Clusterização e classificação de alarmes industriais utilizando word embeddings

Isaac Medeiros * **Diego Cavalcanti** ** **Juan Villanueva** ***

* *Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, PB, (e-mail: isaac.medeiros@alumni.cear.ufpb.br).*
** *Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, PB, (e-mail: diego.cavalcanti@estudante.cear.ufpb.br).*
*** *Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, PB, (e-mail: jmauricio@cear.ufpb.edu.br)*

Abstract: The analysis of industrial alarm content is crucial for detecting and preventing failures in operational processes. Alarms act as an alert system, signaling the operations team about abnormal conditions and potential failures in real-time. However, the excessive generation of records by these systems can make it difficult to identify and respond effectively to critical situations. Therefore, it is essential to develop efficient alarm management strategies to prioritize and group them intelligently. Furthermore, by conducting a thorough analysis of industrial alarm data, it is possible to gain a deeper understanding of operational conditions, recognize recurring patterns, identify trends, and take proactive measures to prevent failures. In this study, data from events and alarms were collected from the SCADA system of a thermoelectric plant located in the state of Paraíba. Natural language processing (NLP) techniques were used for the preprocessing of alarm texts, allowing for the generalization of information by excluding equipment identifiers and low-relevance semantic words. The BERT language model was used for the numerical representation of the texts, and clustering and classification techniques were applied for the efficient grouping of records. This approach not only improves alarm management but also contributes to a safer and more efficient operational environment, which is essential for the industry's sustainability and productivity.

Resumo: A análise do conteúdo de alarmes industriais é de suma importância para a detecção e prevenção de falhas em processos operacionais. Os alarmes funcionam como um sistema de alerta, sinalizando à equipe de operação sobre condições anormais e potenciais falhas em tempo real. Entretanto, a geração excessiva de registros por parte dos sistemas pode dificultar a identificação e a resposta eficaz diante de situações críticas. Por isso, torna-se imprescindível o desenvolvimento de uma gestão eficiente dos alarmes, a fim de priorizá-los e agrupá-los de maneira inteligente. Além disso, ao realizar uma análise minuciosa dos dados de alarmes industriais, é possível obter um entendimento mais aprofundado das condições operacionais, reconhecer padrões recorrentes e identificar tendências, além de agir de forma proativa para evitar falhas. Neste estudo, foram coletados e utilizados dados de eventos e alarmes extraídos do sistema SCADA de uma usina termoelétrica situada no estado da Paraíba. Para o pré-processamento dos textos dos alarmes, foram empregadas técnicas de processamento de linguagem natural (PLN), que permitiram a generalização das informações, excluindo identificadores de equipamentos e palavras semânticas de baixa relevância. Utilizou-se o modelo de linguagem BERT para a representação numérica dos textos, e técnicas de clusterização e classificação foram aplicadas para o agrupamento eficiente dos registros. Essa abordagem não apenas melhora a gestão dos alarmes, mas também contribui para um ambiente operacional mais seguro e eficiente, o que é fundamental para a sustentabilidade e a produtividade da indústria.

Keywords: Industrial Alarms; Alarm Management; Clustering; Natural Language Processing; Word embedding.

Palavras-chaves: Alarmes Industriais; Gerenciamento de Alarmes; Clusterização; Processamento de Linguagem Natural; Incorporação de Palavras.

1. INTRODUÇÃO

A indústria carrega o estigma de ser um ambiente rico em dados e escasso no desenvolvimento de conhecimento, em parte porque o setor é inerentemente conservador sobre seus processos e métodos. Estudos indicam que, embora a indústria de manufatura gere mais dados do que qualquer outro setor da economia, grande parte destes dados não é explorada pelas empresas (van Aardt, 2015).

Os alarmes industriais desempenham uma função de grande importância nas usinas termoelétricas, e que pode ser compreendida sob várias perspectivas.

