LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

COMPARAÇÃO DE MODELOS CLÁSSICOS E QUÂNTICOS DE NLP PARA CLASSIFICAÇÃO DE ACIDENTES ENVOLVENDO HIDROGÊNIO

Lucca Neves[a*], July Macedo[a], Caio Souto Maior[b], Márcio Moura[a], Isis Lins[a]

[a]Universidade Federal de Pernambuco, Recife, Brazil
[b]Universidade Federal de Pernambuco - CAA, Caruaru, Brazil

RESUMO

O gás hidrogênio é uma das novas tecnologias que buscam revolucionar a indústria energética, conforme seu uso se torna cada vez mais comum, é importante estar ciente dos possíveis perigos que um gás extremamente inflamável como o hidrogênio pode causar para o meio ambiente e população caso ocorra um acidente durante seu manuseio ou armazenamento. Inúmeros acidentes históricos comprovam os danos que um incêndio ou explosão causados pelo hidrogênio podem causar. Esse artigo procura explorar meios de analisar o grande volume de relatórios documentando acidentes com hidrogênio e extrair de forma ágil informações que sejam relevantes para a prevenção de acidentes. Através de modelos clássicos e quânticos de Natural Language Processing (NLP), tentamos treinar modelos que consigam classificar possíveis danos ambientais baseando-se em descrições de acidentes com hidrogênio, utilizando como base de treinamento a Base de Dados de Incidentes e Acidentes com Hidrogênio (HIAD) [WEN et al. 2022].

PALAVRAS-CHAVE. Relatórios de acidentes, Natural Language Processing, Computação Quântica

Tópicos:

ABSTRACT

Hydrogen gas is one of the new technologies that wish to revolutionize the energy industry, as its use becomes increasingly common, is important to be aware of the possible risks that an extremely flammable gas, such as hydrogen, may cause to the environment and people in case of an accident during transport or storage. Numerous historical accidents prove the damage that a fire or explosion may cause. This paper tries to explore means of analyzing the massive volume of reports documenting accidents involving hydrogen and quickly extract information that is relevant to accident prevention. Through both classical and quantum models of Natural Language Processing (NLP), we attempted to train models that could classify possible environmental damages based on descriptions of hydrogen accidents, using as a training base the Hydrogen Incidents and Accidents Database (HIAD) [Wen et al. 2022].

KEYWORDS. Accident reports, Natural Language Processing, Quantum Computing

Paper topics:

1. Introdução

O interesse pelo hidrogênio vem aumentando rapidamente ao longo dos últimos anos, seu uso como fonte de energia limpa diminuindo a necessidade de combustíveis fósseis é de interesse

crescente tanto da indústria, quanto da sociedade e governos ao redor do mundo. No entanto, diversas preocupações em relação às suas características, como sua elevada inflamabilidade e a propensão para induzir a degradação dos materiais [Shen et al. 2023] ainda representam barreiras para seu uso mais pleno.

Acidentes causados pela combustão, explosão ou vazamento de gás hidrogênio podem causar danos severos ao meio ambiente e à sociedade. Diversos acidentes famosos evidenciam a devastação que acidentes desse tipo podem causar. Um dos exemplos mais famosos é o incêndio do dirigível alemão Zeppelin Hindenburg, que ao pegar fogo durante o pouso causou a morte de pelo menos 35 pessoas em 1937 [Cadwallader e Herring, 1999]. Evitar acidentes como esse são de extrema importância e com esse objetivo, a Análise de Risco (AR) é uma ferramenta muito útil para ajudar a orientar investimentos e esforços que possam prever e minimizar possíveis acidentes futuros.

Uma das maneiras de se encontrar possíveis causas comuns a acidentes é através da análise de relatórios escritos após os eventos, no entanto a quantidade de documentos disponíveis faz com que uma análise manual seja impraticável [Macêdo et al. 2022]. Para facilitar a análise de informações de interesse, técnicas de NLP podem ser uma ferramenta útil, permitindo uma análise do conjunto de dados de forma ágil e produzindo interpretações possivelmente úteis dos textos. Com o treinamento de modelos de NLP é possível analisar uma grande quantidade de dados e realizar rapidamente a classificação de alguma característica de interesse presente em seu meio, e assim, possivelmente, até prever se existe tal característica em um novo texto, que não existe no conjunto original.

