INSTITUTO FEDERAL DO ESPÍRITO SANTO
CURSO SUPERIOR DE SISTEMAS DE INFORMAÇÃO

CRISTHIAN FONTANA MATTIUZZI

TRIAGEM AUTOMATIZADA DE CURRÍCULOS USANDO BUSCA VETORIAL E MODELOS DE LINGUAGEM DE LARGA ESCALA

Serra
2025

CRISTHIAN FONTANA MATTIUZZI

TRIAGEM AUTOMATIZADA DE CURRÍCULOS USANDO BUSCA VETORIAL E MODELOS DE LINGUAGEM DE LARGA ESCALA

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Sistemas de Informação do Instituto Federal do Espírito Santo, Campus Serra, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Hilário Tomaz Alves de Oliveira

Serra
2025

Dados Internacionais de Catalogação na Publicação (CIP)

M444t 2025

Mattiuzzi, Cristhian Fontana
Triagem automatizada de currículos usando busca vetorial e modelos de linguagem de larga escala / Cristhian Fontana Mattiuzzi - 2025.
35 f.; il.; 30 cm

Orientador: Prof. Dr. Hilário Tomaz Alves de Oliveira.

Monografia (graduação) - Instituto Federal do Espírito Santo, Coordenadoria do Curso de Bacharelado em Sistemas de Informação, 2025.

1. Inteligência artificial. 2. Currículos - triagem. 3. Busca semântica. 4. Modelos de linguagem de grande escala (LLMs). 5. Vetorização. I. Oliveira, Hilário Tomaz Alves de. II. Instituto Federal do Espírito Santo. III. Título.

CDD 004

Bibliotecário: Valmir Oliveira de Aguiar - CRB6/ES 566

CRISTHIAN FONTANA MATTIUZZI

TRIAGEM AUTOMATIZADA DE CURRÍCULOS USANDO BUSCA VETORIAL E MODELOS DE LINGUAGEM DE LARGA ESCALA

Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de Bacharel em Sistemas de Informação, do curso de Bacharelado em Sistemas de Informação do Instituto Federal do Espírito Santo.

Aprovado em 16 de julho de 2025

COMISSÃO EXAMINADORA

Documento assinado digitalmente
gov.br HILARIO TOMAZ ALVES DE OLIVEIRA
Data: 06/08/2025 14:27:10-0300
Verifique em https://validar.iti.gov.br

_____

Hilário Tomaz Alves de Oliveira (Orientador)
Instituto Federal do Espírito Santo
Campus Serra

Documento assinado digitalmente
gov.br BRUNO CARDOSO COUTINHO
Data: 06/08/2025 18:12:37-0300
Verifique em https://validar.iti.gov.br

_____

Bruno Cardoso Coutinho
Instituto Federal do Espírito Santo
Campus Serra

Documento assinado digitalmente
gov.br MATEUS CONRAD BARCELLOS DA COSTA
Data: 07/08/2025 07:58:43-0300
Verifique em https://validar.iti.gov.br

_____

Mateus Conrad Barcellos da Costa
Instituto Federal do Espírito Santo
Campus Serra

AGRADECIMENTOS

A conclusão deste trabalho representa não apenas o encerramento de um ciclo acadêmico, mas também o resultado de uma jornada marcada por desafios, aprendizados e apoio de muitas pessoas às quais sou profundamente grato.

Agradeço ao meu orientador, Prof. Dr. Hilário Tomaz Alves de Oliveira, pela orientação segura, pelos conselhos valiosos e pela confiança depositada neste projeto. Sua experiência e disponibilidade foram fundamentais para a evolução do trabalho.

Aos colegas e amigos que, direta ou indiretamente, contribuíram com sugestões, críticas construtivas e apoio ao longo do curso. Em especial, aos que compartilharam experiências semelhantes e me motivaram a seguir mesmo nos momentos mais difíceis.

À minha mãe, pelo amor incondicional, incentivo constante e compreensão nos momentos de ausência. Sem o seu apoio, este trabalho não teria sido possível.

Agradeço também aos professores do curso, pela dedicação ao ensino e pela inspiração que proporcionaram ao longo da minha formação.

Por fim, dedico este trabalho a todos que acreditam na educação como instrumento de transformação e que me acompanharam nessa caminhada.

RESUMO

O processo de triagem representa uma etapa estratégica no recrutamento e seleção de colaboradores em organizações. Diante disso, as empresas têm buscado soluções que aprimorem seus procedimentos, com o objetivo de torná-los mais assertivos. Nesse contexto, abordagens baseadas em Inteligência Artificial apresentam potencial para contribuir com esse processo. Este trabalho investiga o desempenho de um sistema de triagem de currículos baseado na combinação entre busca vetorial semântica e modelos de linguagem de larga escala (LLMs). O sistema utiliza *embeddings* gerados por modelos de transformação de sentenças para representar semanticamente currículos em um banco vetorial, permitindo a recuperação de documentos similares com base em um currículo de referência. Em seguida, os currículos recuperados são analisados por um LLM, que ranqueia os candidatos mais aderentes ao perfil buscado. O conjunto de dados utilizado nos experimentos é o *Resume Corpus*, composto por diversos currículos da área de tecnologia, avaliados em três versões: original, resumida com o modelo *LLaMA 3.1 8B* e resumida com o modelo *Pegasus*. Os experimentos demonstraram que a estratégia híbrida aumenta a precisão da triagem sem comprometer a eficiência, desde que a compressão textual preserve a densidade semântica. Os resultados apontam a viabilidade do uso combinado de vetorização e LLMs em processos seletivos automatizados, o que pode contribuir para a redução do tempo de análise e maior assertividade na identificação de talentos.

Palavras-chave: Triagem de currículos. Busca semântica. Modelos de linguagem. Vetorização. LLM.

ABSTRACT

The screening process constitutes a strategic phase in the recruitment and selection of employees within organizations. Consequently, companies have increasingly sought solutions to enhance their procedures, aiming to improve their accuracy and effectiveness. In this context, Artificial Intelligence-based approaches have the potential to support and optimize this process. This work investigates the performance of a resume screening system that combines semantic vector search with Large Language Models (LLMs). The system uses embeddings generated by sentence transformation models to semantically represent resumes in a vector database, allowing the retrieval of similar documents based on a reference resume. Then, the retrieved resumes are analyzed by an LLM, which ranks the candidates that best fit the sought profile. The dataset used in the experiments is the Resume Corpus, composed of several resumes from the technology area, evaluated in three versions: original, summarized with the LLaMA 3.1 8B model, and summarized with the Pegasus model. The experiments demonstrated that the hybrid strategy increases screening accuracy without compromising efficiency as long as textual compression preserves semantic density. The results indicate the viability of combining vectorization and LLMs in automated selection processes, which can contribute to reducing analysis time and increasing accuracy in talent identification.

Keywords: Resume screening. Semantic search. Language models. Vectorization. LLM.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Recursos humanos, frequentemente abreviado como RH, refere-se ao departamento ou função de uma organização responsável pela gestão, aquisição, treinamento, avaliação e retenção de funcionários. O papel dessa área está na determinação de como as pessoas são integradas e contribuem para o sucesso de uma empresa (Taylor; Armstrong, 2020). Tradicionalmente, o setor de RH se concentrava em tarefas administrativas, como contratação e folha de pagamento. No entanto, com a evolução do mercado e das organizações, essa função se expandiu para incluir áreas estratégicas como desenvolvimento de talentos, engajamento dos funcionários e alinhamento da força de trabalho com os objetivos da empresa. Os profissionais de RH, portanto, atuam como parceiros estratégicos, facilitadores e mediadores, garantindo que os talentos da organização sejam efetivamente utilizados para alcançar os objetivos desejados (Dessler; Oderich, 2017).

A classificação eficiente de currículos é fundamental para empresas de recursos humanos, pois representa o primeiro filtro em um processo seletivo e determina a qualidade dos candidatos que avançam para as etapas subsequentes (Grupo SERES, 2023). No cenário atual, em que a quantidade de candidatos por vaga cresce devido ao alcance e facilidade das plataformas digitais, a capacidade de identificar rapidamente os candidatos mais qualificados torna-se essencial. Uma classificação eficaz otimiza o tempo de seleção, reduz custos operacionais e aumenta a probabilidade de identificar indivíduos que realmente atendam às necessidades e cultura da empresa. Além disso, ao garantir que os candidatos mais adequados sejam entrevistados, as empresas não apenas melhoram suas chances de fazer contratações bem-sucedidas, mas também fortalecem sua reputação como empregadores desejáveis, atraindo talentos de alta qualidade no futuro (Breaugh, 2008).

A análise manual, além de consumir muito tempo, é suscetível a erros e inconsistências. Para superar esses desafios, diversas abordagens têm sido adotadas pelas empresas. Uma dessas abordagens são os sistemas do tipo *Applicant Tracking System* (ATS), que são comumente usados para filtrar e classificar currículos com base em palavras-chave e critérios predefinidos. No entanto, esses sistemas tradicionais podem descartar candidatos qualificados que não se encaixam exatamente nos critérios estabelecidos. Abordagens mais recentes empregam técnicas de Inteligência Artificial (IA) e Aprendizado de Máquina (AM), permitindo uma análise mais sofisticada e adaptativa dos currículos. Esses métodos podem identificar padrões e correlações não óbvias, proporcionando uma seleção mais precisa e abrangente de candidatos (Chapelle; Schölkopf; Zien, 2011; Fernandez *et al.*, 2020).

Abordagens baseadas em Inteligência Artificial, especialmente aquelas que utilizam Modelos de Linguagem de Larga Escala (LLMs, do inglês *Large Language Models*), têm ganhado destaque por sua capacidade de interpretar currículos para além da literalidade textual

(Bommasani *et al.*, 2021). Diferentemente dos sistemas tradicionais baseados em palavras-chave, essas soluções consideram o significado contextual das experiências profissionais e a relação semântica entre diferentes competências. Técnicas como a vetorização semântica de sentenças, com o uso de modelos como os *Sentence Transformers* (Reimers; Gurevych, 2019), e a aplicação de LLMs da família *LLaMA*, como o *Meta LLaMA 3.1 8B Instruct*, possibilitam análises mais profundas, permitindo a identificação de similaridades entre documentos com base em significado e não apenas em termos exatos.