Em primeiro lugar, os alarmes industriais auxiliam na manutenção de segurança da planta. Usinas termoelétricas possuem processos complexos e potencialmente perigosos, como a queima de combustíveis fósseis e geração de altas temperaturas e pressões. Os alarmes podem detectar condições anormais, como vazamentos, superaquecimentos ou falhas de equipamentos, alertando prontamente os operadores para adoção de medidas apropriadas para evitar acidentes ou mitigar riscos potenciais (Yang and Dziegielewski, 2007).

Em segundo lugar, os alarmes industriais são essenciais para a operação eficiente das usinas termoelétricas. Por meio destes, é possível detectar desvios das condições normais de operação, como variações de temperatura, pressão, por exemplo, que podem indicar ineficiências ou falhas nos equipamentos (Costa et al., 2019).

Os sistemas de alarmes industriais enfrentam desafios devido ao excesso de registros, necessitando de melhores ferramentas de gerenciamento e apresentação. Em (Cai et al., 2019), técnicas de agrupamento foram propostas para agrupar alarmes correlacionados, facilitando a remoção de alarmes redundantes e a identificação da causa raiz.

Este trabalho tem como objetivo o desenvolvimento de uma metodologia para a clusterização de alarmes industriais, por meio de técnicas de aprendizado de máquina. A clusterização permite agrupar alarmes apenas pelo conteúdo do texto, utilizando métodos de Processamento de Linguagem Natural (PLN), sem que haja alguma classificação ou rótulo prévio. Dessa maneira, torna-se possível priorizar os registros, filtrando apenas os que pertencem a determinados rótulos e subsistemas, ou ordenando por grau de importância. Além disso, o método proposto permite a identificação dos subsistemas que possuem falhas mais frequentes ou mais severas, que podem levar a paradas dos motores ou problemas de operação, bem como os que necessitam de mais intervenções de manutenção.

Além deste capítulo de introdução, este trabalho consta de mais 5 seções. Na segunda seção serão abordados os fundamentos teóricos que sustentaram o desenvolvimento da pesquisa; na terceira, será explanado brevemente sobre a base de dados de alarmes utilizada; na quarta, é especificada a metodologia empregada; na quinta seção, serão apresentados os resultados obtidos; e, por fim, na sexta seção algumas conclusões serão discutidas.

2. FUNDAMENTAÇÃO TEÓRICA

Técnicas de aprendizado de máquina são cada vez mais usadas para transformar dados históricos de alarmes em conhecimento acionável em vários setores industriais. Essas abordagens visam abordar desafios como inundações de alarmes e melhorar a tomada de decisão do operador em ambientes industriais complexos (Pirehgalin et al., 2020).

2.1 Word embeddings

Word embeddings são representações vetoriais de palavras que capturam relacionamentos semânticos e sintáticos em um espaço de baixa dimensão (Lebret, 2016). Esses *embeddings* provaram ser muito relevantes para várias tarefas de PLN, incluindo análise de sentimentos, classificação de texto e geração de frases (Suleiman and Awajan, 2018; Lebret, 2016). Diferentes abordagens para criar *word embeddings* incluem métodos tradicionais, estáticos e contextualizados, com modelos como BERT contribuindo significativamente para *embeddings* contextualizados (Neelima and Mehrotra, 2023).

Existem algumas técnicas de *word embeddings* bastante conhecidas, como o *Term Frequency-Inverse Document Frequency* (TF-IDF), *Word2Vec*, entre outras.

Estudos recentes demonstraram a superioridade do BERT sobre métodos tradicionais de classificação de texto como o TF-IDF. O BERT supera o TF-IDF em tarefas de agrupamento de texto, destacando-se em 28 de 36 métricas (Subakti et al., 2021). Sua capacidade de considerar a posição das palavras e o contexto em frases lhe dá uma vantagem sobre o TF-IDF, por exemplo, que não tem essa capacidade.

