Utilizando o Coeficiente de Concordância de Kappa para Avaliar uma Análise de Sentimentos apoiada por IA

Ana Kessilly Chiachio Cerqueira[1], **Melques Santos Paiva**[1], **Danilo Guimarães Souza Azevedo**[1], **Djan Almeida Santos** [1], **Crescencio Lima** [1], **Luis Paulo da Silva Carvalho**[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA)
Campus Vitória da Conquista
45078-300 – Vitória da Conquista – BA – Brasil

kessilly_chiachio@hotmail.com

{melque1703, dan.azevedo82}@gmail.com

{djan.santos, crescencio, luiscarvalho}@ifba.edu.br

Abstract. *The increasing production of textual data on social media platforms makes Sentiment Analysis (SA) a crucial tool for extracting valuable insights. This paper addresses the transformation of SA by the rise of Large Language Models (LLMs), such as Google Gemini. Our central contribution resides in the reliability of applying the AI model in classifying polarity and emotion in YouTube comments. To this end, we employed the Cohen's Kappa Concordance Coefficient to measure the degree of agreement between the LLM and two human evaluators. The results demonstrated Moderate agreement between the AI and the humans, as well as between the human evaluators themselves. The entire process was consolidated into a functional web application, Pulso Emocional.*

Resumo. *O aumento da produção de dados textuais em plataformas de mídias sociais torna a Análise de Sentimentos (AS) uma ferramenta essencial para a extração de insights valiosos. Este artigo aborda a transformação da AS com o surgimento dos Grandes Modelos de Linguagem (LLMs), como o Google Gemini. Nossa principal contribuição é a validação da confiabilidade do modelo de IA na classificação de polaridade e emoção em comentários do YouTube. Para isso, foi empregado o Coeficiente de Concordância de Cohen (Cohen's Kappa) para medir o grau de concordância entre o LLM e dois avaliadores humanos. Os resultados demonstraram uma concordância moderada tanto entre a IA e os humanos quanto entre os próprios avaliadores humanos. Todo o processo foi consolidado em uma aplicação web funcional, Pulso Emocional.*

1. Introdução

Diariamente, gera-se na internet um volume de dados textuais gigantesco, proveniente de comentários, *reviews* e interações em plataformas digitais e redes sociais. A capacidade de extrair *insights* (percepções) dessa fonte de informações tornou-se um diferencial competitivo e um recurso essencial para criadores de conteúdo e grandes entidades. Um dos possíveis caminhos para realizar tal tarefa é a Análise de Sentimentos.

A Análise de Sentimentos, também conhecida como *Opinion Mining* (mineração de opinião), termos definidos por Liu (2012), é o processo computacional de identificar, extrair e analisar informações subjetivas de um texto para determinar o sentimento ou a atitude geral expressa em relação a um produto, tópico ou evento específico. Isto tem potencial para transformar dados não estruturados (texto) em indicadores de satisfação ou insatisfação do usuário.

Com o avanço da Inteligência Artificial (IA), a Análise de Sentimentos sofreu uma transformação. Historicamente uma atividade laboriosa realizada por humanos ou por algoritmos tradicionais de Aprendizado de Máquina, ela tem evoluído significativamente com a ascensão dos *Large Language Models* (LLMs) ou, em português, Grandes Modelos de Linguagem. Estes modelos, treinados em vastos volumes de dados textuais, demonstram uma capacidade aprimorada de processar e contextualizar a linguagem humana, permitindo que a Análise de Sentimentos seja realizada de forma rápida, escalável e com alta precisão (Zhang et al., 2018).

Apesar dos avanços, a confiabilidade de uma análise automatizada de sentimentos por IA ainda levanta questionamentos. Dada a possibilidade de falhas ou ”alucinações”do modelo, a validação externa é crucial (Qi et al., 2025). A etapa de avaliação deste trabalho foi realizada para confrontar os resultados da IA utilizada, a Google Gemini (especificamente, o modelo gemini-2.5-flash[1]), com uma classificação fornecida por avaliadores humanos. Para tal, foi empregado o Coeficiente de Concordância de Kappa, que permite medir o *inter-rater agreement* (grau de acordo interavaliadores), removendo a concordância aleatória (Berry and Jr, 1988).

Por fim, todo o processo de análise e avaliação foi concretizado em uma solução web, o Pulso Emocional, que visa fornecer aos usuários do YouTube uma ferramenta prática para a análise automatizada de sentimentos de comentários de vídeos através da Inteligência Artificial.

