Gustavo Henrique dos Reis – 0030481821041

SEAD – Sistema de Endereçamento Automático Documentos

O estudo de Falcão, Lopes e Souza (2022) analisou o uso de Processamento de Linguagem Natural (PLN) em diversas áreas do conhecimento, destacando sua crescente acessibilidade devido à redução dos custos de dados. O autor destaca que na área de Computação e Linguagem, há um uso significativo de ferramentas de PLN, especialmente transformers como nas plataformas de IA generativas *BERT* e *LSTM*. Ressalta, no entanto, que a Ciência da Informação utiliza pouco as técnicas de PLN, com poucos artigos identificados. O estudo sugere que a Ciência da Informação poderia aproveitar as melhorias de PLN para aprimorar a gestão do conhecimento e a recuperação de informações destacando então o potencial de integração entre PLN e Ciência da Informação para otimizar o uso de informações.

FALCÃO, L. C. DE J.; LOPES, B.; SOUZA, R. R. **Absorção das tarefas de processamento de Linguagem Natural (PLN) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de PLN pela CI**. Em Questão, v. 28, n. 1, p. 13–34, 1 jan. 2022.

A pesquisa de Rodríguez e Bezerra (2020) objetivou analisar textos do Diário Oficial da União (DOU) para o processamento de dados e interesse sobre essa publicação, utilizando Processamento de Linguagem Natural (PLN). Foi utilizada a plataforma NLTK em Python e exploradas diversas bibliotecas de processamento de texto. O experimento focou identificar nomes próprios para correlacionar essas publicações à nomeação de cargos e outras referências. A análise semântica das sentenças permitiu a extração de Nomes de Pessoas por meio das Técnicas de *tokenização e pos-tagging*. A conclusão revelou uma acurácia de 92% na extração de Nomes de Pessoas, considerada aceitável.

RODRÍGUEZ, M. M.; DANTAS BEZERRA, B. L. **Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias).** **Revista de Engenharia e Pesquisa Aplicada**, v. 5, n. 1, p. 67–77, 2020.

O Trabalho de Andrade (2015) teve como principal objetivo realizar uma prova de conceito para a construção de um modelo automático de triagem de denúncias recebidas pela Controladoria-Geral da União (CGU), classificando automaticamente as denúncias possibilitando um ganho efetivo de tempo e recursos. Para a classificação, as denúncias cadastradas no GCU ou enviadas por e-mail ou cartas passam por uma triagem para excluir mensagens irrelevantes. Após essa primeira fase, são analisadas novamente para determinar se são ou não de competência do referido órgão, sendo habilitadas ou não-habilitadas. As habilitadas serão encaminhadas, conforme a matéria, à unidade competente da CGU para apuração, normalmente, uma das coordenações da Secretaria Federal de Controle Interno (SFC) ou da Corregedoria Geral da União (CRG) e serão cadastradas no SGI, recebendo um número de protocolo. Os resultados mostram que os primeiros conjuntos de dados trabalhados obtiveram uma precisão de 0,41. O autor afirma que apesar da melhora alcançada no modelo proposto, os índices de acerto da triagem manual não foram superados. Isso porque, ao avaliar a quantidade de reencaminhamentos de denúncias triadas para uma nova área, percebe-se que esta taxa de erro fica próxima de 10%.

HELENA, P.; ALVES DE ANDRADE, M. **Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU**.

[s.l: s.n.]. Disponível em: -http://www.real.unb.br/jspui/bitstream/10482/21004/1/2015\_Patr%C3%ADciaHelenaMaiaAlvesdeAndrade.pdf>. Acesso em: 22 mar. 2024.