Essa integração entre recuperação semântica e análise contextual é a base de arquiteturas híbridas como o *Retrieval-Augmented Generation* (RAG), que combinam mecanismos de busca vetorial com modelos generativos para sintetizar respostas mais precisas a partir de documentos relevantes (Lewis *et al.*, 2020). No contexto da triagem de currículos, essa abordagem permite recuperar perfis semanticamente próximos com base em um currículo de referência e, em seguida, aplicar LLMs para realizar uma análise comparativa aprofundada. Como demonstrado nos experimentos deste trabalho, essa estratégia combinada mostrou-se eficaz para reduzir perdas causadas por variações de vocabulário ou estrutura textual, promovendo uma triagem mais eficiente e assertiva.

A representação vetorial de textos permite que currículos sejam comparados com base em seu conteúdo semântico, superando as limitações de filtros lexicais rígidos. Essa abordagem viabiliza a identificação de perfis similares mesmo quando a terminologia empregada varia entre os candidatos. Em combinação com LLMs, que realizam análise contextual sobre os documentos mais relevantes, forma-se uma estratégia híbrida que potencializa a assertividade na triagem automatizada.

No presente trabalho, propõe-se o desenvolvimento e a avaliação de um sistema híbrido para triagem de currículos, baseado na integração entre mecanismos de busca vetorial e modelos de linguagem. O sistema recebe como entrada um currículo de referência, geralmente representando um perfil ideal para determinada vaga, e, a partir disso, realiza a consulta no banco vetorial, retornando os documentos semanticamente similares. Em seguida, os currículos recuperados são processados por um LLM, que analisa o conteúdo e gera uma resposta estruturada com base na similaridade contextual, indicando os candidatos mais adequados. O sistema também permite a experimentação com diferentes modelos de vetorização e LLMs, além de suportar versões resumidas dos currículos com o uso de técnicas de sumarização automática.

Para garantir uma avaliação da proposta, foi utilizado o conjunto de dados *Resume Corpus*, uma base de dados multirótulo composta por mais de vinte mil currículos reais da área de tecnologia da informação, cada um associado a uma ou mais categorias profissionais (Jiechieu; Tsopze, 2020). Com base nesse corpus, foram criadas três versões distintas da base de dados: a versão original, uma versão com os currículos resumidos pelo modelo

Pegasus (Zhang *et al.*, 2020) e outra com o modelo *Meta LLaMA 3.1 8B Instruct*. Essas versões permitiram observar o impacto da compactação textual sobre a densidade semântica dos documentos e, por consequência, sobre o desempenho dos modelos de triagem. As reduções de conteúdo foram avaliadas tanto do ponto de vista da busca vetorial quanto da análise contextual, permitindo identificar limites e oportunidades no uso de sumarizações em sistemas reais de triagem.

Dessa forma, o presente trabalho busca contribuir com a análise de técnicas de triagem automática de currículos ao explorar, de forma integrada, modelos de vetorização semântica, LLMs e estratégias de sumarização. O sistema desenvolvido busca promover maior eficiência no processo seletivo, reduzindo o tempo necessário para a triagem e aumentando a precisão na identificação de candidatos mais adequados. Além disso, os experimentos conduzidos fornecem evidências empíricas sobre o desempenho das diferentes combinações de modelos e versões da base de dados, gerando insumos para a mensuração da eficácia dessas tecnologias.

1.1 LIMITAÇÕES

Embora o sistema proposto tenha demonstrado resultados promissores na triagem automática de currículos por meio da combinação entre busca vetorial e LLMs, algumas limitações devem ser reconhecidas. Primeiramente, o enfoque do trabalho está restrito à análise semântica de informações textuais contidas nos currículos, não contemplando fatores subjetivos ou qualitativos, como o alinhamento cultural entre o candidato e a organização contratante. Além disso, por depender de modelos pré-treinados e bases resumidas, o desempenho do sistema pode ser sensível à perda de informações relevantes durante o processo de sumarização. Outro ponto a considerar é a limitação da base de dados utilizada, composta exclusivamente por currículos da área de tecnologia da informação, o que restringe a generalização dos resultados para outros domínios profissionais. É importante ressaltar que o uso de LLMs em plataformas externas impõe restrições quanto à quantidade de *tokens*, custos de operação e questões relacionadas à privacidade dos dados. Por fim, cabe ressaltar que não foi realizado nenhum experimento em ambiente real com especialistas do setor de RH.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver e avaliar um sistema inteligente de triagem automática de currículos, combinando mecanismos de busca vetorial com modelos de transformação de sentenças e análise contextual com modelos de linguagem de larga escala (LLMs).

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral proposto, são definidos os seguintes objetivos específicos:

- Construir um banco vetorial de currículos a partir de *embeddings* gerados por diferentes modelos de *Sentence Transformers*;
- Implementar um sistema de consulta com base em similaridade semântica e análise contextual;
- Avaliar o impacto de diferentes técnicas de sumarização no desempenho da recuperação semântica;
- Integrar e comparar múltiplos LLMs para ranqueamento dos resultados mais relevantes;
- Avaliar a eficácia da estratégia combinada (busca vetorial + LLM) por meio de métricas de acurácia, tomando os rótulos do conjunto de dados Resume Corpus como referência;

1.3 ESTRUTURA DO TRABALHO

Este trabalho está organizado em cinco capítulos. O Capítulo 1 apresenta a introdução, delineando o problema, os objetivos, a motivação e a justificativa da pesquisa. O Capítulo 2 fornece os fundamentos teóricos relacionados à busca semântica, vetorização de textos e modelos de linguagem de larga escala, oferecendo o embasamento técnico necessário. No Capítulo 3, é descrito o sistema proposto, os dados utilizados e a metodologia experimental. O Capítulo 4 apresenta os resultados obtidos e as análises comparativas entre os modelos testados. Por fim, o Capítulo 5 expõe as conclusões, destacando as contribuições do trabalho, suas limitações e propostas para futuras pesquisas.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os fundamentos teóricos que embasam o desenvolvimento do sistema de triagem de currículos proposto neste trabalho. São discutidos os conceitos de vetorização semântica de textos, mecanismos de busca vetorial e os modelos de linguagem de larga escala (LLMs), com ênfase em suas aplicações no contexto de recuperação de informações e análise contextual de documentos textuais.

2.1 VETORIZAÇÃO SEMÂNTICA DE TEXTOS

A vetorização de textos é uma etapa importante na área de processamento de linguagem natural (PLN), permitindo que documentos textuais sejam representados como vetores numéricos em espaços de alta dimensão. Essa representação possibilita a aplicação de operações matemáticas, como cálculo de similaridade e agrupamento, tornando os textos comparáveis de forma quantitativa (Manning; Raghavan; Schütze, 2008).

Modelos mais antigos, como *Bag-of-Words* e *Term Frequency–Inverse Document Frequency* (TF-IDF), baseiam-se apenas na frequência de ocorrência de palavras, ignorando o contexto semântico em que esses termos aparecem. Como alternativa, técnicas mais recentes passaram a utilizar representações densas, também conhecidas como *embeddings*, que capturam relações semânticas entre palavras. Modelos como o *Word2Vec* (Mikolov *et al.*, 2013) e o *GloVe* (Pennington; Socher; Manning, 2014) introduziram representações distribuídas que preservam proximidade semântica entre termos, ainda que operem ao nível de palavras e não considerem o contexto global das frases.

2.2 MODELOS CONTEXTUAIS DE LINGUAGEM

Com o surgimento do modelo *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin *et al.*, 2018), a vetorização de textos deu um salto qualitativo ao permitir que cada palavra fosse representada em função do seu contexto, considerando simultaneamente os *tokens* à esquerda e à direita. Essa abordagem contextual supera as limitações dos *embeddings* estáticos, como *Word2Vec* ou *GloVe*, que atribuem o mesmo vetor para uma palavra independentemente do seu uso.

Modelos posteriores, como RoBERTa (Liu *et al.*, 2019) e DistilBERT, refinaram ainda mais essa abordagem, ampliando o desempenho em tarefas de compreensão textual. Modelos fundamentados no *Sentence Transformers* (Reimers; Gurevych, 2019) baseiam-se nessas arquiteturas, mas são ajustados especificamente para tarefas de similaridade semântica, tornando-os mais adequados para buscas vetoriais e comparação entre documentos inteiros.

O avanço das arquiteturas baseadas no *Transformer* (Vaswani *et al.*, 2017) permitiu a geração de *embeddings* ao nível de sentenças ou parágrafos, capturando relações semânticas

mais amplas e contextuais. Diferentemente de modelos baseados em palavras isoladas, como *Word2Vec*, esses modelos produzem vetores adaptáveis ao contexto textual e são particularmente úteis para tarefas de comparação entre documentos, como na triagem de currículos.

2.3 BUSCA VETORIAL E RECUPERAÇÃO SEMÂNTICA DE INFORMAÇÃO

Sistemas de busca tradicionais, baseados em palavras-chave, operam por correspondência exata ou parcial entre termos, sendo sensíveis a variações lexicais, sinônimos e estrutura textual. Para superar essas limitações, a busca vetorial emergiu como uma alternativa mais robusta, permitindo a recuperação de documentos com base na similaridade de seus vetores semânticos (Johnson; Douze; Jégou, 2019).

O funcionamento desses sistemas baseia-se na criação de um índice vetorial a partir dos *embeddings* dos documentos, utilizando estruturas otimizadas para busca aproximada por vizinhança mais próxima *Approximate Nearest Neighbors* (ANN). Dentre as bibliotecas especializadas nessa tarefa, destacam-se FAISS, Annoy e ChromaDB. Esta última foi adotada no presente trabalho por sua compatibilidade com consultas semânticas em tempo real e integração com ecossistemas de LLMs.

Segundo Johnson, Douze e Jégou (2019), a métrica mais comumente empregada na comparação vetorial é a similaridade do cosseno, embora também sejam utilizados outros métodos, como a distância euclidiana e o produto escalar. A escolha da métrica e da estrutura de indexação impacta diretamente a qualidade e a eficiência das buscas.

Além da triagem de currículos, a busca vetorial tem sido aplicada com sucesso em sistemas de recomendação, motores de busca semânticos, recuperação de artigos científicos e classificação de documentos jurídicos. Um dos principais desafios desses sistemas está na escalabilidade, já que a busca em espaços vetoriais de alta dimensão pode ser computacionalmente custosa. Para mitigar esse problema, algoritmos de busca aproximada, como *Hierarchical Navigable Small World* (HNSW), vêm sendo adotados, permitindo acelerar a recuperação com perda mínima de precisão (Malkov; Yashunin, 2020).

