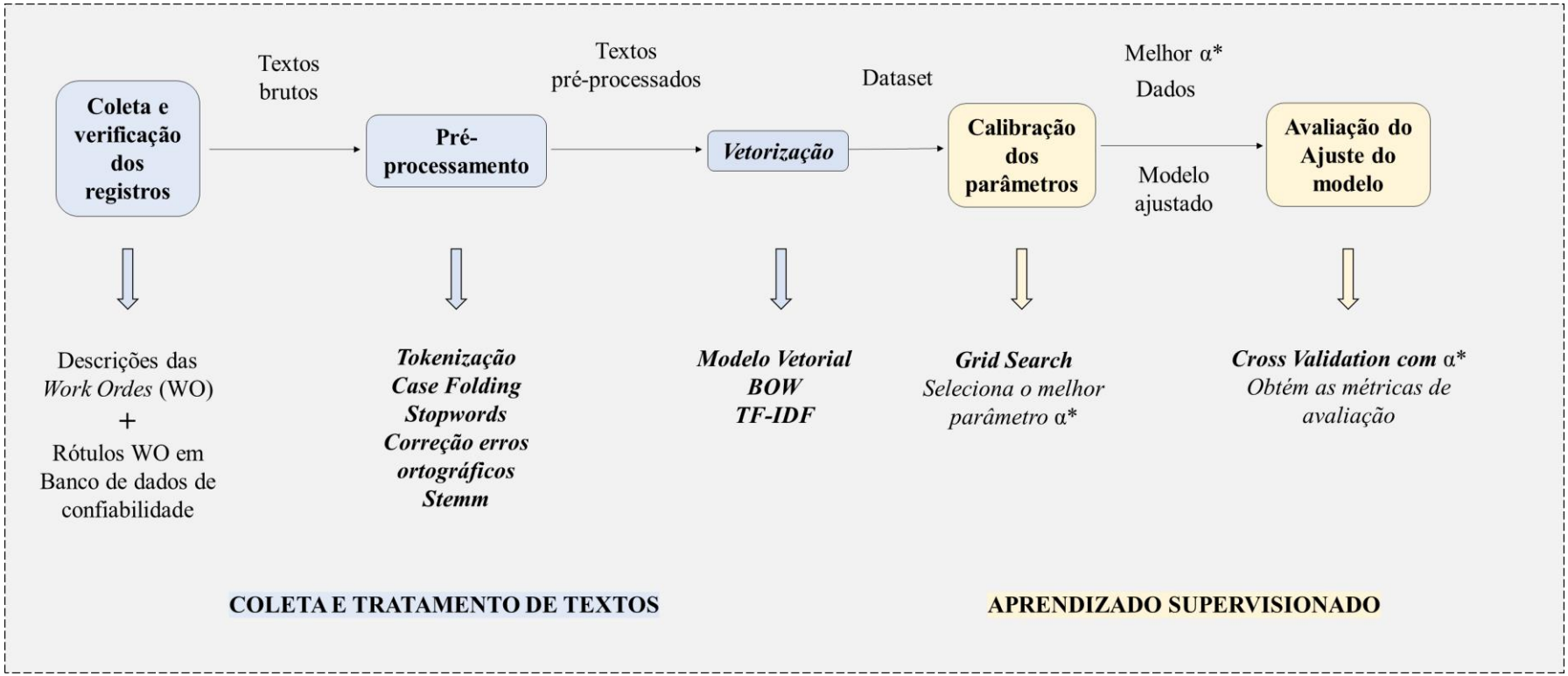


Figura 4 – Fluxograma completo da metodologia de classificação de registros de manutenção

6.1.4.1. Algoritmo de classificação

Para realizar a classificação automática da coleção de registros, dois modelos de classificação *Multinomial Naive Bayes - Multinomial Naïve Bayes* (MNB) e *Complement Naïve Bayes* (CNB) foram selecionados. Os dois algoritmos estão disponíveis na biblioteca *Scikit Learning* (PEDREGOSA *et al.*, 2011) e foram propostos nesta metodologia para a comparação de seus desempenhos, justificando a escolha do melhor algoritmo para a solução do problema.

Ambos os algoritmos são baseados na combinação do Teorema de Bayes e uma suposição ingênua que os atributos do documento são condicionalmente independentes, em que consideram que a distribuição multinomial para representar a ocorrência de uma palavra em cada documento.

No entanto, os algoritmos se diferenciam pelas estimativas do parâmetro $\widehat{\theta_{y_i}}$, em que o algoritmo CNB, descrito em RENNIE *et al.* (2003), pode ser considerado uma adaptação do algoritmo MNB em que as estimativas empíricas costumam ser mais estáveis do que para o MNB. Por esta razão, o algoritmo CNB é especificamente adequado para conjuntos de dados desequilibrados, como é o caso de dados de modo de falha obtidos a partir de uma coleção de registros de manutenção. Assim, esta metodologia propõe testar esta hipótese apresentada por RENNIE *et al.* (2003) que o algoritmo CNB possui regularmente um desempenho superior ao MNB (frequentemente por uma margem considerável) em tarefas de classificação de texto.

Enquanto o parâmetro $\widehat{\theta_{c_i}}$ do MNB pode ser estimado segundo a Equação 7, o CNB estima parâmetros usando todas as amostras de outras classes, em vez das amostras de treinamento da própria classe c_i.

$$\widehat{\theta_{c_i}} = \frac{N_{c_i} + \alpha}{N_c + \alpha n} \qquad \text{Equação 7}$$

Desta forma, o CNB lida com suposições inverídicas como a independência de dados e dados de treinamento desbalanceados, estimando os pesos de limite de decisão para evitar situações tendenciosas em que o classificador prefere involuntariamente uma classe à outra.

O algoritmo CNB aplica estatísticas do complemento de cada classe para calcular os pesos do modelo, conforme apresentado na Equação 8.

$$\widehat{\theta_{c_i}} = \frac{\alpha_i + \sum_{j:y_i \neq c} d_{ij}}{\alpha + \sum_{j:y_i \neq c} \sum_k d_{kj}} \quad \text{Equação 8}$$

Cujo somatório é realizado sobre todos os documentos *j* que não estão na classe c_i, d_{ij} indica o valor TF-IDF do termo no documento, α_i é um hiperparâmetro de suavização como o encontrado no MNB e α é igual ao somatório de todos os α_i.

Então é realizada uma segunda normalização, para lidar com a tendência de documentos mais longos dominarem as estimativas de parâmetros no MNB. Assim, temos a Equação 9.

$$w_{c_i} = \log \widehat{\theta_{c_i}} \quad \text{Equação 9}$$

A regra de classificação é apresentada na Equação 10.

$$\hat{c} = \arg\min_c \sum_i t_i w_{ci} \quad \text{Equação 10}$$

Ou seja, um documento é atribuído à classe que é a menor correspondência do complemento.

Para ambos os modelos, MNB e CNB, o parâmetro testado na calibração foi $\alpha \in \{0, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 0.1 \text{ e } 1\}$. Além disso, o peso das classes foi ajustado pelas probabilidades de ocorrência de cada do modo de falha (que foi obtida previamente no banco de dados de confiabilidade OREDA).

6.1.4.2. Calibração dos modelos

A primeira parte da metodologia do aprendizado consiste em realizar a calibração dos modelos. Selecionado o algoritmo de classificação a ser empregado e os parâmetros a serem ajustados, o *dataset* obtido após as etapas descrita no item 6.1.3 é utilizado para a determinação dos melhores parâmetros do modelo. O procedimento de calibração consiste em testar diferentes combinações de parâmetros específicos de cada modelo de aprendizagem, para otimizar seu desempenho. Nesta metodologia, a calibração é realizada utilizando o algoritmo *Grid Search* contido na biblioteca *Scikit Learning* (PEDREGOSA *et al.*, 2011). O *Grid Search* considera, de uma só vez, todas as combinações dos parâmetros disponibilizado (referente ao algoritmo de classificação escolhido) utilizando o método de validação cruzada e uma métrica para analisar a qualidade dos resultados.

Optou-se por utilizar a validação cruzada *k-fold* como procedimento para fazer devidas avaliações nos modelos e minimizar os efeitos de *overfitting*. Problemas de sobre ajuste (*overfitting*) são ocasionados por considerar apenas uma parte pequena dos dados para testar os resultados. Isto é, o modelo fica muito ajustado aos dados utilizados e perde a capacidade de generalização. Dessa forma, os modelos treinados via cruzada *k-fold* não se ajustariam a apenas a uma pequena amostra, o que tornaria possível prever eficientemente a classificação de novos registros.

No processo de validação cruzada *k-fold*, o conjunto de dados inicial é dividido aleatoriamente em k subconjuntos (F_1, F_2,..., F_k) mutuamente exclusivos. Estes subconjuntos são usualmente denominados *folds* e cada um possui tamanho aproximadamente igual aos demais. O treinamento e o teste são realizados k vezes. A cada iteração i, a partição F_i é utilizada para testar o modelo segundo a métrica de avaliação escolhida. Já as partições restantes, serão utilizadas coletivamente para treinar o modelo de aprendizado (HAN *et al.*, 2012).

Desta forma, cada ciclo *k* corresponde a um conjunto de registros de manutenção escolhidos aleatoriamente. Esta metodologia optou utilizar a função *StratifiedKFold* presente no pacote *Scikit Learning* (PEDREGOSA *et al.*, 2011) para realizar a validação cruzada de forma estratificada. A aplicação desta estratégia permite que cada rodada tenha a mesma proporção de cada classe em relação ao conjunto original, sendo bastante adequada para *dataset* extremamente pequenos (menos que 500 registros).

Da etapa de calibração são obtidos o melhor conjunto de parâmetros e a média da métrica obtida na validação cruzada. Nesta metodologia, adotou-se a métrica F ponderada, apresentada no item 4.2.5.2. Os modelos de classificação e seus respectivos parâmetros foram citados no item 6.1.4.1.

A métrica F ponderada foi selecionada por ser adequada para avaliar a performance da previsão de modelos de classificação multiclasse desbalanceados, pois combina as métricas precisão e *recall*, ao mesmo tempo em que pondera a frequência das classes. Neste processo, a métrica escolhida era calculada a cada rodada, e no final, o resultado considerado consistia no valor médio obtido considerando todas as rodadas.

6.1.4.3. Avaliação do Ajuste do modelo via validação cruzada

Determinados os parâmetros otimizados, é realizada a última etapa da metodologia de classificação. Para avaliar os resultados, os modelos calibrados são

novamente rodados com a validação cruzada *k-fold*, com as mesmas amostras utilizadas na calibração. A ideia é repetir o processo da validação, para verificar a robustez do modelo calibrado.

Nesta etapa, mais métricas serão obtidas a fim de verificar eficiência do modelo com o parâmetro otimizado *versus* o modelo de parâmetros *default* da biblioteca *Scikit Learning* (PEDREGOSA *et al.*, 2011).

Dentre as métricas de avaliação dos modelos de classificação, a etapa de ajuste calcula a acurácia, precisão, recall e medida F (todas ponderadas) para cada rodada. Todas as métricas possuem valores dentro do intervalo [0,1]. Nas situações em que a métrica obtém valores mais próximos de 1, percebe-se que os dados são melhor tratados pelo modelo.

Na etapa de avalição do ajuste, as métricas também são calculadas a cada rodada. Ao fim desta etapa é possível obter os valores médios das métricas. A comparação entre os diferentes algoritmos de classificação, tal como a sua versão *default* e ajustada, permitem identificar, dentre os algoritmos de classificação selecionados, aquele que possui melhor desempenho.

7 ESTUDO DE CASO

Este capítulo apresenta a experimentação e avaliação do procedimento metodológico exposto no Capítulo 6, proposto como solução do problema definido nesta pesquisa.

7.1 Experimentação e resultados

O estudo de caso foi utilizado para testar a aplicabilidade da metodologia proposta. Esta foi descrita e implementada em um framework desenvolvido em linguagem de programação Python utilizando as bibliotecas e os sistema operacional citados no item 6.1.

Para a etapa de coleta de dados, foi elaborada uma rotina Python para reunir os textos dos relatórios (breves e longos), bem como o rótulo catalogado e as probabilidades de modo de falha, criando um conjunto de dados para cada documento em análise. Inicialmente o conjunto de dados coletados consistia de 431 documentos SAP referente a 7 anos (2006-2011 e 2016) de registros manutenção de turbinas utilizadas na E&P de petróleo *offshore*.

A partir da análise dos registros suportada pela ferramenta BR-CHRONOS, observou-se que 12,53% dos registros de manutenção coletados se referiam a registros duplicados ou não se referiam a registros de falha de fato. Assim estes registros foram desconsiderados, permitindo selecionar melhor os dados coletados.

Adicionalmente, verificou-se que após a catalogação manual dos registros considerados segundo a norma citada e o banco de dados OREDA (377 no total, distribuídos conforme a Figura 5) que 15,12% dos documentos foram classificados como modos genéricos ou desconhecidos, dificultando identificar os padrões textuais que representam estes tipos de classe, e que 9.28% dos registros possuíam menos de 10 registros por classe, o que impossibilitou realizar o treinamento de todos os modos de falha existentes segundo a norma para turbina. E portanto, esta parcela de registros também foi desconsiderada.

Verificou-se também ser necessário agrupar três modos de falha de origem vazamento interno ou externo, apenas como a classe “vazamento” para aumentar a quantidade de registros disponível da classe geral e facilitar a classificação. Deste modo, o conjunto total de dados empregado no aprendizado supervisionado consistiu em 265

registros de manutenção previamente e manualmente catalogados por especialistas. Os registros analisados possuíam diferentes modos de falha (Figura 6), distribuídos respectivamente, como leitura anormal do instrumento (AIR – 68,68%), vazamento (24,15%) e pequenos problemas em serviço (SER – 7,17%).

Destes 265 registros SAP considerados, 30 registros continham os textos breve, 265 os textos longos referente a percepção da falha (nota), 33 textos longos quanto a solicitação da correção da falha (ordens) e 39 textos longos da descrição da operação de manutenção em si (operações).

Para exemplificar o conteúdo dos documentos considerados, a Tabela 8 apresenta seis documentos selecionados aleatoriamente dos registros de manutenção obtidos no sistema SAP (Notas, Ordens e Operações) contendo os textos longos e breves.