‌O estudo de Jesus et al. (2023) avaliou o potencial do processamento de Linguagem Natural (PLN) e do *Machine Learning (ML)* para a categorização de trabalhos científicos, comparando duas abordagens: categorização supervisionada e não supervisionado, sendo realizado em duas etapas. Na primeira etapa, um corpus de 37 artigos sobre patrimônio cultural foi categorizado manualmente em duas categorias: A (relacionado ao âmbito digital) e B (não relacionado). Algoritmos de ML foram treinados e testados e a categorização não supervisionada foi realizada usando um algoritmo de clusterização hierárquica. Na segunda etapa, uma estratégia de busca foi utilizada para recuperar mais artigos sobre patrimônio cultural, criando duas subamostras: A1 (450 artigos) e A2 (100 artigos) categorizadas manualmente. Os algoritmos de ML foram novamente treinados e testados, e a categorização não supervisionada foi realizada na subamostra A2. Na primeira etapa, a categorização supervisionada obteve alta acurácia (acima de 90%), enquanto a categorização não supervisionada identificou 5 categorias que representam diferentes aspectos do patrimônio cultural. Na segunda etapa, a categorização supervisionada na subamostra A1 obteve baixa acurácia (cerca de 47%), enquanto na subamostra A2, após refinamento manual, obteve melhor acurácia (cerca de 74%). A categorização não supervisionada na subamostra A2 identificou 15 categorias que representam diferentes temas e subtemas relacionados ao patrimônio cultural. O estudo mostra que as técnicas de PLN e ML são promissoras para a categorização de trabalhos científicos com a categorização supervisionada sendo uma ferramenta útil para automatizar a categorização, porém é dependente da qualidade da amostra de treino e das características das categorias e a categorização não supervisionada podendo ser uma ferramenta útil para explorar um corpus de documentos e identificar novas categorias e padrões.

JESUS, A. F. DE et al. **Processamento de linguagem natural e machine learning na categorização de artigos científicos: um estudo em torno do “patrimônio cultural”**. Revista Ibero-Americana de Ciência da Informação, v. 16, n. 1, p. 167–184, 27 mar. 2023.

O trabalho de Leme (2021) buscou identificar o método com melhor desempenho de precisão e acurácia para classificação supervisionada do rito dos processos de atos de concentração avaliados pelo Conselho Administrativo de Defesa Econômica (CADE) utilizando ferramentas e métodos como Representação Vetorial de Palavras, *Machine Learning, Deep Learning, Redes Neurais, BOW, transformers, LSTM e BERT*. Por meio da plataforma SEI (Serviço Eletrônico de Informações), foram realizados testes com as ferramentas propostas, combinando arquiteturas e treinandos IAs para a classificação. Os autores afirmam que a classificação apresentou uma alta precisão em alguns dos testes, concluindo que, mesmo com o avanço da tecnologia, encontram-se alguns problemas com textos grandes. Afirmam ainda que técnicas mais avançadas não necessariamente apresentam resultados melhores que técnicas mais simples e mostra características de diferentes técnicas que podem ser combinadas em conjunto em uma rede neural, para aumentar a diversidade de tipos de padrões que um modelo pode aprender.

LEME, B. **Classificação automática de documentos de características econômicas para defesa jurídica**. Disponível em: <https://www.teses.usp.br/teses/disponiveis/45/45134/tde-05082021-152340/en.php>. Acesso em: 23 mar. 2024.

O Artigo de Silva e Martins (2019) propõe uma nova abordagem para a classificação automática de documentos (ADC) que integra os métodos *MetaLazy e FaSST kNN*, para a eficiência e a acurácia da classificação. Os autores exploram o aprendizado postergado e a geração de características de coocorrência, criação de novas características de coocorrência, com foco em termos relevantes do documento de consulta com o *MetaLazy* junto com o FaSST kNN, que utiliza técnicas de filtragem e compactação para otimizar o tempo de busca. Segundo os autores, os resultados demonstram que a integração dos metodos traz a uma aceleração do tempo de classificação, com ganhos de desempenho de até 4 vezes em comparação ao original, aumentando também a precisão das classificações em alguns *datasets.*

SILVA, P. H. DA; MARTINS, W. S. **Acelerando o pré-processamento de consultas em sistemas de classificação automática de documentos usando aprendizado postergado**. Disponível em: <https://sol.sbc.org.br/index.php/erigo/article/view/9116/9662>. Acesso em: 23 mar. 2024.

O trabalho de Ferreira (2019) visou a construir um sistema para processamento de dados textuais do Diário Oficial da União (DOU), para extrair, converter, formatar, armazenar e processar dados de diversas fontes para análise e modelos de extração de informação e conhecimento. A arquitetura permitirá integração com módulos de inteligência para replicação do comportamento de interpretação dos dados e identificação de fraudes e anomalias. Para a extração desses dados o autor utilizou a técnicas como *crawling (rastreamento) e Scraping (raspagem de dados),* conceitos de PLN Random Forests, Logistic Regression, Suport Vector Machine. Os resultados descritos pelos autores mostram ser mais precisos na extração de dados derivados de HTML, sem a dependência de outras ferramentas, ao contrário de PDFs.

