UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso de LLMs para a classificação de aspectos em feedbacks de desenvolvedores de software

José Eduardo Athayde Ferrarini

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

José Eduardo Athayde Ferrarini

Uso de LLMs para a classificação de aspectos em feedbacks de desenvolvedores de software

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Heloisa de Arruda Camargo

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Ferrarini, José Eduardo Athayde
Uso de LLMs para a classificação de aspectos em feedbacks de desenvolvedores de software / José Eduardo Athayde Ferrarini ; Profa. Dra. Heloisa de Arruda Camargo. – São Carlos, 2025.
68 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. LLMs. 2. Análise de sentimentos baseada em aspectos. 3. RAG. 4. Feedback de profissionais. 5. Fine-tuning. 6. PLN. I. Carmargo, Heloisa de Arruda, orient. II. Título.

José Eduardo Athayde Ferrarini

Uso de LLMs para a classificação de aspectos em feedbacks de desenvolvedores de software

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Heloisa de Arruda Camargo

Original version

São Carlos

2025

Este trabalho é dedicado a minha família,
por todo o amor e suporte.

AGRADECIMENTOS

Agradeço a Deus,

a Professora Heloísa Camargo, pelo suporte e orientação neste trabalho,

e a todos os professores do ICMC-USP, pela dedicação incansável aos alunos e

Minha esposa Cláudia e minhas filhas Thais e Clara, por tanto amor e apoio incondicional, me ajudando sempre a evoluir e me tornar uma versão melhor de mim mesmo!

"Viver é como andar de bicicleta.
É preciso estar em constante movimento para manter o equilíbrio."
Albert Einstein

RESUMO

FERRARINI, J. E. A. **Uso de LLMs para a classificação de aspectos em feedbacks de desenvolvedores de software**. 2025. 68 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

A evolução dos grandes modelos de linguagem (LLMs) oferece inúmeras oportunidades para realizar processamento de textos, inclusive com modelos especializados na língua Portuguesa, incluindo a análise de sentimentos orientada a aspectos. Este trabalho investiga a aplicação de LLMs e modelos de PLN (SBERT e BerTimbau) na classificação de aspectos em feedbacks de desenvolvedores de software. Para isso, foi gerado um dataset com 5700 feedbacks em português, com auxílio do ChatGPT e depois tratado e complementado com intervenção manual. Os feedbacks do arquivo se distribuem entre 15 aspectos definidos para avaliação de desempenho profissional. O BerTimbau foi refinado (fine-tuning) com esse dataset e comparado com versões base (sem *fine-tuning)*, SBERT e diferentes LLMs. Os resultados indicam que as LLMs apresentam maior capacidade de generalização, com métricas equilibradas (precisão e revocação superiores a 80%), enquanto o BerTimbau com *fine-tuning* mostrou desempenho competitivo em alguns aspectos, mas com respostas mais rígidas e menor generalização. A principal contribuição deste trabalho é demonstrar o potencial do uso de LLMs para refinamento de modelos de linguagem menores para realização de tarefas mais específicas, identificando também sugestões para melhorar os resultados obtidos.

Palavras-chave: Análise de Sentimentos Orientada a Aspectos. LLMs. Feedback Corporativo. Fine-tuning. RAG.

ABSTRACT

FERRARINI, J. E. A. **Applying LLMs to aspect classification in software developers' feedback**. 2025. 68 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

The evolution of Large Language Models (LLMs) offers numerous opportunities for text processing, including specialized models for the Portuguese language and Aspect-Based Sentiment Analysis (ABSA). This work investigates the application of LLMs and NLP models (SBERT and BerTimbau) for aspect classification in software developers' performance feedback. To this end, a dataset of 5,700 feedback entries in Portuguese was generated with the support of ChatGPT and later refined and complemented through manual intervention. The dataset covers 15 aspects defined for professional performance evaluation. BerTimbau was fine-tuned on this dataset and compared with its base version (without fine-tuning), SBERT, and different LLMs. Results show that LLMs provide stronger generalization capabilities with balanced metrics (precision and recall above 80%), while fine-tuned BerTimbau achieved competitive performance in some aspects but exhibited more rigid responses and limited generalization. The main contribution of this study is to demonstrate the potential of using LLMs to refine smaller language models for more specific tasks, while also identifying strategies to further improve results.

Keywords: Aspect-Based Sentiment Analysis. LLMs. Corporate Feedback. Fine-tuning. RAG.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos
LLMs	Large Language Models
IA	Inteligência Artificial
RAG	Retrieval Augmented Generation
PLN	Processamento de Língua Natural
LGPD	Lei Geral de Proteção de Dados Pessoais
GPT	Generative Pre-trained Transformer
ML	Machine Learning
NLP	Natural Language Processing
ABSA	Aspect-Based Sentiment Analysis
CSV	Comma-Separated Values
JSON	JavaScript Object Notation
API	Application Programming Interface
GPU	Graphics Processing Unit
ROC	Receiver Operating Characteristic

SUMÁRIO

1 INTRODUÇÃO

A popularização da utilização das tecnologias de Inteligência Artificial (IA) aumentou a partir do surgimento dos grandes modelos de linguagem -*Large Language Models (LLMs)*, sendo este fato marcado pelo lançamento do ChatGPT em 30 de Novembro de 2022 (OpenAI, 2024), sendo seguido depois pelo surgimento de vários outros LLMs (exemplos de LLMs disponíveis em 2025 no quadro 1).

Estes modelos são treinados com uma grande quantidade de dados e são capazes de interpretar e gerar respostas através do processamento de dados não estruturados, o que oferece ao usuário final muitas possibilidades de utilização a um custo inicial acessível quando comparado ao que seria necessário para treinar modelos de aprendizado de máquina, especialmente em situações onde o conjunto de dados anotados é pequeno (Machado, 2023).

Quadro 1 – LLMs disponíveis em 2025, com e sem código aberto

Modelo	**Empresa**	**Código aberto?**
ChatGPT (GPT-4 / GPT-5)	OpenAI	Não
Gemini (ex-Bard)	Google DeepMind	Não
Claude Sonnet 3/4	Anthropic	Não
LLaMA 2/3	Meta (Facebook)	Sim
Mistral 7B / Mixtral 8x7B	Mistral AI	Sim
Falcon 180B	TII (Emirados Árabes)	Sim
Command R+	Cohere	Não
PaLM 2	Google	Não
ChatGLM-6B	Tsinghua University	Sim
Phi-2	Microsoft	Sim
DeepSeek	DeepSeek	Sim

Fonte: Elaborado pelo autor com base em pesquisa na Internet (30/09/2025).

O surgimento do ChatGPT gerou impacto, interesse e preocupação em muitas pessoas, especialmente pela repercussão na imprensa. Isso ainda foi potencializado pelo fato da Inteligência Artificial causar muito receio por representar um potencial ainda não totalmente conhecido. Pessoas que só tiveram contato com o termo IA através de ficção científica passaram a se preocupar com o tema: Quais empregos serão impactados? Quais atividades serão apenas realizadas de forma automática? As máquinas dominarão as pessoas?

Dentro das empresas, principalmente nas equipes ligadas diretamente ao desenvolvimento de software, a preocupação é em aplicar IA para garantir produtividade, sob o risco de perder espaço para os concorrentes. E várias empresas se adiantaram em promover

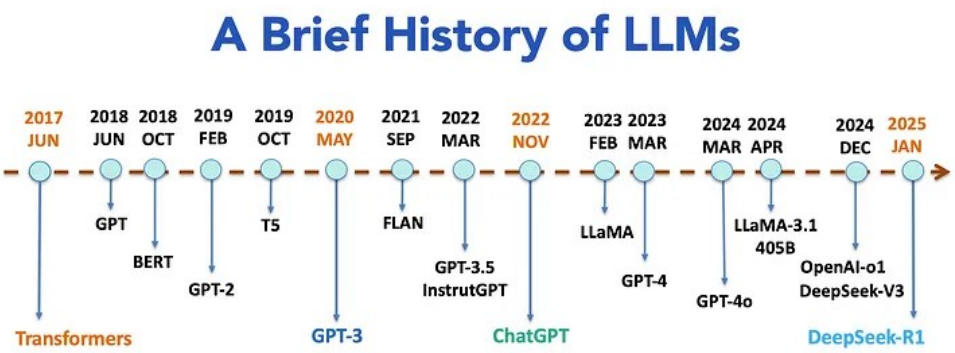

Figura 1 – Principais versões dos LLMs (LMPo, 2025)

aplicação ampla de IA nas suas equipes. De qualquer forma, não há dúvidas do potencial destes modelos em trazer estes ganhos de produtividade. Mas qual a melhor forma de explorá-los? Quais os riscos?

Assim, é importante a realização de estudos acadêmicos e avaliações que ajudem a demonstrar a utilização das LLMs como ferramentas, especialmente num momento onde ocorre o surgimento frequente de novas versões, cada vez mais potentes, numa corrida entre fornecedores em busca da liderança neste mercado (ver figura 1). Manning (2015) diz na conclusão do artigo que este é um período excitante onde a PLN (Processamento de Língua Natural) está no centro tanto do desenvolvimento do aprendizado de máquina quanto na aplicação da resolução de problemas nas empresas, e continua defendendo a importância de trabalhos e estudos com aplicações que ajudem a pensar sobre as línguas, como são processadas e como evoluem, ao invés de apenas se preocupar em buscar eficiência e performance computacional nos modelos.

Então, direcionando o olhar para possibilidades de aplicação, há a análise de sentimentos, que é um campo da PLN e como tal pode se beneficiar bastante da versatilidade das LLMs. Desta forma, há uma contribuição deste trabalho ao avaliar a aplicação de LLMs para realizar análise de textos contendo avaliações de profissionais (*feedbacks*) em língua Portuguesa (que podem ter sido escritos por qualquer pessoa que mantenha relacionamento profissional direto, sendo mais comuns o gestor direto, os colegas e o próprio profissional - como uma auto-avaliação). Para realizar a análise de sentimentos, é muito utilizada uma técnica de identificação de Aspectos (Análise de sentimentos orientada a aspectos (Liu, 2012)). Os aspectos são uma forma de identificar características e partir daí avaliar se são positivas ou negativas.

A motivação em avaliar textos de feedback é importante pelo esforço que esta atividade gera para os gestores de equipes de desenvolvimento de software, uma vez que o processamento destas análises (também conhecida como avaliação de desempenho) é uma das tarefas mais importantes, impopulares e consumidoras de tempo na gestão (Cappelli;

Conyon, 2017, p. 88). Além disso, é recomendado que os textos de feedback contenham uma estrutura de texto mais simples e direta, onde se deve evitar uso rebuscado de linguagem (por exemplo, não é esperado o uso de sarcasmo[1] nestes textos), buscando a maior objetividade possível na comunicação do feedback.

feedback é uma palavra do Inglês, que acabou sendo incorporada inicialmente nas empresas e depois ganhou espaço na própria língua Portuguesa significando:

1. Retorno da informação ou do processo; obtenção de uma resposta;
2. Realimentação;
3. Retroalimentação.

(Michaelis, 2025).