2.4 MODELOS DE LINGUAGEM DE LARGA ESCALA (LLMS)

Modelos de linguagem de larga escala (LLMs, do inglês *Large Language Models*) representam um avanço no campo da inteligência artificial. Baseados em arquiteturas do tipo transformador, esses modelos são treinados com bilhões de parâmetros sobre grandes volumes de dados textuais, sendo capazes de compreender, gerar e manipular linguagem natural com alto grau de sofisticação (Brown *et al.*, 2020).

Dentre os LLMs mais relevantes no cenário atual, destacam-se os modelos da família GPT

(OpenAI), Gemini (Google), LLaMA (Meta) e Falcon (TII). Esses modelos são capazes de realizar tarefas como tradução, sumarização, geração de texto, compreensão contextual, entre outras tarefas envolvendo linguagem natural (Fan *et al.*, 2024).

Ao serem utilizados em conjunto com mecanismos de busca vetorial, os LLMs podem desempenhar um papel complementar: enquanto a busca vetorial identifica documentos semanticamente próximos, o LLM é capaz de realizar uma leitura comparativa mais profunda, considerando relações implícitas, sequência lógica de informações e coesão textual (Bommasani *et al.*, 2021; Fan *et al.*, 2024).

2.5 INTEGRAÇÃO ENTRE BUSCA VETORIAL E MODELOS DE LINGUAGEM

A combinação entre mecanismos de recuperação semântica e LLMs tem sido popularizada por arquiteturas do tipo *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2020). Nessa abordagem, a busca vetorial identifica os documentos mais relevantes para uma consulta, e o LLM é responsável por sintetizar ou gerar uma resposta a partir desses documentos. Essa estratégia permite que o modelo de linguagem opere com um contexto mais restrito e relevante, melhorando a precisão e reduzindo a geração de informações factualmente incorretas, fenômeno conhecido como Alucinação.

Esses modelos podem, por exemplo, interpretar que "desenvolvimento de aplicações web em React e Node.js" é semanticamente próximo de "experiência com sistemas front-end e back-end modernos", mesmo que os termos exatos não coincidam. Tal capacidade é particularmente útil para currículos com terminologias variadas ou que contenham descrições indiretas de habilidades.

No contexto da triagem de currículos, essa integração permite primeiro recuperar perfis semanticamente próximos com base em um currículo de referência e, posteriormente, aplicar o LLM para interpretar os conteúdos recuperados e realizar uma análise comparativa mais profunda.

Além do RAG, outras arquiteturas híbridas têm sido propostas para aprimorar a integração entre mecanismos de recuperação e modelos gerativos, destacando-se o *Fusion-in-Decoder* (FiD) (Lewis *et al.*, 2020) e o RePlug (Shi *et al.*, 2023). Essas abordagens tratam a fusão de múltiplos documentos contextuais de forma mais eficiente e, embora ainda incipientes no contexto da triagem de currículos, representam caminhos promissores para evoluções futuras da presente proposta.

2.6 SUMARIZAÇÃO AUTOMÁTICA DE TEXTO

A sumarização automática de textos (SAT) é uma tarefa da área de PLN, cujo objetivo é criar automaticamente versões compactas de um ou mais textos que preservem seu conteúdo

mais relevante. Essa abordagem é especialmente útil em contextos com limitações de processamento, como o uso de LLMs com restrições de tamanho máximo de *tokens*.

Existem dois tipos principais de sumarização: a extrativa, que seleciona trechos literais do texto original, e a abstrativa, que gera novos textos com base na compreensão do conteúdo. Modelos como o *Pegasus* (Zhang *et al.*, 2020) e o *Meta LLaMA 3.1 8B Instruct* permitem realizar sumarizações abstrativas com alto grau de fidelidade semântica. A utilização dessas técnicas, no presente trabalho, visa testar o impacto da compressão textual na qualidade da triagem automatizada de currículos.

Recentemente, Sarmento e Oliveira (2024) realizaram uma investigação abrangente sobre técnicas de sumarização automática de textos em português do Brasil, com foco na aplicação de abordagens extrativas e abstrativas em artigos jornalísticos. O estudo avaliou desde métodos clássicos de ponderação de sentenças, como *TextRank* e posição das frases, até o uso de modelos de LLMs, como GPT-4o, LLaMA 3 e Gemma, além de modelos ajustados como PTT5 e FLAN-T5. Utilizando três bases de dados amplamente adotadas na literatura (TeMário, CSTNews e RecognaSumm) os autores analisaram o desempenho dos modelos por meio das métricas ROUGE-L e BERTScore. Os resultados apontaram que modelos abstrativos ajustados apresentaram desempenho competitivo e boa qualidade textual, enquanto os LLMs se destacaram em precisão, mas com alto custo computacional.

2.7 TRIAGEM AUTOMATIZADA DE CURRÍCULOS

Historicamente, os sistemas de triagem automatizada iniciaram com o desenvolvimento dos sistemas do tipo *Applicant Tracking Systems* (ATS), cuja lógica principal baseava-se em filtros estáticos de palavras-chave e regras manuais (Huang; Rust, 2019). Embora amplamente utilizados, esses sistemas demonstraram limitações ao ignorar contexto, sinônimos e variações linguísticas comuns em documentos de currículos (Overton, 2017). Com a popularização das técnicas de Aprendizado de Máquina (AM), surgiram abordagens supervisionadas capazes de classificar currículos com base em conjuntos rotulados, como demonstrado por Santos (2022) e outros estudos aplicados (Malik; Fatima; Ahmad, 2020). Mais recentemente, a integração de vetorização semântica e modelos de linguagem de larga escala (LLMs) tem se mostrado promissora, permitindo uma análise contextualizada e adaptável dos currículos (Bommasani *et al.*, 2021).

Além disso, estudos como o de Malik, Fatima e Ahmad (2020) destacam a importância da curadoria de dados e da diversidade dos conjuntos de treinamento. Sistemas que dependem exclusivamente de palavras-chave ou vocabulário fechado tendem a apresentar viés contra candidatos com perfis atípicos, mesmo que possuam competências relevantes. A introdução de modelos semânticos pode mitigar esse problema ao considerar o significado global dos textos, permitindo reconhecer competências formuladas de maneiras distintas.

Diferentes abordagens têm sido propostas na literatura para automatizar a triagem de currículos, com variações nos métodos de vetorização, bases de dados utilizadas e estratégias de classificação ou ranqueamento. A Tabela 1 apresenta um resumo comparativo dos principais trabalhos relacionados, destacando os conjuntos de dados adotados e as tecnologias empregadas em cada estudo.

Tabela 1 – Trabalhos correlatos sobre triagem automatizada de currículos.

Referência	**Dataset**	**Tecnologia / Abordagem**
Luo et al. (2018)	~10.000 currículos privados	ResumeNet: rede neural semi-supervisionada para avaliação da qualidade dos currículos
Bhatia et al. (2019)	Currículos do LinkedIn e genéricos	Parser semântico + BERT; ranqueamento por compatibilidade com 73% de acerto
Santos (2022)	1.340 currículos reais (área de TIC)	TF-IDF + classificadores (Random Forest, Naive Bayes); categorização multiclasse por área
Neto e Saraiva (2023)	Talentos Carreira RH	Word2Vec, Wang2Vec, FastText, GloVe; similaridade do cosseno
Saatçı e Kurt (2024)	~123 vagas e currículos associados	PLN para extração de competências; similaridade de Jaccard e cosseno
Gan et al. (2024)	Dataset próprio com currículos reais	Agentes baseados em LLMs para sumarização, classificação e ranqueamento
Heakl et al. (2024)	13.389 currículos de diversas fontes	Classificação com LLMs (BERT, Gemma 1.1); acurácia de até 92%
Lancetti et al. (2025)	2.164 vagas do LinkedIn (QA)	Vetores binários de *soft skills*; recomendação com filtros e cosine
Este trabalho	Resume Corpus (29.035 currículos TI)	Busca vetorial com *Sentence Transformers*; análise contextual com LLMs (LLaMA); sumarização automática

Fonte: Elaborado pelo autor (2025).

Luo et al. (2018) propuseram o ResumeNet, um modelo baseado em rede neural semi-supervisionada para avaliação da qualidade de currículos. O sistema foi treinado com aproximadamente dez mil documentos reais e avalia aspectos como completude e clareza das informações. O enfoque está na análise da apresentação do currículo como um todo, sem realizar correspondência com uma vaga específica.

Bhatia et al. (2019) desenvolveram um sistema de parsing e ranqueamento de currículos com uso de BERT, focado na correspondência entre documentos e vagas. O modelo realiza extração semântica das seções do currículo (educação, experiência, habilidades) e calcula uma pontuação de compatibilidade. O sistema obteve 73% de precisão, demonstrando a viabilidade da abordagem para triagem inteligente.

Santos (2022) propôs uma metodologia de categorização automática de currículos na área de Tecnologias da Informação e Comunicação (TIC), utilizando um conjunto de currículos reais visando classificá-los em até 20 categorias distintas, como *Desenvolvedor Web*, *Analista de Suporte*, *Administrador de Redes*, entre outras. Os currículos foram processados com técnicas de *Bag-of-Words* e vetorizados com *TF-IDF*, sendo posteriormente classificados com diferentes algoritmos, como *Random Forest* e *Naive Bayes*. O autor destaca, por exemplo, que currículos contendo palavras como “HTML”, “CSS” e “JavaScript” foram

fortemente associados à categoria *Desenvolvedor Web*, enquanto termos como “roteador”, “switch” e “cabeamento” apareceram com frequência em currículos classificados como *Administrador de Redes*. Esse tipo de associação léxica demonstra como modelos estatísticos podem inferir automaticamente a área de atuação de um candidato com base no vocabulário predominante. Ainda segundo Santos (2022), os modelos apresentaram acurácia superior a 90% nas tarefas de classificação binária, demonstrando o potencial da abordagem para a triagem inicial de candidatos em sistemas de recrutamento automatizado.

Esse tipo de solução, embora eficaz em determinados contextos, depende fortemente da engenharia de atributos e de um vocabulário previamente conhecido. Em contraste, sistemas baseados em busca semântica e LLMs, como o proposto neste trabalho, podem ser capazes de inferir semelhanças contextuais mesmo quando os termos usados variam, o que amplia a aplicabilidade e a robustez do sistema frente a currículos mais heterogêneos.