Neste trabalho, a geração de *word embeddings* foi por meio da técnica *Bidirectional Encoder Representations from Transformers* (BERT).

2.2 Modelo de linguagem BERT

O BERT, introduzido pelo Google AI Language em 2018, consiste em um avanço bastante significativo em PLN (Processamento de Linguagem Natural), com um grande impacto nesta área de pesquisa (Gupta, 2024).

Modelos de linguagem pré-treinados como BERT mostraram desempenho notável em tarefas de classificação de texto (Sun et al., 2019). Esses modelos passam por pré-treinamento não supervisionado em grandes corpora de texto, permitindo que sejam ajustados para tarefas específicas com conjuntos de dados menores (Barbon and Akabane, 2022).

O BERT emprega uma abordagem de treinamento bidirecional para a arquitetura *Transformer* (Gupta, 2024). Este modelo utiliza um mecanismo de atenção para aprender relacionamentos contextuais entre palavras em ambas as direções simultaneamente (Kachkou, 2021). O modelo *Transformer*, que forma a base do BERT, consiste em um codificador que lê a entrada de texto e um decodificador que gera previsões (Kachkou, 2021).

2.3 Algoritmo K-means

O algoritmo *K-means* é uma técnica popular de agrupamento de dados amplamente utilizada na mineração de dados (Abhishekkumar and Sadhana, 2017; Martarelli and Nagano, 2019), e permite agrupar dados em k *clusters*, buscando padrões específicos ao conjunto de dados (Mucherino et al., 2009). Em 1967, MacQueen propôs inicialmente o *K-means*, sendo um dos algoritmos de aprendizado supervisionado mais simples, aplicado para resolver o problema do *cluster* bem conhecido (SUN Ji-Gui and Lian-Yu, 2008). Trata-se de um algoritmo de agrupamento por particionamento, cujo método é classificar os objetos de dados fornecidos em k *clusters* diferentes por meio de iterações, convergindo para um mínimo local. Portanto, os resultados dos *clusters* gerados são compactos e independentes.

O algoritmo consiste em duas etapas separadas. A primeira fase seleciona k centroides aleatoriamente, onde o valor inicial de k é inicializado. O próximo passo é inserir cada objeto de dados para o centroide mais próximo (Fahim et al., 2006). A distância euclidiana é geralmente considerada para determinar a distância entre cada objeto de dados e os centros dos *clusters*. Quando todos os objetos de dados estão incluídos em algum *cluster*, a primeira etapa é concluída e um agrupamento inicial é feito. Recalcula-se a média dos *clusters* formados anteriormente. Esse processo iterativo continua repetidamente até que a função critério atinja o valor mínimo.

2.4 Algoritmo SVC com kernel linear

O SVC (*Support Vector Classifier*) com *kernel* linear é um algoritmo de classificação de suporte vetorial que utiliza uma função de *kernel* linear para o mapeamento dos dados de entrada em um espaço de alta dimensionalidade (Takahashi, 2015). Neste espaço, é possível encontrar um hiperplano que separa os dados de duas classes com o mínimo de margem de separação, que se trata da distância entre os pontos de cada classe. Devido ao *kernel* em questão, o hiperplano utilizado na classificação deve estar na condição linear.

A função *kernel* linear é definida como:

$$k(a, b) = a^T b. \tag{1}$$

Pesquisas recentes exploraram a combinação de *embeddings* BERT com vários classificadores para tarefas de classificação de texto. (Cordeiro et al., 2022) compararam algoritmos clássicos como SVM com BERT para classificar comunicações irregulares, descobrindo que BERT obteve o melhor desempenho com 96% de pontuação F1. (Ilic et al., 2022) conduziram uma revisão abrangente de modelos de classificação de texto, comparando abordagens tradicionais como SVM e *Random Forest* com modelos de última geração usando *embeddings* BERT e GPT-2. Seus resultados mostraram que os modelos BERT e GPT-2 tiveram melhor desempenho, com BERT superando ligeiramente o GPT-2 em tarefas de classificação binária. Para problemas multiclasse, o *C-Support Vector Classifier* e o BERT exibiram o melhor desempenho, destacando a eficácia das abordagens baseadas em BERT em vários cenários de classificação de texto.

