

FRAMEWORK DE AVALIAÇÃO DA QUALIDADE DE DIAGNÓSTICOS GERADOS POR IA PARA EXPERIÊNCIA DO CLIENTE: UM ESTUDO DE CASO NA NEOENERGIA

Gustavo Henrique de Almeida^a , Luís Felipe Granello dos Santos^a , Inácio

Rodrigo Nere Dantas^a, Bruna Nascimento Marques^a

^a *Gerência de Tecnologia e Ciência de Dados, Neoenergia*

Resumo

Este artigo apresenta o framework NEOscore, desenvolvido para avaliar a qualidade de textos gerados por modelos de linguagem de grande porte (LLMs) para suportar a estratégia das áreas de Experiência do Cliente na Neoenergia, resumizando os principais pontos comentados pelos clientes em diversos canais e pesquisas de satisfação internas. Combinando métricas semânticas automáticas (como BERTscore e SBERT) com avaliação humana, o framework foi desenvolvido para avaliar critérios como consistência, relevância, concisão e coerência dos textos gerados pelos modelos. Os testes mostraram a capacidade do NEOscore em classificar adequadamente diagnósticos satisfatórios e insatisfatórios, mesmo em casos com alucinações e imprecisões factuais. O framework fornece uma abordagem objetiva para avaliar a qualidade de textos gerados por inteligência artificial permitindo que a empresa explore todo o potencial dos modelos sem comprometer a qualidade das informações.

Palavras-chave: NEOscore;BERTscore;CX

1. INTRODUÇÃO

Com a evolução exponencial dos modelos de linguagem grande (LLMs) tais como os GPTs da OpenAI e o Google Gemini, muitas áreas têm sido impactadas, pois estas ferramentas têm despertado o interesse para diversas atividades estratégicas de gerenciamento de negócios, como é o caso das áreas de Experiência do Cliente (CX) (Rane et al.,2023).

A CX, que se baseia em análises constantes sobre as jornadas, necessidades e satisfação dos clientes, está se tornando cada vez mais relevante em um cenário onde os clientes buscam personalização em suas interações com as empresas.

A interseção entre CX e inteligência artificial tem se tornado cada vez mais evidente, mas a literatura referente à avaliação objetiva da capacidade dos LLMs em desempenhar atividades como analisar sentimentos de clientes, atendimento personalizado automatizado e fornecer *insights* para a melhorias na experiência do cliente, ainda é preliminar (Peruchini et al.,2024; Verma et al.,2023).

Em particular, a análise de *feedback* do cliente é crucial para empresas brasileiras do setor de energia, como a Neoenergia, que passa pela transição para o mercado liberalizado, cenário em que a centralidade no cliente é um fator estratégico cada vez mais relevante. A alta demanda por serviços, a necessidade de resolver problemas rapidamente e a importância de oferecer uma experiência personalizada são fatores que exigem uma análise eficiente e precisa do *feedback* do cliente. Tradicionalmente, a análise de *feedback* envolve processos manuais, o que pode ser demorado e sujeito a vieses.

Neste sentido, uma das aplicações mais promissoras dos LLMs é a geração automática de resumos informativos sobre a experiência dos clientes com os serviços e atendimento prestados, que aqui chamamos de diagnóstico. Esses diagnósticos fornecem *insights* valiosos sobre as necessidades, expectativas e insatisfações dos clientes com base em comentários deixados em uma multiplicidade de canais, como pesquisas de satisfação internas, mídias sociais e canais de reclamação, como o Reclame Aqui e consumidor.gov. Os LLMs podem auxiliar na tomada de

decisão estratégica das empresas, proporcionando uma análise mais rápida, precisa e detalhada do *feedback* dos clientes.

Este estudo visa apresentar o *framework* NEOscore (*Natural-based Evaluation Overall score*), desenvolvido para avaliar a qualidade de diagnósticos gerados por diferentes LLMs a partir de comentários de clientes sobre o atendimento e serviços prestados pela Neoenergia. O objetivo é identificar os pontos fortes e fracos de cada métrica, estabelecendo limiares de qualidade com base em critérios como concisão, coerência, relevância e relevância. Para isso, combinamos métricas automáticas de similaridade semântica, como BERTscore e SBERT, com a avaliação humana realizada por especialistas e analistas das áreas de Experiência do cliente da empresa.