Fornecer feedback tornou-se um recurso muito utilizado para avaliar pessoas, empresas, produtos e serviços. Consiste em uma forma de identificar pontos positivos, negativos e oportunidades de melhoria sobre o que está sendo avaliado, e devido a isso, são necessárias avaliações constantes, e especialmente ao avaliar pessoas, de mais de uma fonte de avaliação.

Apesar de a avaliação dos feedbacks dos profissionais poder ter outros usos (ajustes de salário, por exemplo), o aspecto mais importante diz respeito aos resultados dos profissionais, motivando as pessoas a executar bem seu trabalho, e os feedbacks ganham papel importante em mostrar para as pessoas se estão indo bem, não tão bem e como podem fazer diferente.(Cappelli; Conyon, 2017)

A maneira mais adequada de construir os feedbacks para que sejam efetivos é associar a fatos, números e situações vividas que possam identificar para o profissional o motivo e origem do feedback que está recebendo. Assim, é importante concentrar-se em fatos e não em suposições, tendo foco na mensagem que se deseja transmitir, e abordar o desempenho da pessoa baseado nos objetivos e metas que foram estabelecidos juntos, no início do ciclo de avaliação (Harvard Business Review, 2019, p. 23).

1.1 Propósito

A realização desde trabalho se justifica pela importância de se aprofundar o entendimento sobre a utilização de LLMs em tarefas práticas e cotidianas, e avaliar seu potencial de execução, os riscos e cuidados necessários. Além disso, não foram encontrados trabalhos anteriores que utilizassem LLMs para análise de textos em português no contexto específico da avaliação de feedback de profissionais. A análise proposta neste trabalho

[1] Sarcasmo é uma forma sofisticada de discurso onde o conteúdo dito ou escrito representa o oposto do que significa. No contexto da análise de sentimentos, quando algo positivo é dito, significa na realidade algo negativo, e vice-versa, tornando muito difícil de ser lidado nas análises automáticas (Liu, 2012, p. 52).

traz um componente de multidisciplinaridade, expandindo a utilidade para a área de conhecimento de Administração (desenvolvimento profissional), com a expectativa apoiar gestores de profissionais através de uma discussão prática e útil sobre o tema.

1.2 Objetivo geral

Avaliar a utilização de LLMs para apoiar atividades de avaliação de feedbacks de profissionais em equipes de desenvolvimento de software.

1.2.1 Hipóteses

1. Uma LLM pode ser utilizada para gerar dados de treinamento úteis para refinar um modelo de Processamento de Linguagem Natural apto a realizar uma tarefa de classificação de um conjunto pré-estabelecido de aspectos em textos de feedback de profissionais
2. O modelo com *fine-tuning* irá desempenhar melhor na tarefa específica de classificação de aspectos do que um classificador mais genérico sem *fine-tuning*
3. LLMs apresentarão resultados gerais melhores nas tarefas de classificação de textos que os modelos menores

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Trabalhos Relacionados

Para a fundamentação deste trabalho, foram encontrados estudos que tiveram como principal objetivo analisar a aplicação de métodos para análise e interpretação de textos automática. O foco dos trabalhos é a aplicação e avaliação de algoritmos e técnicas de ciência da computação para tratar os diversos aspectos linguísticos dos textos, o que tipicamente caracteriza a área de estudos de PLN - Processamento de Linguagem Natural (ou numa tradução mais adequada ao Português: Processamento de Língua Natural) (Pardo, 2005). As principais aplicações destes estudos buscavam desenvolver ferramentas para apoio à escrita (tanto ortográfica quanto gramatical) e tradutores automáticos (Pardo, 2005, p. 3), e a complexidade das técnicas aumenta à medida que se aprofunda o nível de conhecimento extraído dos textos. Análises mais básicas de fonética/fonologia, num próximo nível a morfologia, depois sintaxe, depois a semântica até chegar na pragmática/Discurso (Pardo, 2005, p. 4). Desta forma, os trabalhos tem característica multidisciplinar, nas áreas de ciência da computação e linguística, além de outras que podem ter seus conhecimentos utilizados para algumas das técnicas de PLN. Jorge (2022) utilizou as técnicas para avaliação de utilidade em recomendações feitas por usuários de uma plataforma de jogos, enquanto Santos (2023) trabalhou com a classificação de textos jurídicos.

Alguns estudos de PLN passaram a direcionar o objeto para a análise de sentimentos, utilizando inteligência artificial para identificação de sentimentos em textos (em especial a utilização de *transformers*, uma arquitetura de redes neurais). Trabalhos como os de (Liu, 2012), (Freitas; Vieira, 2015), (Taboada, 2016), (Panchendrarajan *et al.*, 2016), (Kauer, 2016) e (Vargas; Pardo, 2018) tem como foco a avaliação ou propostas para a identificação de aspectos e classificação destes dentro dos textos, comparando diferentes implementações e modelos de inteligência artificial. Os trabalhos de Freitas e Vieira (2015) e de Vargas e Pardo (2018) utilizam *datasets* em português, mas apenas o segundo apresenta tratamento de aspectos implícitos. Já o trabalho de Kansaon, Brandão e Pinto (2018) buscou analisar sentimentos diretos e indiretos em textos de *Tweets*[1] em Português.

A partir de 2022, o lançamento do *ChatGPT* da *OpenIA* iniciou uma popularização na utilização de LLMs em tarefas diárias, uma vez que estes modelos são capazes de executar tarefas avançadas de análise e geração de conteúdo sem depender de treinamentos específicos (capacidades de generalização). O interesse causado por este lançamento provocou uma corrida e muitos outros fabricantes lançaram seus próprios LLMs (quadro 1). Machado

[1] *Tweets* são textos curtos limitados a 140 caracteres que eram publicados por usuários da rede social *Twitter*. A partir de 2017 o limite para usuários comuns passou para 280 caracteres. Em 2023 o *Twitter* passou a se chamar *X*.

(2023) já inclui no seu trabalho referência às LLMs na análises como possível uso para tarefas que anteriormente demandaria a aplicação direta de técnicas de PNL. Já os trabalhos de (Ramakrishnan, 2024) e (Silva *et al.*, 2024) são dedicados de forma exclusiva à análise das LLMs. O primeiro, com foco em métodos de aplicação para LLMs, enquanto o segundo já focou na aplicação de LLMs para a identificação de aspectos implícitos e explícitos, em um contexto de críticas gastronômicas para o idioma Português.

Assim, acompanhando a evolução que as técnicas para análises automáticas de textos e especialmente considerando a evolução acelerada dos modelos LLM ainda há poucos trabalhos direcionados à língua Portuguesa, principalmente em comparação do Inglês e ao Chinês. Adicionalmente, não foram encontrados trabalhos que tenham como objetivo a análise de textos de feedback de profissionais, o que representa uma contribuição numa atividade importante para os gerentes de equipes de desenvolvimento de software. A tabela 2 apresenta os trabalhos encontrados com as suas caracterísicas principais.

Quadro 2 – Trabalhos relacionados: Estudos sobre análise de sentimentos e classificação de aspectos.

Trabalhos	Feedback de pessoas	LLM	RAG	PLN	Dataset em Português	Aspectos	
						Explíc.	Implíc.
Pardo (2005)				✓	✓		
Liu (2012)				✓		✓	✓
Freitas; Vieira (2015)				✓		✓	
Taboada (2016)				✓	✓	✓	✓
Panchendrarajan *et al.* (2016)				✓		✓	✓
Kauer (2016)				✓		✓	✓
Vargas; Pardo (2018)				✓	✓	✓	✓
Kansaon; Brandão; Pinto (2018)				✓	✓	✓	✓
Jorge (2022)				✓	✓		
Santos (2023)				✓	✓		
Machado (2023)		✓		✓	✓	✓	✓
Silva *et al.* (2024)		✓			✓	✓	✓
Ramakrishnan (2024)		✓	✓				
Gupta; Ranjan; Singh (2024)		✓	✓	✓			
Gao *et al.* (2024)		✓	✓	✓			
Wu *et al.* (2025)		✓	✓	✓			

Fonte: Elaborado pelo autor (2025).

Legenda:
Feedback de Pessoas: Textos avaliados se referem a feedbacks de pessoas?
LLM: Trabalho utiliza LLMs para avaliar os textos?
RAG: Trabalho utiliza RAG para avaliar os textos?
PLN: Trabalho aplica diretamente os algoritmos de PLN
Dataset em Português: Se o trabalho utiliza textos em língua Portuguesa para análise
Aspectos: Se o trabalho faz mapeamento de aspectos explícitos e/ou implícitos na abordagem

2.2 IA, PLN e LLMs

Apesar da IA ser um objeto de estudo que existe desde os primeiros anos do desenvolvimento computação eletrônica, aparecia para o público em geral mais como um produto de ficção científica. Durante muitas décadas permaneceu num contexto acadêmico e de pesquisa, com utilização práticas em algumas situações e problemas mais específicos.

Uma dessas áreas de IA é a automação de atividades ligadas à linguística, ou seja, atividades relacionadas com interpretação, geração, tradução e classificação de textos. Esta área, conhecida como PLN, possui característica híbrida, atraindo pesquisadores tanto de Ciência da Computação como também de linguística, uma vez que seus trabalhos demandam conhecimento aprofundado em ambas as áreas. Esta é uma das áreas da ciência da computação que por muito tempo desenvolveu estudos incluindo algoritmos de IA para alcançar seus objetivos, desde algoritmos estocásticos até redes neurais, principalmente após o desenvolvimento da arquitetura de *transformers*, que surgiu com este nome em artigo de Vaswani *et al.* (2017), e que foi fundamental para acelerar o desenvolvimento de soluções de PLN.

Desta forma, a partir da década de 2010 (em especial a partir de 2015) o sucesso dos estudos de PLN usando *deep learning*, ou aprendizado profundo, foi marcante e colocou esta área no foco de estudos de inteligência Artificial (Manning, 2015). A partir daí, o desenvolvimento das técnicas de aprendizado profundo junto com a arquitetura de *transformers* viabilizou o desenvolvimento dos grandes modelos de linguagem - *Large Language Models* (LLMs), e estes foram responsáveis por dar visibilidade, popularidade e maior acesso à IA para as pessoas em geral. Afinal, a utilização destes modelos não requer conhecimento prévio, pois um usuário mesmo leigo consegue fazer questionamentos através de uma interface de chat simples e seguir interagindo com os resultados gerados pelo modelo.