3. BASE DE DADOS

Para o desenvolvimento deste trabalho, foi extraída uma amostra de 18.427.862 registros de alarmes e eventos presentes no sistema SCADA (*Supervisory Control and Data Acquisition*) da planta termoelétrica. Os dados foram coletados para os anos de 2019 a 2022, em formato tabular. Os alarmes possuem 3 níveis de severidade:

- Severidade 0: Alarmes que indicam *shutdown* dos motores da usina;
- Severidade 1: Alarmes de processo;
- Severidade 2: Eventos;

Na Tabela 1, a título de exemplo, tem-se uma amostra com os dados extraídos do sistema SCADA.

Tabela 1. Formato do conjunto de dados

Timestamp	Message	Severity
2019-10-29 13:48:18.650	CCP 02: Pump 10P003 Unselected mode	1
2019-10-29 13:57:55.617	CCP 2: D1451 Inactive	2
2019-10-29 14:19:44.840	GCP 29: L0210 2PZL7180 Emergency Stop Air Pressure Low Shut Down	0

Os textos dos alarmes, severidade, limites de valores e demais configurações são feitas no supervisório. Quando se tem como referência uma variável analógica, por exemplo, o nível de água nas caldeiras, é possível monitorar até quatro níveis de alarme: muito baixo, baixo, alto e muito alto. Também é possível monitorar uma variável pela especificação de múltiplas subcondições, através de alarmes discretos, monitorar variáveis digitais pela especificação de alarme em borda de subida ou na borda de descida, dentre outras possibilidades.

4. METODOLOGIA

Na primeira etapa do trabalho foram pré-processados os textos dos alarmes da amostra utilizada. O pré-processamento de texto é uma etapa fundamental no processamento de linguagem natural e na análise de dados, visando melhorar a qualidade e a usabilidade de dados textuais para análise ou modelagem subsequente (de Brito and da Silva Gomes Gomes, 2019). As etapas executadas foram as seguintes (Bird et al., 2009):

- Conversão dos caracteres para minúsculo, buscando a uniformização dos dados;
- Remoção de caracteres indesejados, como pontuação, símbolos, números e caracteres especiais. Dessa forma, é evitada a introdução de ruído no modelo de NLP, como também simplifica o processamento e torna o modelo mais generalizável;
- Remoção de *Stopwords*: eliminação de palavras muito comuns e com pouco significado, como artigos, preposições e pronomes. Tais palavras não contribuem significativamente para a classificação dos alarmes, e podem ser removidas para melhorar a eficiência do processamento;
- Lematização: redução das palavras às suas raízes correspondentes, retirando todas as inflexões;

- Remoção de palavras não pertencentes à língua inglesa: esta etapa foi utilizada para a remoção de *tags* de equipamentos e sensores, que permaneceram após as etapas de pré-processamento anteriores;

Segue um exemplo das etapas de pré-processamento realizadas em uma das mensagens dos alarmes presentes no conjunto de dados utilizado para o desenvolvimento do trabalho que deu origem a este artigo.

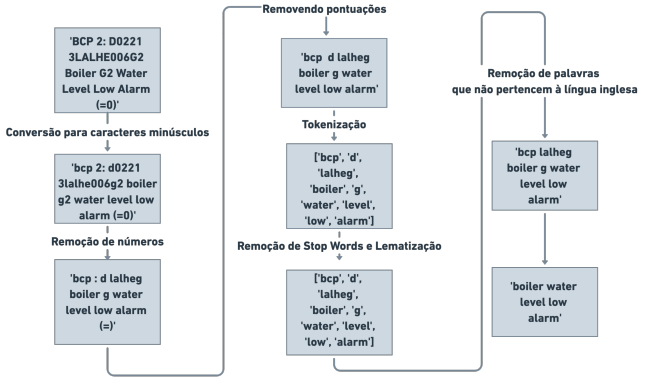

Figura 1. Etapas de pré-processamento das mensagens dos alarmes

Após todas as etapas do tratamento dos dados de texto, as mensagens alarmes foram transformadas em *embeddings* utilizando o modelo pré-treinado *bert-base-nli-mean-tokens* e a biblioteca *sentence-transformers*.