Observa-se, no entanto, que alguns dados foram corrompidos e descaracterizados propositalmente por motivos de confidencialidade. Outra consideração a respeito da Tabela 8 é que a classe denominada na coluna “Rótulo registro” se refere a catalogação manual realizada pelos especialistas, onde o código “AIR” representa o registros classificados como leitura anormal do instrumento, “LK” à vazamento e “SER” pequenos problemas em serviço.

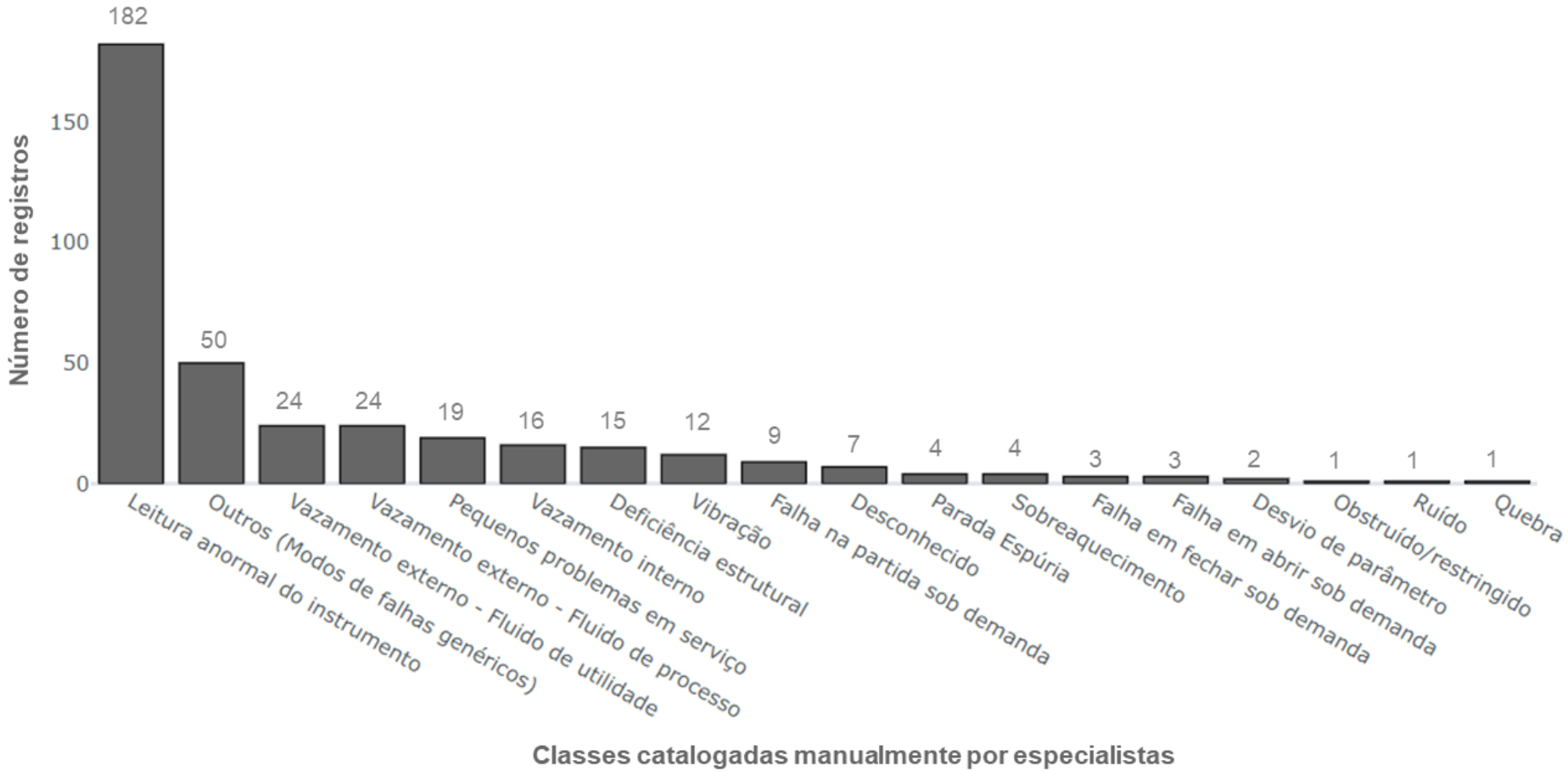


Figura 5 – Distribuição dos registros por classe catalogada manualmente por especialistas

Fonte: Dados da Pesquisa

Tabela 8 – Exemplo de conteúdo de texto breve e longo de cada documento antes da concatenação dos textos e o pré-processamento dos dados

ID Registro	Rótulo Registro	Texto Breve	Texto Longo
12	AIR	-	**Nota:** * 17.04.2009 10:20:27 [redacted] * -Sanar problema de falsa indicação do [redacted] * * -Localização: [redacted] * * -Responsável: [redacted] **Ordem:** [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] **Operação:** [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.
108	LK	-	**Nota:** 26.05.2009 14:46:08 [redacted] OPERANDO COM BAIXA EFICIÊNCIA. * ABERTURA PARA LIMPEZA, INSPEÇÃO, REPARO E POSTE RIOR FECHAMENTO. * NECESSITA ANDAIME. * 26.05.2009 15:12:40 [redacted] * 03.09.2009 07:43:44 [redacted] * O BS.: SUBSTITUIR FEIXE DE TUBOS CONFORME DEFINIDO PELO [redacted] **Ordem:** - **Operação:** -

ID Registro	Rótulo Registro	Texto Breve	Texto Longo
185	SER	-	**Nota:** * 08.01.2008 22:18:58 [redacted] * Recomendações de Inspeção Externa Periódica do [redacted] Separador * de Sucção 2º Estágio. * 09.01.2008 17:12:01 [redacted] * AS MEDIDAS AINDA ESTÃO PENDENTES DE LIBERAÇÃO. FAVOR REPASSAR A NOTA * PARA O [redacted] QDO AS MEDIDAS FOREM DEVIDAMENTE LIBERADAS PELO [redacted]. * 10.01.2008 12:53:22 [redacted] * AS MEDIDAS AINDA NÃO ESTÃO LIBERADAS. ESTÃO SOB A ANÁLISE DO [redacted]. **Ordem:** Recomedações Insp. Ext. [redacted] INSTALAR ALAVANCA NA VÁLVULA DE BLOQUEIO DO [redacted] P´ROXIMO AO BOCAL [redacted] - PISO DO COMPRESSOR. **Operação:** INSTALAR ALAVANCA NA VÁLVULA DE BLOQUEIO DO LIT-[redacted], PRÓXI MO AO BOCAL [redacted] - PISO DO COMPRESSOR. -Item 03 da medida da nota ZR.
348	SER	[redacted] - Falha na solenoide	**Nota:** * 05.01.2016 15:53:06 [redacted] * Falha na solenoide, utilizar [redacted] **Ordem:** - **Operação:** 05.01.2016 15:53:06 [redacted] solenoide em falha, utilizar [redacted]
351	LK	[redacted] Vazamento	**Nota:** * 27.01.2016 16:22:58 [redacted] * Vazamento pelo selo do diafragma do pressostato. necessário substituição. **Ordem:** [redacted] - Vazamento OM SOLICITADA PELO TEC. [redacted] **Operação:** Vazamento pelo selo do diafragma do pressostato. necessário substituição.

ID Registro	Rótulo Registro	Texto Breve	Texto Longo
349	AIR	[redacted] Indicação Espúria	**Nota:** * 09.01.2016 14:59:07 [redacted] * SOLICITADO: [redacted] * APROVADO: [redacted] * * FALHA: O [redacted] ESTÁ OSCILANDO DE 0 A 300ºC. * * ATIVIDADE: VERIFICAR / SANAR FALHA NO [redacted] * * LOCALIZAÇÃO: [redacted] * * --------- DELINEAMENTO --------- * * TIPO DE TRABALHO: À QUENTE * NECESSITA PARADA DO EQUIPAMENTO: TG-A; * PRESSURIZADO: NÃO; * NECESSITA MONTAGEM DE ANDAIME: NÃO; * ELÉTRICO: NÃO. * MÃO DE OBRA NECESSÁRIA: 02 INSTRUMENTISTAS / 02 HORAS. * * LISTA DE TAREFAS: * - VERIFICAR / SANAR FALHA NO [redacted]. **Ordem:** [redacted]-Indicação Espúria Ordem na Base, realizado o serviço e está em processo de scaneamento. **Operação:** 09.01.2016 14:59:07 [redacted] SOLICITADO: [redacted] APROVADO: [redacted] FALHA: O [redacted] ESTÁ OSCILANDO DE 0 A 300ºC. ATIVIDADE: VERIFICAR / SANAR FALHA NO [redacted]. LOCALIZAÇÃO: TG-A --------- DELINEAMENTO --------- TIPO DE TRABALHO: À QUENTE NECESSITA PARADA DO EQUIPAMENTO: [redacted]; PRESSURIZADO: NÃO; NECESSITA MONTAGEM DE ANDAIME: NÃO; ELÉTRICO: NÃO. MÃO DE OBRA NECESSÁRIA: 02 INSTRUMENTISTAS / 02 HORAS. LISTA DE TAREFAS: - VERIFICAR / SANAR FALHA NO [redacted]. Programar troca do sensor/ Apontar [redacted] na OM/

Fonte: Dados da Pesquisa

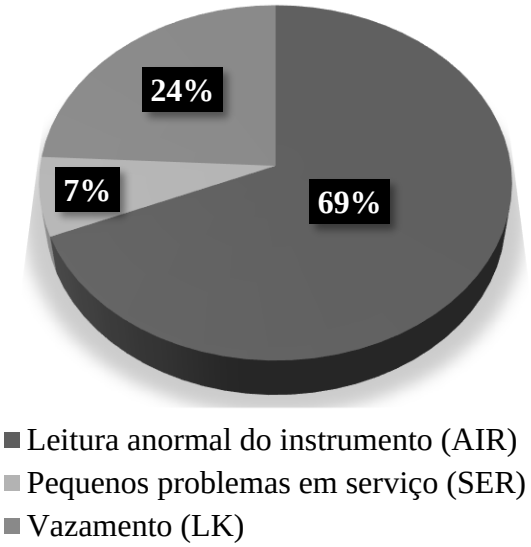


Figura 6 – Percentual de registros de acordo com as classes catalogadas considerada

Fonte: Dados da Pesquisa

A seguir são detalhados os resultados de obtidos nos estágios 2, 3 e 4 da metodologia.

7.1.1. Pré-Processamento dos textos

Após a coleta e verificação dos registros, a segunda etapa consiste em realizar o pré-processamento dos textos dos registros. No caso dos registros SAP os mesmo são compostos por quatro tipos de textos. Um texto breve que se repete a cada nota, ordem e operação equivalente ao mesmo registro, e três textos longos denominados notas (percepção da falha), ordens (solicitação da correção da manutenção) e operações (a descrição da atividade de manutenção em si).

Devido ao fato que cada tipo texto SAP possuir padrões próprios a cada um deles, os textos de cada registro SAP foram pré-processados separadamente para facilitar o processo de reconhecimento de expressões regulares proveniente de cada tipo de texto. Ao final da etapa de pré-processamento, todos os textos pré-processados foram concatenados, sendo denominados antes da etapa de pré-processamento como texto bruto, e após, como texto pré-processado.

No entanto, ressalva-se que os resultados aqui relatados consideram que os textos (tanto os brutos, quanto os pré-processados) são obtidos da concatenação de todos os textos longos e breves advindos dos dados SAP (notas, ordens e operações) que tratam do mesmo registro de manutenção. Assim todos os dados e estatísticas apresentados já

consideram cada registro como a soma dos quatro tipos de textos, isto é, todos formam um único texto. Por exemplo, no caso do registro identificado como 351 (citado na Tabela 8) seu texto bruto é a junção do texto breve e dos textos longos das notas, ordens e operações conforme apresentado na Tabela 9.

Tabela 9 – Exemplo de documento após a concatenação dos textos

ID Registro	Rótulo Registro	Texto Bruto
351	LK	[redacted] – Vazamento. * 27.01.2016 16:22:58 [redacted] * Vazamento pelo selo do diafragma do pressostato. necessário substituição. [redacted] - Vazamento OM SOLICITADA PELO TEC. [redacted]. Vazamento pelo selo do diafragma do pressostato. necessário substituição.

Fonte: Dados da Pesquisa

Porém, existe alguns registros que não possuem os quatro tipos de textos citados, como é o caso do registro identificado na Tabela 8 como 108, no qual só há o texto longo da nota.

O agrupamento dos textos foi considerado pois é realizado de modo similar ao processo de catalogação manual realizada por especialistas, que avaliam o conteúdo de todos os textos de um único registro para obter a informação de modo de falha. A escolha, também se deve ao fato que desta forma é possível lidar com dados faltantes, sem que seja necessário desconsiderar nenhum registro considerado como falha, permitindo que mais registros fossem catalogados.

Adicionalmente, o agrupamento dos textos permite melhorar a confiança nas informações contidas quando comparado a utilização de apenas um único texto (texto breve ou nota ou ordem ou operação), uma vez que nem sempre os autores de uma nota, ordem e operação, referentes a um mesmo registro são realizadas por um mesmo autor.

No entanto, duas desvantagens desta abordagem é que isto também influencia baixa padronização dos textos e na alta variabilidade no tamanho dos registros considerados, uma vez que em algumas situações vários tipos de textos (texto breve, notas, ordens, operações) são considerados e em outras, apenas um.

As estatísticas do quanto do texto original foi reduzido após o fim da etapa de pré-processamento podem ser observadas nas Tabelas 10 e 11.

A partir da análise apresentada na Tabela 10 é possível verificar que o tamanho médio dos textos brutos em relação ao texto pré-processado reduziu 63%. Enquanto as

estatísticas obtidas na Tabela 11, apresentam que o tamanho dos textos brutos em relação aos quartis 25%, 50% e 75% do tamanho dos textos pré-processados variaram em média 68% aproximadamente.

Tabela 10 – Percentual de redução dos textos originais (brutos) após a etapa de pré-processamento

	Menor texto	**Maior texto**	**Tamanho médio**	**Desvio padrão**
Texto bruto	88	1.046	241,15	154,09
Texto pré-proc.	8	621	89,8	885,64
% reduzido	91%	41%	63%	44%

Fonte: Dados da Pesquisa

Tabela 11 – Estatísticas descritivas dos textos originais (brutos) e após a etapa de pré-processamento (pré-proc.)