2. TRABALHOS RELACIONADOS

Definir métricas automáticas abrangentes que tenham correlação com julgamento humano e que considerem diversos aspectos dos textos gerados tais como coerência, precisão factual, fluidez, similaridade semântica, reconhecimento de nuances de sentimento, entre outros, é um dos maiores desafios para avaliação de textos gerados por modelos de linguagem.

No entanto, diversos estudos procuram avançar o conhecimento nesta área, buscando aprimorar as métricas existentes, avaliar sua performance combinando com protocolos de avaliação humana e discutindo aspectos referentes à confiabilidade e eficiência (Nguyen et al. (2024); Von Däniken et al. (2022); He et al. (2022); Eddine et al. (2022)). Isso demonstra que quanto mais estes modelos são incorporados em contextos de negócio, maior é a necessidade de métricas e protocolos de avaliação consistentes.

Podemos dividir as métricas automáticas em dois grupos: as métricas tradicionais em PLN (como ROUGE, BLEU e sobreposição exata de palavras-chave) e as métricas baseadas em transformadores pré-treinados (como BERTscore, MoverScore e SBERT), que vêm ganhando cada vez mais relevância nas tarefas de avaliação de textos gerados por IA.

As métricas tradicionais são extremamente rápidas, mas podem ser pouco confiáveis pois não são robustas para identificar sinônimos e paráfrases, limitando-se à identificação de correspondências exatas entre n-grams, como é o caso do ROUGE. Já as métricas baseadas em

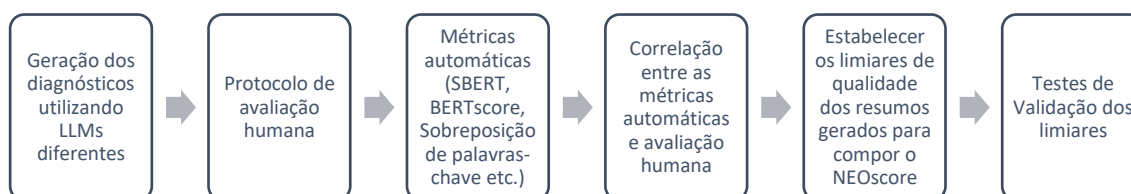
transformadores podem ser mais confiáveis, pois são capazes de identificar similaridades semânticas utilizando representações por *embeddings*, mas podem ser computacionalmente exaustivas e exigirem recursos computacionais que nem sempre estarão disponíveis em todas as aplicações (Eddine et al. (2022)).

Além dos estudos mencionados acima, Nguyen et al. (2024) conduziram uma pesquisa de avaliação humana para avaliar a correlação entre várias métricas automáticas e julgamentos humanos de qualidade de resumo. Eles descobriram que as métricas baseadas em transformadores pré-treinados, como BERTscore e MoverScore, apresentaram correlações mais fortes com julgamentos humanos do que as métricas tradicionais de PNL, como ROUGE e BLEU. Este estudo enfatiza a importância de incorporar a avaliação humana no processo de desenvolvimento e avaliação de métricas automáticas para avaliação da qualidade do resumo.

3.METODOLOGIA

Para definir o *framework* de avaliação dos diagnósticos na Neoenergia, adotamos a estratégia de combinar métricas automáticas de avaliação de sumarização de textos com um protocolo de avaliação humana. As métricas mais correlacionadas ao julgamento humano tiveram peso maior na definição dos limiares de qualidade exigidos para que um resumo automático seja considerado adequado às necessidades específicas do negócio. O esquema geral do desenvolvimento do NEOscore pode ser visualizado abaixo (figura 1).

Figura 1. Esquema geral do desenvolvimento do framework NEOscore.



