

O uso de NLP e data science na indústria de óleo e gás - um case de aquisição de conhecimento sobre integridade e confiabilidade na área submarina

The use of NLP and data science in the oil and gas industry - a case of acquiring knowledge about integrity and reliability in the subsea area

Lucas Augusto Penna

de Carvalho 

Petrobras, Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: lucas.penna@petrobras.com.br

Rene Thiago Capelari

Orlowski 

Petrobras, Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: renethiago@petrobras.com.br

Anderson Barata Custódio 

Petrobras, Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: andersonbc@petrobras.com.br

Italo de Oliveira Matias 

Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: italomatias@puc-rio.br

Fábio Freitas da Silva 

Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: fabio.freitas@puc-rio.br

Marcelo Romeu Gonçalves 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: marceloromeugoncalves@puc-rio.br

Roberto Meireles Acruche 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: acruche@puc-rio.br

Soeli Teresinha Fiorini 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: soeli@puc-rio.br

Whanderley Souza Freitas 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: whanderley_souza@puc-rio.br

Anderson Oliveira da Silva 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: andersonjs@puc-rio.br

Eduardo Shimoda 

Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro-Rio de Janeiro, Brasil.
E-mail: shimoda@puc-rio.br

RESUMO

Um sistema foi desenvolvido para aprimorar a análise de integridade e confiabilidade em equipamentos submarinos, lidando com a complexidade de dados não estruturados de em torno de 120.000 documentos PDF. Esse sistema consiste em dois serviços: um Extrator, que utiliza *Machine Learning* (ML) e expressões regulares para minerar dados de relatórios de inspeção, e uma Interface para visualização e análise dos dados estruturados. A identificação de fontes relevantes e requisitos do sistema foi obtida em reuniões com especialistas submarinos, e, através amostras de relatórios de inspeção, utilizadas para treinar o Extrator. A metodologia envolveu *Support Vector Machines* (SVM) para classificação de tipos de tabelas e regex para extração. Dificuldades surgiram devido à diversidade de estruturas de relatórios, levando a uma abordagem híbrida com ML para identificação de contexto. A interface web resultante oferece funcionalidades como busca avançada, visualização de dados extraídos, análise estatística da confiabilidade e dashboards analíticos. O software contribuiu significativamente para a eficiência das atividades de análise submarina, promovendo a transformação digital na área ao proporcionar uma valiosa base histórica de dados estruturados para apoio à tomada de decisões.

ABSTRACT

A system was developed to improve the analysis of integrity and reliability in subsea equipment, dealing with the complexity of unstructured data of around 120,000 PDF documents. This system consists of two services: an Extractor, which uses Machine Learning (ML) and regular expressions to mine data from inspection reports, and an Interface for viewing and analyzing structured data. Identification of relevant sources and system requirements was obtained in meetings with subsea experts, and, through sample inspection reports, used to train the Extractor. The methodology involved Support Vector Machines (SVM) for classifying table types and regex for extraction. Difficulties arose due to the diversity of reporting structures, leading to a hybrid approach with ML for context identification. The resulting web interface offers features such as advanced search, visualization of extracted data, statistical analysis of reliability and analytical dashboards. The software contributed significantly to the efficiency of underwater analysis activities, promoting digital transformation in the area by providing a valuable historical base of structured data to support decision making.

Palavras-chave:

Área submarina.
Gestão de integridade.
Confiabilidade.
Transformação digital.
Mineração de dados.

Keywords:

Subsea.
Integrity management.
Reliability.
Digital transformation.
Data mining.

Recebido:

08 de janeiro de 2025

Aceito para publicação:

29 de janeiro de 2025

Publicado:

26 de fevereiro de 2025

<https://doi.org/10.70369/f9vk7h11>



1. INTRODUÇÃO

Na As análises de integridade, extensão de vida, confiabilidade e risco em equipamentos e dutos submarinos requerem dados que não são facilmente obtidos, pois muitos deles estão em documentos no formato PDF. Sendo assim, grande parte do tempo de uma análise é tomado por atividades de busca, catalogação e formatação de dados, o que diminui a eficiência do processo. Estudos conduzidos pela Society of Petroleum Engineers (SPE) e por empresas globais de E&P verificaram que o gerenciamento de dados é um grande problema na indústria de óleo e gás e que mais da metade do tempo do engenheiro e geocientista é gasto apenas pesquisando e reunindo dados antes da análise (BRULE, 2015). Os dados não estruturados constituem uma grande parte dos ativos de dados na indústria de petróleo e gás (ASFOOR; HASAN, 2020). Texto ou dados não estruturados compreendem aproximadamente 80% dos dados gerados por muitas áreas em uma organização, e podem estar em várias formas, como documentos, planilhas, mapas, imagens, e-mails e relatórios (KANIMOZHI; VENKATESAN, 2015).