	Q25	**Q50**	**Q75**
Texto bruto	143	191	289
Texto pré-proc.	40	63	103
% reduzido	72%	67%	64%

Fonte: Dados da Pesquisa

Ambos os resultados das Tabelas 10 e 11 corroboram com o fato que a maior parte do conteúdo dos registros é algum tipo de ruído, e precisaram ser removidos. A maioria deles, foram pontuações e caracteres especiais do corpus. Mas além deles, stopwords e expressões regulares que não acrescentavam informações úteis quanto a identificação do modo de falha do equipamento.

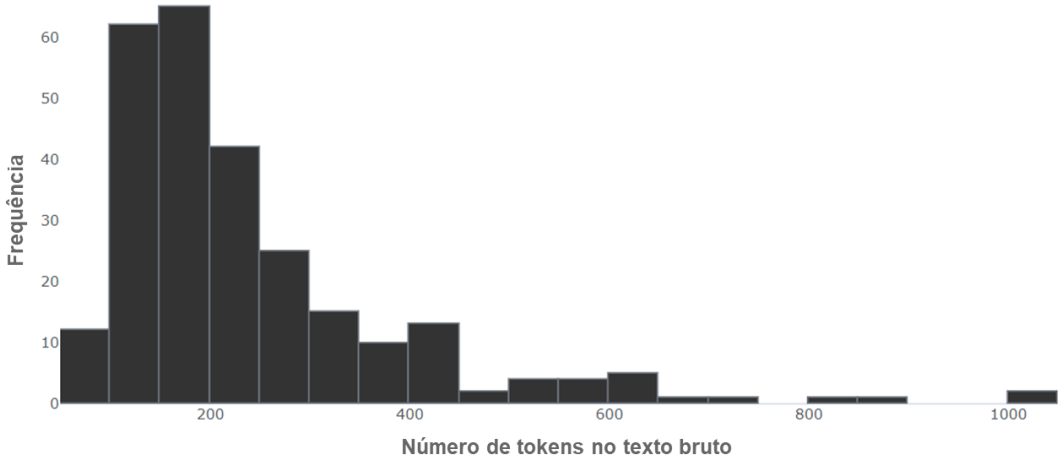
Também é possível observar a partir dos desvios padrão calculados em relação às médias (Tabela 10), que há uma grande variação nos tamanhos dos textos. Como citado anteriormente, este fato pode ser justificado em razão da concatenação dos textos SAP, mas também devido a pessoalidade empregada por cada autor do registro (alguns mais prolixos e outro menos) e em função da baixa padronização dos textos (alguns continham informações mais detalhadas sobre toda a atividade e percepções da falha).

A maneira como os registros são distribuídos em relação ao seu tamanho (isto é, o número de tokens) antes e após o pré-processamento do texto é apresentada nas Figuras 7 e 8, respectivamente.

Avaliando a Figura 7 observa-se que a maioria dos registros originais possuem entre 150 e 200 tokens. Após a etapa de pré-processamento a maior parte dos registros tem entre 0 e 50 tokens, conforme pode ser observado na Figura 8.

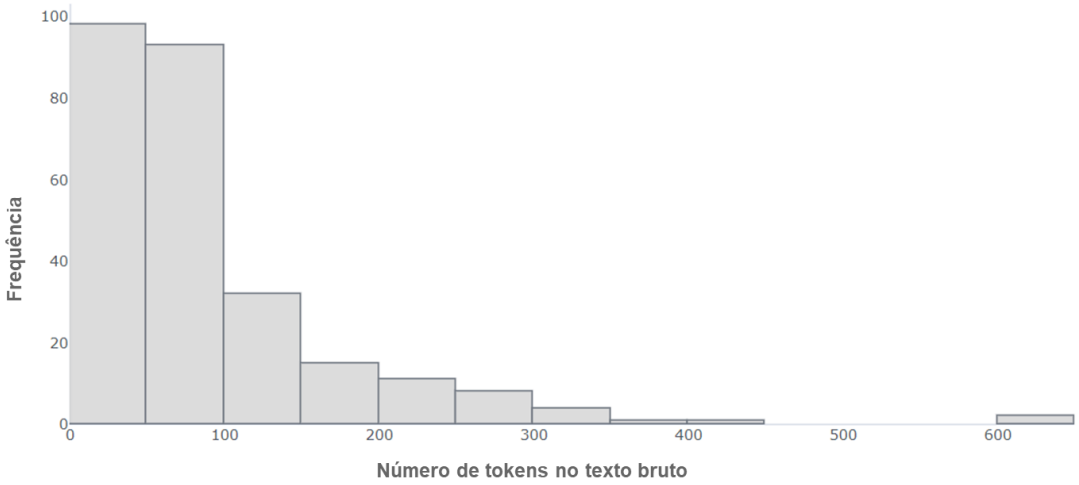
Ademais, analisando o histograma do tamanho dos registros antes e após a etapa de pré-processamento (Figuras 7 e 8) e também a partir das distribuições dos quartis (Tabela 11) observa-se que a mediana do tamanho dos registros brutos é 191, enquanto o dos registros pré-processados é 63.

Figura 7 – Histograma do número de tokens em relação aos textos brutos



Fonte: Dados da Pesquisa

Figura 8 – Histograma do número de tokens em relação aos textos pré-processados



Fonte: Dados da Pesquisa

A Tabela 12 apresenta os exemplos dos registros citados na Tabela 8 após a etapa de pré-processamento, utilizando ou não a técnica *stemming*. São apresentadas as frequências de cada palavra ou termo "*stemm"* contidas em cada registro, além disso são apresentados os tamanhos dos textos brutos e pré-processados e o percentual de texto reduzido ao fim desta etapa.

Tabela 12 – Exemplo dos documentos textuais após a etapa de pré-processamento

ID Reg.	Texto pré-proc. (sem *Stemming*)	Texto pré-proc. (com *Stemming*)	Tam. texto bruto	Tam. texto pre-proc.	% red.
12	sanar[2] problema[2] falsa [3] indicação [3] falha [1]	san [2] problem [2] fals [3] indic [3] falh [1]	237	11	95,36%
108	operando [1] baixa [1] eficiência [1] abertura [1] limpeza [1] inspeção [1] reparo [1] posterior [1] fechamento [1] necessita [1] substituir [1] feixe [1] tubos [1] definido [1]	oper [1] baix [1] efici [1] abert [1] limp [1] inspeç [1] repar [1] posteri [1] fech [1] necessit [1] substitu [1] feix [1] tub [1] defin [1]	218	14	93,58%
185	recomendações [2] inspeção [1] externa [1] periódica [1] separador [1] sucção [1] estágio [1] medidas [3] ainda [2] pendentes [1] liberação [1] favor [1] repassar [1] devidamente [1] liberadas [2] análise [1] insp [1] instalar [1] alavanca [1] válvula [1] bloqueio [1] bocal [1] piso [1] compressor [1]	recomend [2] inspeç[1] extern [1] periód [1] separ [1] sucç [1] estági [1] med [3] aind [2] pend [1] liber [3] favor [1] repass [1] devid [1] anális [1] insp [1] instal [1] alavanc [1] válvul [1] bloquei [1] bocal [1] pis [1] compres [1]	466	29	93,78%

ID Reg.	Texto pré-proc. (sem *Stemming*)	Texto pré-proc. (com *Stemming*)	Tam. texto bruto	Tam. texto pre-proc.	% red.
348	falha [2] solenoide [2]	falh [2] solenoid [2]	83	4	95,18%
349	espúria [2] falha [3] oscilando [1] verificar [2] sanar [2] delineamento [1] trabalho [1] quente [1] necessita [2] parada [1] equipamento [1] pressurizado [1] montagem [1] elétrico [1] obra [1] necessária [1] instrumentistas [1] horas [1] lista [1] tarefas [1] ordem [1] base [1] realizado [1] processo [1] scaneamento [1]	espúr [2] falh [3] oscil [1] verific [2] san [2] deline [1] trabalh [1] quente [1] necessit [2] par [1] equip [1] pressur [1] mont [1] elétr [1] obr [1] necessár [1] instrument [1] hor [1] list [1] taref [1] ord [1] bas [1] realiz [1] process [1] scane [1]	582	31	94,67%
351	vazamento [3] selo [1] diafragma [1] pressostato [1] necessário [1]	Vaz [3] Sel [1] Diafragma [1] Pressostat [1] Necess [1]	136	7	94,85%

Fonte: Dados da Pesquisa

Avaliando os resultados do pré-processamento dos registros apresentados na Tabela 12, observa-se algumas limitações do algoritmo *stemming* utilizado. Por exemplo, a palavra “inspeção” aparece nos registros 108 e 185 e quando é reduzida a seu termo “*stemm”* é representada pelo termo “inspeç”. O fato do termo possuir o c cedilha (ç) impossibilita a generalização de palavras como “inspecionar” e “inspeção” para o radical “inspec”.

O caso ocorre mesmo quando é realizada a remoção prévia de acentos e de caracteres especiais como c cedilha (ç) para normalização. Um teste foi realizado

utilizando a biblioteca Python *Unidecode* (SOLC, 2019) na etapa de pré-processamento. Realizando a remoção com o pacote *Unidecode* e posteriormente aplicando o algoritmo *stemming*, a palavra inspeção é reduzida ao termo "inspeca", o que também impossibilita que as palavras "inspecionar" e "inspeção" sejam representadas pelo termo "inspec". O mesmo pode ser observado para a palavra "sucção", presente no registro 185.

No entanto, observa-se um caso bem sucedido da aplicação do algoritmo *stemming* que demonstra a relevância da normalização das palavras de mesmo radical. Este caso pode ser visualizado no registro 185, as palavras "liberação" e "liberadas" são representadas como dois tokens diferentes com frequência 1 e 2, respectivamente (no caso sem *stemming*. Mas, quando o processo *stemming* é realizado, as palavras são normalizadas para o termo "liber" que possui uma frequência igual 3 neste documento. Neste caso, o processo de *stemming* reduziu um token no documento. Algo que pode ocorrer em outras ocorrências do termo ao longo da coleção, e desta forma, permitir reduzir a dimensionalidade do problema.

Com o auxílio de um algoritmo de nuvem de palavras aplicado ao conjunto de dados que contém todos os registros de manutenção das turbinas, é possível visualizar os termos mais frequentes na coleção. Em que palavras que são representadas em uma fonte de texto maior representam as palavras mais comuns da coleção de registros desta classe. Ao mesmo tempo, palavras representadas pelo mesmo tom de cor possuem a mesma frequência.

Os resultados da aplicação do algoritmo de nuvem de palavras configurado para apresentar as 20 principais palavras mais frequentes após a etapa de pré-processamento para cada modo de falha são mostrados nas Figuras 9, 10 e 11.

Avaliando as palavras mais frequentes da classe vazamento (em relação ao registros catalogados) apresentada na Figura 9, pode se observar que a palavra "vazamento" é de fato a palavra mais frequente desta classe. Por ser uma palavra exclusiva a esta classe, ela também é a palavra mais representativa de registros classificados como modo de falha vazamento. A segunda palavra mais frequente desta classe é a palavra "sanar", um indicativo da ação de manutenção realizada para tratar esse modo de falha. Além disso, observa-se que frequentemente são citados os fluídos vazados. Em que se observa a partir da nuvem desta classe, que "vazamento de água" podem ser mais comuns que "vazamento de óleo" ou "lubrificante". Também é possível inferir que o local mais comum de vazamento neste tipo de equipamento, segundo os

registros avaliados, é no “lado acoplado da bomba de injeção” ou no “selo” dela, seguido de “linha” e “permutador”.

Nos casos dos registros da classe leitura anormal do instrumento (Figura 10), as correlações não são tão óbvias. Palavras-chave como “instrumento”, “sensor”, “alarme”, “indicação” possuem menor frequência quando comparadas aos termos “bomba” e “verificar” que são mais genéricos, por exemplo.

Quanto aos registros classificados como pequenos problemas em serviço (Figura 11), as palavras mais frequentes não se referem exclusivamente a este modo de falha. Além disso, os registros da classe SER são registros mais desuniformes que os da classe AIR e LK. Este fato pode dificultar a classificação automática, ocasionando erros na categorização.

É importante notar a partir da análise das Figuras 9, 10 e 11, que se fosse realizada uma vetorização dos documentos utilizando a abordagem TF, isto poderia levar a erros de classificação, principalmente nos casos das classes leitura anormal do instrumento e pequenos problemas em serviço, em que os termos mais comuns não são de fato exclusivos a estes respectivos modos de falha. Isto é um indicativo que a abordagem TF-IDF pode, de fato, ser uma boa estratégia para registros desse tipo, uma vez que não consideram apenas a frequência da palavra em um registro, mas também o inverso da frequência da palavra em relação a coleção.

Figura 9 – Nuvem de palavras das 20 principais palavras em relação aos registros classificados como vazamento (LK)

Fonte: Dados da Pesquisa

Figura 10 – Nuvem de palavras das 20 principais palavras em relação aos registros classificados como leitura anormal do instrumento (AIR)

Fonte: Dados da Pesquisa

Figura 11 – Nuvem de palavras das 20 principais palavras em relação aos registros classificados como pequenos problemas em serviço (SER)

Fonte: Dados da Pesquisa

De modo a ter uma visão geral da baixa normalização dos registros em comparação com o descrito na NBR ISO 14224, o mesmo algoritmo de nuvem de palavras foi utilizado. As nuvens de palavras elaboradas com a norma podem ser observadas nas Figuras 12, 13 e 14.

Ao avaliar as palavras chaves apresentadas nas descrições dos modos de falha da norma, verifica-se que para o modo de falha vazamento (Figura 12) a norma indica apenas

a palavra vazamento (que descreve a própria falha) e o fluido vazado (água, gás, lubrificante, fluidos de utilidade ou processo) e se o vazamento ocorreu interna ou externamente ao equipamento. No caso do modo de falha leitura anormal do instrumento (Figura 13) a norma indica vários sinônimos para as palavras “leitura” (indicação, leitura, alarme) e “anormal” (falso(a), errado, oscilante, espúrio). Enquanto para o modo de falha pequenos problemas em serviço, a norma indica diversos sinônimos deste tipo de falha, por exemplo: frouxo/afrouxamento, sujeira, desconexão e descoloração. Isto indica que este tipo de falha não possui apenas uma única causa conhecida, o que dificulta a sua classificação em relação aos modos de falha (AIR e LK) que possuem termos muito mais característicos e próprios.