2. Método de Avaliação

Para enfrentar o problema de incerteza e garantir a robustez científica dos resultados da avaliação da IA usada neste trabalho na classificação de comentários do YouTube, foi empregado o Coeficiente de Concordância de Kappa (Cohen, 1960; Brennan and Prediger, 1981; Wan et al., 2015). Este coeficiente é uma fórmula matemática-estatística que mede a concordância entre dois *raters* (avaliadores) que classificam um conjunto de itens em categorias mutualmente exclusivas (Berry and Jr, 1988; Stefanovitch et al., 2022), sendo particularmente útil para dados categóricos, tais como, por exemplo, polaridades de sentimentos: ”POSITIVO”, ”NEGATIVO”, ”NEUTRO”.

Consideramos que o coeficiente de concordância de Kappa se enquadra na validação cruzada de Análise de Sentimentos (IAs vs. Humanos e Humanos vs. Humanos) por (Brennan and Prediger, 1981): (i) ser mais preciso do que um cálculo de simples porcentagem, já que ele ajusta o cálculo para remover casos de concordância que ocorreriam meramente por acaso; e (ii) resolver a necessidade de determinar o nível de confiabilidade e acordo entre os julgamentos, servindo como uma medida robusta de

[1] `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash?hl=pt-br`

validação para a qualidade dos rótulos gerados pela IA em comparação com os rótulos humanos.

No contexto deste trabalho, o Kappa foi aplicado para calcular três relações de concordância: (i) IA vs. Avaliador 1, (ii) IA vs. Avaliador 2 e (iii) Avaliador 1 vs. Avaliador 2 (Concordância Interavaliadores Humana).

O Coeficiente Kappa de Cohen é um valor calculado em escala numérica de 0 até 1, que pode ser interpretado através da Força do Concordância (Landis and Koch, 1977; Munoz and Bangdiwala, 1997), que estabelece os níveis enumerados na Tabela 1.

Tabela 1. Forças de Concordância do Coeficiente de Kappa

Kappa	Força de Concordância
<0.00	Poor (pobre)
0.00-0.20	Slight (leve)
0.21-0.40	Fair (razoável)
0.41-0.60	Moderate (moderado)
0.61-0.80	Substantial (substancial)
0.81-1.00	Almost Perfect (quase perfeita)

Para aproveitar a máxima capacidade de um *LLM*, torna-se imperativa a *Engineering Prompt* (Engenharia de *Prompt*). Este termo é definido como o processo de criação, refinamento e otimização de comandos (*prompts*) fornecidos a modelos de IA generativa para obter respostas mais precisas e relevantes (Ekin, 2023). No contexto da Análise de Sentimentos, esta técnica permite instruir a IA a atuar como um especialista, garantindo que a classificação de polaridade e emoção seja retornada em um formato estrito e processável computacionalmente (Lu and Liang, 2025).

A Figura 1 contém o *prompt* que contextualizou o *LLM*, Google Gemini, para que fosse realizada a classificação de sentimento de um texto.

Para quantificar a confiabilidade do modelo de IA na Análise de Sentimentos, definimos um processo de validação estruturado em etapas, culminando na aplicação do Coeficiente de Concordância de Kappa. Este método, ilustrado pela Figura 2, buscou replicar o rigor de estudos em Engenharia de Software que utilizam especialistas para validar a precisão de métodos automáticos Runeson and Höst (2009).

Para criar as Planilhas de Avaliação, foram minerados os 60 comentários mais relevantes do vídeo ”Fantástico: Futuro das inteligências artificiais e os perigos para a humanidade”[2], do canal do YouTube, G1[3]. Ao minerar comentários mais relevantes e não a totalidade (de comentários) do vídeo, não apenas não extrapolamos limites de acesso da *API* do YouTube[4], também contamos com uma amostra de sentimentos que foi validada e aceita como relevante por outros usuários que ”curtiram”aquelas opiniões.

O processo metodológico iniciou-se com a criação e execução de um *script* em

[2] `https://www.youtube.com/watch?v=D2KIu_yDeJk`

[3] `https://www.youtube.com/@g1globo`

[4] Os limites da *API* do YouTube são baseados em um sistema de cotas diárias, 10000 unidades por dia. Cada operação de API consome um número específico de unidades. Por exemplo, uma busca consome 100 unidades.

```
prompt_sistema_conteudo = f"""
Você é um assistente especializado em avaliar
sentimentalmente trechos de texto.
Você deve categorizar os textos em três
polaridades: NEGATIVO, POSITIVO e NEUTRO.
A sua classificação deve ser uma estrutura JSON
contendo: a polaridade associada ao atributo
'polaridade'.
A sua classificação deve conter SOMENTE o
conteúdo do JSON e NADA mais.
Ou seja, a sua classificação não pode conter
caracteres ou informações que exijam limpeza ou
modificação do JSON.
"""
```

Figura 1. Um *Prompt* para Análise de Sentimentos.