Diante do cenário de uso intensivo de dados não estruturados de documentos, tornou-se necessário utilizar técnicas para auxiliar na manipulação, processamento e, principalmente, extração de informações de interesse presentes nos documentos, como forma de tratar esses dados para gerar informação de valor para o negócio. Com esse intuito, a aplicação da Inteligência Artificial (IA) e suas ferramentas de Mineração de Dados – Hand, David e Adams (2014) e Han *et al.* (2011) – os métodos de Processamento de Linguagem Natural (Natural Language Processing – NLP) Manning *et al.* (2014), além da Ciência de Dados, Van (2016), cresceram no setor de óleo e gás, tratando diversas questões em diferentes domínios, como por exemplo os citados nos artigos Antonio *et al.* (2019), Zhang *et al.* (2021); Pankaj *et al.* (2018); Subrat e Nanda (2020); e Machado *et al.* (2021).

2. OBJETIVO

Este artigo descreve um estudo de caso de dois anos de pesquisa, no setor de óleo & gás, voltado para a aquisição de conhecimento sobre a integridade e confiabilidade de equipamentos e dutos submarinos, através da aplicação de técnicas de aprendizado de máquina, NLP e ciência de dados sob um legado de documentos de inspeção submarina em formato PDF e Excel. Tais documentos representam um histórico de eventos relacionados ao ciclo de vida dos equipamentos e dutos, insumo importante para a gestão da integridade e confiabilidade, bem como o monitoramento do desempenho dos ativos de modo a assegurar sua segurança. Como resultado da pesquisa, foi implementada uma arquitetura para a extração e estruturação dos dados dos relatórios de barco, e, uma ferramenta para a análise e interpretação desses dados estruturados. Inicialmente, será apresentado neste artigo um

embasamento teórico, descreve-se o dataset e a metodologia utilizada na pesquisa. Na sequência, apresenta-se o case com a descrição detalhada de algumas das principais funcionalidades e, por final, as conclusões obtidas.

3. NLP E DATA SCIENCE

Aprendizado de Máquina (AM) é uma área da computação que evoluiu a partir do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em Inteligência Artificial (IA). AM explora a construção e o estudo de algoritmos que podem aprender e fazer previsões a partir dos dados (M ANNILA, 1996). Geralmente, as etapas de AM são compostas de atividades de Mineração de Dados, a qual, por definição, visa obter conhecimento a partir dos dados armazenados, possibilitando algum nível de aprendizado e, consequentemente, resultando em aplicações de suporte à tomada de decisão.

O NLP é uma subárea da computação que engloba por exemplo, notação para expressões regulares, algoritmos e técnicas específicas Manning *et al.* (2014) e compõe grande parte do desenvolvimento das pesquisas neste projeto.

4. METODOLOGIA

A identificação das fontes de dados relevantes, o levantamento de processos/atividades e a definição de requisitos foram obtidos em reuniões com os envolvidos da área submarina da empresa (equipamentos e dutos). Além disso, fontes de informações, diretrizes, processos, normas e documentos foram relacionados ao domínio da área submarina indicados por estes envolvidos.

A empresa do setor de óleo & gás, participante do projeto, disponibilizou relatórios de operações de barco gerados por diversas embarcações que prestam serviço a área submarina. Foram disponibilizadas para análise e extração de dados, uma amostra de 120.000 relatórios. Os arquivos disponibilizados referem-se às inspeções de dutos e equipamentos submarinos dos quais foram gerados por diversas empresas prestadoras de serviço, havendo diferentes padronizações dos documentos, tornando assim o processo de extração de dados ainda mais complexo. Para alguns tipos de serviços de inspeção, por exemplo, foram encontrados cerca de 222 tipos diferentes de tabelas de inspeção na amostra de documentos. A empresa também forneceu diversos outros documentos e planilhas para a realização de análises, como por exemplo avaliação da efetividade do sistema de proteção catódica e registros de falhas.