FERREIRA, H. **Processamento de Linguagem Natural e Classificação de textos em Sistemas Modulares**. [s.l: s.n.]. Disponível em: <https://bdm.unb.br/bitstream/10483/25114/1/2019\_HugoHondaFerreira\_tcc.pdf>.

O trabalho de Oliveira et al. (2022) aborda a importância do processamento automático de textos na era digital, e destaca a necessidade de compreender e extrair informações de uma vasta quantidade de dados textuais gerados diariamente em diversas fontes. O autor descreve o desenvolvimento histórico do Processamento de Linguagem Natural (PLN) e o papel fundamental dos modelos de Aprendizagem Profunda e visa a explorar como as representações textuais podem ser processadas por meio de técnicas de Aprendizagem Profunda para resolver problemas de PLN e tem por objetivo introduzir diferentes formas de representar textos em vetores numéricos, desde abordagens tradicionais como *bag-of-words* até *word embeddings*, exemplificar o uso de redes neurais profundas para resolver tarefas de PLN, como Classificação de Textos e Sumarização de Sentenças com bibliotecas como *Numpy, Keras*, *Scikit-learn, Matplotlib e Seaborn* para construção de gráficos. O trabalho apresenta alguns resultados de trabalhos realizados em parceria com o setor público no contexto da pandemia da Covid-19, destacando o uso de soluções de Processamento de Linguagem Natural (PLN) para lidar com grandes volumes de dados textuais e fornecer insights relevantes. De acordo com os autores os resultados da pesquisa demonstram como as tecnologias de PLN podem ser aplicadas em diferentes contextos, desde a análise de dados da Segurança Pública até o enfrentamento de crises de saúde pública, como a pandemia da Covid-19, contribuindo para tomadas de decisão mais eficientes e melhorias nos serviços prestados à população.

Oliveira, Bárbara Stéphanie Neves, et al. **"Processamento de linguagem natural via aprendizagem profunda."**Sociedade Brasileira de Computação (2022).

O trabalho de Oliveira Junior (2018) teve por objetivos verificar a acurácia da classificação de documentos jornalísticos em português e mensagens de grupos de discussão em inglês em categorias pré-definidas, utilizando características extraídas por meio de técnicas de processamento de linguagem natural (NLP) e alimentando classificadores de redes neurais. Além disso ele realizou uma variedade de testes utilizando diferentes técnicas de geração de vetores (TF-IDF e Doc2Vec) e aplicando-as em diferentes arquiteturas de redes neurais. Os resultados extraídos destacam que o uso de redes neurais podem produzir resultados satisfatórios mesmo com bases de dados compostas por poucos documentos. Além disso foi observado que a escolha dos parâmetros adequados para essas abordagens ainda é um desafio, dado o grande número de opções disponíveis e a escolha entre acurácia e tempo de treinamento.

DE OLIVEIRA JUNIOR, W. R.**O USO DE REDES NEURAIS CONVOLUCIONAIS E TOTALMENTE CONECTADAS PARA A CATEGORIZAÇÃO DE TEXTOS EM IDIOMAS PORTUGUÊS E INGLÊS**. [s.l: s.n.]. Disponível em: <https://ppgia.pucpr.br/pt/arquivos/doutorado/teses/2018/054\_Tese\_Walter%20Ribeiro%20de%20Oliveira%20Junior.pdf>. Acesso em: 10 abr. 2024.

O trabalho de Rocha (2019) teve como objetivo geral implementar um modelo de CNN para classificar documentos de identificação brasileiros. Para a realização da pesquisa foi necessária a criação de um conjunto de dados, a aplicação de técnicas de *data augmentation* e a utilização de frameworks em Python para a implementação do modelo. Os resultados foram que as arquiteturas de redes neurais convolucionais alcançaram aproximadamente 97% de acurácia na classificação dos documentos, com destaque para a arquitetura *C1P1TC2b*, que obteve uma acurácia de 98,59% no momento da validação. O autor observou também que o a análise da matriz de confusão é importante para verificar os casos de falsos negativos e falsos positivos para corrigir possíveis erros de rotulagem no conjunto de dados.

ROCHA, A. N. DA. **Classificação de documentos de identificação com redes neurais convolucionais**. Disponível em: <https://repository.ufrpe.br/handle/123456789/3952>. Acesso em: 12 abr. 2024.