Com a representação vetorial das sentenças pelos *embeddings*, foi inicializado um valor de k *clusters*, para a aplicação do algoritmo K-means.

Devido a grande volumetria de dados utilizados, foram extraídas amostras aleatórias de 10.000 registros por ano para a tarefa de clusterização. Os dados coletados foram de 2019 a 2022, totalizando 40.000 registros.

O K-means foi implementado utilizando o método do coeficiente de *silhouette*. Trata-se de uma técnica amplamente utilizada para avaliar a qualidade dos *clusters*, bem como determinar o número ideal de agrupamentos nas implementações que utilizam o algoritmo K-means (Sai et al., 2017; quan Zhao, 2010). O coeficiente de *silhouette* mede o quão bem um objeto se encaixa em seu próprio *cluster* em comparação com *clusters* vizinhos, com valores variando de -1 a +1, em que valores mais altos indicam que os dados estão melhor agrupados (Sai et al., 2017). Demonstrou-se que esse método proporciona um melhor resultado na avaliação da eficácia do agrupamento em comparação com outras técnicas (quan Zhao, 2010). Buscando também encontrar o número ideal de agrupamentos, foi utilizado o método de *Elbow*, porém os resultados obtidos foram bastante insatisfatórios, uma vez que alarmes com conteúdos muito distintos foram incluídos nos mesmos grupos.

As etapas seguintes concentraram-se em transformar o aprendizado não-supervisionado em supervisionado: classificar uma nova amostra de dados utilizando os rótulos obtidos na clusterização. Após a limpeza dos dados para a remoção de valores nulos, o conjunto de dados foi dividido em treino e teste (80% e 20%, respectivamente).

A classificação foi realizada utilizando o *LinearSVC*, que integra um conjunto de modelos de máquina de vetor de suporte (*SVM*), um modelo de aprendizado de máquina supervisionado que soluciona problemas de classificação de grupos. No SVC (*Support Vector Classifier*) foi utilizado um *kernel* linear, principalmente devido à sua eficiência e ter produzido resultados consistentes.

5. RESULTADOS

No processo de clusterização, foram realizadas várias iterações para obtenção do valor de k mais adequado. Com isso, foi encontrado um valor ótimo de 249 *clusters* utilizando o método do coeficiente de *silhouette*, conforme Figura 2.

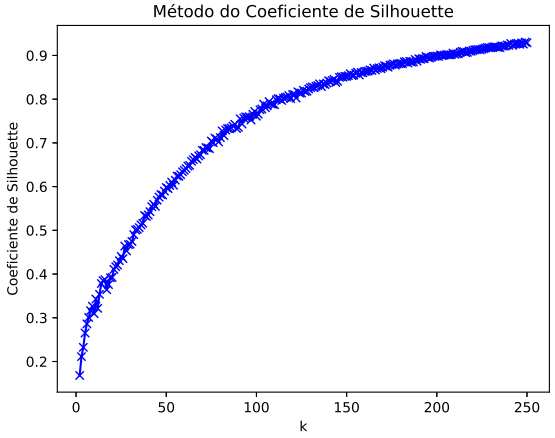

Figura 2. Número ótimo de *clusters* - Método da silhouette

Após o processo de clusterização, foram feitas algumas validações nos resultados, principalmente para verificar se os alarmes foram agrupados corretamente na mesma categoria. No *cluster* de rótulo 4, por exemplo, foram agrupados os registros referentes ao sistema de condensado da planta, especificamente envolvendo as caldeiras. Na Tabela 2 é possível visualizar algumas mensagens dos alarmes pertencentes ao ***cluster* 4**.