As atividades de pesquisas desenvolvidas, até o momento da escrita do artigo, foram organizadas em três grupos:

- Obtenção de conhecimento utilizando técnicas de Processamento de Linguagem Natural;
- Análise e identificação de objetos em dados não estruturados por meio de Modelos Supervisionados de Aprendizado de Máquina;
- Ciência de dados e modelos estatísticos.

A abordagem baseada em aprendizado de máquina consistiu na identificação de contextos baseado em expressões regulares por um algoritmo classificador. Esta abordagem iniciou sua implementação pela seleção da biblioteca para extração de sentenças dos textos dos relatórios. Uma vez definida a biblioteca que faria a extração de sentenças, foi reunido um conjunto de relatórios, de onde foram extraídas as sentenças que continham os termos de interesse. Estas sentenças foram separadas em dois grupos, as que expressavam um sentido negativo em relação ao termo e as que expressavam um sentido afirmativo. Para geração do modelo de classificação das sentenças foram testados e comparados quatro algoritmos classificadores. Os algoritmos e suas respectivas acurácias são demonstradas na Tabela 1.

Tabela 1 – Algoritmos de classificação e suas acurácias.

Algoritmo	Acurácia
Regressão logística	0.95
KNN	0.93
Random Forest	0.97
MLP	0.91

Fonte: elaborado pelos autores.

Para se obter valor da acurácia foi utilizado a técnica de cross validation [24]. A acurácia final é dada pela média das acurácias alcançadas durante as iterações. Como demonstrado na Tabela 1, embora todos os algoritmos tenham alcançado resultados satisfatórios, o algoritmo Random Forest teve desempenho melhor, sendo então o escolhido para a classificação de sentenças.

Foi utilizada uma abordagem baseada em aprendizado de máquina, bem como foi criada a visualização dos termos presentes no relatório na qual teve o sentido do contexto classificado como afirmativo. Além dos termos, também foi criada a visualização dos contextos na qual os termos estão inseridos.

Durante o desenvolvimento do projeto de pesquisa, as técnicas de AM e MD foram utilizadas desde a criação de conjunto de dados para avaliar amostras com características que melhor identificam o tipo de relatório de inspeção e a classe de tabelas de medição, quanto o uso de algoritmos e técnicas de modelo supervisionado em tarefas de classificação automática em objetos de PDF e Excel.

5. RESULTADOS

5.1. Medições de potencial eletroquímico

Os relatórios de barco trazem uma informação importante que são as medições de PE (potencial eletroquímico), representando uma medida da condição de proteção do equipamento submarino instalado com relação à mecanismos de corrosão. Os valores medidos e os locais do equipamento onde são realizadas as medições são dados relevantes para avaliação da efetividade do sistema de proteção catódica. Com o intuito então de otimizar a quantidade de pontos medidos e o tempo de utilização da embarcação no serviço, foi necessário a extração de tabelas de PE e armazenamento dos seus dados de maneira estruturada para as análises dos especialistas da área submarina. No estudo, diversos tipos de tabelas PE foram identificados e estruturados, ou seja, seus dados extraídos, organizados e persistidos em banco de dados. Até o atual momento, foram identificados 222 modelos de tabelas no formato PDF e Excel, sendo a grande parte passível de estruturação.

A Figura 1 apresenta, através de busca usando a ferramenta computacional desenvolvida, a existência de relatórios que possuem tabelas de potencial. A coluna "Potencial" pode ter os seguintes valores: "Não" – quando não há tabela de PE no relatório; "Sim" – quando há tabela de PE no relatório e "Sim*" – quando há tabela de PE e existem valores que estão fora da faixa de proteção padrão. Quando existe a tabela, pode-se clicar sobre o link para ver e editar os dados das medições, conforme Figura 2. Na coluna "Ações" da Figura 1, pode-se clicar nos ícones para visualizar o relatório por completo, ver todas as imagens dele e ver todos os comentários (contexto) associados com os termos relevantes encontrados, visualizados na coluna "Termos". Assim, obteve-se um valioso conjunto de dados de PE para a análise dos especialistas.

Figura 1. Visualização dos relatórios com tabelas de PE e alerta de relatórios com PE fora da faixa padrão (Sim*).