Neto e Saraiva (2023) exploraram o uso de embeddings em português (Word2Vec, FastText, Wang2Vec, GloVe) para medir a similaridade entre currículos e descrições de vagas. O sistema desenvolvido foi aplicado em dados reais da consultoria Talentos Carreira RH e demonstrou a viabilidade do uso de modelos semânticos no idioma português.

Saatçı e Kurt (2024) propuseram uma aplicação prática de triagem com técnicas de PLN voltadas à extração de competências profissionais. Os currículos foram comparados com vagas usando medidas de similaridade semântica como Jaccard e cosseno. O trabalho se destaca por sua ênfase em usabilidade e baixo custo computacional.

Gan et al. (2024) apresentaram um framework baseado em agentes inteligentes com LLMs, capazes de realizar múltiplas tarefas automaticamente, como leitura, classificação e sumarização de currículos. A abordagem destaca-se por sua arquitetura autônoma e modular, com capacidade de adaptação a diferentes contextos.

Heakl et al. (2024) introduziram o benchmark ResumeAtlas, uma base com mais de 13 mil currículos rotulados para avaliação de modelos de classificação baseados em LLMs, como BERT e Gemma. O estudo alcançou acurácia de até 92% e reforça o potencial dos modelos de linguagem em tarefas de triagem supervisionada.

Lancetti et al. (2025) propuseram o sistema TalentJobRadar, com foco em soft skills e cargos da área de qualidade de software. A proposta combina vetores binários de competências e um radar de similaridade baseado na similaridade do cosseno, oferecendo uma interface visual interativa e filtros personalizados para a recomendação de perfis.

2.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Os conceitos apresentados neste capítulo fornecem embasamento teórico necessário para a construção do sistema proposto. A integração entre vetorização semântica, busca vetorial e análise contextual com LLMs representa uma abordagem inovadora no campo da triagem de currículos, alinhando-se às tendências contemporâneas de automação e inteligência artificial no setor de recursos humanos.

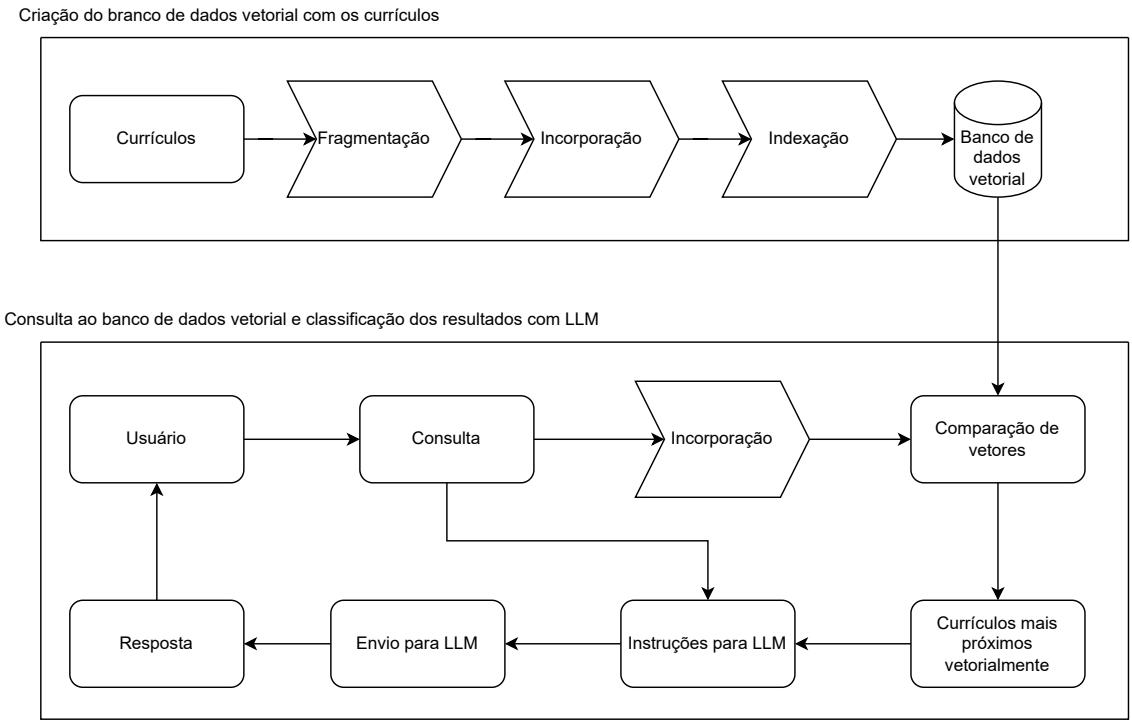
3 DESENVOLVIMENTO

Este capítulo apresenta os detalhes do sistema proposto, abordando sua arquitetura funcional, os dados utilizados e os experimentos conduzidos para validação. São descritos o fluxo de interação entre usuário e sistema, os componentes utilizados na construção do banco vetorial e a metodologia experimental aplicada.

3.1 SISTEMA PROPOSTO

O funcionamento do sistema proposto baseia-se na integração entre um mecanismo de busca semântica em um banco de dados vetorial e um modelo de linguagem de grande escala (LLM). O fluxo de uso inicia-se com a submissão de um currículo pelo usuário, por meio do componente de interação no formato de *chat*. Após o envio, o sistema converte o conteúdo textual do documento em vetores de alta dimensão utilizando um modelo de transformador de sentenças (especificado no parâmetro "modelo"). A Figura 1 ilustra o fluxo de execução do sistema proposto.

Figura 1 – Fluxo geral do sistema proposto.



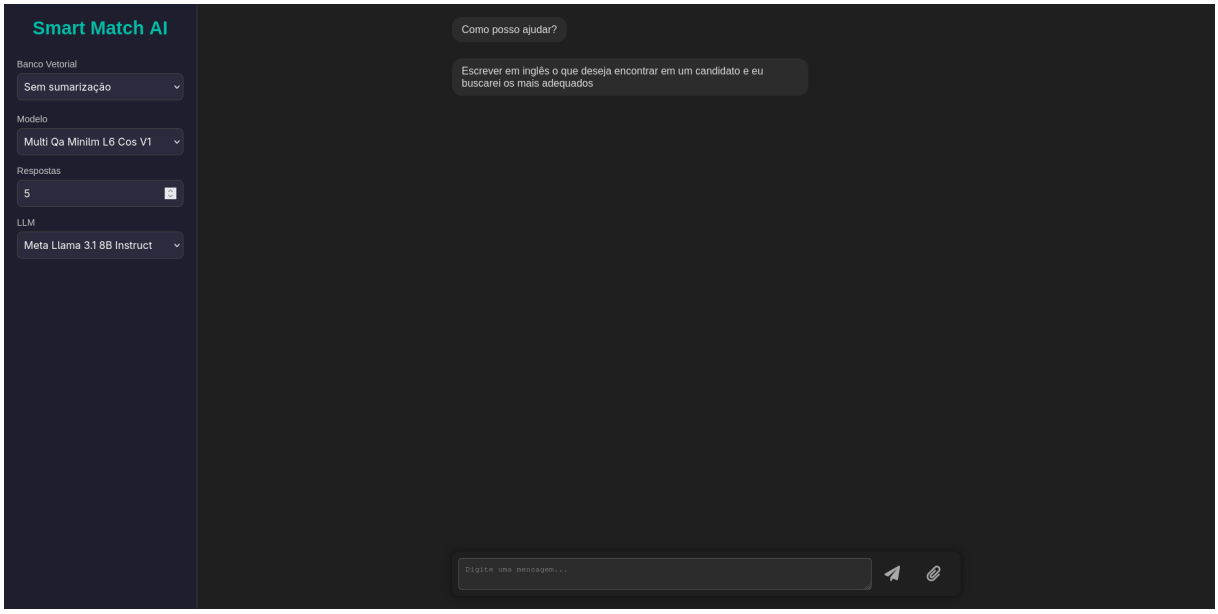
Fonte: Elaborado pelo autor (2025).

A Figura 1 apresenta o fluxo geral do sistema desenvolvido, dividido em duas etapas principais: a construção do banco de dados vetorial e o processo de consulta e análise contextual. Na primeira etapa, os currículos são processados por meio de fragmentação, geração de *embeddings* e indexação, resultando em um banco vetorial otimizado para recuperação semântica. Na segunda etapa, o usuário fornece um currículo de referência,

que também é vetorizado e comparado com os documentos armazenados. Os currículos mais similares, com base na proximidade vetorial, são então enviados a um modelo de LLM, juntamente com instruções específicas. O LLM realiza uma análise contextual dos documentos recuperados e retorna uma resposta estruturada com os candidatos mais aderentes ao perfil buscado.

Após a vetorização do currículo de entrada, ele é então comparado com os currículos previamente armazenados previamente no banco de dados vetorial (ChromaDB), retornando os documentos mais semanticamente semelhantes conforme a quantidade definida pelo parâmetro "respostas". Uma vez recuperados os documentos mais relevantes, esses dados são repassados como entrada para o LLM escolhido (definido no parâmetro "LLM"), que realiza a análise contextual e gera uma sugestão final sobre o candidato mais apropriado com base nas informações extraídas. Esse processo combina a busca vetorial com a capacidade interpretativa de modelos de linguagem, resultando em uma resposta precisa e contextualizada. A Figura 2 e Figura 3 ilustram a tela de interação do usuário com o sistema proposto.

Figura 2 – Sistema proposto: visão inicial do sistema.