Os LLMs são chamados "grandes modelos de linguagem"pois são treinados com uma quantidade muito grande de dados. São bilhões e até trilhões de parâmetros utilizados para a aprendizagem utilizando redes neurais com transformers. Esta quantidade de parâmetros permite aos modelos aprenderem em profundidade e detalhes a utilização da língua e este conhecimento permite a aplicação destes modelos em aplicações como tradução automática, geração de textos, produção de resumos e análise de sentimentos (OpenAI, 2024).

Junto com esta ampla disponibilização de recursos de IA, também surgiram grandes preocupações relacionadas à utilização ética, seus riscos e impactos sociais decorrentes da sua popularização.

2.2.1 Adaptação das LLMs para tarefas específicas

Os resultados obtidos das LLMs podem ser ajustados para atender melhor a tarefas específicas (Ramakrishnan, 2024). Há três técnicas mais comumente utilizadas para isso:

2.2.1.1 Prompting

Consiste no uso mais trivial das LLMs, através da submissão ao modelo de uma pergunta (*prompt*). É necessário levar em consideração que LLMs são modelos pré-treinados com documentos e dados, embarcando nestes modelos possibilidades de vieses e uma limitação quanto à atualização das informações presentes no modelo (Ramakrishnan, 2024). Os custos para adaptar uma LLM dependerão do modelo escolhido e do grau de assertividade exigido pela tarefa.

2.2.1.2 Retreivel-Augmented Generation (RAG)

Em algumas situações, apenas submeter uma consulta (*prompt*) não é o mais adequado. Informações geradas depois da data de treinamento dos modelos não estarão disponíveis para serem incluídas nas respostas (Ramakrishnan, 2024). Pode ser necessário prover informações atuais, ou em outros casos, incluir informações que são privadas do negócio e, portanto, não fazem parte do conjunto de dados utilizados no treinamento do modelo. A técnica RAG consiste em incluir no *prompt* informações atualizadas ou proprietárias, fazendo com que o modelo as utilize para prover as respostas. RAG provou ser efetiva para reduzir as taxas de alucinação[2], bastando para isso solicitar no prompt que o modelo cite os documentos fontes utilizados por ele para gerar a resposta, o que facilita a verificação humana de algum erro na resposta. Tanto na técnica RAG quanto a de *prompting* não alteram a estrutura da LLM treinada, apenas ajustam as perguntas de entrada (prompt) para melhorar a elucidação da resposta gerada (Ramakrishnan, 2024, p. 53).

2.2.1.3 Refinamento de instruções

Esta técnica se adequa a situações onde é necessário processar informações específicas em domínios de negócios especializados (legislações ou bases médicas, por exemplo). Neste caso, o refinamento implica em ajustar pesos internos do modelo e pode resultar em esforço computacional significativo para o retreinamento do modelo, especialmente em LLMs grandes (Ramakrishnan, 2024).

É possível aumentar a efetividade de "raciocínio"e solução de problemas das LLMs usando uma estratégia conhecida como "consulta em cadeia de pensamento"(chain-of-thought prompting) (Ramakrishnan, 2024). Ao invés de limitar a consulta a uma única

[2] alucinação é um termo usado para descrever um fenômeno onde um modelo cria respostas incorretas ou inconsistentes com os dados reais

questão, são oferecidos passos intermediários, simulando como um humano seguiria para construir o raciocínio e chegar à resposta.

2.2.2 Diferentes formas de aplicar RAG

A evolução desta técnica vem apresentando um número crescente de opções para aplicação de RAG nos projetos. Gupta, Ranjan e Singh (2024) mostram que há um conjunto grande de domínios onde a acurácia dos fatos e contextos são exigidas. RAG se mostra eficaz em melhorar a acurácia através da recuperação de informação relevante e a partir daí gerar respostas embasadas nestes dados. Neste trabalho Gupta; Ranjan; Singh indica diferentes técnicas RAG para uso em modelos baseados em texto, imagens, audio ou multimodais.

Para (Gao *et al.*, 2024), RAG surgiu como uma solução promissora para incorporar conhecimento a partir de base de dados externas, aumentando a acurácia e credibilidade na geração de respostas, especialmente para tarefas de conhecimento intensivo, além de permitir a utilização contínua de dados atualizados além da integração de informações de domínio específico. Neste trabalho os autores revisam artigos sobre o assunto e compilam a progressão dos paradigmas de RAG, agrupando as abordagens em "Naive RAG"(RAG ingênuo ou simples), "Advanced RAG"(RAG avançado) e "Modular RAG"(RAG Modular), também trazendo detalhes sobre a estrutura principal dos frameworks RAG, que se dividem em três partes: Retrieval (recuperação), generation (geração) e augmentation (aumentação).

Já (Wu *et al.*, 2025) indica que RAG é uma forma de resolver problemas presentes nos LLMs, como as alucinações, falta de atualização de conteúdos e ausência de conhecimentos para domínios específicos. Neste trabalho os autores revisam técnicas de RAG, detalhando diferentes tarefas e aplicações.

2.2.3 Como avaliar os resultados obtidos pela LLMs?

Avaliar se o resultado gerado pelas LLMs é aceitável pode exigir avaliar múltiplas dimensões, incluindo atualidade da resposta, ausência de erros, relevância, tom e clareza. Humanos podem ser utilizados para fazer esta tarefa, contudo é caro e lento. Uma prática usada atualmente é combinar avaliação por humanos com a utilização de outras LLMs para avaliar os resultados gerados (Ramakrishnan, 2024).

2.2.3.1 Utilizar IA versus Realizar a tarefa manualmente

Identificar vantagem na utilização de Inteligência Artificial ao avaliar feedbacks de profissionais em um time de desenvolvimento ágil é o principal objetivo deste trabalho. (Ramakrishnan, 2024, p. 54) propõe uma equação para esta avaliação:

$$C1 + C2 + C3 < CBAU$$

- C1 = Custos de utilização da solução GenAI
- C2 = Custos para adaptar GenAI para a tarefa
- C3 = Custos para detectar e corrigir erros na GenAI
- CBAU = Custos atuais para realização da tarefa (*Business as Usual (BAU)*)

Segundo Ramakrishnan também é importante considerar riscos de eventuais erros nas respostas geradas pela GenAI. São aceitáveis para a aplicação proposta? Afinal, os modelos LLMs se eximem de responsabilidade sobre os conteúdos gerados, cabendo o risco de eventual erro unicamente a quem utiliza os modelos. Desta forma, se os custos de preparação, utilização, verificação e correção das LLMs são mais baixos que os custos atuais (BAU), então a tarefa é uma boa candidata para construção de um piloto.

2.2.3.2 Riscos de resultados alterados por atualizações nas LLMs

Desenvolver uma aplicação baseada em LLMs exigirá um processo de testes mais completo e recorrente do que apenas para realizar o seu lançamento. Será necessário acompanhamento e validação dos resultados constantes, especialmente se for utilizada uma LLM comercial, pois estas podem sofrer atualizações que alterem o funcionamento de *prompts*, fazendo até alguns já validados anteriormente pararem de funcionar (Ramakrishnan, 2024, p. 55).

2.3 Análise de sentimentos

Análise de Sentimentos é uma área de pesquisa em computação relacionada com processamento de linguagem natural, com o objetivo de analisar opiniões, sentimentos, avaliações, atitudes e emoções em relação a um produto, serviço, organização, indivíduo, problema, evento (Machado, 2023, p. 21). Se posiciona entre a linguística e a ciência da computação, e visa identificar os sentimentos contidos em textos (Taboada, 2016). Liu (2012, p. 7) diz que análise de sentimento (também chamada de mineração de opiniões) é "um campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções de pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, assuntos e seus atributos, e que este campo de estudo representa um problema de grande dimensão."

A análise de sentimentos pode ser aplicada em diferentes profundidades de escopo (Machado, 2023). Quando é realizada em documentos inteiros, o objetivo é identificar

opiniões no texto que determinem a polaridade do documento na sua totalidade, identificando se é uma opinião positiva ou negativa. Já quando é aplicada a cada sentença, busca encontrar a polaridade da sentença, identificando se a mesma é objetiva ou subjetiva, se contém alguma forma de opinião, e apenas após este processo é realizada a análise da sentença em busca da sua polaridade.(Kauer, 2016) também tem como objetivo em seu trabalho determinar a polaridade através da utilização de aspectos. Além de determinar a polaridade dos sentimentos, há outras aplicações práticas, como mostrado por (Jorge, 2022), no qual o autor realiza a previsão de utilidade de avaliações de usuários de jogos. Panchendrarajan *et al.* (2016) estuda a identificação automática de aspectos em textos de avaliações sobre restaurantes, uma mesma linha realizada por (Silva *et al.*, 2024), sendo que este último realiza a identificação de aspectos explícitos e implícitos em críticas gastronômicas em Português. Já (Liu, 2012), realiza a identificação e classificação de emoções e opiniões dentro dos textos, aplicando também análise de subjetividade.

2.3.1 Aplicações de Análise de Sentimentos

Liu (2012, p. 8-9) diz que com o crescimento das mídias sociais em seus múltiplos formatos, indivíduos e organizações passaram a utilizar estes conteúdos disponibilizados para tomada de decisões, como por exemplo a compra de um produto, antes influenciada por amigos ou parentes, agora pode utilizar revisões de outros consumidores ou discussões em fóruns públicos sobre o produto. Organizações podem minerar opiniões públicas disponíveis ao invés de precisarem organizar pesquisas, questionários e grupos focais. Além destes dados externos, muitas organizações também utilizam seus dados internos, como por exemplo, feedback de consumidores coletados a partir de emails e das centrais de antendimento, além de pesquisas próprias.

Além destas aplicações práticas nas organizações, Liu (2012 p. 9-10) também cita a utilização de análise de sentimentos em pesquisas e estudos como:

- a predição de performance em vendas;
- a utilização de revisões para criar *ranking* de produtos e mercadorias;
- as opiniões publicadas em *blogs* e no *Twitter* sobre a liga norte-americana de futebol e as respectiva relação com apostas realizadas;
- a predição de resultados de eleições baseado em sentimentos extraídos de comentários do *Twitter*;
- a proposta de método para prever volumes de comentários em *blogs* políticos;
- a utilização de dados do *Twitter*, revisões e blogs sobre cinema para prever receitas de bilheteria para filmes;

- Análise de emails para estudar diferenças emocionais nos gêneros;
- o mapeamento de emoções em romances e contos de fada;
- Utilização de sentimentos e humores no *Twitter* para prever variações no mercado de ações.

Estes exemplos mostram a amplitude de aplicações e como há potencial de impactos através da aplicação de análise de sentimentos.