Fonte: elaborado pelos autores.

Na Figura 2, os valores das medições (colunas de medida e menor valor absoluta da tabela) fora da faixa de proteção catódica estão marcados em vermelho, conforme indicado, servindo de alertas para a tomada de ações e podem ser descartados pelo especialista que realiza a análise, caso julgue que não estejam corretos ou sejam irrelevantes, não sendo esses valores posteriormente considerados em buscas ou cálculos. Com isso, o especialista faz um saneamento na base de dados, mantendo somente o que é de interesse.

Figura 2. Visualização e edição de valores de potencial.

Tabelas de Potencial Eletroquímico

Tabela 10 – Medição de potencial eletroquímico						
Pontos de medição		1ª medida (mV)	2ª medida (mV)	Menor valor absoluto (mV)	Fotos	
Conexão 1 MSP x Tramo C016A3	Flange do duto	-925	-923	-923		
	Flange	-922	-925	-922		4
	Estojo	-927	-928	-927		6
1ª e 2ª abraçadeira com duas boias	Abraça	-601	-602	-601		15
	Abraça	-606	-603	-603		17
3ª Abraçadeira com uma boia	Abraça	-601	-602	-601		23
	Colar de ancoragem	-819	-819	-819		45
Conexão 2 Tramos C016A3 x C016A4	Flange	-820	-819	-819		47
	Flange	-820	-819	-819		48
	Estojo	-820	-820	-820		49

Local	1ª Medição	2ª Medição	Menor Medição	Foto	Ignorar medição	Editar
Conexão 1 MSP x Tramo C016A3	-925 mV	-923 mV	-923 mV	4	☐	Editar
Conexão 1 MSP x Tramo C016A3	-922 mV	-925 mV	-922 mV	5	☐	Editar
Conexão 1 MSP x Tramo C016A3	-927 mV	-928 mV	-927 mV	6	☐	Editar
1ª e 2ª abraçadeira com duas boias	-601 mV	-602 mV	-601 mV	15	☒	Editar
1ª e 2ª abraçadeira com duas boias	-606 mV	-603 mV	-603 mV	17	☒	Editar
3ª abraçadeira com uma boia	-601 mV	-602 mV	-601 mV	23	☒	Editar
Colar de ancoragem	-819 mV	-819 mV	-819 mV	45	☐	Editar

Fonte: elaborado pelos autores.

5.2. Histórico MPE (ANM)

Para uma análise mais detalhada e correlacionada das informações das inspeções de PE ao longo do tempo, foi criada uma funcionalidade que apresenta o histórico de medições de potencial eletroquímico, para equipamentos do tipo ANM (Árvore de Natal Molhada). Nesta funcionalidade, apresentada na Figura 3, é possível exibir para cada poço:

- Se houve inspeções de onde foi realizada a medição de PE;
- Se houve troca de subequipamentos no poço;
- Se foi citado o termo skid de anodo no relatório;
- A última medição de potencial (valor mais positivo);
- O nível de desgaste dos anodos (pior caso) na última inspeção;
- A tag e tempo de operação do subequipamento presente no poço com maior tempo de operação;
- (data da inspeção – última data da instalação do subequipamento);
- O link para o relatório da última inspeção.

Na guia “Detalhes”, pode-se visualizar todos os relatórios anteriores das inspeções naquele poço.

Figura 3. Visualização do histórico MPE.

Poço	Relatório MPE?	Trocou Subeqto?	Cita Skid Anodos?	Última Inspeção	Potencial Última Medição (mV)	Potencial Dentro Faixa (-500 ~ -1000)?	Desgaste Anodos (última inspeção)	Tag (subeqto com maior tempo Operação)	Anos (subeqto com maior tempo Operação)	Relatório (última inspeção)	Detalhes
	Sim	Não	Não	07/12/2021	-996	Sim	MÉDIO		6.11	SKBI21-109.pdf	
	Sim	Sim	Não	07/12/2021	-989	Sim	MÉDIO		6.40	SKBI21-108.pdf	
	Sim	Sim	Não	07/12/2021	-989	Sim	MÉDIO		5.72	SKCab21-107.pdf	

Fonte: elaborado pelos autores.