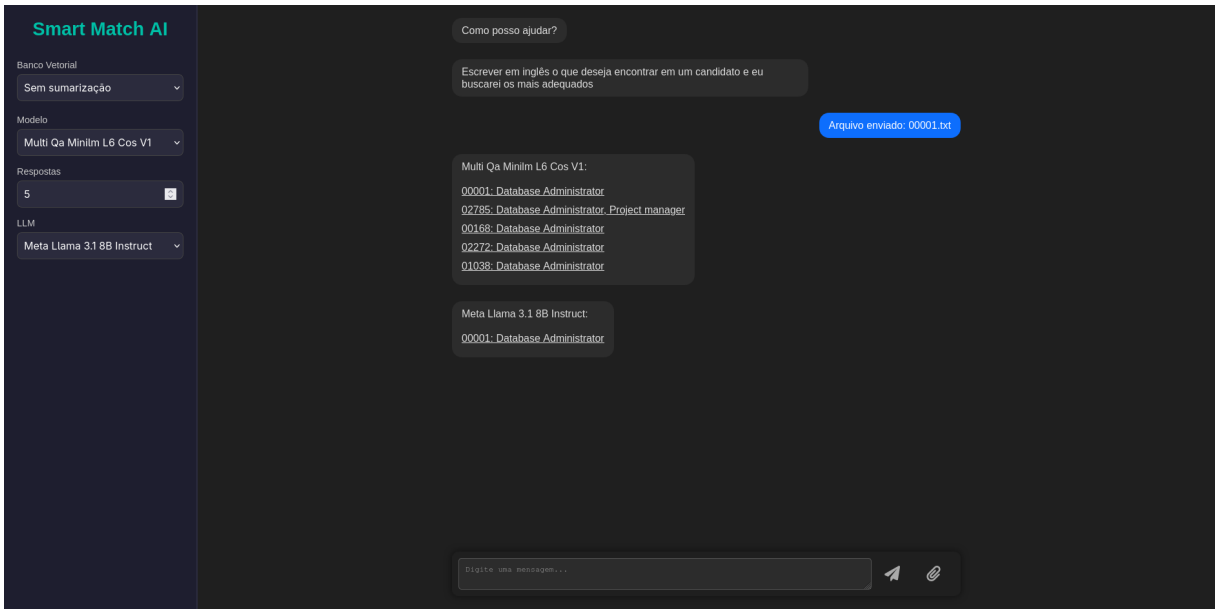
Fonte: Elaborado pelo autor (2025).

A escolha dos parâmetros de controle influencia diretamente os resultados apresentados no *chat*, permitindo ao usuário ajustar a busca de acordo com diferentes combinações de banco de dados, modelos de vetorização e modelos de linguagem.

3.2 BASE DE DADOS

Para avaliar o desempenho do sistema proposto, experimentos foram realizados utilizando a base de currículos desenvolvida por Jiechieu e Tsopze (2020). Esse conjunto de dados

Figura 3 – Sistema proposto: envio de arquivo e respostas do sistemas.



Fonte: Elaborado pelo autor (2025).

é comumente usado para tarefas de classificação multirótulo de currículos, no qual cada currículo pode estar associado a mais de uma categoria profissional, como “Desenvolvedor Web”, “Desenvolvedor Python”, entre outras categorias. Os documentos são apresentados em formato textual não estruturado, refletindo com razoável fidelidade a diversidade de formatações encontradas em currículos reais.

O conjunto de dados é estruturado por meio de arquivos emparelhados, sendo um com extensão *.txt*, que contém o conteúdo textual completo de um currículo, e outro com extensão *.lab*, que armazena os rótulos ocupacionais associados àquele currículo. Essa organização facilita a vinculação direta entre o texto e suas respectivas categorias profissionais. A Figura 4 ilustra um exemplo desses arquivos.

O conjunto de rótulos presentes nos arquivos *.lab* é composto por 10 categorias profissionais distintas, refletindo diferentes especializações na área de tecnologia da informação. São elas: *Database_Administrator*, *Software_Developer*, *Systems_Administrator*, *Project_manager*, *Security_Analyst*, *Web_Developer*, *Network_Administrator*, *Front_End_Developer*, *Java_Developer* e *Python_Developer*. Como se trata de um problema de classificação multirótulo, um mesmo currículo pode estar associado a mais de uma dessas categorias.

A base de dados possui um total de 59.566 arquivos. Para garantir a qualidade dos dados, foi desenvolvido um algoritmo na linguagem de programação Python para verificar a presença de conteúdo nos arquivos de rótulo com extensão *.lab*. Arquivos que apresentavam rótulos vazios foram excluídos juntamente com os arquivos de texto correspondentes, resultando em um total de 58.070 arquivos válidos. Considerando que cada currículo é composto por dois arquivos (um texto e um rótulo), obteve-se uma base inicial composta por 29.035

Figura 4 – Currículo estruturado de um desenvolvedor com os rotulos de *Web Developer*, *Software Developer* e *Front-End Developer*.

Front-End Web Developer
Web Developer — Philadelphia, PA
Status: Authorized to work in the US for any employer

Work Experience

Front-End Web Developer – Freelance
Philadelphia, PA — Feb 2018 to Present
Developed **Fansí Schmansí**, responsive site with menu and Google Maps API.
• fansischmansi.ml
Developed **Syzygy**, one-page app with task list, clock, SoundCloud & Unsplash API.
• syzygy.ml
Developed **Film Scope**, film catalogue using OMDB API.
• Film Scope

Data Analyst – Maurice H. Kornberg School of Dentistry
Philadelphia, PA — Oct 2016 to Sep 2018
• Restructured employee archive (Excel, HTML, JS)
• Managed 45,000 patient charts (Axium)
• Tracked student progress (Excel, Word)

Web Designer and Research Intern – Drexel University
Mechanical Eng. and Mechanics Dept — May to Sep 2015
• Created lab experiments page (HTML, CSS, Photoshop)

Data Organizer – Community College of Philadelphia
Jan to May 2015
• Scanned and indexed files in online system

Education

Natural Science – Temple University
Philadelphia, PA — May 2018

Skills

Git, HTML, JS, Bootstrap, CSS, React, UI, UX, Photoshop, Sketch, InDesign, Command Line, jQuery, HTML5, CSS3

Links

github.com/debxdev
debthornton.ml

Groups

Association for Computing Machinery (Jan–May 2018)
Kappa Delta Pi (Sep 2017–May 2018)

Additional Information

Interests: BJJ, MMA, swimming, yoga, acoustic guitar, Reddit, beer, marine biology, travelling

Fonte: Elaborado pelo autor (2025).

currículos. A soma dos caracteres de todos os textos dessa base totalizou 200.814.853, formando assim a base de dados utilizada nos experimentos.

Com o intuito de avaliar o desempenho do sistema utilizando versões resumidas dos currículos, duas novas bases derivadas foram criadas. A segunda base foi construída utilizando a plataforma SambaNova Cloud[1], especializada em computação para inteligência artificial, com acesso a modelos de linguagem avançados. Para esse processo, foi utilizado o modelo *Meta LLaMA 3.1 8B Instruct*, que oferece uma das maiores capacidades de requisições diárias disponíveis na plataforma (28.800 requisições por dia). Esse modelo foi empregado para gerar resumos abstrativos dos currículos presentes na primeira base. O resultado foi uma nova base com o mesmo número de currículos (29.035), mas com uma total de 45.779.030 caracteres, representando aproximadamente 22,8% do volume

[1] <https://sambanova.ai/>

original. Adicionalmente, foi gerada uma terceira versão da base utilizando o modelo de sumarização abstrativa *Pegasus*, aplicado localmente por meio da biblioteca Hugging Face[2], com o modelo *google/pegasus-xsum*. O processo foi realizado sobre os mesmos 29.035 currículos da base original. Como resultado, obteve-se uma base com 5.194.230 caracteres, o que representa aproximadamente 2,6% do volume textual original. Essa redução mais agressiva permitiu avaliar os impactos da compressão extrema sobre o desempenho dos modelos de busca semântica.

3.3 METODOLOGIA EXPERIMENTAL

Os experimentos realizados neste trabalho foram desenvolvidos utilizando a linguagem Python, na versão 3.9. A partir da análise dos rótulos presentes na base de dados, foram selecionadas 10 ocupações distintas com maior representatividade. Para a construção do conjunto de avaliação, foram separados 100 currículos para cada uma dessas ocupações, totalizando 1.000 currículos exclusivos, que não foram incluídos nos bancos vetoriais utilizados nas buscas. Como consequência, todas as versões do banco de dados vetorial passaram a conter 28.035 documentos.

Experimento 1: Avaliação dos modelos de transformação de sentenças

O primeiro experimento teve como objetivo identificar, de forma comparativa, quais modelos de *Sentence Transformers* apresentavam melhor desempenho na tarefa de recuperação semântica, considerando três versões da base de dados: a base original, a base resumida com auxílio de um modelo LLM (*Meta LLaMA 3.1 8B Instruct*) e a base resumida com o modelo *Pegasus*. Para isso, foram testados treze modelos distintos, incluindo variantes como *Multi QA MiniLM L6 cos v1*, *Multi QA Mpnet base dot v1*, *All MiniLM L6 v2*, *Distiluse Multilingual Base v1 e v2*, *All Mpnet base v2*, entre outros. Na Tabela 2, são apresentadas as principais características desses modelos, como dimensionalidade, suporte linguístico e tarefa principal para a qual foram otimizados.

Cada modelo foi aplicado individualmente nas três bases de dados mencionadas, resultando na criação de múltiplos bancos vetoriais indexados no ChromaDB. A avaliação foi realizada utilizando os 1.000 currículos reservados exclusivamente para teste. Esses currículos foram empregados como consultas, e diferentes quantidades dos top-n resultados mais similares foram testadas: 1, 5, 10, 20 e 30 documentos por consulta.

A métrica adotada para avaliar o desempenho dos modelos baseou-se na comparação entre os rótulos dos currículos recuperados e os do currículo de referência. A partir disso, foi calculada uma razão média de coincidência entre os rótulos, normalizada pelo número de documentos retornados.

[2] <https://huggingface.co/>

Tabela 2 – Modelos de *embeddings* e suas características.

Nome do Modelo	**Dimensão**	**Idiomas Suportados**	**Tarefa Principal**
Multi QA MiniLM L6 Cos V1	384	Inglês	Busca Semântica
Multi QA Mpnet Base Dot V1	768	Inglês	Busca Semântica
All MiniLM L6 V2	384	Inglês	Similaridade de Sentenças
Distiluse Base Multilingual Cased V1	512	15 idiomas (PT, EN, ES, etc.)	Similaridade Multilíngue
Distiluse Base Multilingual Cased V2	512	50+ idiomas	Similaridade Multilíngue
All Distilroberta V1	768	Inglês	Similaridade de Sentenças
Paraphrase Albert Small V2	768	Inglês	Detecção de Parafrase
Paraphrase Multilingual MiniLM L12 V2	384	50+ idiomas	Similaridade Multilíngue
Paraphrase MiniLM L3 V2	384	Inglês	Detecção de Parafrase
Multi QA Distilbert Cos V1	768	Inglês	Busca Semântica
Paraphrase Multilingual Mpnet Base V2	768	50+ idiomas	Similaridade Multilíngue
All MiniLM L12 V2	384	Inglês	Similaridade de Sentenças
All Mpnet Base V2	768	Inglês	Similaridade de Sentenças

Fonte: Elaborado pelo autor com base em dados disponíveis na plataforma Hugging Face (2025).