Tabela 2. *Cluster* 4 - Sistema de condensado

Message	***Cluster***
BCP 2: D0244 3LAHHHE006J2 Boiler J2 Water Level High/High Alarm (=0)	4
BCP 2: D0292 3LAHHHE006P2 Boiler P2 Water Level High/High Alarm (=0)	4
BCP 2: D0220 3LAHHHE006G2 Boiler G2 Water Level High/High Alarm (=0)	4

Os registros do *cluster* 4 corresponderam a um total de **2713 alarmes**.

No *cluster* 8, por exemplo, foram agrupados os alarmes relacionados ao sistema de radiadores da planta, conforme Tabela 3.

Tabela 3. *Cluster* 8 - Radiadores da planta

Message	***Cluster***
CCP 02: Radiator Fan 44HE003A2E2 Unselected mode	8
CCP 02: Radiator Fan 39HE003L2Q2 Unselected mode	8
CCP 02: Radiator Fan 43HE003L2Q2 Unselected mode	8
CCP 02: Radiator Fan 5HE003L2Q2 Unselected mode	8

Os registros do *cluster* 3 totalizaram 698 alarmes.

Conforme é possível observar nas duas tabelas, a clusterização resultou em uma distribuição bem definida dos alarmes entre os grupos. Os textos dos alarmes de cada *cluster* possuem conteúdos semelhantes, e fazem referência aos mesmos subsistemas da planta.

Uma vez realizada a etapa de validação com os demais grupos que foram gerados, o passo seguinte foi de classificar novas amostras de dados por meio dos *clusters* encontrados.

Após o treinamento do modelo de classificação, a acurácia obtida foi de 99,8%. O classificador foi aplicado em uma amostra de 40.000 registros, de forma análoga ao processo de clusterização.

Após a classificação, para o *cluster* 0, foram obtidas mais mensagens dos alarmes para este grupo, em comparação à etapa anterior de clusterização. Como é possível visualizar na Tabela 4, os dados foram classificados de forma consistente, pois os textos mencionam as caldeiras e, dessa forma, pertencem ao sistema de condensado da planta.

Tabela 4. *Rótulo* 4 - Sistema de Condensado

Message	**Rótulos - Classificação**
BCP 2: D0292 3LAHHHE006P2 Boiler P2 Water Level High/High Alarm (=0)	4
BCP 2: D0244 3LAHHHE006J2 Boiler J2 Water Level High/High Alarm (=0)	4
BCP 2: D0220 3LAHHHE006G2 Boiler G2 Water Level High/High Alarm (=0)	4
BCP 2: D0196 3LAHHHE006D2 Boiler D2 Water Level High/High Alarm (=0)	4

Para uma melhor visualização dos erros entre os *clusters* (amostra clusterizada) e dos rótulos preditos pelo algoritmo de classificação (amostra classificada), foi gerado um gráfico que contém a distribuição das predições por *cluster*, conforme Figura 3. Utilizou-se uma função do tipo $y = x$ para representar os rótulos preditos no eixo y e os *clusters* no eixo x, destacando os rótulos que divergiram dos valores reais.