Figura 2. Fluxo de Validação Estruturado do Coeficiente de Kappa

Python para coletar dados por meio da *API* do YouTube, resultando na mineração e geração de uma planilha mestra de comentários.

Em seguida, um segundo *script* automatizou a preparação de duas tabelas idênticas contendo os comentários e o espaço reservado para as classificações de polaridade dos dois avaliadores humanos independentes.

Um terceiro *script* foi implementado para realizar a Análise de Sentimentos automatizada. Este *script* acessou a API do Gemini e, utilizando um *prompt* especificamente calibrado (Engenharia de *Prompt*), processou a mesma amostra de comentários para gerar uma planilha com a análise de polaridade da Inteligência Artificial.

Por fim, com os resultados consolidados (avaliações humanas e da IA), um quarto *script* foi executado para aplicar o Coeficiente de Kappa de Cohen. Esta etapa final foi crucial para medir a concordância inter-avaliadores e a concordância entre a IA e cada avaliador.

2.1. Participantes da Avaliação (IAs e Humanos)

Entre os autores deste artigo, enquanto a autora principal se responsabilizou por minerar as mensagens, preparar as planilhas de avaliação e consolidar os resultados da avaliação, outros dois foram selecionados como avaliadores. Eles são profissionais com experiência em desenvolvimento de software e áreas relacionadas: (i) O **Avaliador 1** é formado em Sistemas de Informação (2016), trabalha na área como Especialista/Analista de Sistemas desde 2013, enquanto (ii) o **Avaliador 2** possui formação em Engenharia da Computação (2021) e atua como desenvolvedor desde 2018.

Vale salientar que, por decisão prévia, visando evitar viés nas suas respostas e diagnósticos, os dois co-autores selecionados como avaliadores não tiveram contato com a pesquisa iniciada e preparada pela autora principal do estudo. Eles foram engajados posteriormente em dois momentos: (i) para realizar o preenchimento das Planilhas de Avaliação e (ii) posteriormente para coleta de suas considerações finais e revisão deste artigo.

2.2. Resultados de Concordância

O valor do Kappa calculado para a comparação entre a IA, Google Gemini, e os Avaliadores Humanos resultou nos valores exibidos na Figura 3, que mostra a captura da saída de texto do *script* Python utilizado para automatizar o cálculo.

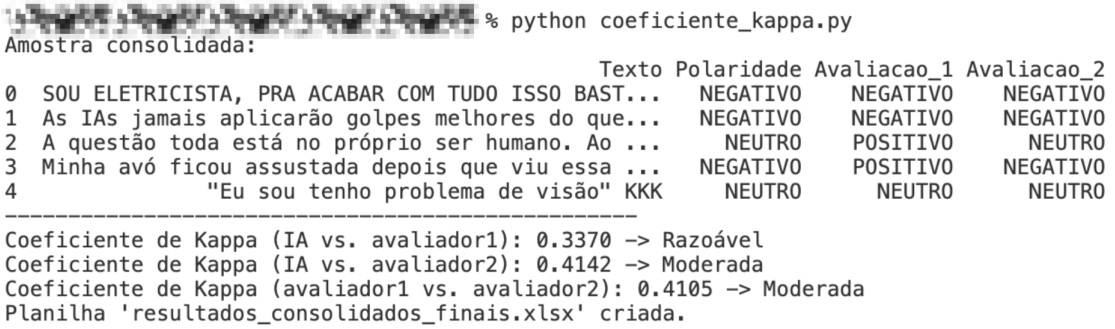

Figura 3. Captura da saída de resultado no terminal.

Os resultados demonstraram que a IA obteve uma concordância **Razoável** com um dos avaliadores e **Moderada** com o outro avaliador. Entre os próprios avaliadores a força de concordância foi **Moderada**.

Embora não tenha ocorrido para nenhum dos casos um nível superior/perfeito de concordância (vide Tabela 1), consideramos os níveis alcançados satisfatórios. Esta conclusão se baseia no fato de que, mesmo entre os avaliadores humanos, a concordância atingida foi **Moderada**, reforçando que a tarefa de classificação de sentimentos em linguagem natural é inerentemente subjetiva. Portanto, o *LLM* demonstrou um desempenho adequado ao se alinhar ao padrão de subjetividade humana.