2.3.2 Identificação de Aspectos

A análise por aspectos é um modelo de análise de sentimentos que busca determinar as opiniões sobre aspectos específicos de um produto, serviço ou qualquer outra entidade. Um aspecto consiste numa característica comum ao conjunto de elementos em análise, que pode ajudar a classificar este indivíduo dentro do grupo. Por exemplo, para telefones celulares, exemplos de aspectos que podem ser usados seriam "tamanho da tela", "tempo de duração da bateria"e "quantidade de memória". Para realizar esta análise, é essencial identificar quais são as entidades mencionadas e quais os seus respectivos aspectos ou características. A partir daí, a polaridade de cada aspecto é determinada (Machado, 2023, p. 23). Além disso, a própria extração dos aspectos dos textos de forma automática é foco de alguns trabalhos, como por exemplo em (Costa; Pardo, 2020).

Os aspectos podem surgir no texto tanto de forma implícita quando explícita. Em alguns tipos de texto, em determinadas culturas, os aspectos implícitos existem em aproximadamente 30% dos textos (Panchendrarajan *et al.*, 2016). Há também muitas vezes a necessidade de agregar entidades em um mesmo aspecto num contexto em específico (ou são sinônimos ou muito próximos), como por exemplo, "preço"e "custo".

Um outro desafio encontrado é o idioma para a análise. A maior parte dos trabalhos e estudos é realizado em Inglês ou Chinês. Em seu trabalho, Machado (2023) localizou uma quantidade pequena de trabalhos em Português, sendo que apenas dois *datasets* publicados com identificação de aspectos implícitos.

2.4 Avaliação de feedback de funcionários

Esta é uma atividade muito desafiadora para o gestor pelo impacto que pode representar no desenvolvimento dos profissionais e pala ausência de um processo objetivo e direto para gerar os feedbacks. Para avaliar um profissional, o gestor precisa reunir informação de escuta, convivência, documentos e observação dos profissionais enquanto desempenham suas atividades, além dos resultados alcançados. Em algumas empresas, estas avaliações são semestrais, a aí o desafio para o gestor também passa a ser o de lembrar de tantas informações sobre cada pessoa do time. Este feedback, normalmente

feito semestralmente ou anualmente é conhecido como **Feedback Formal**. Por isso é recomendado aos gestores a prática do **Feedback Constante**, que acontece de maneira frequente ou *ad hoc* (Harvard Business Review, 2019, p. 12-13). Este último é muito importante para garantir intervenções adequadas, permitindo ajustes necessários para o funcionário atingir suas metas e reforço positivo nos aspectos que estão sendo corretamente executados.

É importante que ambos os tipos de feedback sejam utilizados, uma vez que o feedback formal representa um fechamento de ciclo, resumindo todas as avaliações durante o período e gera uma reflexão fundamental para gerar as expectativas e objetivos para o próximo ciclo. Já o feedback constante é fundamental para garantir que o funcionário não será surpreendido com o feedback do final do ciclo, além de permitir que correções sejam realizadas ao longo do período, em busca de melhorar o atingimento das metas (Harvard Business Review, 2019, p. 14).

2.4.1 Documentação dos feedbacks dentro das organizações

É fundamental a organização de um histórico dos profissionais, com pontos e observações registradas ao longo do período de avaliação, sendo muito importante, principalmente em situações onde há queda de desempenho e a necessidade de tomada de decisão sobre a manutenção de uma pessoa na equipe. Assim, alguns pontos importantes para construir esta documentação [3] (Harvard Business Review, 2019, p. 112-113):

- Registrar data e detalhes do que aconteceu
- Evitar julgamentos, dando foco aos fatos ocorridos
- As anotações devem ser feitas preferencialmente no mesmo dia em que foi dados feedback a alguém
- Anexar emails ou anotações destaquem realizações, tendo sido estas testemunhadas ou elogios fornecidos por outras pessoas
- Para problemas de desempenho, além de documentar o problema é importante incluir o que foi acordado como próximos passos, ações planejadas, prazos e resultados esperados
- Solicitar feedback de outras pessoas (outros colegas, clientes, outros gestores etc.) que estejam em posição de avaliar o profissional.

[3] É importante ter em mente que o registro de histórico dos profissionais precisa respeitar as leis de privacidade e proteção de dados (no Brasil, a LGPD - Lei Geral de Proteção de Dados Pessoais, sendo necessário que os gestores consultem os setores jurídicos das empresas.

- Solicitar relatórios informais dos próprios profissionais sobre a sua evolução, eventuais problemas, ajustes e ações tomadas

3 METODOLOGIA

A abordagem escolhida foi a análise comparativa dos resultados obtidos pela execução de um mesmo conjunto de dados por diferentes modelos, realizando a comparação entre LLMs e modelos PLN no contexto da classificação de aspectos. Neste trabalho foi feita a análise do resultado das execuções dos modelos SBERT, BerTimbau (que neste trabalho também recebeu um refinamento, ou *fine-tuning* customizado), e LLMs (ChatGPT 4 e 5, Claude Sonnet e Gemini). Para comparar os resultados foram calculados os indicadores:

- **Acurácia (*Accuracy*):** calcula a proporção de previsões corretas, ou sobre o total de casos. Indica o desempenho global do modelo. Neste trabalho, um acerto significa que o modelo corretamente identificou um aspecto que está de fato presente no feedback, ou que acertamente não indicou um aspecto que não está presente no texto.

- **Precisão (*Precision*)**: calcula, a partir dos casos identificados como positivos pelo modelo, quais de fato eram positivos. Indica a confiabilidade das previsões positivas feitas pelo modelo. Neste trabalho uma precisão maior indica que o modelo erra menos quando indica que um aspecto está presente em um feedback.

- **Revocação / sensibilidade (*Recall* / *Sensitivity*):** mede a capacidade do modelo de encontrar todas as ocorrências positivas. Uma alta sensibilidade indica que o modelo responde poucos "falsos negativos", ou seja, o modelo tem poucos casos onde deixa de indicar um aspecto que estava presente no feedback.

- **Especificidade (*Specificity*):** calcula a capacidade do modelo em indicar corretamente que um aspecto não está presente em um feedback. Ou seja, o modelo indica corretamente que um aspecto não está presente em um feedback.

- **F1-Score:** é a média harmônica entre Precisão e Revocação (*Precision* e Recall). Mostra o equilíbrio do modelo entre a precisão e revocação, e é útil quando o foco é observar se os modelos evitam de forma mais equilibrada os falsos negativos e falsos positivos.

- **Youden's J:** dá uma medida de desempenho global do modelo, utilizando Sensibilidade e Especificidade. Mede a capacidade do modelo em distinguir entre classes positivas e negativas:

```
J = Sensibilidade + Especificidade - 1
```

A motivação desta análise foi discutir a possibilidade de utilizar modelos que tenham modelos de custo, disponibilidade e principalmente que permitam garantir uma maior privacidade e segurança sobre dados sensíveis que possam precisar ser processados como conteúdo RAG.

3.1 Uso de RAG neste trabalho

A Utilização de RAG neste trabalho se justifica por:

- Utilizar informações específicas de uma organização sobre o processo de avaliação. No caso do experimento, os aspectos escolhidos e sua definição podem ser ajustados de acordo com o interesse da empresa. Isso é especialmente útil para quando se usa as LLMs para processar estas informações

- Utilizar informações específicas sobre objetivos dos profissionais e dados de contexto do projeto para tornar a classificação de aspectos mais acurada. Estes dados são sensíveis e não devem ser processados por um modelo de LLM aberto, conforme a LGPD (Lei Geral de Proteção de Dados Pessoais - Lei nº 13.709/2018)

3.2 Aspectos

Para permitir uma comparação para a classificação dos aspectos, foi pré-determinada uma lista de aspectos para os feedbacks analisados. Os aspectos foram incluídos em um arquivo no formato json, para serem utilizado como informação adicional (RAG) no processamento dos textos de feedback. Foram adicionados 15 aspectos ao arquivo, sendo eles:

Quadro 3 – Aspectos considerados na análise de feedbacks

Categoria	**Aspectos**
Comportamentais	Colaboração, Comunicação eficaz, Empatia, Relacionamento interpessoal, Resiliência.
Técnicos e de desempenho	Conhecimentos Técnicos, Qualidade das entregas, Gestão do tempo, Pontualidade, Organização.
Pessoais e de atitude	Autoconfiança, Autonomia, Proatividade, Adaptabilidade, Inovação e Criatividade.

Fonte: Elaborado pelo autor (2025).

Exemplo de um dos aspectos:

```
{
    "aspecto": "Colaboração",
    "definicao": "Capacidade de trabalhar em equipe, contribuindo
                  ativamente com os colegas, compartilhando
                  conhecimento e ajudando a alcançar objetivos
                  comuns.",
    "sinonimos": ["Trabalho em equipe", "Cooperação",
                  "Parceria", "Apoio mútuo"]
 }
```

Cada aspecto, além do nome que o identifica, recebe uma definição e um conjunto de sinônimos. Aqui neste trabalho não se pretende afirmar que esta lista de aspectos represente o conjunto mais adequado, nem que seja uma lista definitiva para ser usada em análises de feedback. Tampouco a descrição dada e os sinônimos propostos são frutos de estudo especializado para sua definição. Assim, apenas representa um conjunto de aspectos que fazem sentido ao contexto, a partir da visão deste autor, e que servem neste estudo como estrutura para comparação dos resultados entre os diferentes modelos.

3.3 Documentos e *datasets* utilizados

Os documentos abaixo foram utilizados com o objetivo de equalizar algumas definições para os modelos terem o mesmo ponto de partida para realizar a classificação dos feedbacks.

3.3.1 Documento de expectativas por cargo

Documento contendo a definição dos cargos para a empresa. Neste documento, para cada cargo, haverá uma descrição das expectativas que devem ser atendidas para o profissional atender adequadamente ao cargo ocupado. Usado na geração das frases de feedback para o Dataset de refinamento do BerTimbau.

3.3.2 Dataset Aspectos

Arquivo tipo ".json"que traz uma definição para cada um dos 15 aspectos utilizados neste trabalho. Também traz uma lista de sinônimos. Este arquivo foi usado para gerar o Dataset para refinamento do modelo BerTimbau (*fine-tuning*) e também como RAG durante a execução dos experimentos com todos os modelos.

3.3.3 Dataset Feedbacks

Arquivo tipo ".json"contendo 25 feedbacks usados para avaliar os modelos.