Experimento 2: Avaliação dos LLMs na seleção dos currículos mais adequados

O segundo experimento teve como foco avaliar a capacidade de modelos de linguagem de grande escala (LLMs) em selecionar, entre os currículos retornados pela busca vetorial, aquele mais apropriado e os cinco mais relevantes com base em um currículo de referência.

Para esse experimento, foi utilizada a base de dados composta por currículos resumidos com o modelo *Meta LLaMA 3.1 8B Instruct*, combinada ao modelo vetorial que apresentou melhor desempenho no experimento anterior. A partir de cada currículo de referência, os top-10 currículos mais semelhantes foram recuperados do banco vetorial e utilizados como entrada para o LLM.

A interação com os modelos foi realizada por meio de requisições via API na plataforma SambaNova Cloud, respeitando os limites de tokens impostos por cada modelo. Foram utilizados quatro modelos da família *LLaMA* para essa tarefa: *Meta-Llama-3.2-1B-Instruct*, *Meta-Llama-3.1-8B-Instruct*, *Meta-Llama-3.3-70B-Instruct* e *Llama-4-Scout-17B-16E-Instruct*. Cada modelo recebeu um *prompt* contendo o currículo de referência, os currículos candidatos retornados pela busca vetorial e instruções específicas para selecionar os currículos mais apropriados, retornando a resposta em formato estruturado.

A avaliação de desempenho considerou a sobreposição entre os rótulos do currículo de referência e os currículos selecionados pelos LLMs, utilizando o mesmo critério adotado

no primeiro experimento. O número de currículos avaliados por consulta foi limitado em função das restrições de *tokens* da plataforma utilizada.

O *prompt* foi estruturado de forma clara e objetiva, orientando o modelo a retornar apenas os identificadores dos currículos mais similares em formato JSON. Na Tabela 3 é apresentado o *prompt* utilizado neste experimento.

Tabela 3 – Estrutura do *prompt* utilizado para o modelo LLM.

Conteúdo do Prompt
I will provide you with one base resume and 10 other resumes. Your task is to compare the 10 resumes to the base resume and identify the 5 most similar ones based on content, skills, experience, and overall relevance. Each resume will have a unique identifier in the format: “resume_id: <ID>”. At the end, return a JSON object containing only the 5 most similar resume IDs, sorted from most to least similar. Format the output strictly like this: { “most_similar_resume_ids”: [“<ID1>”, ..., “<ID5>”] } Do not include any explanation, only the JSON response. Here is the base resume: <user_submitted_entry_resume> And here are the 10 other resumes: — <resume_id> <resume_content> —

Fonte: Elaborado pelo autor (2025).

3.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou os aspectos técnicos e experimentais do sistema de triagem automática de currículos proposto. Foram descritos o funcionamento geral da arquitetura, a base de dados utilizada e os procedimentos de avaliação adotados. A integração entre mecanismos de busca vetorial e modelos de linguagem foi detalhada, com ênfase na vetorização dos currículos, na recuperação semântica dos currículos mais relevantes e na análise contextual realizada pelos LLMs. Além disso, foram explicadas as diferentes versões da base de dados (original e resumidas) e a metodologia empregada nos dois experimentos conduzidos: avaliação dos modelos de transformação de sentenças e análise comparativa entre LLMs. Esses elementos fornecem o embasamento necessário para a interpretação dos resultados, os quais são discutidos no capítulo seguinte.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados obtidos a partir dos experimentos descritos na Metodologia Experimental. As análises consideram o desempenho de diferentes modelos de transformação de sentenças e modelos de linguagem de larga escala (LLMs) na tarefa de identificação de currículos mais adequados com base em similaridade semântica e interpretação contextual.

4.1 EXPERIMENTO 1 - AVALIAÇÃO DOS MODELOS DE *EMBEDDINGS*

Na Tabela 4, Tabela 5 e Tabela 6 são apresentados os resultados deste primeiro experimento considerando os currículos originais, a versão resumida do currículo pelo modelo *Llama 3 8B* e pelo modelo *Pegasus*, respectivamente. Considerando as três versões dos textos dos currículos e os treze modelos de *embeddings* analisados, foram consideradas trinta e nove combinações distintas. Os resultados demostram o desempenho médio da acurácia na tarefa de recuperação semântica, medido em diferentes quantidades de documentos (Top-n) retornados, sendo avaliados os valores de n iguais a 1, 5, 10, 20 e 30.

Tabela 4 – (Base original) Acurácia dos modelos considerando diferentes quantidades de retornos.

Modelo	**Top-1**	**Top-5**	**Top-10**	**Top-20**	**Top-30**
Multi QA MiniLM L6 Cos V1	**77,557**	73,128	72,044	71,386	71,172
Multi QA Mpnet Base Dot V1	77,205	73,227	72,316	71,424	71,069
All MiniLM L6 V2	77,113	71,743	70,664	69,737	69,315
Distiluse Base Multilingual Cased V1	76,990	72,300	71,466	70,209	69,731
Distiluse Base Multilingual Cased V2	75,943	70,927	69,414	68,362	67,928
All Distilroberta V1	74,815	69,724	68,709	67,634	67,056
Paraphrase Albert Small V2	73,770	69,219	68,116	67,321	66,812
Paraphrase Multilingual MiniLM L12 V2	73,707	69,170	67,818	67,093	66,531
Paraphrase MiniLM L3 V2	73,595	68,500	67,589	66,659	66,058
Multi QA Distilbert Cos V1	73,295	70,546	69,771	69,443	69,051
Paraphrase Multilingual Mpnet Base V2	73,497	68,364	67,031	66,338	66,046
All MiniLM L12 V2	71,330	69,331	68,484	67,600	67,230
All Mpnet Base V2	64,108	61,469	60,781	60,773	60,355

Fonte: Elaborado pelo autor (2025).

Os resultados obtidos usando a base de dados original demonstram que os modelos *Multi QA MiniLM L6 Cos V1*, *Multi QA Mpnet Base Dot V1* e *All MiniLM L6 V2* apresentaram os melhores desempenhos, especialmente para consultas com poucos resultados. Esse desempenho pode estar relacionado à arquitetura otimizada desses modelos para tarefas de busca semântica.

Analisando os resultados obtidos na base de dados sumarizada com auxílio do modelo *Llama 3.1 8B*, observou-se uma leve queda geral na acurácia, atribuída à perda de informações durante o processo de sumarização. Ainda assim, os modelos previamente destacados continuaram entre os mais eficazes, mantendo padrões consistentes de desempenho.

Tabela 5 – (Base resumida *LLaMA 3.1 8B*) Acurácia dos modelos considerando diferentes quantidades de retornos.

Modelo	Top-1	Top-5	Top-10	Top-20	Top-30
Multi QA MiniLM L6 Cos V1	71,635	**71,822**	71,203	70,330	70,024
Multi QA Mpnet Base Dot V1	71,875	70,644	70,153	69,714	69,451
All MiniLM L6 V2	71,812	69,779	69,137	68,726	68,398
Distiluse Base Multilingual Cased V1	69,185	67,758	67,784	67,370	67,076
Distiluse Base Multilingual Cased V2	68,135	65,987	65,290	64,637	64,324
All Distilroberta V1	68,413	67,970	67,320	66,805	66,563
Paraphrase Albert Small V2	65,992	65,977	66,034	65,505	65,319
Paraphrase Multilingual MiniLM L12 V2	65,015	64,617	64,318	63,422	63,105
Paraphrase MiniLM L3 V2	65,487	65,064	64,522	63,875	63,477
Multi QA Distilbert Cos V1	70,665	70,883	70,658	69,870	69,314
Paraphrase Multilingual Mpnet Base V2	68,505	68,113	67,497	66,811	66,656
All MiniLM L12 V2	66,772	66,364	65,897	65,465	65,229
All Mpnet Base V2	67,375	67,211	66,759	66,029	65,805

Fonte: Elaborado pelo autor (2025).

A versão da base de dados resumida usando o modelo *Pegasus* apresentou os piores desempenhos gerais entre as três versões avaliadas. Esse resultado evidencia que a técnica de sumarização empregada comprometeu significativamente a densidade semântica dos textos, refletindo negativamente na eficácia dos modelos de transformação de sentenças. Tal impacto pode ser atribuído à elevada taxa de compressão observada: o volume textual da base foi reduzido para 5.194.230 caracteres, correspondendo a aproximadamente 2,6% do total original. Essa redução extrema, embora eficiente do ponto de vista de compactação, resultou em perda de conteúdo informativo relevante para as tarefas de recuperação semântica.

Tabela 6 – (Base resumida *Pegasus*) Acurácia dos modelos considerando diferentes quantidades de retornos.

Modelo	Top-1	Top-5	Top-10	Top-20	Top-30
Multi QA Mpnet Base Dot V1	**59,050**	54,061	50,801	48,211	47,146
All MiniLM L6 V2	58,758	56,469	54,378	51,921	50,262
Multi QA MiniLM L6 Cos V1	56,460	54,999	53,385	50,885	49,337
Distiluse Base Multilingual Cased V1	59,895	54,961	51,975	48,910	47,412
All Distilroberta V1	55,802	52,629	51,129	49,650	48,713
Multi QA Distilbert Cos V1	47,988	52,259	51,078	49,577	48,520
Paraphrase Albert Small V2	52,638	49,940	48,276	46,286	45,105
Paraphrase Multilingual MiniLM L12 V2	51,800	48,732	47,342	46,026	45,132
Distiluse Base Multilingual Cased V2	53,045	49,130	47,336	45,419	44,419
All MiniLM L12 V2	51,473	49,810	48,557	46,694	45,448
Paraphrase Multilingual Mpnet Base V2	48,890	47,338	45,700	44,379	43,640
All Mpnet Base V2	50,842	48,173	46,632	45,223	44,506
Paraphrase MiniLM L3 V2	46,540	43,968	43,129	42,613	42,415

Fonte: Elaborado pelo autor (2025).

4.2 EXPERIMENTO 2 - AVALIAÇÃO DOS LLMS

Este segundo experimento analisou o desempenho de diferentes LLMs na tarefa de selecionar, a partir dos 10 currículos mais semelhantes retornados pelo ChromaDB, o currículo mais adequado (Top-1) e os cinco mais relevantes (Top-5). A base de dados utilizada foi a

versão resumida com *LLaMA 3.1 8B Instruct*, combinada ao modelo vetorial *Multi QA MiniLM L6 Cos V1*. É importante ressaltar que apesar da versão original da base de dados ter apresentado melhores resultados no primeiro experimento, ela não foi utilizada neste experimento devido ao tamanho dos currículos ultrapassar a limitação do número máximo de *tokens* permitido pelos LLMs usados neste experimento. Na Tabela 7 são apresentados os resultados deste segundo experimento.