3.1 GERAÇÃO DOS RESUMOS

3.1.1 Datasets

Foram utilizados 15 comentários de clientes sobre os canais digitais da empresa, coletados por meio de pesquisas de satisfação internas. Os comentários foram selecionados aleatoriamente, abrangendo comentários de clientes satisfeitos e insatisfeitos com a experiência proporcionada pelo aplicativo, site, WhatsApp e Agência Virtual da empresa.

O número de comentários foi escolhido para facilitar a etapa de avaliação humana, para que os avaliadores comparem as informações presentes nos comentários originais com as informações ressaltadas pelos diagnósticos, identificando possíveis inconsistências e mensurando quão bem os modelos transmitem as ideias sem omitir informações relevantes.

3.1.2 Pré-processamento

Os textos utilizados em todos os experimentos foram devidamente pré-processados, com remoção de stopwords, tokenização, remoção de acentuação e de pontuação.

3.1.3 Modelos e prompt utilizados

Foram utilizados cinco modelos diferentes para comparação, todos configurados com abordagens zero-shot. Os LLMs escolhidos foram os gemini-1.5-flash e gemini-1.5-pro da Google; GPT-3.5-turbo e GPT-4.0, da OpenAI; e o Microsoft Copilot. Além dos modelos de linguagem, utilizamos um diagnóstico humano como *baseline*, para comparar a performance dos modelos.

"Com base nos comentários abaixo, forneça um feedback executivo, conciso, sobre a experiência dos clientes com os canais digitais da empresa. O output deve ser um parágrafo único."

3.2 PROTOCOLO DE AVALIAÇÃO HUMANA

Para avaliação humana dos resumos gerados, 20 especialistas e analistas de experiência do cliente deram notas de 1 (completamente insatisfatório) a 5 (completamente satisfatório), utilizando escala Likert, para os critérios consistência, relevância, concisão e coerência para cada diagnóstico. O intuito da avaliação humana é compará-la com métricas automáticas de avaliação de geração textual, estabelecendo um *framework* completo que auxilie na definição de limiares de qualidade dos diagnósticos, para que estes sejam aderentes às necessidades da empresa, bem como elucidar possíveis falhas e limitações das métricas utilizadas para esta aplicação.

Os participantes tiveram acesso aos 15 comentários de clientes com opiniões diversas sobre os canais digitais da empresa, e durante o processo de avaliação o nome dos modelos utilizados foi ocultado para diminuir viés de confirmação, evitando que a percepção dos avaliadores fosse influenciada pelo conhecimento prévio sobre o desempenho dos modelos. Essa técnica de ocultação do nome dos modelos ajuda a garantir que a avaliação seja mais imparcial e objetiva, focando nos critérios estabelecidos para a análise da qualidade dos diagnósticos. Os comentários foram selecionados com base na diversidade de opiniões sobre diferentes canais.

Descrição dos critérios

Consistência: Um bom diagnóstico é consistente com os dados de entrada, ou seja, mantém as informações mais importantes dos comentários dos clientes, sem distorcer ou omitir dados relevantes.

Coerência: Um diagnóstico coerente apresenta uma estrutura lógica e clara, com informações interligadas de forma natural e compreensível.

Relevância: Um bom diagnóstico é relevante para a experiência do cliente, ou seja, foca nas informações mais importantes e úteis para a tomada de decisão e a melhoria da experiência do cliente.

Concisão: Um bom diagnóstico é conciso, evitando frases repetitivas e linguagem prolixa, transmitindo a informação de forma clara e direta.

3.3 MÉTRICAS AUTOMÁTICAS

Neste estudo, escolhemos métricas amplamente utilizadas para avaliar quais tem a maior correlação com os resultados da avaliação humana e com potencial para compor o *framework* de avaliação da qualidade dos diagnósticos gerados pelos LLMs. Dentre as métricas selecionadas estão: SBERT, BERTScore, sobreposição exata de palavras-chave, sobreposição de palavras-chave por similaridade semântica e entropia.

Para garantir a qualidade de resumos gerados para diagnóstico em experiência do cliente é imprescindível que o resumo mantenha as informações relevantes presentes nos comentários dos clientes, sem omissões e alterações no sentido original dos *feedbacks*.