Ao realizar uma comparação entre as palavras mais frequentes das classes (segundo os registros avaliados) com os termos contidos nas descrições de cada modo de falha segundo a NBR ISO 14224, é possível observar a falta de padronização dos registros em relação a mesma.

Além disso, nota-se que os registros das classes vazamento estão provavelmente mais de acordo com a norma que os registros de leitura anormal do instrumento, seguido da classe pequenos problemas em serviço. Aliás, observa-se que na classe pequenos problemas em serviço nenhuma das palavras-chaves da norma está entre as mais frequentes nos registros avaliados.

Figura 12 – Nuvem de palavras das principais palavras-chave contidas nas descrições da norma para o modo de falha vazamento (LK)

Fonte: NBR ISO 14224 (ABNT, 2011)

Figura 13 – Nuvem de palavras das principais palavras-chave contidas nas descrições da norma para o modo de falha leitura anormal do instrumento (AIR)

Fonte: NBR ISO 14224 (ABNT, 2011)

Figura 14 – Nuvem de palavras das principais palavras-chave contidas nas descrições da norma para o modo de falha pequenos problemas em serviço (SER)

Fonte: NBR ISO 14224 (ABNT, 2011)

7.1.2. Vetorização dos textos

Após a etapa de pré-processamento (etapa 2), os textos pré-processados são vetorizados para então serem utilizados como entradas dos modelos de classificação.

Ao fim da etapa da segunda etapa da metodologia proposta, os dados pré-processados consistiam em 715 *features* (termos).

A matriz termo-frequência representada deste estudo de caso consistiu em uma matriz de tamanho 265x715, com 2686 ocorrências de valores não nulos e de esparcidade igual a 1,42.

7.1.3. Classificação dos registros de manutenção

Após as etapas de pré-processamento e vetorização dos documentos, é realizada a categorização dos documentos com base em dois algoritmos de ML nos conjuntos de dados, MNB e CNB. Ambos os modelos foram calibrados com as respectivas probabilidades a priori de os registros serem classificados em relação a cada classe.

As probabilidades foram obtidas no banco de dados de confiabilidade OREDA e são apresentadas na Tabela 13.

Tabela 13 – Probabilidades de ocorrência dos modo de falha avaliados

Modo de falha	**Probabilidade %**
Pequenos problemas em serviço (SER)	7,09
Leitura anormal do instrumento (AIR)	10,40
Vazamento (LK)	15,19

Fonte: Dados da pesquisa

Ambas as etapas da classificação (calibração e de avaliação do ajuste do modelo) utilizaram a validação do conjunto de dados. Assim, a fim de garantir a divisão dos conjuntos na validação cruzada em ao menos na proporção 1:4, o *Stratified K Fold* foi aplicado em 5 *folds*. Desta forma, a cada rodada o conjunto de treinamento representava 80% (212 registros) dos dados e o teste 20% (53 registros). Assim, o conjunto de treinamento continha aproximadamente 144 registros rotulados como leitura anormal do instrumento (AIR), 50 de vazamento (LK) e 18 de pequenos problemas em serviço (SER) e no conjunto de teste, 38, 14 e 1 registro, respectivamente.

Nas Figuras 16 e 17 são apresentados os resultados da etapa de calibração para ambos os algoritmos de ML propostos na metodologia.

Figura 15 – Resultado da validação cruzada na etapa de calibração do modelo MNB

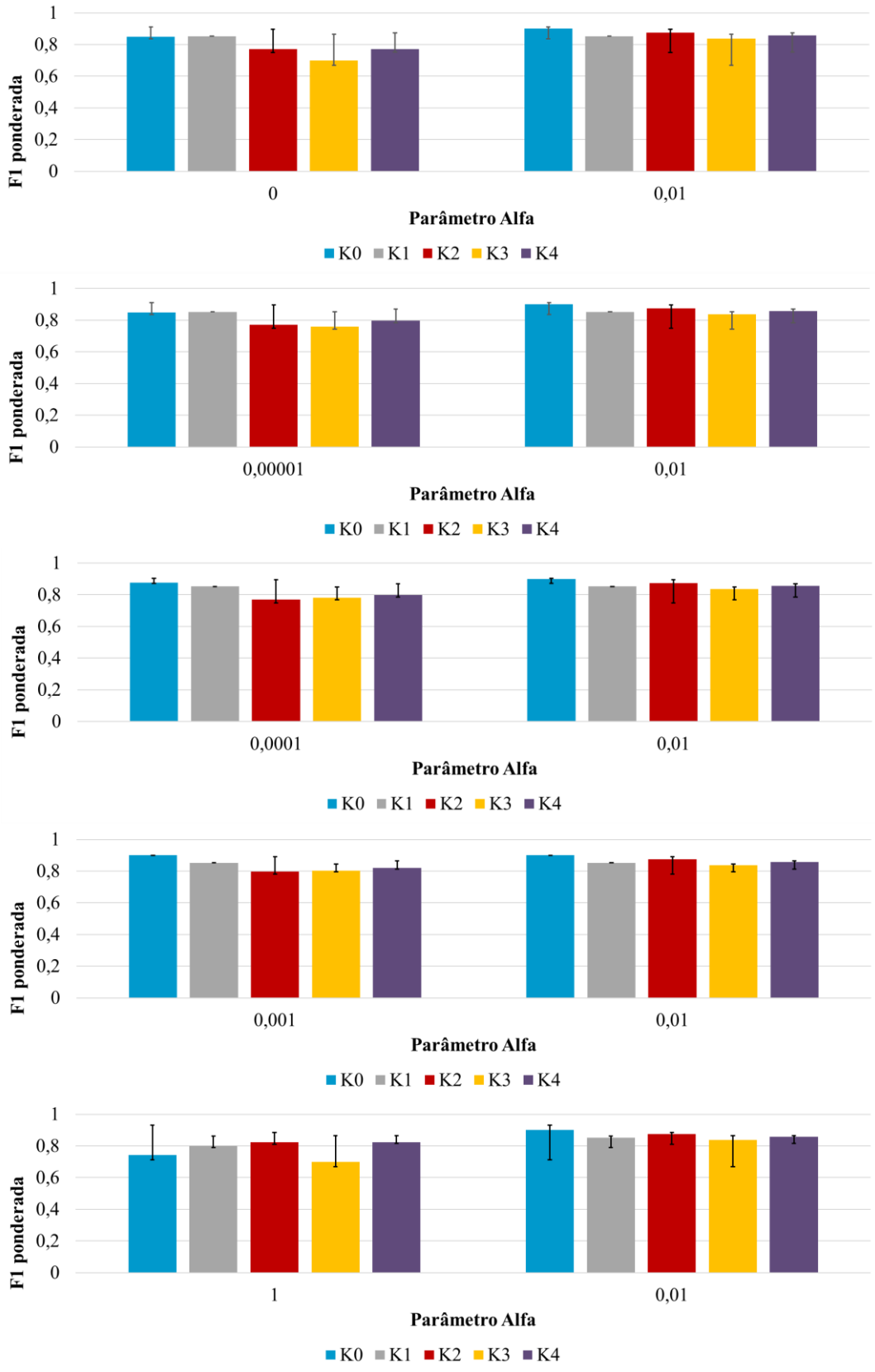
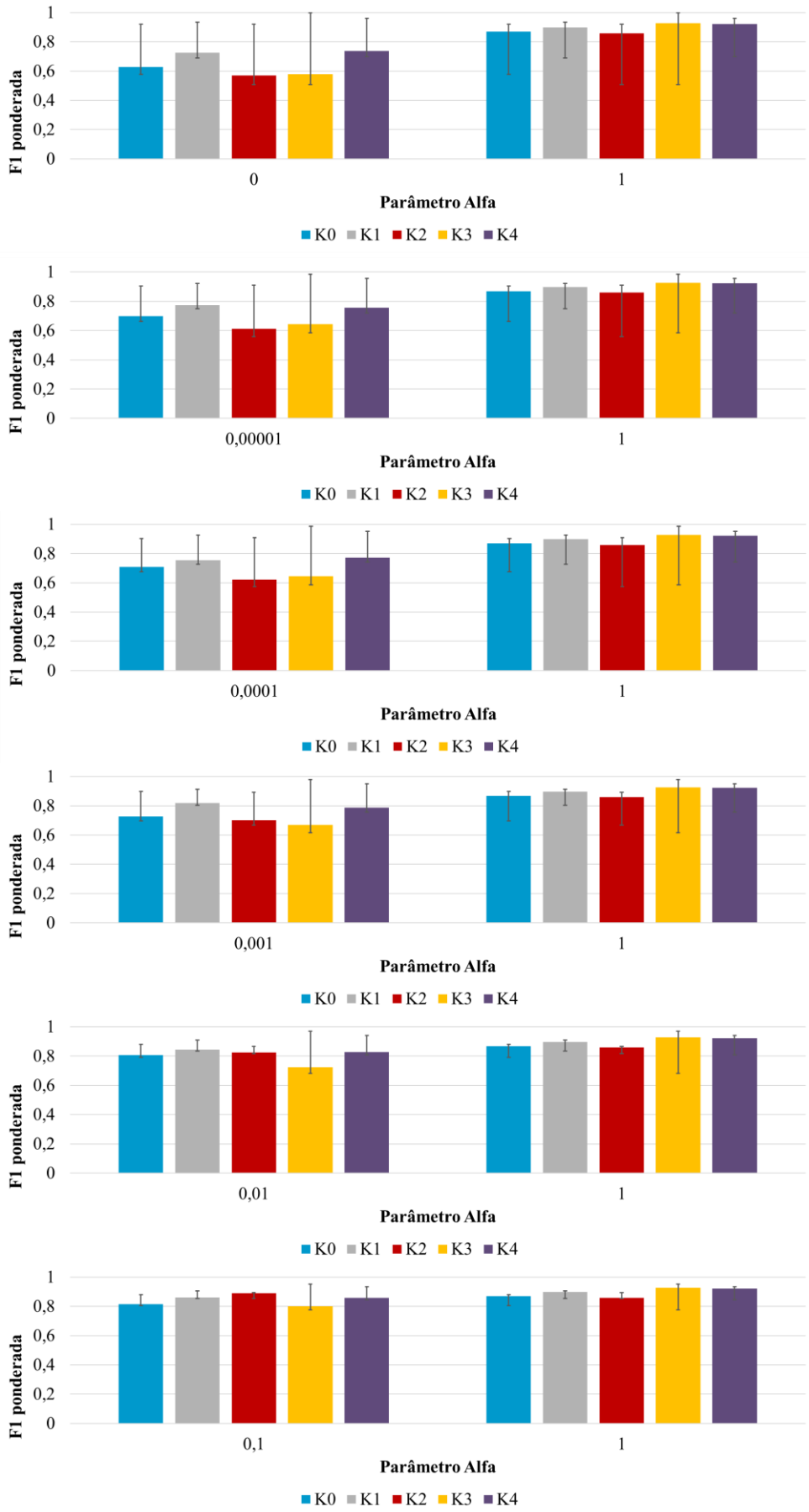


Figura 16 – Resultado da validação cruzada na etapa de calibração do modelo CNB



As Figuras 16 e 17 apresentam os parâmetros alfa testados em comparação ao melhor parâmetro determinado para cada modelo. Além disso a barra de erros apresentada nas figuras representa o desvio médio do desempenho dos modelos em relação a cada rodada para todos os alfas testados.

A partir da Figura 15 é possível observar que para o modelo MNB o parâmetro alfa igual a 0,01 supera a métrica F1 ponderada em relação a todas as outras métricas testadas em todas as rodadas realizadas. Assim, o modelo MNB será ajustado com o parâmetro alfa igual a 0,01.

Nos resultados da calibração do modelo CNB (Figura 16) nota-se que o parâmetro alfa igual a 1,0 (o mesmo alfa utilizado no CNB na versão *default*) supera a métrica F1 ponderada em relação a todas as outras métricas testadas em todas as rodadas realizadas. Desta forma, o melhor modelo CNB já é o modelo CNB é o próprio modelo *default*, com alfa igual a 1,0.

Na segunda etapa da fase de classificação, os modelos ajustados com os parâmetros otimizados são novamente rodados com a validação cruzada *k-fold* com as mesmas amostras utilizadas na calibração. O processo de validação foi realizado novamente a fim de verificar a robustez dos modelos calibrados (MNB e CNB), e comparar os desempenhos dos modelos ajustados em relação as suas versões *default*.

A Figura 17 apresenta os *boxplots* que foram gerados a partir da acurácia, precisão, recall e métrica F ponderadas em relação a distribuição dos registros em relação a cada classe, e obtidas da validação cruzada realizada na etapa de avaliação de ajuste do modelo. Observa-se a partir da análise realizada na Figura 16, que o modelo CNB ajustado é o próprio modelo com parâmetro default. Assim, a Figura 17 apresentará um único resultado para o modelo CNB.