3. Aplicando a Análise de Sentimentos apoiada por Inteligência Artificial

Com a conclusão da fase de validação e a obtenção de um nível de concordância satisfatório no Coeficiente de Kappa, demonstramos a confiabilidade do modelo de Análise de Sentimentos baseado em IA. Isto forneceu a base necessária para a etapa de aplicação prática e desenvolvimento da ferramenta, Pulso Emocional.

O Pulso Emocional transcende a simples classificação de polaridade de sentimentos, englobando outras funcionalidades. A Tabela 2 traz as funcionalidades atreladas a outras áreas de atuação da Análise de Sentimentos conforme elencadas nos trabalhos de Sharma et al. (2025) e Islam et al. (2024) e que foram incluídas no Pulso Emocional.

Tabela 2. Principais funcionalidades de análise de sentimentos

Tarefa de Análise de Sentimentos	Resumo sobre a tarefa
Análise de Sentimentos em Nível de Documento	Determina o sentimento geral (positivo, negativo ou neutro) de um documento completo, geralmente associado a uma única entidade.
Detecção de Emoções	Identifica emoções específicas além da polaridade básica, como alegria, raiva ou frustração.
Resumo de Opinião	Gera resumos representativos a partir de múltiplas opiniões sobre um mesmo tema.

O Pulso Emocional se tornou uma aplicação web que segue uma arquitetura dividida entre *front-end* (interface de usuário) e *back-end* (serviços de processamento e manipulação de dados) desenvolvida com o objetivo de transformar dados não estruturados de plataformas como o YouTube em informação gerencial acionável. A Figura 4 ilustra a arquitetura do Pulso Emocional.

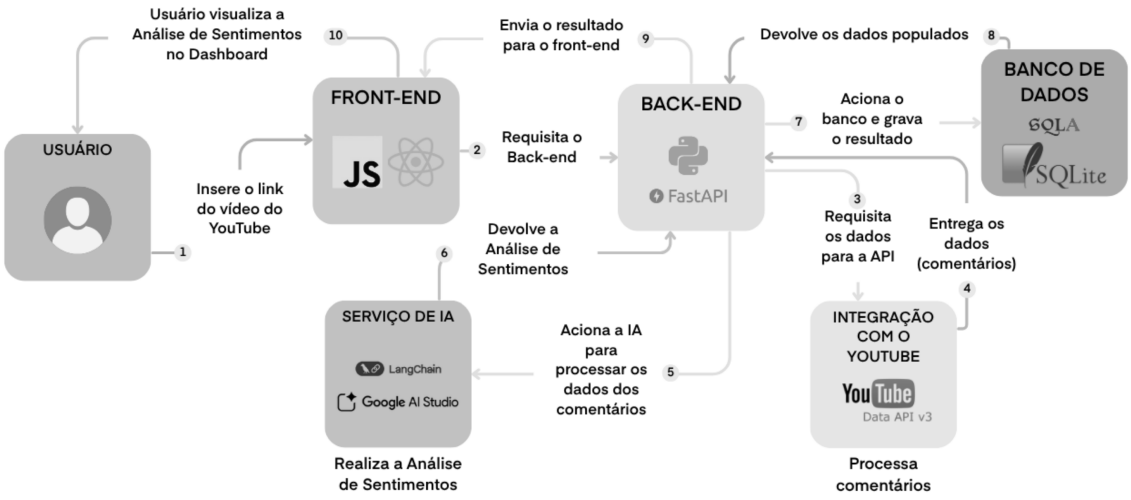

Figura 4. Arquitetura do Pulso Emocional.

O fluxo se inicia com o *front-end* (React/JavaScript), que envia a requisição para o *back-end*. Este, por sua vez, atua como agente integrador, buscando dados brutos através da YouTube Data API[5] e enviando os textos dos comentários para o Serviço de IA, Google AI Studio[6], gerenciado pela biblioteca LangChain[7], para a classificação de polaridade,

[5] https://developers.google.com/youtube/v3?hl=pt-br
[6] https://aistudio.google.com/
[7] https://www.langchain.com/

resumo e emoção. Os resultados processados são persistidos no Banco de Dados SQLite[8], garantindo a integridade e rastreabilidade da análise, antes de serem recuperados pelo *back-end* e devolvidos ao *front-end* para visualização pelo usuário.