Como queremos avaliar a performance de modelos para sumarização abstrativa de textos, é altamente esperado que o resumo gerado não contenha exatamente as mesmas palavras utilizadas pelos clientes, mas sim utilize sinônimos e paráfrases sem perder as nuances presentes nos comentários. Métricas de similaridade semântica e precisão factual desempenham um papel fundamental na mensuração de quão bem os LLMs conseguem reter os pontos levantados pelos clientes, permitindo uma avaliação precisa da qualidade de cada resumo.

3.3.1 BERTScore

O BERTscore é uma métrica de similaridade semântica introduzida por Zhang et al. (2020), que utiliza *embeddings* contextuais gerados pelo transformador pré-treinado BERT (*Bidiretional Encoder Representations from Transformers*) (Devlin et al., 2019). Essa métrica tem sido amplamente utilizada para avaliar tarefas de geração textual como sumarização de textos e tradução automática, em que a avaliação é feita a partir da comparação dos textos gerados por modelos de IA e textos de referência. Os valores de precisão (PBERT), recall (RBERT) e f1-score (FBERT) se baseiam na similaridade de cosseno e são medidas úteis de fluência e coerência dos textos, pois levam em conta o contexto geral das frases que os compõem.

O score RBERT calcula a média da maior similaridade entre cada palavra no resumo de referência e todas as palavras do resumo gerado, enquanto o score PBERT calcula a média da

maior similaridade entre cada palavra no resumo gerado e todas as palavras no resumo de referência. Já o FBERT combina a precisão e recall do BERTscore em uma única métrica, para quantificar o tradeoff entre PBERT e RBERT. Para o cálculo das métricas utilizamos o módulo BERTscorer, da biblioteca python bert_score.

A abordagem proposta pela utilização do BERTscore se sobressai quando comparada a métricas tradicionais como BLEU e ROUGE, pois leva em consideração o contexto em que as frases aparecem, o que é crucial ao avaliar textos gerados por LLMs, que não necessariamente irão reproduzir exatamente os comentários dos clientes, mas sim sintetizar as informações principais de forma estruturada.

3.3.2 Sentence-BERT (SBERT)

Sentence-BERT é uma modificação do modelo neural BERT, capaz de extrair relações semânticas significativas entre frases (Reimers et al., 2019). Este modelo gera *sentence embeddings* que podem ser comparadas por similaridade de cosseno, levando em consideração o contexto em que aparecem. Para nosso estudo, usamos SBERT para avaliar possíveis indícios de redundância nos diagnósticos, a partir das métricas de média (m-SBERT) e variância (v-SBERT) da similaridade entre pares de frases de um mesmo resumo. Também utilizamos a média ponderada de similaridade (w-SBERT), dando pesos maiores para a similaridade entre frases consecutivas.

Textos com frases muito similares entre si podem ser redundantes e não abarcar os principais pontos levantados pelos clientes, o que é indesejável para nossa aplicação. Valores altos de média da similaridade entre as frases de um resumo sugere que o texto apresenta informações repetidas, enquanto alta variância indica que o texto é diverso, mais abrangente e complexo.

A avaliação da redundância através do SBERT nos ajudará a identificar resumos que são concisos, informativos e com poucas repetições de ideias, contribuindo para a avaliação da qualidade geral dos diagnósticos gerados pelos LLMs e, consequentemente, para a melhoria da experiência do cliente na Neoenergia.

3.3.3 Sobreposição de palavras-chave (*Keyword Coverage*)

A sobreposição de palavras-chave é uma boa estratégia para avaliar se o resumo consegue reter as informações principais presentes nos comentários dos clientes. Esta avaliação é crucial para garantir a confiabilidade dos modelos e monitorar a relevância dos diagnósticos gerados por inteligência artificial. Neste estudo, utilizamos duas abordagens de sobreposição: sobreposição exata (KCr) e sobreposição semântica baseada em limiar (KCs).