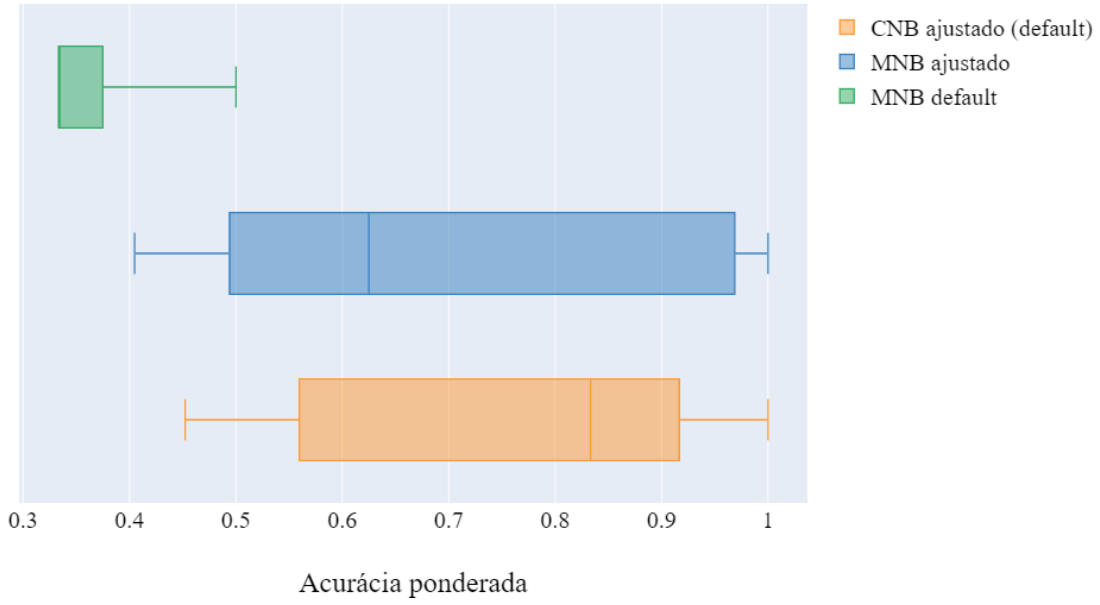
Ademais, os valores médios e o desvio padrão de cada métrica considerando todas as rodadas na etapa de avaliação dos modelos são apresentados na Figura 18.

A Figura 17 indica como as métricas escolhidas variam em cada *fold* da validação cruzada para ambos os classificadores avaliados. Analisando a Figura 18, é possível identificar a variabilidade dos resultados dessas métricas no procedimento de validação cruzada, em todos os modelos apresentados na figura. Tal fato é um indicativo que uma certa dependência das amostras selecionadas para cada *fold*. Este comportamento observado é esperado, considerando o pequeno conjunto de dados utilizado para testar os resultados.

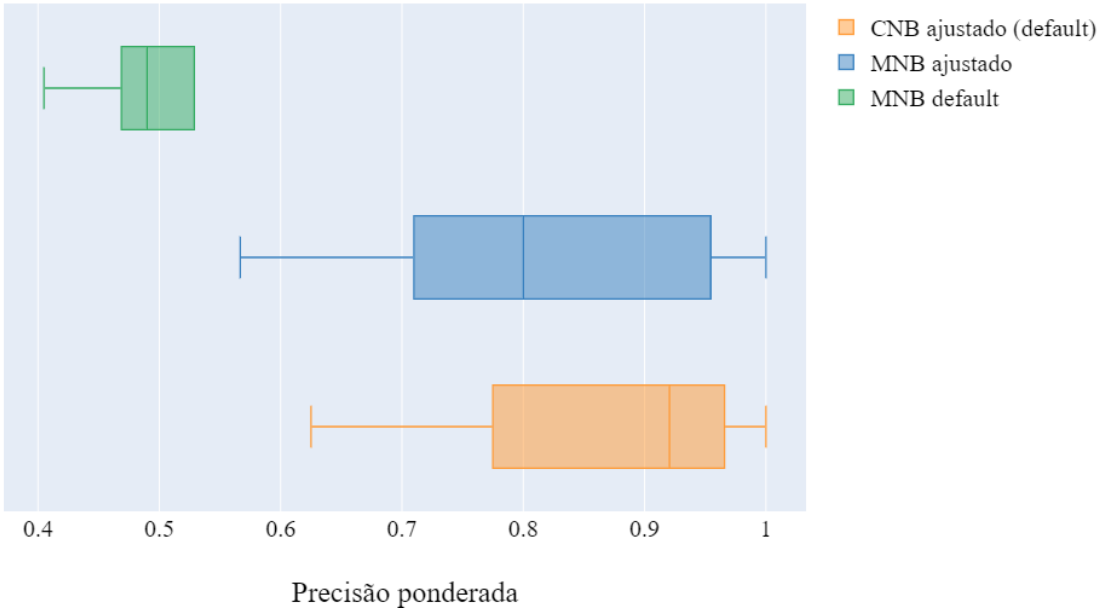
Também é possível inferir a partir dos resultados apresentados na Figura 17, o modelo MNB ajustado em relação a sua versão com parâmetro *default* da biblioteca *Scikit Learning* (PEDREGOSA *et al.*, 2011) teve desempenho superior em relação a todas as métricas de avaliação. No entanto, mesmo a versão ajustada do MNB não conseguiu superar o desempenho do modelo CMN.

Figura 17 – Resultados do ajuste do modelo obtidos por validação cruzada do conjunto de teste para CNB e MNB com parâmetros defaults

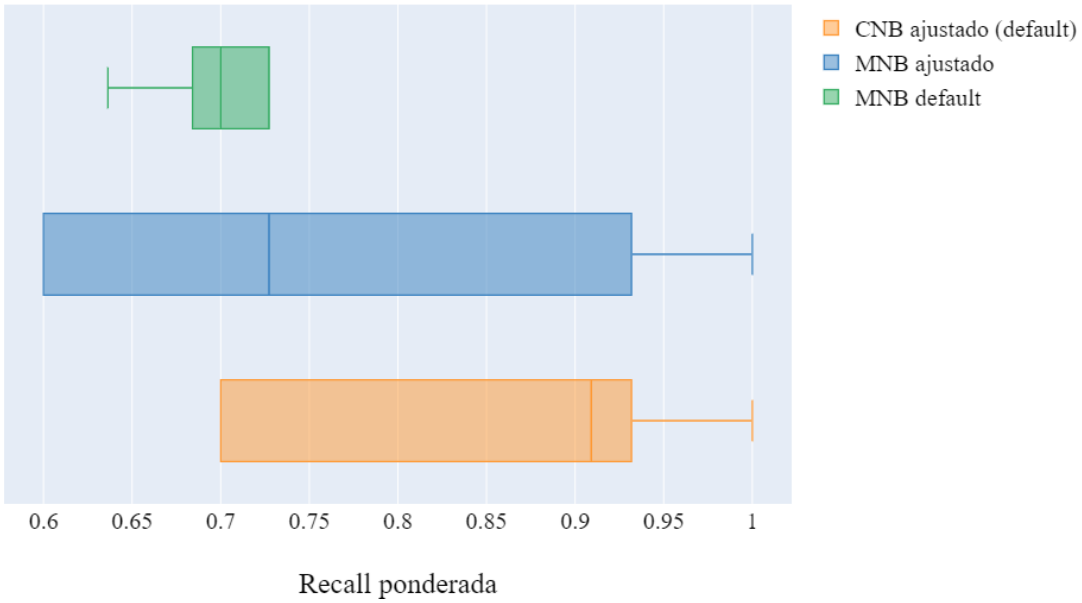
(a) Acurácia balanceada



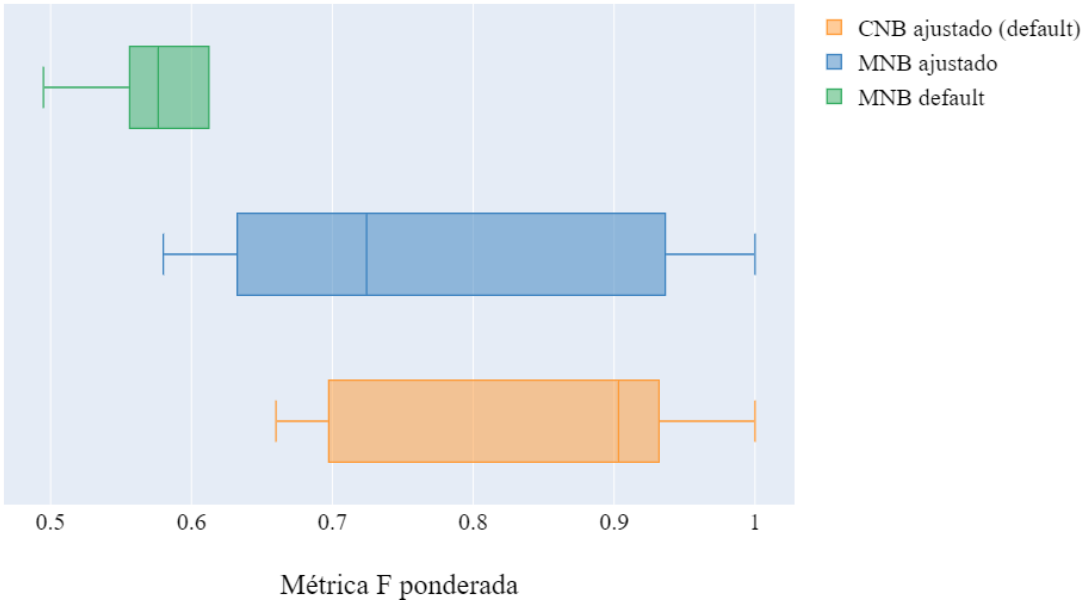
(b) Precisão ponderada



(c) *Recall* Ponderada



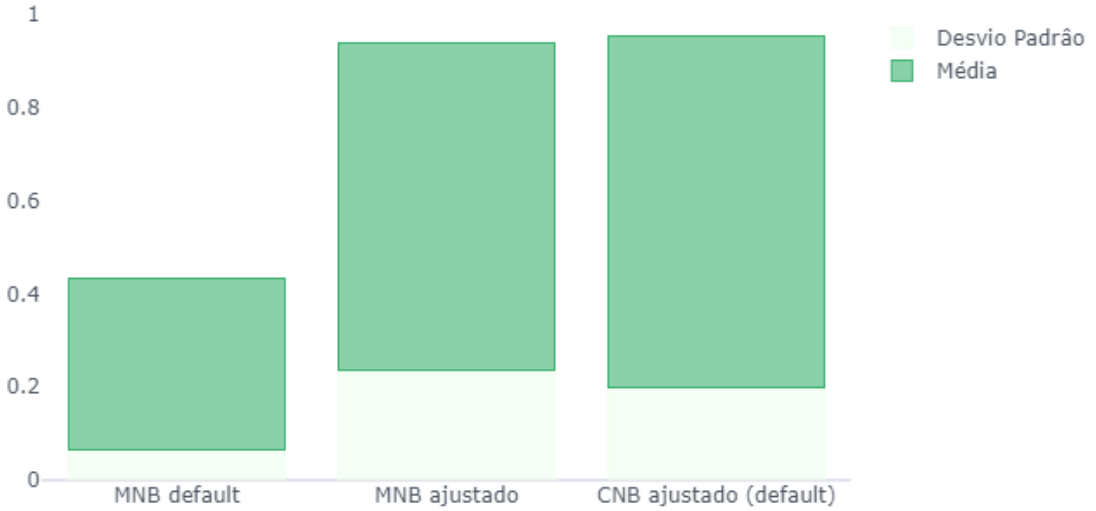
(d) métrica F ponderada



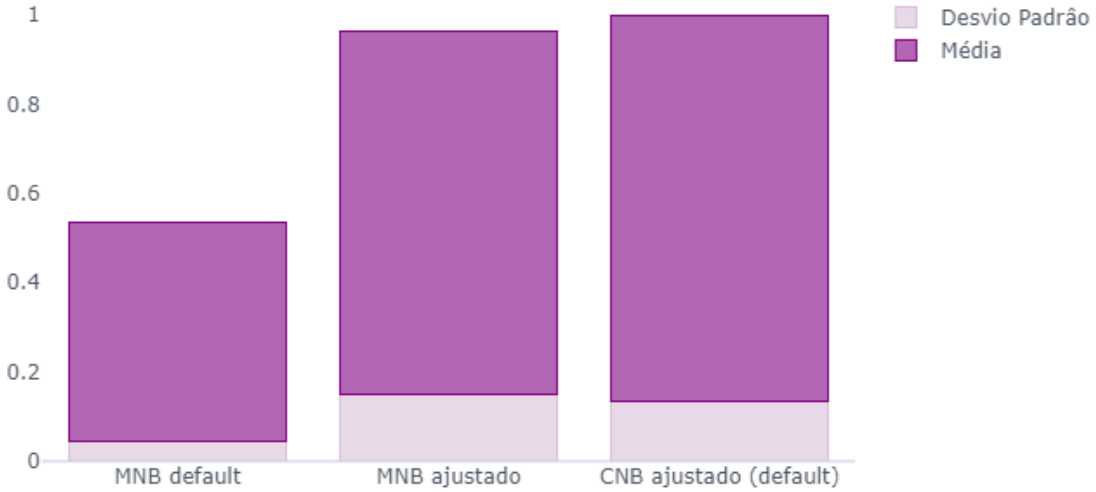
A partir da análise dos resultados apresentados na Figura 18, nota-se que o modelo MNB *default* obteve as menores médias em relação a todas as métricas avaliadas, mas também o menor desvio padrão. Comparando o desempenho dos modelos CNB ajustado (*default*) e o modelo ajustado do MNB, as médias obtidas foram bem próximas. O modelo CNB obteve médias superiores (aproximadamente 80%) em todas as métricas, exceto a métrica de acurácia que foi 75,4%. Enquanto o modelo MNB ajustado teve médias em torno de 77% para todas as métricas, exceto a acurácia (70%). No entanto, o modelo MNB ajustado teve maior desvio em relação aos 3 modelos comparados.