Disponibilizamos um vídeo demonstrativo do Pulso Emocional realizando a Análise de Sentimentos de um vídeo através do link:

`https://youtu.be/OAgBEQaN7t0`

3.1. Opinião dos Participantes sobre o Pulso Emocional

Considerando que os dois participantes do nosso estudo possuem experiência em desenvolvimento de software, foi solicitado que avaliassem o Pulso Emocional e emitissem uma opinião sobre o resultado final e sugerissem a implementação de outra funcionalidades, considerando aquelas elencadas na Tabela 2.

O Avaliador 1 disse: ”Sobre o vídeo, achei muito bom e direto ao ponto. Deu pra entender o que a ferramenta faz. Também gostei da interface gráfica, achei limpa e dentro dos padrões atuais dos sites mais conhecidos que geralmente a gente acessa. Sobre as funcionalidades, seria interessante implementar a Análise de Sentimentos Multilíngue, a Detecção de Spam de Opinião e Análise de Sentimentos em nível de frase”.

O Avaliador 2 disse: ”A interface ficou intuitiva, fácil de entender, de usar e bem bonita. Sobre as funcionalidades não implementadas, as que me chamaram mais a atenção foram: Detecção de spam de Opinião (Acho que esse pode ir um pouco além. Exemplo: em um vídeo famoso sobre ”adultizacao”[9], é comentado que muitos criminosos usam de comentários disfarçados em redes sociais para divulgação de conteúdo infantil), Análise de Sentimentos em nível de frase e Extração e Padronização em Tempo”.

Visando a transparência e a replicabilidade dos achados, todos os artefatos utilizados na avaliação, os *scripts* de cálculo do coeficiente Kappa, as planilhas com as classificações preenchidas e o código-fonte do software resultante foram disponibilizados publicamente em um repositório no GitHub através dos links:

`https://github.com/kessillychiachio/PulsoEmocionalAvaliacao/`

`https://github.com/kessillychiachio/PulsoEmocional/`

4. Trabalhos Correlatos

Para a fundamentação teórica e seleção dos trabalhos correlatos, realizou-se uma busca sistemática nas bases Google Scholar[10] e IEEE Xplore[11]. Foram aplicadas as strings de busca ’Análise de Sentimentos’, ’Sentiment Analysis YouTube LLM’ e ’Cohen’s Kappa Sentiment Analysis’, com um recorte temporal entre 2018 e 2025 para garantir o alinhamento com o estado da arte. A seleção dos artigos discutidos baseou-se nos critérios de:

[8] `https://sqlite.org/`

[9] Adultização é a exposição de crianças e adolescentes a comportamentos, responsabilidades e experiências adultas antes da maturidade necessária. Esse fenômeno pode se manifestar de diversas formas, como a hipersexualização precoce e a imposição de responsabilidades excessivas.

[10] `https://scholar.google.com.br`

[11] `https://ieeexplore.ieee.org/Xplore/home.jsp`

(i) relevância em relação ao uso de LLMs para processamento de linguagem natural e (ii) aplicação do Coeficiente Kappa como métrica de validação. Dado que a integração entre LLMs e Análise de Sentimentos em comentários de mídias sociais é um tema emergente, priorizamos estudos que exploram a subjetividade e a confiabilidade desses modelos.

Islam et al. (2024) revisaram arquiteturas de Aprendizado Profundo, apontando que modelos como LSTM e GRU superam o Machine Learning tradicional, mas ainda falham em nuances como sarcasmo. Enquanto Islam propõe arquiteturas híbridas complexas para mitigar isso, nosso trabalho investiga se o uso de Large Language Models (LLMs) pré-treinados, especificamente o Google Gemini, consegue lidar com essas sutilezas linguísticas sem a necessidade de treinar modelos do zero, democratizando o acesso à análise semântica avançada.

Buscemi and Proverbio (2024) compararam o desempenho de LLMs (ChatGPT 3.5, ChatGPT 4, Gemini PRO e LLaMA2 7b) em cenários ambíguos, notando inconsistências e vieses no Gemini Pro. Diferentemente deste estudo, que comparou IAs entre si em cenários controlados, nosso trabalho foca na validação do Gemini contra a interpretação humana (o consenso entre avaliadores). Buscamos não apenas identificar falhas, mas quantificar estatisticamente a confiabilidade da IA em dados reais e não estruturados.

Chamid et al. (2024) demonstraram a eficácia do Coeficiente Kappa de Cohen para medir a consistência entre avaliadores em reviews de marketplace, atingindo concordância quase perfeita (0.893). Nosso artigo adota a mesma metodologia estatística (Kappa) para garantir rigor científico. Contudo, ao aplicá-lo em comentários de redes sociais (YouTube) em vez de avaliações de produtos estruturadas, nossos resultados de concordância moderada evidenciam a maior complexidade e subjetividade da linguagem informal.