As palavras-chave dos resumos foram extraídas utilizando o módulo *yake*, enquanto as palavras-chave dos comentários foram selecionadas pelo mesmo método e foram validadas por avaliação humana, priorizando termos considerados imprescindíveis para a aplicação desejada, tais como “*aplicativo*”, “*WhatsApp*”, “*site*”, “*login*”, “*atendimento*” etc.

Para calcular as métricas referentes à sobreposição exata (KCr), primeiro definimos uma função de indicação, tal que:

$$f_r(w_i, x_j) = \begin{cases} 1, & \text{se } w_i = x_j \\ 0, & \text{se } w_i \neq x_j \end{cases} \quad (1)$$

, em que w_i é uma palavra-chave do resumo gerado e x_j é uma palavra-chave dos comentários originais.

Assim, podemos definir:

$$KCr(precision) = \frac{1}{nw} \sum_{i=0}^{nw} f_r(w_i, x_j) \quad (2)$$

$$KCr(recall) = \frac{1}{nx} \sum_{i=0}^{nw} f_r(w_i, x_j) \quad (3)$$

em que nw é o número de palavras-chave do resumo gerado e nx é o número de palavras-chave do texto original. Os valores de precisão e recall variam de 0 a 1, sendo 1 a correspondência exata entre todas as palavras-chave de ambos os textos. Uma limitação dessa abordagem é que caso o resumo utilize sinônimos a correspondência não será contabilizada, mesmo que a ideia transmitida seja a mesma e esteja condizente com os comentários originais dos clientes.

Para superar esta limitação, utilizamos uma abordagem de sobreposição semântica, especialmente relevante para avaliar tarefas de sumarização abstrativa e, neste caso, pode ser

um bom indício de quão bem um resumo capta os tópicos mais comentados pelos clientes levando em consideração o contexto das frases, o que é crucial para aplicações em CX. Para isso, o primeiro passo foi representar os comentários originais e os diagnósticos gerados utilizando *embeddings* para cada palavra-chave. Utilizamos o modelo paráfrase-distilroberta-v2 para representação por *embeddings* contextuais. Depois, calculamos a similaridade de cosseno de cada par de palavras-chave, considerando como correspondências pares de palavras-chave com similaridade acima de um limiar estabelecido. Portanto, podemos definir a função de indicação como

$$f_s(w_i, x_j) = \begin{cases} 1, & \text{se } \text{sim_cos}(embedding(w_i), embedding(x_j)) \geq t \\ 0, & \text{se } \text{sim_cos}(embedding(w_i), embedding(x_j)) < t \end{cases} \quad (4)$$

, em que t é o limiar de similaridade necessário para considerar uma correspondência. Neste caso, testamos diversos limiares empiricamente e o limiar adotado foi de 0.7.

Em alguns casos, uma única palavra-chave do resumo gerado pode ser semanticamente similar a várias palavras dos textos originais e, como queremos apenas uma correspondência por palavra-chave, consideramos a correspondência com a maior similaridade, de modo que:

$$KCs(precision) = \frac{1}{nw} \sum_{i=0}^{nw} f_s(w_i, x_j) \quad (5)$$

$$KCs(recall) = \frac{1}{nx} \sum_{i=0}^{nw} f_s(w_i, x_j) \quad (6)$$

A precisão e o recall da sobreposição de palavras-chave podem ser boas medidas para avaliar a relevância dos resumos, a partir da cobertura de tópicos abordados pelos clientes que estão presentes nos resumos, garantindo que os insights gerados sejam de fato baseados na percepção dos clientes.

3.3.4 Entropia de Shannon e diversidade lexical

A entropia de Shannon é uma métrica amplamente utilizada em Teoria da Informação (Shannon, 1948), e no contexto de análise de resumos gerados por IA pode ser utilizado para

avaliar a diversidade e riqueza lexical dos textos. Textos mais diversos e lexicalmente complexos possuem valores mais elevados de entropia, enquanto textos mais simples e até mesmo repetitivos apresentam valores mais baixos, sendo um bom indicador da redundância interna dos diagnósticos gerados.

$$H(X) = -\sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (7)$$

, onde $p(x_i)$ é a probabilidade de ocorrência de uma palavra x_i .