Tabela 7 – (Base resumida com *Meta LLaMA 3.1 8B Instruct* e modelo vetorial *Multi QA MiniLM L6 Cos V1*) Acurácia das respostas usando LLMs.

Modelo	**Top-1**	**Top-5**
Meta LLaMA 3.1 8B Instruct	**70,297**	69,788
LLaMA 4 Scout 17B 16E Instruct	57,312	53,913
Meta LLaMA 3.3 70B Instruct	57,940	56,189
Meta LLaMA 3.2 1B Instruct	56,187	57,216

Fonte: Elaborado pelo autor (2025).

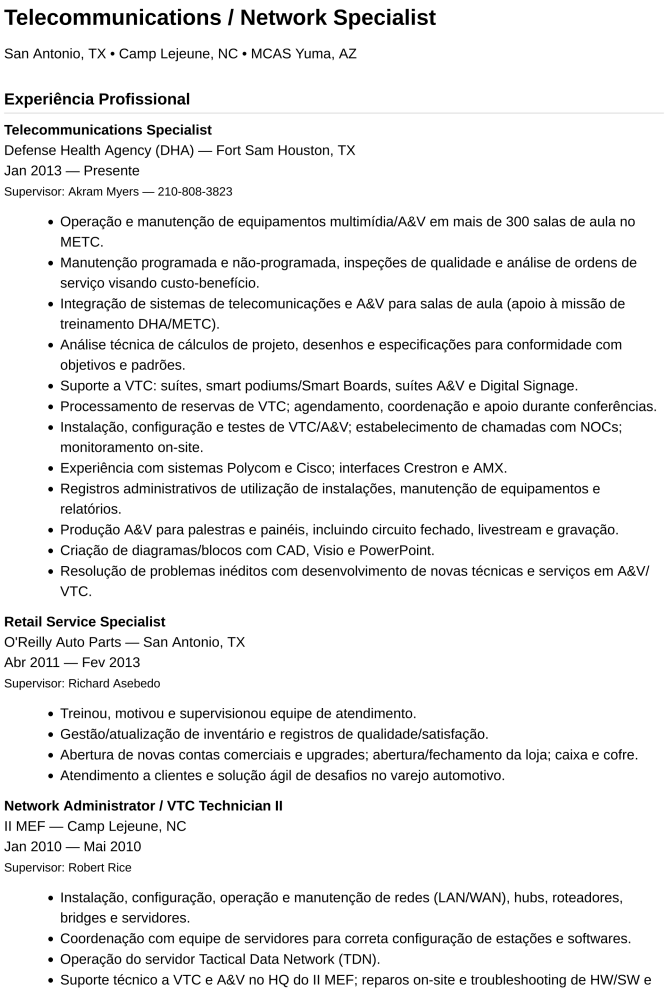
O modelo *Meta LLaMA 3.1 8B Instruct* destacou-se com a melhor acurácia média, tanto na seleção do currículo mais adequado quanto dos cinco mais relevantes. Esse resultado sugere que, além de ser utilizado para sumarização, esse modelo apresenta boa capacidade de análise e comparação contextual. Modelos de maior porte, como o *Meta LLaMA 3.3 70B Instruct*, não superaram o desempenho do modelo de 8 bilhões de parâmetros, indicando que apenas o aumento de escala não garante resultados superiores.

Por outro lado, os resultados também sugerem que modelos LLM podem contribuir de forma complementar, auxiliando na tarefa de ranqueamento e filtragem dos resultados, mesmo quando aplicados sobre versões resumidas dos currículos. A estratégia combinada, vetorização com modelos compactos e eficientes, seguida de interpretação com LLMs, revelou-se promissora para sistemas de apoio à triagem de currículos, desde que se preserve um nível mínimo de riqueza informacional nos textos processados.

Embora os experimentos tenham demonstrado resultados promissores, também foram identificadas classificações incorretas. Um caso exemplar é mostrado na Figura 5 e Figura 6, que apresentam o currículo estruturado de um profissional com experiência em telecomunicações e administração de redes, originalmente rotulado como *Network Administrator*. O sistema, entretanto, o classificou como *Web Developer*, *Python Developer* e *Software Developer*. Essa divergência decorre, possivelmente, de menções no texto a tecnologias e práticas comuns ao desenvolvimento de software, como scripts, integração de sistemas e uso de ferramentas de automação, gerando sobreposição semântica com perfis de programação.

Entre os fatores que podem ter contribuído para alguns dos erros de classificação identificados nos experimentos estão:

Figura 5 – Currículo estruturado de um desenvolvedor com o rótulo de *Network Administrator* (Página 1).



Telecommunications / Network Specialist

San Antonio, TX • Camp Lejeune, NC • MCAS Yuma, AZ

Experiência Profissional

Telecommunications Specialist
Defense Health Agency (DHA) — Fort Sam Houston, TX
Jan 2013 — Presente
Supervisor: Akram Myers — 210-808-3823

- Operação e manutenção de equipamentos multimídia/A&V em mais de 300 salas de aula no METC.
- Manutenção programada e não-programada, inspeções de qualidade e análise de ordens de serviço visando custo-benefício.
- Integração de sistemas de telecomunicações e A&V para salas de aula (apoio à missão de treinamento DHA/METC).
- Análise técnica de cálculos de projeto, desenhos e especificações para conformidade com objetivos e padrões.
- Suporte a VTC: suítes, smart podiums/Smart Boards, suítes A&V e Digital Signage.
- Processamento de reservas de VTC; agendamento, coordenação e apoio durante conferências.
- Instalação, configuração e testes de VTC/A&V; estabelecimento de chamadas com NOCs; monitoramento on-site.
- Experiência com sistemas Polycom e Cisco; interfaces Crestron e AMX.
- Registros administrativos de utilização de instalações, manutenção de equipamentos e relatórios.
- Produção A&V para palestras e painéis, incluindo circuito fechado, livestream e gravação.
- Criação de diagramas/blocos com CAD, Visio e PowerPoint.
- Resolução de problemas inéditos com desenvolvimento de novas técnicas e serviços em A&V/VTC.

Retail Service Specialist
O'Reilly Auto Parts — San Antonio, TX
Abr 2011 — Fev 2013
Supervisor: Richard Asebedo

- Treinou, motivou e supervisionou equipe de atendimento.
- Gestão/atualização de inventário e registros de qualidade/satisfação.
- Abertura de novas contas comerciais e upgrades; abertura/fechamento da loja; caixa e cofre.
- Atendimento a clientes e solução ágil de desafios no varejo automotivo.

Network Administrator / VTC Technician II
II MEF — Camp Lejeune, NC
Jan 2010 — Mai 2010
Supervisor: Robert Rice

- Instalação, configuração, operação e manutenção de redes (LAN/WAN), hubs, roteadores, bridges e servidores.
- Coordenação com equipe de servidores para correta configuração de estações e softwares.
- Operação do servidor Tactical Data Network (TDN).
- Suporte técnico a VTC e A&V no HQ do II MEF; reparos on-site e troubleshooting de HW/SW e

Fonte: Elaborado pelo autor (2025).

- **Granularidade limitada dos rótulos** — categorias amplas como *Software Developer* abrangem atividades técnicas de outros domínios;

- **Ruído semântico** — termos compartilhados entre áreas distintas (linguagens, frameworks, ferramentas);

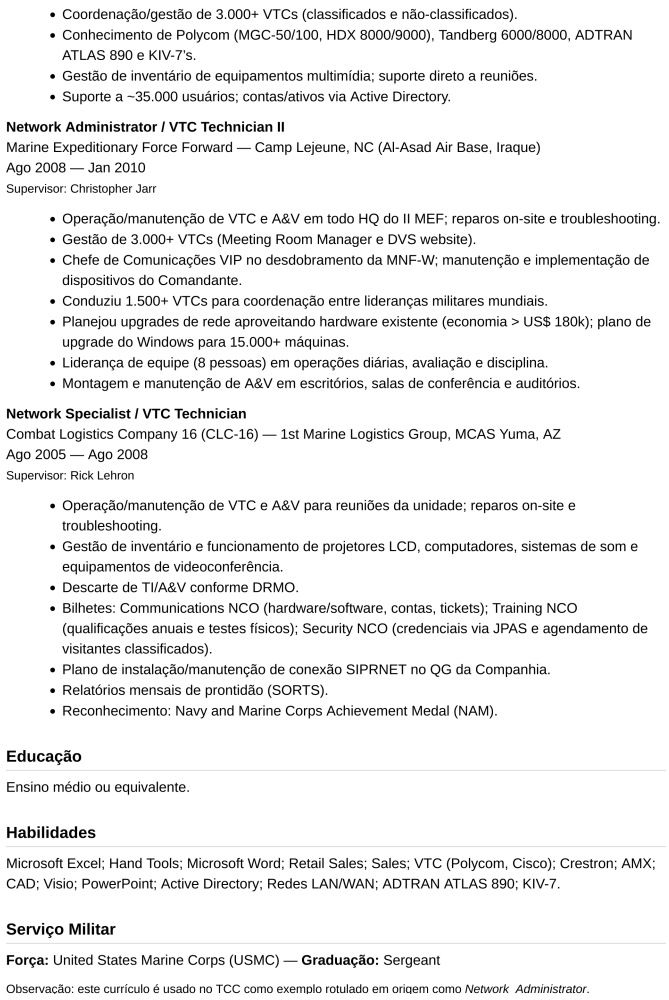
- **Ausência de filtros pós-busca** — a análise contextual não eliminou categorias irrelevantes após a recuperação vetorial.

Esses casos afetam diretamente as métricas de acurácia e evidenciam a necessidade de estratégias adicionais, como:

- Modelos de *embeddings* treinados especificamente para o domínio de currículos;

- Etapas adicionais de classificação supervisionada após o ranqueamento ou estratégias de reordenação;

Figura 6 – Currículo estruturado de um desenvolvedor com o rótulo de *Network Administrator* (Página 2).



- Coordenação/gestão de 3.000+ VTCs (classificados e não-classificados).
- Conhecimento de Polycom (MGC-50/100, HDX 8000/9000), Tandberg 6000/8000, ADTRAN ATLAS 890 e KIV-7's.
- Gestão de inventário de equipamentos multimídia; suporte direto a reuniões.
- Suporte a ~35.000 usuários; contas/ativos via Active Directory.