Figura 18 – Resultados do ajuste do modelo obtidos por validação cruzada do conjunto de teste para CNB e MNB com parâmetros defaults

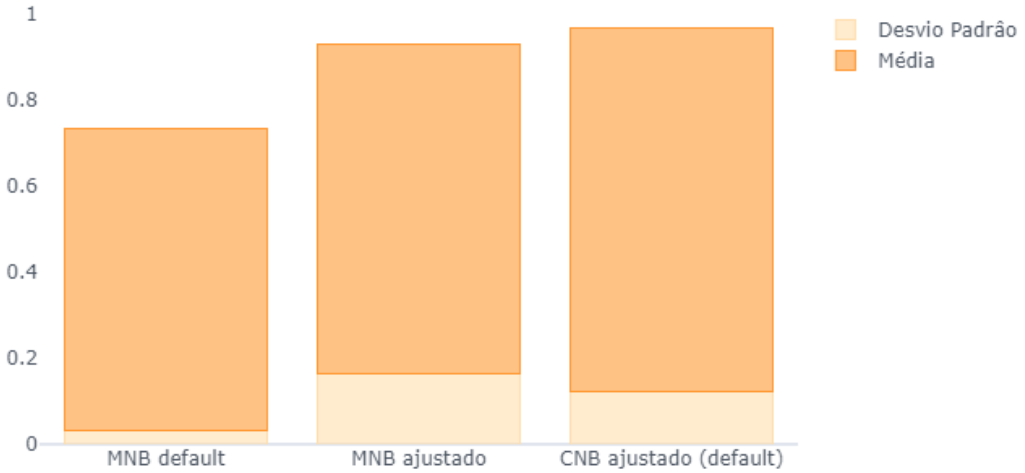
(a) Acurácia balanceada



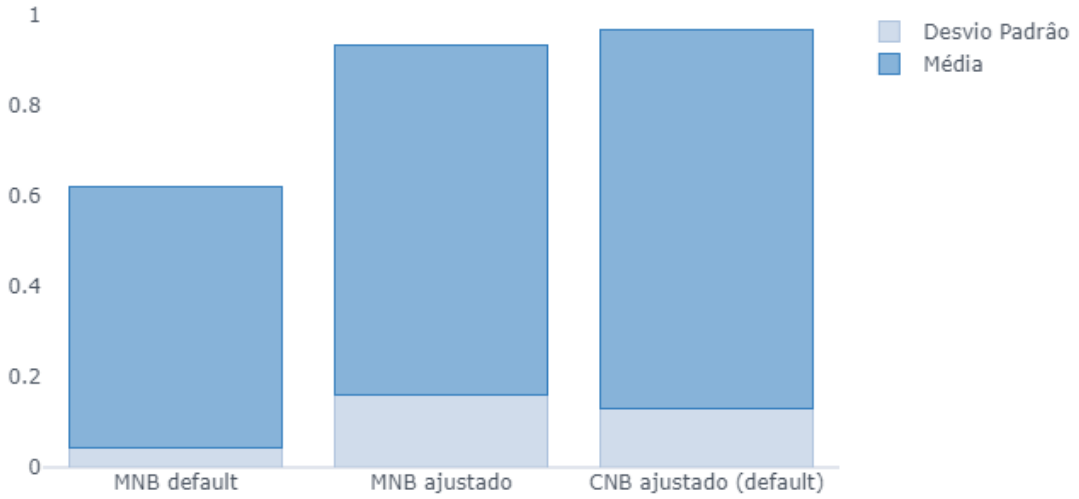
(b) Precisão ponderada



(c) *Recall* Ponderada



(d) métrica F ponderada



A partir desta análise das Figuras 18 e 19, observa-se que em comparação aos dois modelos de ML selecionados nesta metodologia, o modelo CNB foi considerado o melhor modelo para resolver o problema apresentado nesta pesquisa.

As Figuras 20 e 21 apresentam a matriz confusão com a média dos resultados de predição com base na validação cruzada do modelos ajustados MNB e CNB.

Figura 19 – Resultado médio da predição das classes segundo o modelo MNB ajustado

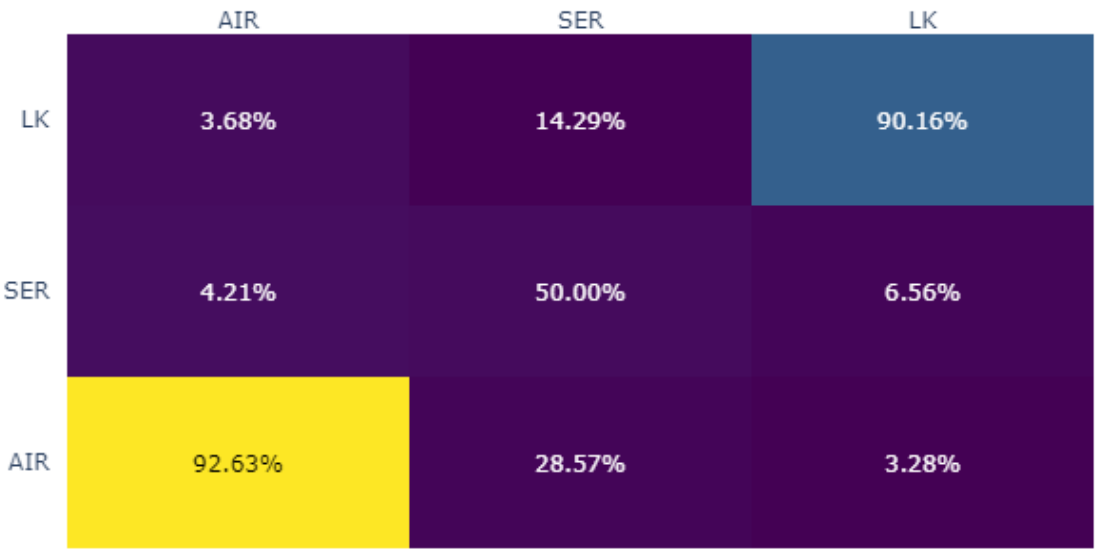
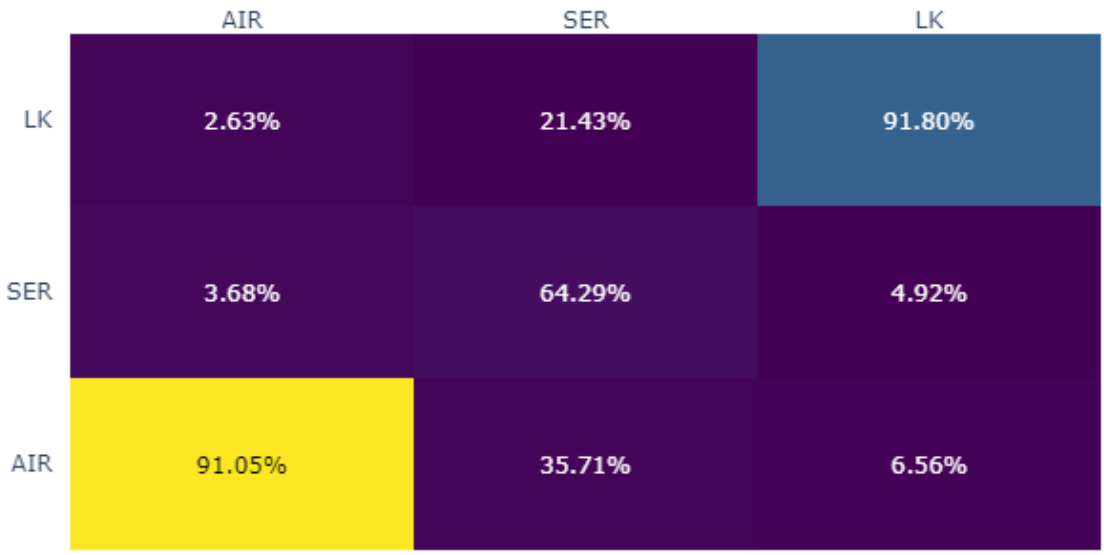


Figura 20 – Resultado médio da predição das classes segundo o modelo CNB ajustado (*default*)



De um total de 53 registros no conjunto de teste da validação, 70,24% foram classificados corretamente pelo método MNB ajustado com 92,63% de acertos da classe leitura anormal do instrumento (AIR), 90,16% da classe vazamento e 50,00% da classe pequenos problemas em serviço (SER).

Em relação ao método CNB ajustado, o método consegui classificar corretamente as classes 75,40% das vezes, com 91,05% de acertos da classe leitura anormal do instrumento (AIR), 64,29% da classe vazamento e 91,80% da classe pequenos problemas em serviço (SER).

Com base nos resultados apresentados nesta seção, é possível concluir que os resultados obtidos na classificação dos registros de manutenção foram satisfatórios mesmo com algumas limitações da abordagem proposta. Uma delas foi a variabilidade das métricas na validação cruzada, devido ao pequeno conjunto de dados utilizados testar os resultados. Outra limitação observada foi a capacidade dos modelos em prever as classes do modo de falha pequenos problemas em serviço. No entanto, é possível observar que esta abordagem possui grande potencial de aplicação.

8 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo encerra esta dissertação apresentando as principais conclusões obtidas do desenvolvimento desta pesquisa. São também apresentadas as contribuições deste estudo e possíveis estratégias para aperfeiçoamento do modelo proposto em trabalhos futuros.

8.1 Considerações Finais

Dados de confiabilidade fornecem informações valiosas sobre a natureza e frequência de ativos e possuem grande importância para a atividade de produção de petróleo, visto que viabilizam a otimização dos processos de manutenção e produção, reduzindo de perdas e custos relacionados as atividades de manutenção, tempo de inatividade não programado, avarias inesperadas, dentre outros. Devido a sua relevância, quanto maior o volume e a qualidade dos dados catalogados, mais valiosas são as informações fornecidas sobre a natureza e frequência de ativos e mais assertivas poderão ser as decisões sobre a operação e manutenção de um equipamento ou sistema.

Neste contexto, o estudo de técnicas de mineração e processamento de linguagem natural que buscam classificar automaticamente registros de manutenção em relação à natureza da falha, principalmente, quanto a textos escritos na língua portuguesa é ainda incipiente e precisa ser aprimorado.

Usualmente, a tarefa de catalogação é realizada manualmente por um especialista, que avalia cada registro com o objetivo de identificar o respectivo modo de falha. Além de trabalhosa, a exploração manual é realizada de forma subjetiva, considerando apenas o julgamento de especialistas, o que pode induzir uma análise tendenciosa sujeita a muitos erros, afetados não apenas por aspectos técnicos, mas também por outros não técnicos (BLANCO-M *et al.*, 2019).

Assim, em consonância com o processo de digitalização da indústria O&G e motivada pela crescente inclinação do setor por sistemas automatizados, o objetivo desta pesquisa foi investigar as possibilidades de automatização da tarefa de catalogação de dados de confiabilidade e, especificamente, explorar métodos tradicionais no desenvolvimento de um modelo para a mineração de registros de manutenção escritos em português e a sua classificação automática em relação aos modos de falha de equipamentos.

A metodologia proposta objetivou reduzir a subjetividade e aumentar a velocidade do processo de catalogação de dados de confiabilidade. Ademais, o estudo tem caráter inovador quando considerada a aplicação das técnicas de mineração de textos, NLP e ML em registros de manutenção escritos na língua portuguesa.

A metodologia foi aplicada a um estudo de caso com dados de texto de manutenção de turbinas a gás em atividades de E&P de petróleo de uma empresa brasileira. Para aplicação da metodologia, um algoritmo foi desenvolvido em Python, utilizando as bibliotecas NLTK e *Scikit Learn*.

Na etapa de pré-processamento, é notória a importância da identificação e remoção dos ruídos, principalmente as informações que não se referem de fato a falha, como a descrição dos procedimentos necessários para a manutenção em si. Da análise realizada, verificou-se que mesmo textos pré-processados com tamanho pequeno podem ser corretamente classificados, e que alguns textos longos não foram capazes de fornecer boas informações para uma correta classificação. Isto é, não há uma relação que comprova que o tamanho do texto pode influenciar positivamente na classificação, mas sim a sua qualidade ou o quão relevantes as palavras que representam os documentos de fato são relevantes para identificar o modo de falha. Consequentemente, há uma a necessidade do refinamento do conjunto dos dados brutos para obtenção de resultados mais confiáveis. Ademais, observa-se que a qualidade dos resultados das etapas seguintes dependeu muito da etapa de pré-processamento.

Na etapa de classificação, nota-se que a abordagem metodológica é promissora. A avaliação de ambos algoritmos de aprendizado de máquina escolhidos para a classificação de texto obtiveram resultados aceitáveis. Os resultados obtidos mostram que o classificador CMB obteve melhores resultados quando comparado ao MNB, o que era esperado uma vez que esse algoritmo é particularmente adequado para conjuntos de dados desequilibrados, como é o caso deste problema. A comparação do desempenho de ambos os algoritmos de ML corrobora com a hipótese de RENNIE *et al.* (2003) que CNB possui melhor performance em relação ao modelo multinomial clássico (MNB) para tarefas de classificação de texto. No entanto, um problema detectado na etapa de classificação é que os valores das métricas utilizadas, tanto na fase de calibração quanto na etapa de avaliação do ajuste do modelo, apresentaram resultados instáveis durante a validação cruzada para ambos os classificadores. A variabilidade das medidas encontradas pode ser trabalhada com a investigação de outras estratégias para seleção dos atributos e melhor ajuste dos parâmetros dos modelos.

A escassez de dados catalogados disponíveis dificulta a realização de estudos mais robustos na área de confiabilidade e nos processos de gestão da manutenção. Este fator também limita uma validação mais ampla da metodologia proposta, empregando dados reais de registros de manutenção envolvendo todos os modos de falha possíveis de um equipamento. Ademais, a falta de padronização dos registros é uma situação recorrente, e que atrapalha tanto o aprendizado supervisionado realizado nesta pesquisa quanto a abordagem não supervisionado, como realizado por CHEN; NAYAK (2007). Caso o conjunto de dados disponíveis fosse maior e respeitasse os padrões previstos na norma NBR ISO 14224, a catalogação automática seria facilitada. Desta forma seria possível haver o mínimo de alterações nas métricas avaliadas na validação cruzada e obter resultados melhores na classificação em geral.

Apesar dos fatores anteriormente citados pode-se concluir que os resultados na classificação foram bastante satisfatórios quando considerada a complexidade do problema estudado e o número limitado de registros avaliados.

Finalmente, a seguir são elencadas as principais conclusões obtidas nesta pesquisa, corroboradas pelos resultados da aplicação da metodologia proposta no estudo de caso proposto.

- Mesmo uma metodologia com abordagem simplista, utilizando técnicas clássicas de mineração de textos e aprendizado de máquina supervisionado, tem grande potencial de ser aplicada à catalogação de dados modos de falha. E, possivelmente, auxiliar na catalogação de outros dados qualitativos de R&M, como por exemplo a categorização do mecanismo de falha e dos itens manuteníveis de um equipamento.
- O emprego da metodologia para classificar textos não estruturados é viável, ainda que testado em um número limitado de registros ou mesmo em situações que o conteúdo dos registros tem baixa qualidade ou pouco conteúdo relevante. A implementação do procedimento metodológico permitiu classificar automaticamente os documentos e obter um nível moderado de precisão.
- Classificar completamente os modos de falha de equipamentos são tarefas extremamente desafiadoras, principalmente para uma quantidade limitada de recursos (especialistas, tempo necessário para realização da tarefa, entre outros). Observou-se que alguns dos principais desafios da tarefa são a alta frequência de dados incompletos; erros de ortográficos/digitação; abreviação de termos; o uso de vocábulos escritos também em inglês; e, a presença de frases que acrescentam pouco significado para identificação do modo de falha nestes relatórios, mas que poderiam ser utilizadas para obtenção de outros

tipos de dados de confiabilidade, como a identificação do mecanismo de falha, do item manutenível, recursos homem-hora, dentre outros.