2. Metodologia

Para classificar a presença, ou não, de danos ambientais em descrições de acidentes envolvendo hidrogênio, seis modelos clássicos de NLP e um modelo quântico foram treinados. Os modelos clássicos combinam algoritmos de aprendizado de máquina (ML) com modelos de linguagem, enquanto o modelo quântico utiliza um ansatz específico.

Primeiramente, todas as etapas foram implementadas em Python, os textos foram filtrados da base de dados do HIAD 2.0, utilizando a biblioteca Pandas, focando na coluna ‘Environmental damage’ (Danos ambientais). Foram identificadas 64 instâncias, sendo 12 com danos ambientais e 52 sem danos ambientais. Cada texto passou por um pré-processamento, através de funções fornecidas pela biblioteca NLTK, que incluiu a remoção de pontuações e *stop words* (i.e., palavras que podem ser removidas sem gerar perda de informação e contexto), seguido de um resumo por TextRank para reduzir cada texto a uma única sentença. Adicionalmente, para o modelo quântico, cada texto resumido foi ainda mais reduzido para 20 palavras selecionadas aleatoriamente.

Para comparação da performance dos modelos, três conjuntos de dados foram criados. O primeiro conjunto continha apenas os textos pré-processados. O segundo conjunto incluía textos pré-processados com informações adicionais do HIAD 2.0, como ‘Tipo de liberação’, ‘Substância liberada’, ‘Quantidade de liberação’, ‘Fonte de ignição’, ‘Explosão de alta pressão’ e ‘Tipo de chama’. O terceiro conjunto era similar ao segundo, mas com *data augmentation* nas instâncias com danos ambientais, equilibrando a proporção de textos (12 instâncias com danos ambientais e 52 instâncias sem danos ambientais) através da substituição de palavras por sinônimos.

Os modelos clássicos consistiam na combinação de três algoritmos de ML com dois modelos de linguagem, resultando em seis combinações. Os algoritmos de ML utilizados foram Naive Bayes, Random Forest e Support Vector Machine (SVM). O Naive Bayes [Zhang 2004] utiliza o Teorema de Bayes para calcular a probabilidade de as características pertencerem às classes, considerando-as independentes. O Random Forest [Breiman 2001] cria múltiplas árvores de decisão a partir de subconjuntos variados dos dados, decidindo a classe com base na maioria dos votos das árvores. O SVM [Chang e Lin 2011]separa as classes com uma reta, maximizando a distância dos pontos mais próximos de cada classe.

Os modelos de linguagem utilizados foram Bag-of-Words (BoW) [Silva et al. 2018] e Term Frequency – Inverse Document Frequency (TF-IDF)[Chen 2024]. O BoW cria uma matriz contando o número de ocorrências de cada palavra em cada sentença. O TF-IDF calcula a frequência da palavra na sentença multiplicada pelo logaritmo natural do inverso da frequência de documentos onde a palavra aparece. O modelo quântico utilizou o IQPAnsatz, um método que transforma um diagrama de strings (representação matemática da sentença) em um circuito quântico. Em outras palavras, os textos pré-processados foram convertidos pelos modelos matemáticos BoW e TF-IDF, que transformam textos em matrizes numéricas para análise pelos modelos clássicos de NLP. Enquanto o IQPAnsatz transformou diagramas de strings em circuitos quânticos para o modelo quântico.

Os três conjuntos de dados descritos anteriormente foram divididos em treino, teste e validação. Portanto, tanto os modelos clássicos quanto o modelo quântico foram treinados, testados e validados em três cenários. A seguir serão apresentados os principais resultados e discussões.