3.4 Avaliação da classificação de aspectos: comparação

Esta avaliação tem por objetivo comparar resultado de uma LLM com resultados obtidos por outros modelos de PLN, em uma atividade específica: A identificação/classificação de aspectos dentro de textos de feedback. Também é pretendido demonstrar a viabilidade de se gerar feedbacks em língua Portuguesa para criar o dataset de treinamento utilizado no *fine-tuning*. A comparação foi realizada entre:

- SBert
- BerTimbau base
- BerTimbau com *fine-tuning*
- LLM

3.4.1 Dataset para *fine-tuning* de classificação

Para realizar o *fine-tuning* do BerTimbau foi utilizado o modelo "*neuralmind/bert-base-portuguese-cased*". Não foi possível localizar, durante a elaboração deste trabalho, um dataset contendo frases ou conteúdos de feedback relacionados à profissionais de desenvolvimento de software. Assim, foi necessário criar um dataset para este fim, utilizando o ChatGPT 4.1, entre Julho e Agosto de 2025 (OpenAI, 2024). O arquivo foi formado a partir da execução sucessiva de solicitações via prompt para o ChatGPT. Foram gerados blocos de frases (geralmente 30, 40 ou 50) de feedback para cada combinação ("aspecto"x "sentimento"), pedindo que o ChatGPT respeitasse a definição do aspecto contida no arquivo "aspectos.json". Exemplo de PROMPT aplicado para a construção do dataset:

```
"Gerar mais dados para adicionar ao arquivo
"Aspect_Classification_fine_tuning.csv"
para o aspecto "Relacionamento interpessoal"
(usando como referência dados deste aspecto
do arquivo "aspectos.json") criar novos e
identicos 90 feedbacks "Negativo",
e cada linha criada terá coluna label = 1"
```

Através da execução sucessiva deste prompt, para cada aspecto, foram gerados 2700 frases no total (15 aspectos x 2 sentimentos (positivo ou negativo) x 90 feedbacks).

Depois foram executados prompts solicitando a criação de feedbacks que não tivessem relação com um determinado aspecto:

```
"Gerar mais dados para adicionar ao arquivo
```

```
"Aspect_Classification_fine_tuning.csv" que não seja relacionado
com o aspecto "Relacionamento interpessoal" (usando como
referência dados deste aspecto do arquivo "aspectos.json")
criar novos e identicos 90 feedbacks "Negativo", e cada
linha criada terá coluna label = 0"
```

Com isso, mais 2700 feedbacks foram criados, contudo com a label = 0, indicando que a frase de feedback não está relacionada com um determinado aspecto. Tamanho final total do dataset = 5400 exemplos. (arquivo "Aspect_Classification_fine_tuning.csv")

Após os feedbacks gerados, foi criado um script Python para tratamento dos dados, responsável por realizar as seguintes validações:

1. **Remoção de duplicatas:** O Script varre o dataset inteiro e identifica Feedbacks idênticos. Para tratar, as ações executadas foram:

    - feedbacks duplicados foram ajustados manualmente pelo autor, através de substituição de palavras por sinônimos, ou pequena alteração na forma de escrever, mantendo o sentido do conteúdo do feedback gerado anteriormente via LLM; ou

    - frase do feedback foi totalmente reescrita pelo autor. Nestes casos, geralmente a frase foi criada totalmente manualmente pelo autor; ou

    - remoção de feedbacks duplicados, substituindo os mesmos por novos feedbacks gerados pela LLM (ChatGPT 4.1)

2. **Verificação de equilíbrio do Dataset:** O Script verifica se todas as combinações de classificação de feedbacks estão com a mesma quantidade de amostras (no caso 180 feedbacks para cada combinação de aspecto com cada label).

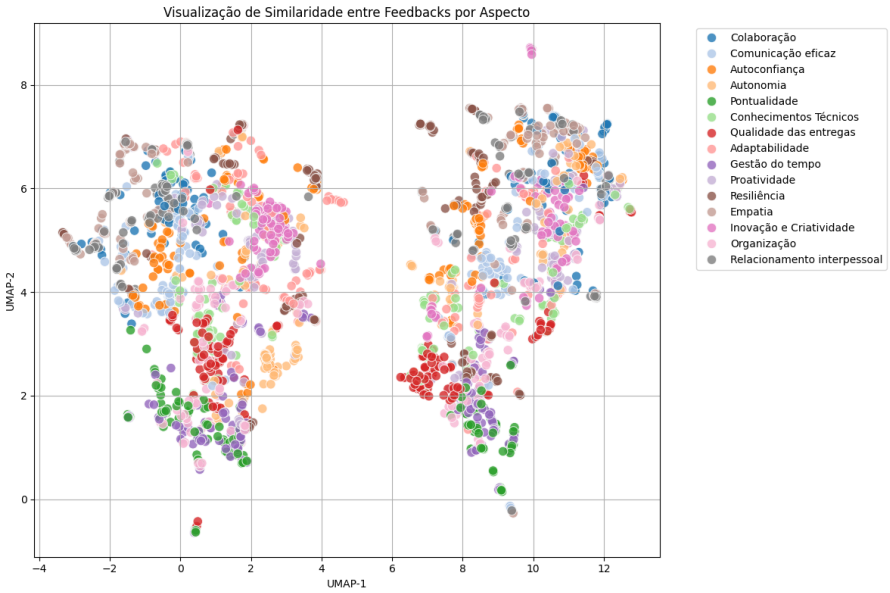

Figura 2 – Similaridade dos feedbacks - Dataset de *fine-tuning*

3.4.2 Resultados do fine-tuning - BerTimbau - label binário

A estratégia do label binário busca aumentar a capacidade do modelo em identificar positivamente aspectos que existam no texto (label=1), assim como aumentar a capacidade do modelo de identificar quando um aspecto não está relacionado ao feedback (label=0). Resumidamente o *fine-tuning* treina o modelo para identificar frases que se relacionam com um aspecto assim como frases que não se relacionam com o aspecto.

Como neste trabalho foram considerados 15 aspectos, o Dataset de treinamento possui a mesma quantidade de exemplos positivos (label = 1) para cada um dos aspectos. Da mesma forma, existe a mesma quantidade de exemplos negativos (label = 0), onde o aspecto indicado não está presente no respectivo feedback.

Foram executados alguns experimentos de *fine-tuning*, conforme resultados apresentados a seguir:

Execução inicial já atinge um resultado bastante positivo.

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.456800	0.225034	0.924074	0.923854
2	0.235900	0.251454	0.929630	0.929490
3	0.202100	0.252146	0.935185	0.935148

Tabela 1 – Resultado de *fine-tuning* com 3 épocas e learning rate de 1e-5

No segundo experimento, a redução do learning rate fez o resultado ao final de 3 épocas ficar abaixo do obtido no primeiro experimento de *fine-tuning*.

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.672000	0.610042	0.746296	0.744013
2	0.603900	0.503263	0.770370	0.769561
3	0.560900	0.472678	0.785185	0.785138

Tabela 2 – Resultado de *fine-tuning* com 3 épocas e learning rate de 1e-6

E a execução escolhida, por alcançar melhor acurácia e métrica F1: Redução do learning rate combinada com aumento do treinamento para 5 épocas.

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.544400	0.307022	0.883333	0.882996
2	0.319700	0.247092	0.909259	0.909064
3	0.268400	0.238568	0.935185	0.935185
4	0.225800	0.257799	0.929630	0.929629
5	0.200000	0.250984	0.938889	0.938887

Tabela 3 – Resultado de *fine-tuning* com 5 épocas e learning rate de 5e-6.

3.4.3 Conjunto de feedbacks usados para testar os modelos

Para a execução do experimento, foram escritos 25 feedbacks pelo próprio autor. Como o *fine-tuning* foi realizado com a maioria dos feedbacks criados por LLM, o objetivo de ter feedbacks manuais no teste final foi evitar um possível viés que melhorasse os resultados do BerTimbau FT artificialmente, pela tendência da escrita por LLM ficar mais "similar"com os feedbacks do treinamento do modelo. Além disso, esta estratégia simula o cenário imaginado para uma aplicação real deste cenário: Ter apoio de um LLM para gerar dados de refinamento, para depois usar no processamento de feedbacks elaborados por humanos. Ou seja, os textos podem ser menos diretos, podem conter subjetividade e principalmente estilos de redação diferentes. Estes feedbacks também continham diferentes tamanhos (quantidade de sentenças), e foram construídos de forma que, cada aspecto tivesse pelo menos 4 sentenças onde estivessem presentes.

3.4.4 Execução do Experimento

aspectos.json: Arquivo contém os 15 aspectos, passados como RAG no código para complementar o processamento dos dados pelo script

feedbacks.json: Contém 25 feedbacks não-rotulados utilizados como massa de testes para este trabalho.

analise_feedbacks_comparada.py: Script que carrega os 3 modelos para comparação: SBERT, BerTimbau base e o BerTimbau FT, obtido neste trabalho realizando o refinamento (*fine-tuning*).

Para fins de análise de dados, os modelos foram executados com limiar bem baixo 0.1, possibilitando que a avaliação e filtro fosse realizado posteriormente, durante a própria análise dos dados.

Num segundo momento, os mesmos arquivos aspectos.json e feedbacks.json foram incluídos num prompt e executados em diferente LLMs com o seguinte prompt:

```
Leia cada um dos feedbacks contidos no arquivo "feedbacks.json", e
para cada um deles, avalie quais aspectos estão relacionados com este
feedback, considerando a lista de aspectos contida no arquivo
"aspectos.json". Quando encontrar uma relação, explique a razão do
aspecto estar presente no feedback.
```

Para o experimento foram utilizadas as LLMs:

- ChatGPT 5
- ChatGPT 4.1 (através do copilot)
- Claude Sonnet 4
- Gemini 2.5 flash

Os resultados foram compilados e analisados com ajuda de planilha eletrônica disponibilizada no GitHub no arquivo feedbacks_aspectos.xls.

3.4.5 *Scripts* gerados para execução dos experimentos

`arquivo: analise_feedbacks_comparada.py`

Para o desenvolvimento deste trabalho foi criado um script em python, com ajuda do GitHub Copilot usando os modelos *GPT 4.1* e *Claude Sonnet 4*, responsável por executar uma pipeline para carregar os dados de "aspectos.json"e "feedbacks.json", e executar para cada feedback análise pelos modelos SBERT, BerTimbau base e BerTimbau FT, gravando o resultado final da probabilidade/similaridade no arquivo "comparacao.json".

`arquivo: TCC_FT_BT_ClassificaAspectos.ipynb`

Também foi criado um notebook que foi executado no google colab, para fazer uso do recurso de GPU, para realizar o *fine-tuning* que gerou o modelo BerTimbau FT. Este notebook foi também criado com ajuda do GitHub Copilot usando o modelo *Claude Sonnet 4* e também o google colab usando modelo *Gemini*

4 AVALIAÇÃO EXPERIMENTAL

Neste trabalho foi realizada a comparação de resultados de classificação de aspectos para os modelos SBERT, BerTimbau, BerTimbau com *Fine-Tuning*, e LLMs.

SBERT = "sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2"

BerTimbau base = "neuralmind/bert-base-portuguese-cased"

BerTimbau *fine-tuning* = "FT_BT_ClassificaAspectos", *fine-tuning* realizado neste trabalho, a partir do BerTimbau base para língua portuguesa e dataset também construído neste trabalho.