5.3. Análise de confiabilidade

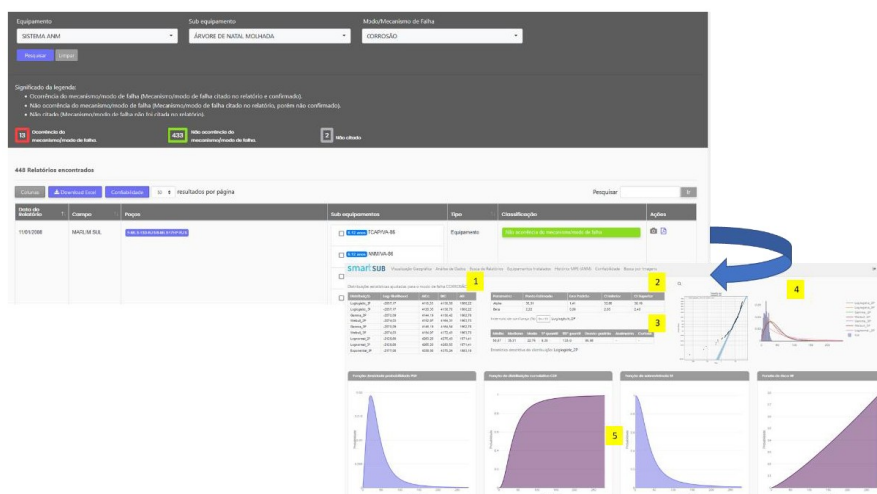
Analisar o histórico das inspeções realizadas ao longo da vida útil dos equipamentos e dutos submarinos e os modos/mecanismos de falha identificados é um importante insumo para a análise de confiabilidade. Sendo assim, as seguintes atividades foram realizadas:

- Obtidos os dados do sistema de informação, através de tabelas do banco de dados de inspeções e do histórico de processos;
- Realizada a identificação dos relatórios por local de instalação, poço ou campo, e tipo de equipamento;

- Realizada a vinculação dos subequipamentos instalados aos relatórios em função da data da inspeção;
- Determinado o tempo de operação do subequipamento quando da inspeção (data da inspeção - data da última instalação do subequipamento);
- Identificado os indícios de falha através da mineração de dados nos relatórios usando-se técnicas de ML;
- Implementado um processo para validação dos indícios de falha por especialistas da área submarina, compondo assim um banco de dados de registros de falha;
- Implementada uma funcionalidade de cálculo estatístico dos registros do banco de dados de falha, armazenando em banco os parâmetros estatísticos das análises realizadas e possibilitando a realização de cálculos estatísticos através de uma calculadora.

Com base na implementação dessas funcionalidades e na busca pela existência de dados de falha nos relatórios, criou-se a interface apresentada na Figura 4 do módulo de confiabilidade. A opção “Confiabilidade” exibe estatísticas associadas para cada distribuição (média de falhas, mediana, moda), gráfico de probabilidades e histograma, bem como as curvas de densidade.

Figura 4. Interface do módulo de confiabilidade.



Fonte: elaborado pelos autores.

Na Figura 4 pode ser observado:

1. Múltiplas distribuições que são ordenadas de acordo com seu ajuste;
2. Possibilidade de seleção no nível de confiança para a análise sendo realizada;
3. Parâmetros estatísticos da distribuição: média de falhas, mediana, moda, etc.;
4. aderência das distribuições pelo gráfico de probabilidades e histograma;
5. Visualização das curvas de densidade:
 - a. Função de densidade de probabilidade (PDF),
 - b. Função de distribuição cumulativa (CDF),
 - c. Função sobrevivência (SF),
 - d. Função de risco (HF).

6. CONSIDERAÇÕES FINAIS

Esse artigo apresentou uma pesquisa na área de integridade e confiabilidade de equipamentos e dutos submarinos no setor de óleo & gás. Sob uma base de relatórios de barco em PDF e Excel foi realizada: i) identificação e estruturação dos dados mais relevantes dos relatórios e suas tabelas de eventos, potencial eletroquímico e manuseio, reduzindo o tempo gasto dos especialistas envolvidos nessa atividade; ii) recuperação de imagens das inspeções de equipamentos e dutos submarinos e suas legendas, provendo uma base para análise dos mecanismos de falhas e outros problemas relevantes; iii) busca de relatórios por termos de interesse e obtenção de seus contextos para auxiliar na análise de integridade; iv) geração de alertas dos relatórios com o potencial eletroquímico fora da faixa esperada para a avaliação da efetividade do sistema de proteção catódica; v) funcionalidade de histórico de medições de potencial eletroquímico para equipamento do tipo ANM. Também foi desenvolvido um módulo de confiabilidade que provê dados para análises quantitativas da probabilidade de ocorrência de falhas. Nesse módulo, é possível realizar análises estatísticas importantes do histórico operacional de equipamentos em uso, insumo importante para o planejamento das inspeções, possibilitando a otimização da periodicidade e do escopo dos planos de inspeção.