- A metodologia proposta pode beneficiar a tarefa de catalogação, reduzindo a subjetividade da análise e contribuindo para a velocidade do processo de catalogação. Dessa forma, mais dados podem ser catalogados e os possíveis problemas nos ativos podem ser melhor estudados. Adicionalmente, a classificação imparcial dos registros reduz significativamente o esforço necessário na tarefa de validação final realizada pelo especialista.
- A enorme lacuna de trabalhos que abordam especificamente o problema apresentado utilizando soluções de mineração de textos e aprendizado supervisionado, mesmo com os avanços da digitalização e um cenário de crescimento de aplicação das técnicas de NLP e aprendizado profundo, corrobora para a hipótese que documentos textuais contendo valiosas informações podem estar sendo subutilizados pela indústria. Tal fato destaca ainda mais a importância de trabalhos de pesquisa nesta área e demonstram o grande impacto positivo que o uso adequado destes dados pode trazer para a indústria.
- Como apresentado por SALO *et al.* (2019) esta pesquisa também corrobora para justificar o enorme potencial das técnicas de inteligência artificial, especialmente para instalações mais antigas onde há uma a grande quantidade de dados históricos não estruturados, mas com enorme valor para compreensão dos eventos de falha.

8.2 Contribuições do trabalho

A principal contribuição obtida nesta pesquisa foi propor e implementar um método que possibilita a catalogação automática de registro em modos de falha, envolvendo o uso de técnicas de mineração de textos e aprendizado de máquina supervisionado.

Embora o método proposto se baseie nos modelos clássicos de representação de documentos (vetorial) em uma BOW e utilizando as técnicas TF-IDF e testando apenas dois tipos de classificadores, um aspecto relevante é o fato desta ser a primeira abordagem conhecida para a mineração de textos de manutenção escritos em português que tenta classificar os registros automaticamente quanto aos modos de falha utilizando técnicas de aprendizado supervisionado.

Devido a utilização das bibliotecas já existentes no Python para aplicações de *Data Science*, a implementação da metodologia proposta é facilitada, sendo esta outra interessante contribuição da pesquisa.

Igualmente, outro significante aspecto deste trabalho é que mesmo se tratando de um problema muito difícil e utilizando modelos simplificados, é possível obter resultados positivos e tão satisfatórios quanto os apresentados no Capítulo 7. Isto corrobora para a hipótese que há sim valor em obter informação a partir de textos não estruturados e em formato livre. Principalmente, em se tratando de dados tão valiosos quanto os de confiabilidade e quando há tão poucos recursos necessários dedicados exclusivamente para a atividade fim.

Um ponto limitante do trabalho é a falta de dados disponíveis para treinar todas as classes, o que dificultou avaliar o desempenho dos classificadores quando considerado todos os modos de falha possíveis de um único tipo de equipamento. Observa-se que quanto mais classes forem treinadas mais difícil o problema se torna e mais propenso os classificadores estarão ao erro, podendo não obter tão bons resultados como os do estudo de caso apresentado. Adicionalmente, outro fator limitante é a ausência ou baixa uniformização dos textos, que caso fosse realizada com base nas normas vigentes, auxiliaria a reduzir os erros de classificação.

Considerando tudo o que foi exposto neste subitem e ao longo desta dissertação, mesmo com as limitações e dificuldades do problema, o desenvolvimento de um modelo capaz de classificar os registros com razoáveis resultados já representa um grande avanço no processo de automatização da tarefa de catalogação de dados de confiabilidade, sobretudo para o setor de O&G.

8.3 Sugestões para Pesquisas futuras

Ainda há muito a desenvolver e ampliar nessa metodologia, especialmente em etapas críticas que têm um impacto significativo no resultado. Usando a abordagem atual, é possível continuar realizando experimentos para refinar o modelo, identificar erros e testar em um volume muito maior de dados. Algo interessante a se fazer no futuro é avaliar e comparar diferentes abordagens em cada etapa da metodologia.

Uma estratégia possível para lidar com a limitação de dados é utilizar também os próprios exemplos de descrição da norma NBR ISO 14224 para treinar o algoritmo. Adicionalmente, outra sugestão para pesquisa futura é desenvolver uma taxonomia ou

sistema de registro que seja mais propenso a ter seus dados aproveitados e interpretados por um modelo de previsão, dado que o potencial destes dados é enorme e que caso eles fossem melhor registrados, facilitaria todo o processo de catalogação.

Outra abordagem a ser avaliada, é melhorar a etapa de validação dos resultados da classificação cruzando informações do banco de dados monitoramento do equipamento (como os dados PI) para tentar validar automaticamente algumas classes categorizadas. Para algumas determinadas classes é possível verificar indicativos que o equipamento falhou. Por exemplo, em casos que o classificador categoriza um registro como "vibração" é possível investigar se, durante o período de percepção da falha, os valores de vibração estavam anormais.

As pesquisas futuras poderão ser concentradas na melhoria da etapa de pré-processamento, como o uso de um *Thesaurus* – um dicionário de sinônimos de palavras, siglas, acrônimos e abreviação de palavras, para resolver problemas de vocabulário (homonímia e sinonímia). Além disso, poderão ser testadas outras abordagens para a seleção de atributos, que não a puramente estatística como o TF-IDF, considerando também a informação semântica e gramática das palavras.

Outras possibilidades são testar outros modelos mais sofisticados para representação de documentos (como uma rede neural para representação de texto) e outros algoritmos de aprendizado de máquina como árvore de decisão, floresta aleatória, *K nearest neighbor* ou redes neurais para a classificação dos registros.

9 REFERÊNCIAS BIBLIOGRÁFICAS

7GRAUS © 2011 - 2020. **Sinônimos.com.br - dicionário de sinônimos online**. Disponível em: <https://www.sinonimos.com.br/busca.php?q=fluido>. Acesso em: 4 abr. 2020.

ABNT. **ABNT NBR ISO 14224:2011 Indústrias de petróleo e gás natural — Coleta e intercâmbio de dados de confiabilidade e manutenção para equipamentos**. . Rio de Janeiro: Associação Brasileira de Normas Técnicas. Disponível em: <https://www.abntcolecao.com.br/ufrj/>. , 2011

AGGARWAL, Charu C. **Data Classification: Algorithms and Applications**. First Edit ed. [S.l.]: Chapman & Hall/CRC, 2014.

AHMAD, Rosmaini; KAMARUDDIN, Shahrul. An overview of time-based and condition-based maintenance in industrial application. **Computers and Industrial Engineering**, v. 63, p. 135–149, 2012. Disponível em: <http://dx.doi.org/10.1016/j.cie.2012.02.002>.

AKHMEDJANOV, Farit M. **Reliability databases: state-of-the-art and perspectives**. **Report R-1235, Risö National Laboratory, Roskilde, Denmark**. [S.l: s.n.], 2001. Disponível em: <http://www.risoe.dk/rispubl/VEA/veapdf/ris-r-1235.pdf>.

AL-ALWANI, Mustafa A. et al. **From Data Collection to Data Analytics: How to Successfully Extract Useful Information from Big Data in the Oil & Gas Industry?** . Bali: [s.n.], 2019.

ALVARES, Reinaldo Viana; GARCIA, Ana Cristina Bicharra; FERRAZ, Inhaúma. **STEMBR: A stemming algorithm for the Brazilian Portuguese language**. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2005)**. [S.l: s.n.], 2005.

AN, Dawn; KIM, Nam H; CHOI, Joo-ho. Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. **Reliability Engineering and System Safety**, v. 133, p. 223–236, 2015. Disponível em: <http://dx.doi.org/10.1016/j.ress.2014.09.014>.

ANTONIAK, M et al. **Natural Language Processing Techniques on Oil and Gas Drilling Data**. . SPE Intelligent Energy International Conference and Exhibition held in Aberdeen, United Kingdom, 6-8 September 2016: [s.n.], 2016.

ARIF-UZ-ZAMAN, Kazi et al. Extracting failure time data from industrial maintenance records using text mining. **Advanced Engineering Informatics**, v. 33, p. 388–396, 2017. Disponível em: <http://dx.doi.org/10.1016/j.aei.2016.11.004>.

ARMITAGE, Chris J. Streamlined RCM Process for Drilling Equipment. 2003.

ARUMUGAM, Sethupathi et al. **Revealing Patterns within the Drilling Reports Using Text Mining Techniques for Efficient Knowledge Management**. . SPE Eastern Regional Meeting held in Canton, Ohio, USA, 13-15 September 2016: [s.n.], 2016.

ARUMUGAM, Sethupathi; RAJAN, Shebi; GUPTA, Sanjay. **Augmented Text Mining for Daily Drilling Reports using Topic Modeling and Ontology**. . SPE Western Regional Meeting held in Bakersfield, California, USA, 23 April 2017. This: [s.n.], 2017.

ASFOOR, Hasan; KASKAS, Walid; ARAMCO, Saudi. **Harnessing the Power of Natural Language Processing and Fuzzy Theory to Improve Oil and Gas Data Management Efficiency**. . SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition held in Bali, Indonesia, 29-31 October 2019: [s.n.], 2019.

AZADEH, A; EBRAHIMIPOUR, V; BAVAR, P. A fuzzy inference system for pump failure diagnosis to improve maintenance process : The case of a petrochemical industry. **Expert Systems With Applications**, v. 37, n. 1, p. 627–639, 2010. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2009.06.018>.

BENDELL, Tony. An Overview of Collection, Analysis, and Application of Reliability Data in the Process Industries. **IEEE Transactions on Reliability**, v. 37, n. 2, p. 132–137, 1988.

BEVILACQUA, Maurizio; BRAGLIA, Marcello; MONTANARI, Roberto. The classification and regression tree approach to pump failure rate analysis. **Reliability Engineering and System Safety**, v. 79, p. 59–67, 2003.

BIRD, Steven; LOPER, Edward; KLEIN, Ewan. **Natural Language Processing with Python**. [S.l.]: O'Reilly Media Inc., 2009.

BLANCO-M, Alejandro et al. A Text-Mining Approach to Assess the Failure Service History. **Energies**, v. 12, n. 10, p. 1–20, 2019.

BORTOLINI, Rafaela; FORCADA, Núria. Analysis of building maintenance requests using a text mining approach : building services evaluation services evaluation. **Building Research & Information**, v. 0, n. 0, p. 1–11, 2019. Disponível em:

<https://doi.org/10.1080/09613218.2019.1609291>.

BRAGLIA, Marcello et al. Data classification and MTBF prediction with a multivariate analysis approach. **Reliability Engineering and System Safety**, v. 97, p. 27–35, 2012. Disponível em: <http://dx.doi.org/10.1016/j.ress.2011.09.010>.

BRYNJOLFSSON, Erik; MCAFEE, Andrew. Big Data : The Management Review. **Harvard Business Review**, n. October, p. 1–12, 2012. Disponível em: <http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf>.

CASTIÑEIRA, David et al. Machine Learning and Natural Language Processing for Automated Analysis of Drilling and Completion Data. 2018.

CHEN, Lin; NAYAK, Richi. A case study of failure mode analysis with text mining methods. 2007, Gold Coast: Australian Computer Society, 2007. p. 49–60.

CHO, J J; RICE, K D; PHILLIPS, R G. Improving service reliability for drilling and evaluation operations using an optimized RCM strategy. **Proceedings of the Annual Offshore Technology Conference**, v. 1, p. 222–229, 2013. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84897426180&partnerID=40&md5=a3abcec78e76537486d5247e67b0c922>.

CHOI, Inhwan et al. Production availability for new subsea production systems with seabed storage tanks. **The 2013 World Congress on Advances in Structural Engineering and Mechanics (ASEM13)**, p. 662–672, 2013.

COLOMBO, Danilo et al. **Discovering Patterns within the Drilling Reports using Artificial Intelligence for Operation Monitoring**. . Offshore Technology Conference Brasil held in Rio de Janeiro, Brazil, 29–31 October 2019: [s.n.], 2019.

CORVARO, Francesco et al. Reliability, Availability, Maintainability (RAM) study, on reciprocating compressors API 618. **Petroleum**, v. 3, n. 2, p. 266–272, 2017. Disponível em: <http://dx.doi.org/10.1016/j.petlm.2016.09.002>.

DANTAS, Maria Aldilene et al. Modelo de regressão Weibull para estudar dados de falha de equipamentos de sub-superfície em poços petrolíferos. **Produção**, v. 20, n. 1, p. 127–134, 2010.

DEVANEY, Mark et al. **Preventing failures by mining maintenance logs with case-based reasoning**. . 59th Meeting of the Society for Machinery Failure Prevention Technology (MFPT-59), 2005 PREVENTING: [s.n.], 2005.

DHAMODHARAVADHANI, S; GOWRI, R; RATHIPRIYA, R. Unlock Different V's of Big Data Analytics. **International Journal of Computer Sciences and Engineering**, v. 6, n. 4, p. 183–190, 2018.

EBECKEN, Nelson Francisco Favilla; LOPES, Maria Celia Santos; COSTA, Myrian Cristina de Aragão. Mineração de textos. **Sist. Intel. Fundam. e Apl.** Barueri, São Paulo, Brasil: Editora Manola Ltda., 2005. .

EDWARDS, Brett; ZATORSKY, Michael; NAYAK, Richi. **Clustering and Classification of Maintenance Logs using Text Data Mining**. . Australasian Data Mining Conference 2008, November 2008, Adelaide, Australia: [s.n.], 2008.

EISINGER, Siegfried; CLAVÉ, Nicolas. **Offshore & onshore reliability data (OREDA®) collection – OREDA JIP status**. . Houston: Siegfried Eisinger, DNVGL, Norway, OREDA JIP Project Manager Nicolas Clavé, Total, France, OREDA JIP Steering Committee Chairman. , 2018

EXPROSOFT. **WellMaster RMS (Reliability Management System)**. Disponível em: <https://www.exprosoft.com/products/wellmaster-rms/>.