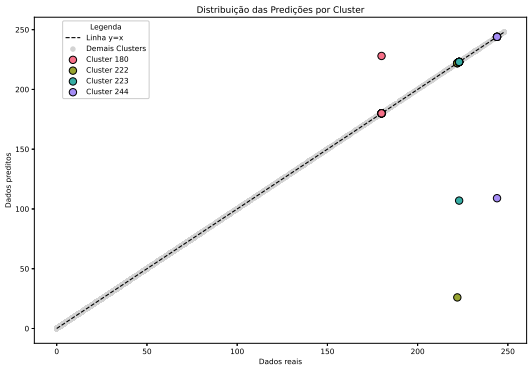

Figura 3. Distribuição das predições por *cluster*

Por meio da Figura 3, foi possível verificar que apenas 4 dos 249 rótulos obtidos pela classificação apresentaram divergências em relação à amostra clusterizada. Para cada um destes, foram calculadas as taxas de erro, correspondentes à diferença entre os dados obtidos pela clusterização e os dados preditos pela tarefa de classificação. Na Tabela 5 é possível visualizar as taxas de erro por rótulo.

Apesar dos rótulos da Tabela 5 terem apresentado divergências entre os valores verdadeiros e os preditos, a volumetria de registros que pertencem a estes grupos corresponderam a apenas 1.70% de toda a amostra classificada.

Tabela 5. Taxas de erro - *Clusters* e rótulos (classificação)

Cluster	Taxa de Erro (%)
180	11.11
222	25.00
223	12.50
244	20.00

6. CONCLUSÃO

A partir da aplicação da metodologia proposta, foi possível agrupar os alarmes apenas pelo conteúdo da mensagem, e com resultados consistentes, de modo que os registros pertencentes ao mesmo *cluster* apresentaram alta similaridade entre si no conteúdo, bem como divergiram de diferentes grupos obtidos, conforme o esperado. O pré-processamento de texto contribuiu diretamente para os melhores resultados, uma vez que foram eliminadas as identificadores dos equipamentos, bem como outros detalhes do texto, tornando o modelo de classificação generalizável. O agrupamento de alarmes que possuem conteúdos semelhantes, independentemente de envolver equipamentos distintos, permite uma série de análises e melhorias no processo de gestão destes registros. Por intermédio dos rótulos associados aos alarmes, é possível determinar os grupos mais ofensores da planta e aumentar a eficiência operacional. A operação e manutenção pode planejar e executar suas atividades de maneira mais eficaz, otimizando a manutenção preventiva e evitando paradas não programadas dos motores ou da planta. Além disso, é possível priorizar os alarmes de determinados grupos, assegurando que alarmes mais críticos sejam atendidos prontamente, diante do grande número de registros administrados pelos operadores em tempo real.

AGRADECIMENTOS

Os autores agradecem à empresa EPASA - Centrais Elétricas da Paraíba por todo o suporte e fornecimento de insumos para o desenvolvimento deste trabalho. Também agradecem à UFPB (Universidade Federal da Paraíba) pelo apoio fornecido para a construção da pesquisa.

REFERÊNCIAS

Abhishekkumar, K. and Sadhana (2017). Survey on k-means clustering algorithm. URL https://api.semanticscholar.org/CorpusID:212560786.

Barbon, R. and Akabane, A.T. (2022). Analise de performance dos modelos gerais de aprendizado de maquina pre-treinados: Bert vs distilbert. *Anais Estendidos do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC Estendido 2022)*. URL https://api.semanticscholar.org/CorpusID:252665685.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Cai, S., Zhang, L., Palazoglu, A., and Hu, J. (2019). Clustering analysis of process alarms using word embedding. *Journal of Process Control*. URL https://api.semanticscholar.org/CorpusID:203137397.

Cordeiro, F., de Andrade Lira Rabelo, R., and Moura, R.S. (2022). Classification of irregularity communications in public ombudsmen using supervised learning algorithms. *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2022)*. URL https://api.semanticscholar.org/CorpusID:255894490.

Costa, A.S.P.d., Gonçalves, J.M.S., Neta, H.O., Alves, D.F.R., Lourenço, E., Franceschi, E., Dariva, C., Araujo, V., Venceslau, A., Leite, M.S., and Borges, G.R. (2019). Application of near-infrared for online monitoring of heavy fuel oil at thermoelectric power plants. part i: Development of chemometric models. *Industrial Amp; Engineering Chemistry Research*, 58, 15681–15692. doi:10.1021/acs.iecr.9b02107.

de Brito, P.F. and da Silva Gomes Gomes, L.P. (2019). Desenvolvimento do módulo de pre-processamento da ferramenta sentimentall. *Revista Singular - Engenharia, Tecnologia e Gestão*. URL https://api.semanticscholar.org/CorpusID:127133548.