Network Administrator / VTC Technician II
Marine Expeditionary Force Forward — Camp Lejeune, NC (Al-Asad Air Base, Iraque)
Ago 2008 — Jan 2010
Supervisor: Christopher Jarr

- Operação/manutenção de VTC e A&V em todo HQ do II MEF; reparos on-site e troubleshooting.
- Gestão de 3.000+ VTCs (Meeting Room Manager e DVS website).
- Chefe de Comunicações VIP no desdobramento da MNF-W; manutenção e implementação de dispositivos do Comandante.
- Conduziu 1.500+ VTCs para coordenação entre lideranças militares mundiais.
- Planejou upgrades de rede aproveitando hardware existente (economia > US$ 180k); plano de upgrade do Windows para 15.000+ máquinas.
- Liderança de equipe (8 pessoas) em operações diárias, avaliação e disciplina.
- Montagem e manutenção de A&V em escritórios, salas de conferência e auditórios.

Network Specialist / VTC Technician
Combat Logistics Company 16 (CLC-16) — 1st Marine Logistics Group, MCAS Yuma, AZ
Ago 2005 — Ago 2008
Supervisor: Rick Lehron

- Operação/manutenção de VTC e A&V para reuniões da unidade; reparos on-site e troubleshooting.
- Gestão de inventário e funcionamento de projetores LCD, computadores, sistemas de som e equipamentos de videoconferência.
- Descarte de TI/A&V conforme DRMO.
- Bilhetes: Communications NCO (hardware/software, contas, tickets); Training NCO (qualificações anuais e testes físicos); Security NCO (credenciais via JPAS e agendamento de visitantes classificados).
- Plano de instalação/manutenção de conexão SIPRNET no QG da Companhia.
- Relatórios mensais de prontidão (SORTS).
- Reconhecimento: Navy and Marine Corps Achievement Medal (NAM).

Educação

Ensino médio ou equivalente.

Habilidades

Microsoft Excel; Hand Tools; Microsoft Word; Retail Sales; Sales; VTC (Polycom, Cisco); Crestron; AMX; CAD; Visio; PowerPoint; Active Directory; Redes LAN/WAN; ADTRAN ATLAS 890; KIV-7.

Serviço Militar

Força: United States Marine Corps (USMC) — **Graduação:** Sergeant

Observação: este currículo é usado no TCC como exemplo rotulado em origem como *Network Administrator*.

Fonte: Elaborado pelo autor (2025).

- Ajuste de *prompts* para reforçar distinções entre funções com habilidades sobrepostas.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o desenvolvimento e a avaliação de um sistema de triagem automática de currículos, combinando mecanismos de busca semântica com modelos de linguagem de larga escala (LLMs). A proposta se baseia na integração entre vetorização usando *Sentence Transformers* e análise contextual com LLMs, aplicada sobre uma base de currículos reais e rotulados.

Os experimentos conduzidos demonstraram que a escolha do modelo de vetorização influencia diretamente a eficácia do sistema na tarefa de recuperação semântica. Modelos otimizados para busca, como o *Multi QA MiniLM L6 Cos V1*, apresentaram os melhores desempenhos na maioria dos cenários avaliados. Além disso, foi possível constatar que a densidade semântica dos textos influencia no desempenho do sistema: versões excessivamente resumidas, como aquelas geradas com o modelo *Pegasus*, comprometeram a qualidade dos resultados.

Por outro lado, a utilização de LLMs como etapa complementar de ranqueamento mostrou-se eficaz. O modelo *Meta LLaMA 3.1 8B Instruct*, além de ser utilizado na sumarização dos currículos, obteve os melhores resultados na seleção contextual dos candidatos mais adequados, evidenciando a viabilidade de incorporar LLMs mesmo em bases reduzidas, desde que a compressão mantenha as informações mais relevantes.

Dessa forma, pode-se concluir que a estratégia híbrida proposta é tecnicamente viável e pode ser interessante para o desenvolvimento de ferramentas de apoio à triagem automática de currículos. O equilíbrio entre compressão textual, qualidade semântica e custo computacional deve, contudo, ser considerado em futuras implementações em escala real.

As principais contribuições deste trabalho incluem:

- Proposição de uma arquitetura híbrida para triagem de currículos, combinando busca vetorial e LLMs;
- Avaliação sistemática do impacto da sumarização dos currículos na recuperação semântica;
- Criação de uma base de dados derivada com versões resumidas dos currículos, útil para futuros estudos;
- Comparação entre diversos modelos de transformação dos currículos e LLMs.

Apesar dos resultados obtidos, algumas limitações devem ser consideradas:

- O conjunto de dados é limitado a currículos da área de tecnologia da informação, o que pode restringir a generalização dos resultados;

- A avaliação depende de rótulos ocupacionais fixos, não contemplando nuances mais subjetivas do processo de seleção;

- Restrições de *tokens* e limitações de custo impediram o uso de LLMs de maior porte com mais exemplos por consulta.

- Não foram avaliados LLMs comerciais e potencialmente mais eficazes, como os modelos GPT-4o da OpenAI.

Como continuidade deste trabalho, sugerem-se as seguintes linhas de pesquisa futura:

- Aplicação da abordagem proposta a currículos de outras áreas do conhecimento, ampliando a generalização do sistema;

- Desenvolvimento de uma interface interativa com *feedback* do usuário para refinamento do ranqueamento;

- Emprego de técnicas de aprendizado ativo (*Active Learning*) para ajustar os modelos com base em preferências reais dos recrutadores;

- Avaliação do uso de LLMs com contexto expandido, como modelos com janelas longas (GPT-4o, Claude 3 ou Gemini), para lidar com textos maiores sem necessidade de sumarização;

- Estudo sobre o impacto da tradução automática de currículos multilíngues na recuperação semântica.

- Realização de estudos de validação com especialistas da área de Recursos Humanos, a fim de avaliar a adesão dos resultados gerados pelo sistema em contextos reais de recrutamento. A participação de profissionais humanos pode revelar aspectos subjetivos e qualitativos que não são capturados por métricas automáticas, contribuindo para o refinamento do sistema e aumento de sua confiabilidade prática.

REFERÊNCIAS

BHATIA, K. et al. End-to-end resume parsing and ranking system. *arXiv preprint*, 2019. Disponível em: <https://arxiv.org/abs/1910.03089>.

BOMMASANI, R. et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

BREAUGH, J. A. Employee recruitment: Current knowledge and important areas for future research. *Human Resource Management Review*, v. 18, n. 3, p. 103–118, 2008.

BROWN, T. et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. *Semi-supervised learning*. [S.l.]: MIT Press, 2011.

DESSLER, G.; ODERICH, C. L. *Administração de recursos humanos*. [S.l.]: Pearson, 2017.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

FAN, W. et al. A survey on rag meeting llms: Towards retrieval-augmented large language models. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2024. p. 6491–6501.

FERNANDEZ, A. et al. An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, v. 6, n. 2, p. 99–120, 2020.

GAN, Z. et al. A survey of llm-based agents. *arXiv preprint*, 2024. Disponível em: <https://arxiv.org/abs/2401.08315>.

Grupo SERES. *Por que fazer a triagem de currículos de forma automática no RH?* 2023. Acesso em: 28 out. 2024. Disponível em: <https://www.gruposeres.com.br/triagem-de-curriculos/>.

HEAKL, X. et al. Resumeatlas: An llm-based resume understanding benchmark. *arXiv preprint*, 2024. Disponível em: <https://arxiv.org/abs/2406.18125>.

HUANG, M.-H.; RUST, R. T. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, SAGE Publications, v. 61, n. 4, p. 15–42, 2019.

JIECHIEU, K. F. F.; TSOPZE, N. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 2020. Disponível em: <https://doi.org/10.1007/s00521-020-05302-x>.

JOHNSON, J.; DOUZE, M.; JÉGOU, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, IEEE, v. 7, n. 3, p. 535–547, 2019.

LANCETTI, W. et al. Talentjobradar: Advanced data-driven recommendations for in-demand qa soft skills and career opportunities. In: *Anais do Simpósio Brasileiro de Sistemas de Informação (SBSI)*. Recife, PE: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/34348>.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2020. v. 33, p. 9459–9474.

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

LUO, X. et al. Resumenet: A learning-based resume quality assessment model with semi-supervised learning. *arXiv preprint*, 2018. Disponível em: <https://arxiv.org/abs/1810.02832>.

MALIK, M.; FATIMA, N.; AHMAD, T. A survey of machine learning techniques for resume screening. *International Journal of Advanced Computer Science and Applications*, v. 11, n. 5, p. 861–867, 2020.

MALKOV, Y. A.; YASHUNIN, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 42, n. 4, p. 824–836, 2020.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

NETO, M. G. de S.; SARAIVA, F. Resume analysis in portuguese using word embeddings: Development of a decision support system for candidate selection. In: *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. São Paulo, SP: SBC, 2023. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/25765>.

OVERTON, S. *ATS Optimization: The Good, The Bad, and The Ugly*. 2017. Disponível em: <https://www.jobscan.co/blog/ats-optimization/>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2019. p. 3982–3992.

SAATÇI, N.; KURT, H. Resume screening with natural language processing: A practical application using semantic similarity. *Alphanumeric Journal*, v. 12, n. 2, 2024. Disponível em: <https://www.alphanumericjournal.com/media/Issue/volume-12-issue-2-2024/resume-screening-with-natural-language-processing-nlp.pdf>.

SANTOS, L. V. d. *Categorização automática de currículos da área de TIC utilizando aprendizado de máquina*. Dissertação (Dissertação (Mestrado em Informática)) — Instituto Federal do Espírito Santo, Vitória, ES, 2022.

SARMENTO, M.; OLIVEIRA, H. T. A. d. Sumarização automática de artigos de notícias em português: da extração à abstração com abordagens clássicas e modelos neurais. In: *Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. [s.n.], 2024. Disponível em: <https://github.com/laicsiifes/benchmark_ptbr_summ>.

SHI, W. et al. Replug: Retrieval-augmented generation with frozen language models. *arXiv preprint arXiv:2201.11692*, 2023. Disponível em: <https://arxiv.org/abs/2201.11692>.

TAYLOR, S. W.; ARMSTRONG, M. *Armstrong's Handbook of Human Resource Management Practice: A Guide to the Theory and Practice of People Management*. [S.l.]: Kogan Page, 2020.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

ZHANG, J. et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*, 2020.