3. Resultados

Para os modelos clássicos todos os testes foram realizados com o mesmo conjunto de teste, contendo seis sentenças sem danos ambientais e uma com danos ambientais. Já o conjunto de teste do modelo quântico foi avaliado por meio de uma validação cruzada de 5 vezes em todos os conjuntos de dados.

De maneira consistente, tanto nos modelos clássicos quanto no modelo quântico, o primeiro e segundo conjuntos de textos (sem *data augmentation* em ambos, com informações extras do HIAD no segundo) obtiveram acurácias médias de 85,71%, prevendo de maneira correta todas as instâncias que não possuíam danos ambientais. No entanto, de maneira comum a todos os testes, classificavam de maneira incorreta as instâncias que possuíam danos ambientais.

O terceiro conjunto de dados, com suas instâncias aumentadas para resolver o desequilíbrio das classes, foi o único conjunto a obter resultados diferentes. No modelo quântico a acurácia média foi de 77,13%, apesar do valor mais baixo, o resultado obtido foi mais balanceado, com apenas uma das cinco rodadas errando a classificação da instância com danos ambientais. Nos modelos clássicos, porém, apenas um dos seis modelos, Naive Bayes com Bag-of-Words, acertou a classificação da instância com danos ambientais, enquanto todos os outros modelos sofreram do mesmo problema dos conjuntos anteriores.

4. Conclusão

Com esses resultados, é possível perceber uma limitação na habilidade dos modelos em discernir a presença de danos ambientais a partir das sentenças utilizadas. E apenas no terceiro conjunto, que teve uma de suas categorias aumentadas, alguns dos modelos conseguiram classificar a presença de danos ambientais de forma correta. Essa melhora sugere que técnicas de aumento de dados pode contribuir na robustez dos modelos.

O modelo quântico ainda apresenta limitações importantes. A acurácia média de 77,13% no conjunto de dados balanceado é inferior à dos modelos clássicos, indicando que há espaço para melhorias na precisão geral do QNLP. Além disso, a necessidade de reduzir os textos para 20 palavras devido as limitações do modelo quântico podem resultar em perda de informação crítica, afetando negativamente o desempenho do modelo.

Por outro lado, os resultados destacam o potencial do QNLP em abordar problemas que envolvem desequilíbrios de classe, uma limitação significativa observada nos modelos clássicos. A capacidade do modelo quântico de fornecer uma classificação mais equilibrada, mesmo com uma acurácia média geral menor, sugere que técnicas de QNLP podem oferecer uma abordagem mais robusta em cenários onde os dados são altamente desbalanceados.

Referências

BREIMAN, L. Random Forests. Machine Learning, 2001. v. 45, n. 1, p. 5–32.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011. v. 2, n. 3, p. 27:1–27:27.
CHEN, L.-C. An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus. Data & Knowledge Engineering, set. 2024. v. 153, p. 102322.

L. C. Cadwallader e J. S. Herring. Safety Issues with Hydrogen as a Vehicle Fuel. Idaho Falls, ID: [s.n.], 1999.

Macêdo, J. B. et al. Identification of risk features using text mining and BERT-based models: Application to
an oil refinery. Process Safety and Environmental Protection, fev. 2022. v. 158, p. 382–399.

SILVA, F. B. *et al*. Graph-based bag-of-words for classification. Pattern Recognition, fev. 2018. v. 74, p. 266–285.

Shen, Y. et al. Preliminary hazard identification for qualitative risk assessment on onboard hydrogen storage and supply systems of hydrogen fuel cell vehicles. Renewable Energy, ago. 2023. v. 212, p. 834–854.

Wen, J. X. et al. Statistics, lessons learned and recommendations from analysis of HIAD 2.0 database. International Journal of Hydrogen Energy, maio. 2022. v. 47, n. 38, p. 17082–17096.

ZHANG, H. The Optimality of Naive Bayes. [S.l.]: [s.n.], 2004. Disponível em: <https://api.semanticscholar.org/CorpusID:8891634>.