3.4 VALIDAÇÃO DO FRAMEWORK

Validamos a capacidade das métricas automáticas de avaliar a qualidade dos diagnósticos realizando análises de correlação entre cada métrica e os critérios de avaliação humana, utilizando coeficientes de correlação de Pearson e Spearman. Definimos nosso framework de avaliação NEOscore atribuindo pesos maiores para as métricas mais fortemente correlacionadas ao julgamento humano, e definindo limiares de qualidade. Essa abordagem garante que o framework seja preciso e reflita a importância relativa de cada métrica na avaliação da qualidade dos diagnósticos. O framework é flexível, permitindo a aplicação do protocolo de avaliação humana, o que prioriza aspectos específicos do contexto de cada negócio e garante que aspectos de qualidade imprescindíveis nos textos tenham peso maior na classificação de um diagnóstico como satisfatório ou insatisfatório.

Posteriormente, realizamos diversos testes para validar a robustez do framework, introduzindo diagnósticos com alucinações, imprecisões factuais, redundâncias, além de resumos gerados a partir de um grande volume e de comentários e com diferentes prompts.

4.RESULTADOS

4.1 PERFORMANCE DOS MODELOS

Para avaliar se os valores obtidos pelas métricas estão coerentes, utilizamos como baseline um diagnóstico escrito por um analista, e calculamos todas as métricas para este resumo. Nossos principais resultados podem ser observados na tabela 1.

Métrica	Gemini 1.5-flash	Gemini 1.5-pro	GPT 3.5-turbo	GPT 4.0	Copilot	Humano
PBERT	0,716	0,721	0,710	0,735	0,745	0,721
RBERT	0,687	0,670	0,679	0,658	0,686	0,676
FBERT	0,701	0,695	0,694	0,695	0,714	0,698
m-SBERT	0,417	0,368	0,405	0,342	0,316	0,358
v-SBERT	0,048	0,059	0,052	0,052	0,041	0,045
w-SBERT	0,489	0,229	0,373	0,238	0,508	0,295
KCr(precision)	0,207	0,333	0,239	0,405	0,400	0,378
KCr(recall)	0,158	0,224	0,145	0,197	0,237	0,224
KCs(precision)	0,552	0,667	0,717	0,676	0,756	0,600
KCs(recall)	0,421	0,447	0,434	0,329	0,447	0,355
Entropia	6,212	5,835	6,070	5,698	5,991	5,889

Tabela 1. Resultados das métricas automáticas.

4.1.1 BERTscore e sobreposição de palavras-chave

Um bom resumo deve sintetizar os principais pontos dos comentários dos clientes sem distorcer o sentido original. Para avaliar a similaridade semântica e a relevância dos resumos gerados, utilizamos as métricas de BERTscore e sobreposição de palavras-chave.

Todos os modelos apresentaram bons resultados no BERTscore, com precisão acima de 0.71 e recall acima de 0.65, indicando que todos conseguiram capturar as ideias principais dos comentários dos clientes. No entanto, algumas diferenças sutis podem ser observadas entre os modelos, como no caso do diagnóstico gerado pelo gpt-4.0.

O diagnóstico gerado pelo GPT-4.0 apresentou a segunda maior precisão (0.73) e o pior recall (0.65), sugerindo que, apesar de sintetizar informações relevantes, ele pode estar omitindo alguns tópicos importantes dos comentários dos clientes. A análise da sobreposição de palavras-chave por similaridade semântica (KCs) corrobora a ideia de que o GPT-4.0 está omitindo tópicos importantes, com uma precisão semântica de 0.67 e um recall de 0.36, o mais baixo entre os modelos. Além disso, este foi o diagnóstico com maior número de correspondências exatas de palavras-chave (KCr), com 41% de precisão, indicando que este foi o diagnóstico que mais compartilhou palavras-chave com os comentários originais.