LLMs = ChatGPT 4.1, ChatGPT 5.0, Claude Sonnet 4 e Gemini 2.5 flash

A execução do teste foi dividida em duas etapas práticas: Na primeira, usando script python, foram obtidos os resultados de similaridade gerados pelos modelos SBERT, BerTimbau e BerTimbau FT. A partir dos valores obtidos foi realizado uma análise para identificar o limiar mais equilibrado para determinar quando o modelo classifica um aspectos como presente em um feedback. Assim, através de um segundo script python, foi determinado qual seria o melhor limiar para ser considerado como ponto de classificação de cada modelo.

Para os modelos SBERT e BerTimbau, a análise ROC dos resultados obtidos sugeriu os seguintes limiares:

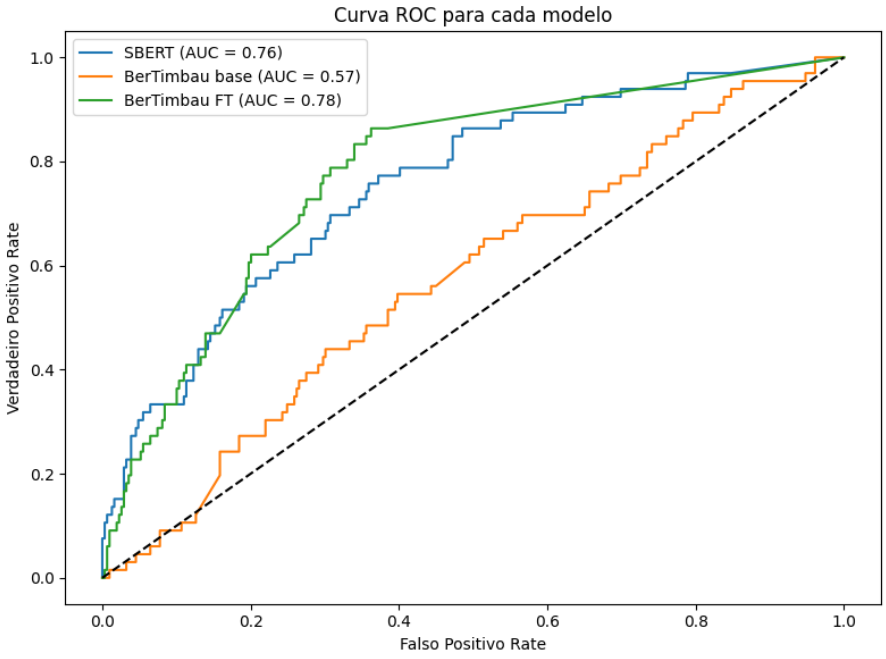

Figura 3 – Análise ROC

Melhor limiar para SBERT: 0.33 (Youden J = 0.41)

Melhor limiar para BerTimbau base: 0.47 (Youden J = 0.14)

Melhor limiar para BerTimbau FT: 0.15 (Youden J = 0.51)

Assim os limiares foram aplicados para considerar quando assumir que o modelo classificou um aspecto como existente para um feedback. A otimização feita baseada no indicador de Yonden J foi usado pois busca o equilíbrio entre os verdadeiros positivos (VP) e verdadeiros negativos (VN). O BerTimbau base não demonstrou ser adequado para a classificação dos aspectos, pois não conseguiu diferenciar suficientemente para a tarefa, trazendo sempre todos os aspectos com valores muito próximos, mas foi uma referência para a comparação com os resultados obtidos pelo modelo BerTimbau FT (com *fine-tuning*).

4.1 Resultados consolidados por modelo

Identificação dos aspectos na amostra de feedbacks: Consolidado dos LLMs (ChatGPT 5, ChatGPT 4.1, Claude 4 e Gemini)	Label						Acurácia	Precisão	Revocação/ Sensibilidade	Especificidade	F1-Score	Youden's J
	1	0	VP	VN	FP	FN						
Colaboração	16%	84%	12	78	6	4	90%	67%	75%	93%	71%	68%
Comunicação Eficaz	24%	76%	21	73	3	3	94%	88%	88%	96%	88%	84%
Autoconfiança	16%	84%	11	81	3	5	92%	79%	69%	96%	73%	65%
Autonomia	16%	84%	12	83	1	4	95%	92%	75%	99%	83%	74%
Pontualidade	16%	84%	13	82	2	3	95%	87%	81%	98%	84%	79%
Conhecimentos Técnicos	20%	80%	17	72	8	3	89%	68%	85%	90%	76%	75%
Qualidade das entregas	16%	84%	15	81	3	1	96%	83%	94%	96%	88%	90%
Adaptabilidade	16%	84%	13	82	2	3	95%	87%	81%	98%	84%	79%
Gestão do tempo	16%	84%	13	80	4	3	93%	76%	81%	95%	79%	76%
Proatividade	24%	76%	10	75	1	14	85%	91%	42%	99%	57%	40%
Resiliência	16%	84%	16	82	2	0	98%	89%	100%	98%	94%	98%
Empatia	16%	84%	15	80	4	1	95%	79%	94%	95%	86%	89%
Inovação e Criatividade	16%	84%	3	83	1	13	86%	75%	19%	99%	30%	18%
Organização	16%	84%	6	81	3	10	87%	67%	38%	96%	48%	34%
Relacionamento interpessoal	20%	80%	15	74	6	5	89%	71%	75%	93%	73%	68%

Figura 4 – Análise resultados BerTimbau FT

LLMs: apresentam métricas mais equilibradas na maioria dos aspectos (acurácia, precisão e recall geralmente $> 80\%$). O ponto forte das LLMs é a capacidade de generalização, lidando melhor com as variações nos conteúdos dos feedbacks. Isso se reflete em resultados mais estáveis mesmo com diferentes formas de e estilos de escrita, e com textos de feedback com tamanhos variados.

Identificação dos aspectos na amostra de feedbacks: SBERT	Label						Acurácia	Precisão	Revocação/ Sensibilidade	Especificidade	F1-Score	Youden's J
	1	0	VP	VN	FP	FN						
Colaboração	16%	84%	4	16	5	0	80%	44%	100%	76%	62%	76%
Comunicação Eficaz	24%	76%	5	16	3	1	84%	63%	83%	84%	71%	68%
Autoconfiança	16%	84%	4	11	10	0	60%	29%	100%	52%	44%	52%
Autonomia	16%	84%	3	10	11	1	52%	21%	75%	48%	33%	23%
Pontualidade	16%	84%	2	14	7	2	64%	22%	50%	67%	31%	17%
Conhecimentos Técnicos	20%	80%	2	16	4	3	72%	33%	40%	80%	36%	20%
Qualidade das entregas	16%	84%	2	9	12	2	44%	14%	50%	43%	22%	-7%
Adaptabilidade	16%	84%	3	10	11	1	52%	21%	75%	48%	33%	23%
Gestão do tempo	16%	84%	3	11	10	1	56%	23%	75%	52%	35%	27%
Proatividade	24%	76%	5	12	7	1	68%	42%	83%	63%	56%	46%
Resiliência	16%	84%	4	15	6	0	76%	40%	100%	71%	57%	71%
Empatia	16%	84%	4	13	8	0	68%	33%	100%	62%	50%	62%
Inovação e Criatividade	16%	84%	4	12	9	0	64%	31%	100%	57%	47%	57%
Organização	16%	84%	2	15	6	2	68%	25%	50%	71%	33%	21%
Relacionamento interpessoal	20%	80%	4	15	5	1	76%	44%	80%	75%	57%	55%

Figura 5 – Análise resultados SBERT

SBERT: desempenho mais inconsistente, com acurácia média menor (muitos aspectos na faixa de 50–67%). Em geral, revocação é razoável, mas com precisão baixa, o que tende a gerar falsos positivos.

Identificação dos aspectos na amostra de feedbacks: BerTimbau fine-tunning	Label						Acurácia	Precisão	Revocação/ Sensibilidade	Especificidade	F1-Score	Youden's J
	1	0	VP	VN	FP	FN						
Colaboração	16%	84%	4	13	8	0	68%	33%	100%	62%	50%	62%
Comunicação Eficaz	24%	76%	6	13	6	0	76%	50%	100%	68%	67%	68%
Autoconfiança	16%	84%	3	9	12	1	48%	20%	75%	43%	32%	18%
Autonomia	16%	84%	4	9	12	0	52%	25%	100%	43%	40%	43%
Pontualidade	16%	84%	4	16	5	0	80%	44%	100%	76%	62%	76%
Conhecimentos Técnicos	20%	80%	4	15	5	1	76%	44%	80%	75%	57%	55%
Qualidade das entregas	16%	84%	3	19	2	1	88%	60%	75%	90%	67%	65%
Adaptabilidade	16%	84%	4	10	11	0	56%	27%	100%	48%	42%	48%
Gestão do tempo	16%	84%	4	15	6	0	76%	40%	100%	71%	57%	71%
Proatividade	24%	76%	4	8	11	2	48%	27%	67%	42%	38%	9%
Resiliência	16%	84%	4	13	8	0	68%	33%	100%	62%	50%	62%
Empatia	16%	84%	4	12	9	0	64%	31%	100%	57%	47%	57%
Inovação e Criatividade	16%	84%	4	18	3	0	88%	57%	100%	86%	73%	86%
Organização	16%	84%	3	15	6	1	72%	33%	75%	71%	46%	46%
Relacionamento interpessoal	20%	80%	2	10	10	3	48%	17%	40%	50%	24%	-10%

Figura 6 – Análise resultados BerTimbau *fine-tuning*

BerTimbau FT: acerta bem em alguns aspectos (como por exemplo: Qualidade das entregas e Inovação e Criatividade), mas falha em outros (Proatividade e Relacionamento interpessoal). Tem revocação alta em alguns casos, mas à custa de baixa precisão. Apresentou diferença significativa em relação ao modelo base (**BerTimbau base**) na tarefa de classificação dos feedbacks. Tende a dar respostas mais binárias, com valores ou muito próximos de 0 ou muito próximos de 1 (entre 0.8 ou 0.98). Esse comportamento indica que o processo de *fine-tuning* aplicado ao modelo criou um classificador mais decisivo, mas também muito atrelado aos padrões observados nos dados de treino. Assim, quando encontra um padrão familiar, o modelo responde de forma assertiva, porém tende a descartar totalmente o que difere do padrão aprendido. Indicando uma redução na capacidade de generalizar para qualquer tipo de texto. Assim, demonstra excelente capacidade para idenficar aspectos principais, mas devido a este padrão de resposta mais rígido ("tudo ou nada"), ignora nuances secundárias.

5 CONCLUSÕES

As duas hipóteses testadas foram confirmadas com os experimentos realizados neste trabalho. A hipótese 1:

"Uma LLM pode ser utilizada para gerar dados de treinamento úteis para refinar um modelo de Processamento de Linguagem Natural apto a realizar uma tarefa de classificação de um conjunto pré-estabelecido de aspectos em textos de feedback de profissionais"

Os experimentos realizados neste trabalho confirmaram esta hipótese.