Os desafios computacionais foram apresentados, sendo testado alguns modelos de IA/NLP para viabilizar a implementação das necessidades do cliente, principalmente na extração dos dados não estruturados, ocasionados pela grande variedade de modelos de relatórios e suas tabelas. Cada vez mais os conceitos da Ciência de Dados, Aprendizagem de Máquina/Inteligência Artificial e o uso de técnicas de Processamento de Linguagem Natural se consolidam no projeto de pesquisa, transformando dados, até então pouco acessíveis, em informações relevantes para o negócio. O software desenvolvido gerou mais eficiência nas atividades de busca, catalogação e formatação de dados necessários para a análise dos especialistas, contribuindo assim, na transformação digital da área submarina, provendo uma ferramenta com uma valiosa base histórica de dados, agora estruturados, com funções para análise e auxílio na tomada de decisão.

Trabalho originalmente apresentado no Congresso ROG em 2024. Acesse [aqui](#).

REFERÊNCIAS

- Antonio, Daniel, Sadlier, Andreas, & Joseph Winston. (2019). Leveraging Oil and Gas Data Lakes to Enable Data Science Factories. <https://doi.org/10.2118/196452-MS>
- Asfoor, Hasan. (2020). Applying Data Science Techniques to Improve Information Discovery in Oil And Gas Unstructured Data. <https://doi.org/10.2523/IPTC-20236-Abstract>
- Brule, M. R. (2015). The data reservoir: How big data technologies advance data management and analytics in e&p. <https://doi.org/SPE-173445-MS>
- Han, Jiawei, Jian Pei, & Micheline Kamber. (2011). Data mining: concepts and techniques. Elsevier. <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>

Hand, David J., & Adams, Niall M. (2014). Data mining. Wiley StatsRef.

<https://doi.org/10.1002/9781118445112.stat06466.pub2>

Kanimozhi KV, & Venkatesan M. (2015). Unstructured data analysis-a survey. International Journal of Advanced Research in Computer and Communication Engineering, 4(3).

<https://doi.org/10.17148/IJARCCCE.2015.4354>

M Annila, Heikki. (1996). Data mining: machine learning, statistics, and databases. 2-9.

<https://doi.org/10.1109/SSDM.1996.505910>

Machado, Rodrigo Chamusca, Leite, Fabbio, Xavier, Cristiano, Albuquerque, Alberto, Lima, Samuel, & Gustavo Carvalho. (2021). Development of Failure Prediction Models for Subsea Blowout Preventers Using Data Analytics and AI. <https://doi.org/10.4043/31027-MS>

Manning, Christopher D. (2014). The Stanford Core NLP natural language processing toolkit. 55-60.

<https://doi.org/10.3115/v1/P14-5010>

Pankaj, Piyush, Geetan, Steve, MacDonald, Richard, Shukla, Priyavrat, Sharma, Abhishek, Menasria, Samir, Xue, Han, & Tobias Judd. (2018). Application of Data Science and Machine Learning for Well Completion Optimization. <https://doi.org/10.4043/28632-MS>

Subrat, Nanda. (2020). Using Data Science to Enable Condition Based Maintenance and Classification for Offshore Assets. <https://doi.org/10.2118/202758-MS>

Van Der Aalst, Wil. (2016). Data science in action. Process mining. Springer.

<https://link.springer.com/book/10.1007/978-3-662-49851-4>

Zhang, Hongbao, Zeng, Yijin, Liao, Lulu, Wang, Ruiyao, Hou, Xutian, Feng, Jiangpeng, & Amol Mulunjkar. (2021). How to Land Modern Data Science in Petroleum Engineering.

<https://doi.org/10.2118/205689-MS>