Fahim, A., Salem, A., Torkey, F.A., and Ramadan, M. (2006). An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, 7, 1626–1633.

Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of bert and its transformative impact on natural language processing. . . . - *Informatics. Economics. Management*. URL https://api.semanticscholar.org/CorpusID:268206609.

Ilic, E., García-Martínez, M., and Pastor, M.S. (2022). A review of text classification models from bayesian to transformers. In *Swiss Text Analytics Conference*. URL https://api.semanticscholar.org/CorpusID:259121342.

Kachkou, D.I. (2021). Language modeling and bidirectional coders representations: an overview of key technologies. URL https://api.semanticscholar.org/CorpusID:234195800.

Lebret, R. (2016). Word embeddings for natural language processing. URL https://api.semanticscholar.org/CorpusID:63947950.

Martarelli, N.J. and Nagano, M.S. (2019). Agrupamento de dados mistos baseados no algoritmo k-means: Uma revisão sistemática da literatura. *Atas da conferência Ibero-Americana WWW/Internet 2019*. URL https://api.semanticscholar.org/CorpusID:214402019.

Mucherino, A., Papajorgji, P., and Pardalos, P.M. (2009). Clustering by k-means. URL https://api.semanticscholar.org/CorpusID:115167921.

Neelima, A. and Mehrotra, S. (2023). A comprehensive review on word embedding techniques. *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 538–543. URL https://api.semanticscholar.org/CorpusID:258219905.

Pirehgalin, M.F., Weiß, I., and Vogel-Heuser, B. (2020). Causal inference in industrial alarm data by timely clustered alarms and transfer entropy. *2020 European Control Conference (ECC)*, 2056–2061. URL https://api.semanticscholar.org/CorpusID:219933670.

quan Zhao, X. (2010). Clustering validity analysis based on silhouette coefficient. *Journal of Computer Applications*. URL https://api.semanticscholar.org/CorpusID:64280482.

Sai, L.N., Shreya, M.S., Subudhi, A.A., Lakshmi, B.J., and Madhuri, K.B. (2017). Optimal k-means clustering method using silhouette coefficient. *International Journal of Applied Research on Information Technology and Computing*, 8, 335–344. URL https://api.semanticscholar.org/CorpusID:188187032.

Subakti, A., Murfi, H., and Hariadi, N. (2021). The performance of bert as data representation of text clustering. *Journal of Big Data*, 9. URL https://api.semanticscholar.org/CorpusID:246654622.

Suleiman, D. and Awajan, A.A. (2018). Comparative study of word embeddings models and their usage in arabic language applications. *2018 International Arab Conference on Information Technology (ACIT)*, 1–7. URL https://api.semanticscholar.org/CorpusID:85496869.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*. URL https://api.semanticscholar.org/CorpusID:153312532.

SUN Ji-Gui, L.J. and Lian-Yu, Z. (2008). Clustering algorithms research. *Journal of Software*, 19(1), 48.

Takahashi, C.C. (2015). Mapeamento explícito como kernel em aprendizado de máquinas de vetores de suporte. URL https://api.semanticscholar.org/CorpusID:209992029.

van Aardt, D. (2015). More data is only useful if it leads to more wisdom. *IT in Manufacturing. URL: https://www. instrumentation. co. za/8423a (Acessado em 01-10-2021)*.

Yang, X. and Dziegielewski, B. (2007). Water use by thermoelectric power plants in the united states1. *JAWRA Journal of the American Water Resources Association*, 43, 160–169. doi:10.1111/j.1752-1688.2007.00013.x.