O (OpenAI, 2024) foi utilizado para gerar uma massa de dados de feedbacks, com conteúdo relacionado a aspectos previamente definidos. Esta massa de dados foi então utilizada para realizar um *fine-tuning* de um modelo BerTimbau base, já baseado na língua Portuguesa.

O treinamento realizado melhorou a capacidade deste modelo para realizar a tarefa específica de classificação para o conjunto de aspectos proposto neste trabalho. A comparação direta entre o BerTimbau Base e o BerTimbau FT evidencia o impacto do processo de *fine-tuning*: O modelo base retorna scores muito próximos entre si, praticamente atuando de forma aleatória. Já o modelo refinado diferencia de maneira assertiva os aspectos, apresentando scores elevados para casos relevantes.

Apesar da melhora nos resultados com o *fine-tuning*, ainda foram observadas limitações, como a atribuição de scores idênticos para múltiplos aspectos em alguns feedbacks mais longos e complexos, o que pode exigir uma estratégia mais aprimorada de embedding ou, como foi testado neste estudo, o processamento em separado por sentença.

hipótese 2:

"O modelo com *fine-tuning* irá desempenhar melhor na tarefa específica de classificação de aspectos do que um classificador mais genérico sem *fine-tuning*"

É possível confirmar esta hipótese, pois o modelo com refinamento do BerTimbau foi capaz de superar o SBERT em vários aspectos avaliados neste trabalho. Apesar de ser mais assertivo e com maior dificuldade de generalização que o SBERT, demonstrou eficiência maior na classificação dos aspectos. Além disso, estratégias de melhoria no dataset de treinamento tendem a melhorar mais ainda os resultados obtidos, como por exemplo, ajustes e melhorias no dataset de treinamento.

hipótese 3: **"LLMs apresentarão resultados gerais melhores nas tarefas de classificação de textos que os modelos menores"**

Esta hipótese também se confirmou, conforme explicado na próxima sessão (5.1), que apresenta os resultados gerais .

5.1 Resultado geral

Melhores resultados por Aspecto X Indicador	Acurácia	Precisão	Revocação/ Sensibilidade	Especificidade	F1-Score	Youden's J
Colaboração	LLMs	LLMs	SBERT	LLMs	LLMs	SBERT
Comunicação Eficaz	LLMs	LLMs	BTFT	LLMs	LLMs	LLMs
Autoconfiança	LLMs	LLMs	SBERT	LLMs	LLMs	LLMs
Autonomia	LLMs	LLMs	SBERT	LLMs	LLMs	LLMs
Pontualidade	LLMs	LLMs	BTFT	LLMs	LLMs	LLMs
Conhecimentos Técnicos	LLMs	LLMs	LLMs	LLMs	LLMs	LLMs
Qualidade das entregas	LLMs	LLMs	LLMs	LLMs	LLMs	LLMs
Adaptabilidade	LLMs	LLMs	BTFT	LLMs	LLMs	LLMs
Gestão do tempo	LLMs	LLMs	BTFT	LLMs	LLMs	LLMs
Proatividade	LLMs	LLMs	SBERT	LLMs	LLMs	SBERT
Resiliência	LLMs	LLMs	SBERT	LLMs	LLMs	LLMs
Empatia	LLMs	LLMs	SBERT	LLMs	LLMs	LLMs
Inovação e Criatividade	BTFT	LLMs	SBERT	LLMs	SBERT	SBERT
Organização	LLMs	LLMs	SBERT	LLMs	LLMs	BTFT
Relacionamento interpessoal	LLMs	LLMs	SBERT	LLMs	LLMs	LLMs

Figura 7 – Comparação LLMs x SBERT x BerTimbau FT

LLMs - Apresentaram melhor performance geral na maioria dos aspectos. Quase sempre liderando os indicadores de Acurácia, precisão e F1-score), confirmando serem mais robustas nos contextos gerais e que conseguem "compreender"os feedbacks com maior diversidade linguística.

SBERT - Vence em revocação/sensibilidade, mostrando-se interessante para contextos onde a prioridade é não perder casos positivos. Também mostrou resultados interessantes no indicador Youden J, o mostra equilíbrio entre sensibilidade e especificidade.

BerTimbau FT (BTFT) - Obteve alguns destaques pontuais (Comunicação Eficaz, Adaptabilidade, Organização), mas teve melhor resultado na especificidade, ao evitar mais falsos positivos para alguns aspectos. Isso demonstra que o BerTimbau com

fine-tuning classifica bem em contextos específicos, mas perde no desempenho geral para os LLMs.

5.1.1 Recorte do resultado geral SBERT x BerTimbau FT

Melhores resultados por Aspecto X Indicador, SBERT x Bertimbau FT	Acurácia	Precisão	Revocação/ Sensibilidade	Especificidade	F1-Score	Youden's J
Colaboração	SBERT	BTFT	SBERT	BTFT	BTFT	BTFT
Comunicação Eficaz	SBERT	BTFT	SBERT	BTFT	BTFT	BTFT
Autoconfiança	SBERT	SBERT	SBERT	BTFT	BTFT	BTFT
Autonomia	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Pontualidade	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Conhecimentos Técnicos	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Qualidade das entregas	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Adaptabilidade	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Gestão do tempo	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Proatividade	SBERT	SBERT	SBERT	BTFT	BTFT	BTFT
Resiliência	SBERT	BTFT	SBERT	BTFT	BTFT	BTFT
Empatia	SBERT	BTFT	SBERT	BTFT	BTFT	BTFT
Inovação e Criatividade	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Organização	BTFT	BTFT	BTFT	BTFT	BTFT	BTFT
Relacionamento interpessoal	SBERT	SBERT	SBERT	BTFT	BTFT	BTFT

Figura 8 – Comparação SBERT x BerTimbau FT

A comparação apenas entre os resultados nos indicadores do SBERT com o BerTimbau mostram que este domina a maior parte dos indicadores (precisão, especificidade, F1-Score e Youden J, em quase todos os aspectos. O destaque do SBERT aparece de forma mais pontual em alguns aspectos como Colaboração, Autoconfiança, Proatividade, Resiliência e Relacionamento Interpessoal), mas mesmo assim menos consistente de forma geral, uma vez que mesmo nestes aspectos perde em relação a especificidade, F1-Score e Youden J.

Este resultado sugere que, após o *fine-tuning* o BerTimbau realmente se ajustou melhor a este domínio de feedbacks, superando o SBERT em equilíbrio entre acertos e erros.

5.1.1.1 Destaques dos modelos

SBERT no desempenho de Revocação/sensibilidade em alguns aspectos, indicando tendência de capturar mais positivos, mas também acaba classificando junto falsos positivos. Este é um ponto interessante para cenários onde não se deseja correr o risco de identificar pontos relevantes, como por exemplo, identificar problemas ou desvios mais graves de comportamento).

BerTimbau FT já demonstra uma clara vantagem em precisão e especificidade, o que significa que quando classifica algo, tende a estar correto. Além disso, o melhor Youden J em todos os aspectos mostra maior equilíbrio entre revocação e especificidade. Ou seja, o *fine-tuning* de mais assertividade ao modelo para este domínio específico.

5.2 Limitações deste trabalho

Para a execução do experimento, os 25 feedbacks foram escritos pelo próprio autor (não foram gerados por LLM, como foram a maior parte dos feedbacks utilizados para o *fine-tuning* do BerTimbau).

Os 25 feedbacks usados para avaliar os modelos foi uma amostra pequena, e um pouco desbalanceada (como algumas sentenças podem ser associadas a mais de um aspecto, isso gerou uma quantidade de labels desigual em alguns aspectos. Isso pode ter gerado um viés na análise de aspectos em específico, contudo não parece colocar em risco a conclusão geral sobre as características e os impactos do *fine-tuning*, assim como a capacidade das LLMs, que conseguem compensar a limitação de dados por terem sido treinadas com grande exposição a textos diversos.

5.3 Estratégias e sugestões identificadas para melhorar os resultados

Para o escopo deste TCC, a comparação do resultado entre modelos com e sem *fine-tuning* já foi suficiente. Contudo, durante os estudos, foram percebidas algumas possíveis oportunidades para melhorar ainda mais os resultados obtidos, mas que não cabiam serem explorados aqui. Portanto seguem abaixo para registro e sugestões para outros estudos:

5.3.1 SBERT + BerTimbau com FT

Considerando os resultados obtidos neste contexto de feedbacks pelos modelos avaliados, uma abordagem que pode ser testada seria combinar SBERT com o refinamento do BerTimbau. Assim seria possível aproveitar a capacidade de identificação de contextos secundários do SBERT, com a posterior aplicação do BerTimbau para identificar mais assertivamente os resultados mais significativos.

5.3.2 Análise por frase para o BerTimbau com FT

O *fine-tuning* do BerTimbau deixou o modelo mais rígido e assertivo, contudo, quando o feedback é mais longo, com várias frases e contextos (como nos feedbacks 9 e 10), apresentou resultados iguais para todos os aspectos. Uma possível estratégia seria segmentar feedbacks mais longos por sentenças (como foi realizado no experimento para os feedbacks 9 e 10) e depois consolidar os resultados destas sentenças isoladas para avaliar o feedback completo.

5.3.3 Aumentar e diversificar o dataset do *fine-tuning*

O fato do modelo BerTimbau com FT ser mais assertivo para o que é mais familiar ao que foi treinado sugere que quanto maior o dataset melhor será a capacidade de lidar com diferentes formas de escrita de frases e contextos. A diversificação também pode trazer benefícios. Ou seja, usar diferentes LLMs para gerar mais frases de feedback, buscando diferentes padrões de escrita podem ajudar na identificação de aspectos em diferente textos.

5.3.4 Utilizar LLM *on-premises*

Os resultados mostram como as LLMs são capazes de generalizar e responder questões em textos complexos e subjetivos melhor do que os modelos menores. Mesmo havendo oportunidades de melhoria no modelo com *fine-tuning* testado neste trabalho, isso exigirá esforço significativo. Diante disso é possível que a decisão seja por utilizar uma LLM. Neste caso, para evitar o problema de expor dados corporativos sensíveis num modelo de rede aberto, há a opção de alguns LLMs que permitem instalação local, dentro da infraestrutura da empresa.

5.3.4.1 Instalação *on-premises* de uma LLM

A solução passa por selecionar um modelo de LLM que permita criar uma instância local (*on-premises*), como por exemplo alguma versões do LLaMa, Mistral, Falcon, Hugging Face e DeepSeek. Neste cenário, com a LLM contida na infraestrutura da empresa, mitigasse o risco do processamento dos dados sensíveis da empresa.

5.3.4.2 Processamento dos dados corporativos pela LLM

A opção mais simples e de menor custo é utilizar RAG para incluir os dados corporativos e assim direcionar as respostas da LLM para usar aquele conjunto de dados para dar as respostas. É o ideal quando o objetivo é utilizar esta LLM para mais de um propósito ou múltiplas tarefas.