Por fim, Chalkias et al. (2023) exploraram sentimentos em comentários educacionais no YouTube utilizando métodos baseados em léxico (VADER/TextBlob), apontando limitações na detecção de ironia. Evoluímos essa abordagem ao substituir dicionários léxicos estáticos por Inteligência Artificial Generativa.

A contribuição do presente trabalho reside na validação empírica e estatística do uso de LLMs, especificamente o Google Gemini, como uma alternativa robusta aos métodos tradicionais de análise de sentimentos em ambientes de alta subjetividade, como o YouTube. Diferente das abordagens baseadas em léxicos estáticos ou arquiteturas de aprendizado profundo que exigem treinamento intensivo, esta pesquisa foca na confiabilidade da IA generativa frente ao julgamento humano. Ao aplicar o Coeficiente Kappa de Cohen para mensurar essa concordância em dados não estruturados e informais, o estudo não apenas preenche uma lacuna sobre a eficácia de modelos pré-treinados em nuances linguísticas, mas também estabelece um parâmetro de rigor metodológico para a substituição ou auxílio da rotulagem manual por processos automatizados de inteligência artificial.

5. Ameaças à Validade

Nesta seção, discutimos as ameaças à validade do nosso trabalho.

Em relação à **Validade de construto**, baseamos nosso trabalho na comparação

de opiniões de IAs e Humanos. Escolhemos os avaliadores humanos entre os coautores deste trabalho, porque (i) estavam disponíveis para participarem do estudo e (ii) queríamos contar com a opinião deles como especialistas em desenvolvimento de software, para avaliar tecnicamente o Pulso Emocional. Todavia, profissionais de outras especialidades (*e.g*, psicologia) talvez pudessem nos fornecer um universo de opiniões mais diverso tanto durante a avaliação quanto sobre o software concretizado.

Considerando a **validade interna**, confeccionamos um *prompt* para realizar a classificação sentimental de mensagens extraídas do YouTube. Posteriormente, utilizamos a classificação para conduzir nossa avaliação. Todavia, por questões de prazos para a conclusão deste trabalho, somente realizamos um único ciclo de interação com os avaliadores sem refinar o *prompt*. Possivelmente, a realização de ciclos extras de avaliação seguidos de refinamento do *prompt* poderia ser benéfico no sentido de atingir uma força melhor de concordância.

Sobre a **validade externa**, nossos achados estão restritos ao conjunto de mensagens que extraímos de um canal do YouTube (Figura 2). Neste caso, não cobrimos mensagens de uma variedade maior de canais e, portanto, consequentemente, de repertório sentimental de um público alvo maior.

Relativo à **validade de conclusão**, contar com dois avaliadores nos proporcionou algumas vantagens: (i) dois avaliadores são suficientes para alcançar uma boa precisão de avaliação via Kappa, e (ii) isso ajudou a reduzir nossa carga de trabalho, pois nos proporcionou um fluxo mais controlado de percepções e sugestões. No entanto, é recomendado que se tenha mais avaliadores para refinar nossas conclusões. Nesse caso, aplicar Kappa para lidar com múltiplos avaliadores (Conger, 1980) (Berry and Jr, 1988) pode mitigar tal ameaça.

6. Conclusão

Este artigo apresentou a concepção e avaliação do uso da Análise de Sentimentos apoiada por um *Large Language Model* (LLM), no caso, o gemini-2.5-flash. O trabalho demonstrou a viabilidade de utilizar LLMs e a Engenharia de *Prompt* para automatizar tarefas de classificação de textos conforme sua polaridade de sentimentos. Um dos pontos focais da nossa contribuição foi a validação do modelo de IA frente à subjetividade da linguagem, utilizando o rigor estatístico do Coeficiente de Concordância de Kappa. A obtenção de um resultado de concordância **Moderada** entre a IA e os avaliadores humanos ratificou a confiabilidade do modelo para a tarefa de análise, legitimando seu uso no produto final, o software web, Pulso Emocional.

Reforçamos outra vantagem do uso de LLMs: a avaliação de comentários informais do YouTube, que contêm jargões, gírias e ambiguidades. Entendemos que, sem o uso de LLM, potencialmente, tais fenômenos linguísticos poderiam afetar negativamente a precisão da Análise de Sentimento. Todavia, a Google Gemini consegue processar bem este tipo de comentários. Além disso, a aplicação do Coeficiente Kappa, que revelou que mesmo entre os avaliadores humanos a concordância não foi perfeita, reforçou que a IA se comportou de forma satisfatória em um domínio inerentemente complexo.