_____. **WellMaster RMS User Group Meeting**. . [S.l: s.n.]. , 2018

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, GREGORY SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An Overview. **Adv. Knowl. Discov. Data Min.** [S.l.]: American Association for Artificial Intelligence, 1996. p. 1–34.

FELDMAN, Ronen; DAGAN, Ido. **Knowledge Discovery in Textual Databases (KDT)**. **International Conference on Knowledge Discovery and Data Mining (KDD)**. [S.l: s.n.], 1995. Disponível em: <http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf>.

FELDMAN, Ronen; SANGER, James. **Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. New York, New York, USA: Cambridge University Press, 2007. Disponível em: <file:///C:/Users/User/Downloads/fvm939e.pdf>.

GASPERIN, Caroline Varaschin; LIMA, Vera Lúcia Strube De. **Fundamentos do processamento estatístico da linguagem natural**. . [S.l: s.n.], 2001. Disponível em: <http://www3.pucrs.br/pucrs/files/uni/poa/facin/pos/relatoriostec/tr021.pdf>.

GHAZVINIAN, A.; NEJATI, H. R.; SAEMI, M. Reliability and uncertainty of prediction of dynamic elastic constants in reservoir rock. **Journal of Canadian Petroleum**

Technology, v. 51, n. 3, p. 198–204, 2012.

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, p. 1796–1801, 2018.

GUNAY, H Burak; SHEN, Weiming; YANG, Chunsheng. Text-mining building maintenance work orders for component fault frequency. v. 3218, 2019.

GUO, Lijie et al. Criticality evaluation of petrochemical equipment based on fuzzy comprehensive evaluation and a BP neural network. **Journal of Loss Prevention in the Process Industries**, v. 22, p. 469–476, 2009. Disponível em: <http://dx.doi.org/10.1016/j.jlp.2009.03.003>.

HAMEED, Z.; VATN, J.; HEGGSET, J. Challenges in the reliability and maintainability data collection for offshore wind turbines. **Renewable Energy**, v. 36, n. 8, p. 2154–2165, 2011. Disponível em: <http://dx.doi.org/10.1016/j.renene.2011.01.008>.

HAN, Jiawei et al. **Data Mining Concepts and Techniques**. [S.l: s.n.], 2012.

HOFFIMANN, Julio et al. **Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing**. . SPE Annual Technical Conference and Exhibition held in Dallas, Texas, 24-26 September 2018: [s.n.], 2018.

HOTHO, Andreas; ANDREAS, Nürnberger; PAASS, Gerhard. A Brief Survey of Text Mining. p. 1–37, 2005.

HUANG, Anna. Similarity measures for text document clustering. **New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings**, n. April, p. 49–56, 2008.

ISO. **ISO 14224:2016 Petroleum, petrochemical and natural gas industries — Collection and exchange of reliability and maintenance data for equipment**. . [S.l: s.n.]. Disponível em: <https://www.iso.org/standard/64076.html>. , 2016

KARDEC, Alan; NASCIF, Julio. **Manutenção. Função Estratégica**. 3ª Edição ed. Rio de Janeiro: QualityMark, 2009. Disponível em: <https://www.passeidireto.com/arquivo/21862338/manutencao-funcao-estrategica-3-ed-alan-kardec-julio-nascif->.

KORDE, Vandana; MAHENDER, C Namrata. Text Classification and Classifiers: A

Survey. **International Journal of Artificial Intelligence & Applications**, v. 3, n. 2, p. 85–99, 2012.

LEWIS, David D; RINGUETTE, Marc. A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches. **In Proceeding of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR`94)**, p. 1–14, 1994.

LIU, Ruonan et al. Artificial intelligence for fault diagnosis of rotating machinery : A review. **Mechanical Systems and Signal Processing**, v. 108, p. 33–47, 2018. Disponível em: <https://doi.org/10.1016/j.ymssp.2018.02.016>.

LOPES, Maria Célia Santos. **MINERAÇÃO DE DADOS TEXTUAIS UTILIZANDO TÉCNICAS DE CLUSTERING PARA O IDIOMA PORTUGUÊS Maria**. 2004. 191 f. Universidade Federal do Rio de Janeiro, 2004.

MA, Zheren et al. **Applications of Machine Learning and Data Mining in SpeedWise® Drilling Analytics: A Case Study**. . Abu Dhabi International Petroleum Exhibition & Conference held in Abu Dhabi, UAE, 12-15 November 2018: [s.n.], 2018.

MAHASIVABHATTU, Kalicharan et al. **Engineering Data Management Using Artificial Intelligence Digitizing Paper Drawings**. . Houston, Texas, USA, 6 – 9 May 2019: [s.n.], 2019.

MAMMAN, Saidu; ANDRAWUS, Jesse A.; IYALLA, Ibiye. Improving the Reliability of Subsea Valves. **Spe**, p. 1–7, 2009.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Modern Information Retrieval (2nd edition)**. 40 W. 20 St. New York, NY. United States: Cambridge University Press, 2008. v. 53.

MARZEC, Mateusz; UHL, Tadeusz; MICHALAK, Dariusz. Verification of text mining techniques accuracy when dealing. **DIAGNOSTYKA**, v. 15, n. 3, p. 51–57, 2014.

MCCALLUM, Andrew; NIGAM, Kamal. **A Comparison of Event Models for Naive Bayes Text Classification**. **Proc. of the AAAI-98 Workshop on Learning for Text Categorization**. Madison, USA: [s.n.], 1998.

MENGUE, Denis Carlos; SELLITO, Miguel Afonso. CONFIABILIDADE PARA UMA BOMBA CENTRÍFUGA PETROLÍFERA MAINTENANCE STRATEGY BASED ON RELIABILITY FUNCTIONS FOR AN OIL CENTRIFUGAL PUMP Denis Carlos

Mengue * E-mail : denismengue@bol.com.br Miguel Afonso Sellitto * E-mail : sellitto@unisinos.br 1 INTRODUÇÃO. **Revista Produção Online**, v. 13, n. 2, p. 759–783, 2013.

MILANA, Diletta et al. Natural Language Understanding for Safety and Risk Management in Oil and Gas Plants. 2019, Abu Dhabi: Society of Petroleum Engineers (SPE), 2019. p. 1–14.

MOBLEY, R. Keith. **An Introduction to Predictive Maintenance**. 2nd Editio ed. [S.l.]: ButterWorth Heinemann, 2002. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B9780750675314X50003>.

MONTEIRO, Danielle de Oliveira et al. **Relatório técnico parcial 3 - Análise e modelagem de incertezas para gerenciamento e otimização de produção de plataformas offshore fase II**. . Rio de Janeiro, RJ - Brasil: [s.n.], 2020.

MUKHERJEE, Saikat; CHAKRABORTY, Amit. **Automated Fault Tree Generation : Bridging Reliability with Text Mining**. . [S.l: s.n.], 2007.

NASERI, Masoud; BARABADY, Javad. System-Reliability Analysis by Use of Gaussian Fuzzy Fault Tree: Application in Arctic Oil and Gas Facilities. **Oil and Gas Facilities**, n. June, p. 85–96, 2015.

NORVIG, Peter. **How to Write a Spelling Corrector**. Disponível em: <https://norvig.com/spell-correct.html>. Acesso em: 17 ago. 2020.

NOSHI, Christine; SCHUBERT, Jerome. A Brief Survey of Text Mining Applications for the Oil and Gas Industry. 2019, Beijing: International Petroleum Technology Conference, 2019. p. 1–13.

NTNU. **Reliability Data Sources**. Disponível em: <https://www.ntnu.edu/ross/info/data>. Acesso em: 28 jan. 2019.

OREDA. **History & About pages (OREDA)**. Disponível em: <https://www.oreda.com/>. Acesso em: 28 jan. 2019.

OREDA COMPANIES. **Offshore Reliability Data Handbook (Topside)**. 4th Editio ed. [S.l.]: OREDA Participants, 2002.

ORENGO, Viviane Moreira; HUYCK, Christian. A stemming algorithm for the Portuguese language. **Proceedings - 8th Symposium on String Processing and**

Information Retrieval, SPIRE 2001, p. 186–193, 2001.

PALMIERI, Luis et al. Petroleum Facilities Reliability Assement Model. **2007 SPE Latin American and Caribbean Petroleum Engineering Conference held in Buenos Aires, Argentina**, p. 1–9, 2007.

PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Scikit-learn: Machine Learning in Python Fabian. **Journal ofMachine Learning Research**, v. 12, p. 2825–2830, 2011.

PENNEL, Mike; HSIUNG, Jeffrey; PUTCHA, V B. Detecting Failures and Optimizing Performance in Artificial Lift Using Machine Learning Models. **SPE Western Regional Meeting**, 2018.

POLETTINI, Nicola. The Vector Space Model in Information Retrieval - Term Weighting Problem Local Term-Weighting. **Entropy**, p. 1–9, 2004.

PORTER, M. F. An algorithm for suffix stripping. **Program: electronic library and nformation systems**, v. 40, n. 3, p. 211–218, 2006.

PRIYADARSHY, Satyam et al. **Framework for Prediction of NPT causes using Unstructured Reports Data Engineering Challenges and extracton from Reports**. . Offshore Technology Conference held in Houston, Texas, USA, 1–4 May 2017: [s.n.], 2017.

PYTHON SOFTWARE FOUNDATION. **About Python**. Disponível em: <https://www.python.org/about/>. Acesso em: 17 ago. 2020.

RENNIE, Jason D M et al. **Tackling the Poor Assumptions of Naive Bayes Text Classifiers**. . [S.l: s.n.], 2003.

SALGADO, Marcia de Fatima Platilha. **Aplicação de Técnicas de Otimização na Engenharia de Confiabilidade**. . Belo Horizonte: [s.n.]. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUOS-8CDG4N?show=full>. , 2008

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 2019, OE 2019**. SPE Offshore Europe Conference and Exhibition held in Aberdeen, UK, 3-6 September 2019: [s.n.], 2019a.

SALO, Erik; MCMILLAN, David; CONNOR, Richard. **Work Orders - Value from**

Structureless Text in the Era of Digitisation. . SPE Offshore Europe Conference and Exhibition held in Aberdeen, UK, 3-6 September 2019: [s.n.], 2019b.

SALTON, G; WONG, A; YANG, C S. A Vector Space Model for Automatic Indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, 1975.

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, p. 159–167, 1996.

SANTOS, Nilis Adriano dos; SELLITO, Miguel Afonso. Estratégia De Manutenção E Aumento Da Disponibilidade De Um Posto De Compressão De Gases Na Indústria Petrolífera. **Revista Produção Online**, v. 16, n. 1, p. 77–103, 2016.

SCIENCEDIRECT. **Publicações Citando OREDA**. Disponível em: <https://www.sciencedirect.com/search/advanced?qs=OREDA&show=100&sortBy=date>.

SIDAHMED, Mohamed; COLEY, Christopher J; SHIRZADI, Shawn. **Augmenting Operations Monitoring by Mining Unstructured Drilling Reports**. . Texas: [s.n.], 2015.

SINGH, G K; KAZZAZ, Saleh Ahmed Al. Induction machine drive condition monitoring and diagnostic research — a survey. **Electric Power Systems Research 64**, v. 64, p. 145–158, 2003.

SOLC, Tomaz. **Unidecode 1.1.1**. Disponível em: <https://pypi.org/project/Unidecode/>. Acesso em: 18 ago. 2020.

SPARKE, Simon J. SPE-174539-MS Global Well Completion Reliability Databases and their Use in Well Design. **Society of Petroleum Engineers (SPE) Well Integrity Symposium**, p. 9, 2015.

SUN, Liping et al. Study on Maintenance Strategy for Fpso Offloading System Based on Reliability Analysis. **Omae-2016**, n. 2003, p. 1–7, 2016.

TAN, Ah-hwee. **Text Mining : The state of the art and the challenges Concept-based**. **Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases**. [S.l: s.n.], 1999. Disponível em: <http://www.mendeley.com/research/text-mining-state-art-challenges-3/>.

TAN, Pang-Ning et al. **Introduction to Data Mining**. 1 st Editi ed. [S.l.]: Addison-Wesley Professional, 2005.

TIAN, Kun et al. **Data Mining of Hidden Danger in Enterprise Production Safety and Research of Hidden Danger ' s Model Conversion**. . International Petroleum Technology Conference held in Beijing, China, 26 – 28 March 2019: [s.n.], 2019.

UCHEREK, Jared et al. **Auto-Suggestive Real-Time Classification of Driller Memos into Activity Codes Using Natural Language Processing**. . IADC/SPE International Drilling Conference and Exhibition held in Galveston, Texas, 3–5 March 2020: [s.n.], 2020.

WANG, Changyong; ZHANG, Honghuan; DUAN, Menglan. Reliability Analysis on Subsea X-tree Tubing Hanger. v. 2, n. 2, p. 51–56, 2012.

WOOD GROUP. **What We Do: Products and Services. Digital Solutions: iQRA**. Disponível em: <https://www.woodgroup.com/what-we-do/view-by-products-and-services/digital-solutions/iqra>.

WOOD GROUP INTETECH. **iQRA Database**. Disponível em: <https://www.iqra-database.com/>.

WORDNET. **PULO - WordNet PT**. Disponível em: <http://wordnet.pt/>.

WU, Wenkuang et al. **Retrieving Information and Discovering Knowledge from Unstructured Data Using Big Data Mining Technique: Heavy Oil Fields Example**. . Kuala Lumpur, Malaysia, 10–12 December 2014.: [s.n.], 2014.

WUTTKE, Régis André; SELLITTO, Miguel Afonso. Cálculo da Disponibilidade e da Posição na Curva da Benheira de uma Válvula de Processo Petroquímico. **Revista Produção Online**, v. VIII, 2008.

ZANGL, G; OBERWINKLER, C P. Predictive data mining techniques for production optimization. 2004, [S.l: s.n.], 2004. p. 2393–2398.

ZHANG, Harry. **The optimality of Naive Bayes**. **Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004**. [S.l: s.n.], 2004.

ZHANG, Tong et al. Industrial text analytics for reliability with derivative-free optimization. **Computers and Chemical Engineering**, v. 135, p. 106763, 2020.

Disponível em: <https://doi.org/10.1016/j.compchemeng.2020.106763>.

ZIO, E. Reliability engineering: Old problems and new challenges. **Reliability Engineering and System Safety**, v. 94, p. 125–141, 2009.