Uma outra opção de implementação é realizar um retreinamento destas LLMs (*fine-tuning*), de forma geral, um processo similar ao realizado neste trabalho para o

modelo BerTimbau. Contudo é preciso considerar o custo computacional, uma vez que estes modelos possuem 7, 13 e até 70 bilhões de parâmetros, enquanto que o BerTimbau possui em torno de 110 milhões de parâmetros.

REFERÊNCIAS

CAPPELLI, P.; CONYON, M. J. What do performance appraisals do? **ILR Review**, v. 71, n. 1, p. 88–116, 2017. ISSN 0019-7939. Doi: 10.1177/0019793917698649. Disponível em: https://doi.org/10.1177/0019793917698649.

COSTA, R. W. M.; PARDO, T. A. S. Métodos baseados em léxico para extração de aspectos de opiniões em português. **null**, 2020.

FREITAS, L. A. D.; VIEIRA, R. Exploring resources for sentiment analysis in portuguese language. *In*: **2015 Brazilian Conference on Intelligent Systems (BRACIS)**. [*S.l.: s.n.*], 2015. p. 152–156.

GAO, Y. *et al.* **Retrieval-Augmented Generation for Large Language Models: A Survey**. 2024. Disponível em: https://arxiv.org/abs/2312.10997.

GUPTA, S.; RANJAN, R.; SINGH, S. N. **A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions**. 2024. Disponível em: https://arxiv.org/abs/2410.12837.

Harvard Business Review. **A Arte de Dar Feedback: Um Guia Acima da Média**. Rio de Janeiro: Sextante, 2019. ISBN 978-85-431-0731-8.

JORGE, G. A. Z. Trabalho de Conclusão de Curso (MBA Inteligência Aritificial e BigData), **Prevendo a utilidade de comentários em português brasileiro de jogos no site Steam**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022. Acesso em: 01 mar. 2025. Disponível em: https://bdta.abcd.usp.br/directbitstream/ac40baae-6351-49d9-8bdf-edfdb04c01c2/Germano%20Ant%C3%B4nio%20Zani%20Jorge_TCC_MBA_GERMANO_JORGE_COMPLETO_VERSAOFINAL_173931.pdf.

KANSAON, D. P. K.; BRANDãO, M. A.; PINTO, S. A. d. P. P. Análise de sentimentos em tweets em português brasileiro. **Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, 2018.

KAUER, A. U. **Análise de sentimentos baseada em aspectos e atribuições de polaridade**. 2016. Tese (Dissertação de mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: http://hdl.handle.net/10183/140910.

LIU, B. **Sentiment Analysis and Opinion Mining**. 1. ed. [*S.l.: s.n.*]: Springer Cham, 2012. XIV, 167 p. (Synthesis Lectures on Human Language Technologies). ISSN 1947-4040. ISBN 978-3-031-01017-0.

LMPO. **A Brief History of LLMs: From Transformers (2017) to DeepSeek R1 (2025)**. 2025. Acesso em: 10 mar. 2025. Disponível em: https://medium.com/@lmpo/a-brief-history-of-lmms-from-transformers-2017-to-deepseek-r1-2025-dae75dd3f59a.

MACHADO, M. T. **Methods for identifying aspects in opinion texts in Portuguese: the case of implicit aspects and their typological analysis**. 2023. Tese (Thesis) — Universidade de São Paulo, São Carlos, 2023. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-16012024-151720/.

MANNING, C. D. Last words: Computational linguistics and deep learning. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 41, n. 4, p. 701–707, dez. 2015. Disponível em: https://aclanthology.org/J15-4006/.

MICHAELIS. **Dicionário brasileiro da Língua portuguesa**. 2025. Acesso em: 27 jan. 2025. Disponível em: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/feedback/.

OPENAI. **ChatGPT: Assistente de inteligência artificial**. 2024. Disponível em `https://chat.openai.com/`. Modelo GPT-4, acesso em: 9 mar. 2025.

PANCHENDRARAJAN, R. *et al.* Implicit aspect detection in restaurant reviews using cooccurence of words. *In*: . Association for Computational Linguistics, 2016. (Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis), p. 128–136. Disponível em: https://aclanthology.org/W16-0421/.

PARDO, T. A. S. **Métodos para análise discursiva automática**. 2005. Tese (Doutorado em Ciências de Computação e Matemática Computacional - Instituto de Ciências Matemáticas e de Computação) — Universidade de São Paulo, São Carlos, 2005. Acesso em: 01 de Março de 2025.

RAMAKRISHNAN, R. A practical guide to gaining value from llms. **MITSloan Management Review - Winter 2025 Issue**, v. 66, p. 49–55, 2024. Disponível em: https://sloanreview.mit.edu/article/a-practical-guide-to-gaining-value-from-llms/.

SANTOS, K. H. R. d. Trabalho de Conclusão de Curso (MBA Inteligência Aritificial e BigData), **Classificação de textos de petições jurídicas com base em técnicas de processamento de línguas naturais**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023. Acesso em: 01 mar. 2025. Disponível em: https://bdta.abcd.usp.br/directbitstream/10cbe4f4-36c7-4c97-bd82-1477ec3348b4/Kelsen%20Henrique%20Rolim%20dos%20Santos.pdf.

SILVA, L. H. N. *et al.* Identificação de aspectos explícitos e implícitos em críticas gastronômicas em português: avaliando o potencial dos llms. **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**, p. 176–181Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/31129.

TABOADA, M. Sentiment analysis: An overview from linguistics. **Annual Review of Linguistics**, v. 2, 02 2016.

VARGAS, F. A.; PARDO, T. A. S. Conference Paper, **Aspect Clustering Methods for Sentiment Analysis**. Springer-Verlag, 2018. 365–374 p. Disponível em: https://doi.org/10.1007/978-3-319-99722-3_37.

VASWANI, A. *et al.* Attention is all you need. *In*: GUYON, I. *et al.* (ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

WU, S. *et al.* **Retrieval-Augmented Generation for Natural Language Processing: A Survey**. 2025. Disponível em: https://arxiv.org/abs/2407.13193.

ANEXOS

ANEXO A – ARQUIVOS E CÓDIGOS-FONTE GERADOS

Os arquivos e códigos-fonte construídos durante a elaboração deste trabalho estão disponíveis no endereço:

https://github.com/EduardoFerrarini/TCC_IABigData_Ferrarini

Conteúdo dos arquivos

AnaliseResultados.zip

- feedbacks_aspectos.xlsx
  Planilha de análise de resultados das execuções dos modelos

CodigosFonte.zip
- analise_feedbacks_comparada.py
  Arquivo código Python que executa os modelos SBERT, BerTimbau base e BerTimbau FT, utilizando Datasets feedback.json e aspectos.json. Grava resultado de saida como comparacao.json.

- analiseROC.py
  Script python para gera análise ROC e encontrar melhor limiar para ser usado para cada modelo

- SBERT_valida.py
  Script para validação dos dados gerados no dataset de treinamento para o fine-tuning

- TCC_FT_BT_ClassificaAspectos.ipynb
  Notebook de treinamento do modelo BerTimbau, utilizando o dataset AspectClassification_fine_tuning.csv. Resultado é o modelo refinado. Foi executado no google colab para fazer uso de GPU.

- ValidaDadosScatterPlot.py
  Script python para validar os dados e treinamento, verificando distribuição por similaridade

Dataset.zip

- AspectClassification_fine_tuning.csv
  5700 feedbacks para os 15 aspectos e label binária.

- aspectos.json
  definição dos 15 aspectos utilizada para execução do trabalho

- comparacaoLabel.json
  Arquivo de output da execução do analise_feedbacks_comparada.py,
  alterado para incluir label de classificação e é utilizado
  pelo script analiseROC.py

- feedbacks.json
  Arquivo contendo os 25 feedbacks utilizados para testar SBERT,
  BerTimbau base, BerTimbau FT, ChatGPT 4, ChatGPT 5, Claude
  e Gemini

ANEXO B – DATASET: *FINE-TUNING*

Para a realização de um fine-tuning nas implementações utilizando modelos de IA para PLN pré-treinados para língua Portuguesa (BerTimbau) foi necessário criar um conjunto de dados, basicamento composto por frases de feedback, direcionadas a um perfil de profissional ("Dev JR", "Dev PL"ou "Dev SR"), e também direcionado a um aspecto: 'Colaboração', 'Comunicação eficaz', 'Autoconfiança', 'Autonomia', 'Pontualidade', 'Conhecimentos Técnicos', 'Qualidade das entregas', 'Adaptabilidade', 'Gestão do tempo', 'Proatividade', 'Resiliência', 'Empatia', 'Inovação e Criatividade', 'Organização' ou 'Relacionamento interpessoal'

Este arquivo de feedback foi gerado com apoio da ferramenta GitHub Copilot utilizando modelo ChatGPT 4.1 em Julho de 2025 (OpenAI, 2024). O arquivo foi formado a partir da execução sucessiva de solicitações via prompt para o ChatGPT. Foram gerados 30 frases de feedback para cada combinação ("aspecto"x "cargo"x "sentimento"), resultando num arquivo com 2.700 linhas.

Exemplo de PROMPT aplicado para construção da base:

"Gerar mais dados para adicionar ao arquivo "BerTimbau_Sentiment_fine_tuning_filtrado.csv"Para o aspecto "Relacionamento interpessoal"(no arquivo "aspectos.json") criar novos e identicos 30 feedbacks 'Negativo"para cargo "Dev SR", considerando especificação de "Dev SR"em "expectativasPorPerfil.json". "

Após os feedbacks gerados, foi criado um script Python para tratamento dos dados, responsável por realizar as seguintes validações:

1. **Remoção de duplicatas:** O Script varre o dataset inteiro e identifica Feedbacks idênticos. Para tratar, as ações executadas foram:

   - feedbacks duplicados foram ajustados manualmente pelo autor, através de substituição de palavras por sinônimos, ou pequena alteração na forma de escrever, mantendo o sentido do conteúdo do feedback gerado anteriormente via LLM; ou
   - frase do feedback foi totalmente reescrita pelo autor. Nestes casos, geralmente a frase foi criada totalmente manualmente pelo autor; ou
   - remoção de feedbacks duplicados, substituindo os mesmos por novos feedbacks gerados pela LLM (ChatGPT 4.1)

2. **Verificação de equilíbrio do Dataset:** O Script verifica se todas as combinações de classificação de feedbacks estão com a mesma quantidade de amostras (no caso

180 feedbacks). Ou seja, garante que há 180 feedbacks para cada "aspecto"x "label". Quando o feedback tem label = 1 significa é que um exemplo de sentença que contém o feedback. Quando o feedback tem label = 0 significa que foi gerado para que não tivesse relação com o aspecto.

`arquivo destino: Aspect_Classification_fine_tunning.csv`