Este trabalho abre algumas propostas para futuras investigações e aprimoramentos. Além daqueles apontados pelos participantes da avaliação (Seção 3.1), a autora principal escolhe estes (Tabela 2):

- **Análise de Sentimentos Baseada em Aspecto e Multimodalidade**: A atual funcionalidade de Análise de Sentimentos Geral pode ser expandida para implementar a Análise de Sentimentos Baseada em Aspecto, permitindo identificar o sentimento exato em relação a tópicos específicos citados nos comentários (exemplo: qualidade de áudio do vídeo e da edição). Além disso, a Análise Multimodal, que combina texto, imagem e áudio, pode ser explorada para fornecer uma compreensão mais rica do conteúdo;
- **Detecção Avançada de Ironia e Sarcasmo**: Embora o LLM tenha demonstrado bom desempenho, a detecção de ironia e sarcasmo permanece um dos desafios mais importantes na Análise de Sentimentos. Um trabalho com refinamento de prompt poderia melhorar ainda mais estas métricas;
- **Avaliação *Inter-LLM* com Kappa**: Embora o presente estudo tenha validado a IA contra dois avaliadores humanos, trabalhos futuros poderiam aplicar o Coeficiente Kappa para comparar a concordância entre diferentes *LLMs (como Gemini vs. ChatGPT)* e também com mais avaliadores.
- **Métricas de Desempenho de Classificação**: Além do Coeficiente Kappa, que foca na concordância interavaliadores, trabalhos futuros podem expandir a avaliação estatística através do cálculo de métricas de desempenho como *Acurácia, Precisão, Revocação (Recall) e F1-Score*. A utilização dessas métricas, baseadas em uma Matriz de Confusão, permitiria quantificar o erro do modelo em categorias específicas (ex: identificar se a IA tende a confundir 'Neutro' com 'Negativo') e comparar o desempenho do *LLM* com algoritmos de *Machine Learning* tradicionais.

Por fim, este trabalho contribui para a metodologia de pesquisa, provando que é possível utilizar métricas de confiabilidade interavaliadores para atestar o rigor dos resultados de um *LLM* em aplicações que automatizam atividades normalmente realizadas por agente humanos, tais como a Análise de Sentimentos. Também forneceu uma ferramenta funcional e validada para a comunidade de criadores de conteúdo, o Pulso Emocional.

6.1. Nosso uso de IA

Além do uso da Google Gemini como agente de automação de Análise de Sentimentos, durante a condução deste trabalho, também contamos com o apoio das seguintes IAs e/ou softwares baseados e IA:

- **GitHub Copilot**[12]: Como uma ferramenta de IA integrada ao Visual Studio Code (VS Code), auxiliou diretamente na produtividade da codificação;
- **Figma Make**[13]: Utilizamos o Figma para geração de protótipos funcionais. Após desenhar a interface, empregamos o Figma Make, uma IA do figma, para criar interfaces (com o código) a partir de descrições;
- **ChatGPT**[14]: Utilizamos como assistência na revisão gramatical e sumarização de conteúdos redigidos pela autora.

[12]https://github.com/features/copilot?locale=pt-BR
[13]https://www.figma.com/make/
[14]https://www.chatgpt.com/

Referências

Berry, K. J. and Jr, P. W. M. (1988). A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48(4):921–933.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Buscemi, A. and Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis.

Chalkias, I., Tzafilkou, K., Karapiperis, D., and Tjortjis, C. (2023). Learning analytics on youtube educational videos: Exploring sentiment analysis methods and topic clustering. *Electronics*, 12(18):3949.

Chamid, A. A., Widowati, and Kusumaningrum, R. (2024). Labeling consistency test of multi-label data for aspect and sentiment classification using the cohen kappa method. *Ingénierie des Systèmes d'Information*, 29(1):161–167.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322.

Ekin, S. (2023). Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices.

Islam, M., Kabir, M., Ghani, N. A., Zamli, K., Zulkifli, N., Rahman, M. M., and Moni, M. (2024). "challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach". *Artificial Intelligence Review*, 57.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Lu, K. and Liang, H. (2025). Ncl-nlp at semeval-2025 task 11: Using prompting engineering framework and low rank adaptation of large language models for multi-label emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Munoz, S. R. and Bangdiwala, S. I. (1997). Interpretation of kappa and b statistics measures of agreement. *Journal of Applied Statistics*, 24(1):105–112.

Qi, S., Gui, L., He, Y., and Yuan, Z. (2025). A survey of automatic hallucination evaluation on natural language generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. In *Empirical Software Engineering*, volume 14, pages 131–164.

Sharma, N. A., Ali, A., and Kabir, M. A. (2025). A review of sentiment analysis: tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, 19:351–388.

Stefanovitch, N. et al. (2022). Resources and experiments on sentiment classification using polarity labels. In *LREC (Language Resources and Evaluation Conference) 2022*.

Wan, T., Jun, H., Pan, W., Hua, H., et al. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai archives of psychiatry*, 27(1):62.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Minibiografia dos autores

Ana Kessilly Chiachio Cerqueira (IFBA): Especialista em Desenvolvimento Web pelo Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA, 2025) e graduada em Ciência da Computação pelo Centro Universitário Estácio Ribeirão Preto (2023). Possui trajetória profissional de três anos no desenvolvimento de software, com ênfase em arquitetura de sistemas web. Sua atuação técnica concentra-se na implementação de soluções escaláveis, com foco em pesquisa e desenvolvimento voltados à integração de Inteligência Artificial em Softwares como Serviço (SaaS).

Melques Santos Paiva (IFBA): Desenvolvedor de Software com mais de 7 anos de experiência em desenvolvimento web, com foco em back-end utilizando PHP e Go. É graduado em Engenharia de Computação pela Faculdade Independente do Nordeste e atualmente cursa pós-graduação em Desenvolvimento Web pelo IFBA. Atualmente, é Senior Software Developer na DB1 Global Software, onde lidera o desenvolvimento e manutenção de integrações com gateways de pagamento de alta escala. Sua atuação envolve sistemas que processam milhões de transações diárias, além de melhorias de performance e monitoramento.

Danilo Guimarães Souza Azevedo (IFBA): Graduado em Sistemas de Informação pelo Instituto Federal da Bahia (IFBA), Campus Vitória da Conquista-BA. Especialista em Desenvolvimento Web pelo IFBA. Atua profissionalmente como especialista em sistemas no mercado de varejo.

Crescêncio Lima (IFBA): Possui graduação em Ciência da Computação pela Universidade Estadual do Sudoeste da Bahia (2005), pós graduação em Engenharia de Software pela Faculdade Boa Viagem (2008), mestrado em Ciência da Computação pela Universidade Federal de Pernambuco (2011) e doutorado pelo programa de pós-graduação da Universidade Federal da Bahia (PGCOMP-UFBA). Atua como professor colaborador do Programa de Pós-Graduação em Engenharia de Sistemas e Produtos (PPGESP). Nos anos de 2019-2023 foi coordenador do Curso de Pós-graduação Lato Sensu em Desenvolvimento Web (PGDW) do IFBA campus Vitória da Conquista e atualmente é vice-coordenador do PGDW.

Djan Almeida Santos (IFBA): Professor do Instituto Federal da Bahia, campus Vitória da Conquista. Doutor em Ciência da Computação pela Universidade Federal da Bahia (2023), Mestre em Ciências Ambientais com Área de Concentração em Meio Ambiente e Desenvolvimento e linha de pesquisa em modelagem computacional pela Universidade Estadual do Sudoeste da Bahia (2013), especialista em Administração de Sistemas de Informação pela Universidade Federal de Lavras - MG, especialista em Desenvolvimento de Sistemas Web pela Faculdade de Tecnologia e Ciência (FTC) e Bacharel em Ciências da Computação pela Universidade Estadual de Santa Cruz (2003). Atuou na área da Ciência da Computação com ênfase em Engenharia de Software, compreensão de programas, sistemas configuráveis, modelagem matemática e simulação computacional.

Luis Paulo da Silva Carvalho (IFBA): Graduado em Informática pela Universidade Católica do Salvador (2002), Mestrado em Sistemas e Computação pela UNIFACS (2012) e Doutorado em Ciência da Computação pela UFBA (2020). Atualmente é professor do IFBA em Vitória da Conquista. Atuou por 16 anos na Indústria de Software como Programador, Analista, Arquiteto de Sistema e Líder Técnico. Tem experiência na área de

Ciência da Computação, com ênfase em Arquitetura de Sistemas de Computação, atuando principalmente nos temas: dispositivos móveis, telecomunicação, automação, sistemas embarcados, sistemas distribuídos, inteligência artificial, bancos de dados, sistemas web, visualização de software, arquitetura orientada a serviços, padrões de projeto, análise e modelagem de sistemas, IHM, Inteligência Artificial e simulação computacional.