Os valores de KCr foram menores que os observados para sobreposição semântica (KCs) para todos os modelos, o que era um resultado esperado já que estamos tratando de sumarizações

abstrativas, ou seja, os modelos podem não estar utilizando as palavras-chave exatamente iguais às dos comentários, mas utilizam sinônimos e paráfrases sem perda semântica, o que é crucial para que um resumo seja satisfatório.

Comparando as métricas BERTscore e KCs, observamos que todos os modelos são consistentes em termos de precisão, com exceção do diagnóstico gerado pelo gemini-1.5-flash, que apresentou KCs muito inferior à precisão BERTscore (0.55 e 0.72, respectivamente). Essa discrepância pode comprometer a relevância do diagnóstico, pois ele mantém o sentido original dos temas tratados pelos clientes, porém não necessariamente traz as informações mais importantes. Além disso, o limiar de similaridade adotado para o cálculo das métricas KCs foi de 0.7, o que pode ser considerado um limiar suficientemente alto para capturar a correspondência de sinônimos, reforçando a ideia de que o diagnóstico gerado pelo gemini-1.5-flash pode ser menos relevante que os demais.

4.1.2 SBERT e redundância

As métricas SBERT foram utilizadas para identificar possíveis redundâncias nos diagnósticos, uma vez que foi implementada para a comparação da similaridade semântica entre as diferentes frases que compõem um diagnóstico. Os modelos em geral apresentaram médias de similaridade relativamente baixas, com o diagnóstico gerado pelo gemini-1.5-flash tendo a maior média (0.41) e o copilot com a menor média (0.32), sugerindo que este é o diagnóstico com maior diversidade de frases, em termos semânticos.

4.1.3 Entropia e diversidade lexical

A entropia neste caso pode ser entendida como uma medida da diversidade lexical do texto, contribuindo para avaliação da riqueza do vocabulário utilizado. Resumos com baixa entropia podem apresentar muitas palavras repetitivas e com baixa diversidade, enquanto textos com maior entropia demonstram maior riqueza lexical, sendo uma medida direta de qualidade do texto gerado. Como todos os diagnósticos apresentaram valores de entropia comparáveis à entropia do diagnóstico humano (5.88), é possível inferir que todos os modelos foram capazes de gerar resumos com a complexidade e diversidade lexical adequadas ao contexto de negócio.

De forma geral, os modelos de linguagem tiveram performance comparável e em alguns casos até melhor que o diagnóstico humano, demonstrando a capacidade promissora destes modelos para geração de diagnósticos em experiência do cliente.

4.2 CORRELAÇÕES COM AVALIAÇÃO HUMANA

Após análise das possíveis correlações entre as avaliações humanas e as métricas automáticas utilizando os métodos de Pearson e Spearman, encontramos relações significativas entre as variáveis para cada critério avaliado (consistência, coerência, concisão, relevância). Métricas como RBERT, SBERT, entropia e KCs demonstraram boa correlação com julgamento humano, demonstrando que podem ser robustas para compor o framework. Os correlogramas podem ser observados no apêndice B ao final deste artigo.

4.2.1 Concisão/Redundância

As métricas que apresentaram correlações mais fortes com este critério foram m-SBERT e precisão de KCs e KCr. A precisão da sobreposição de palavras-chave demonstrou correlações positivas com as notas dadas pelos avaliadores humanos para a concisão dos resumos, sugerindo que diagnósticos com alta precisão focam nos aspectos mais relevantes dos comentários, diminuindo a inclusão de informações desnecessárias. Já o m-SBERT apresenta correlações negativas, indicando que resumos com frases muito similares entre si tendem a ser redundantes e pouco concisos, sendo provavelmente mais longos que o necessário.

Um exemplo é o diagnóstico gerado pelo gemini-1.5-flash, que teve a menor média de concisão segundo avaliação humana (3.43) e a pior precisão semântica KCs (0.55), além de ter a maior média de similaridade entre frases (0.41), sugerindo que este resumo é o com mais indícios de redundância, comprometendo a concisão necessária para um bom diagnóstico executivo.

4.2.2 Coerência

Correlações positivas foram encontradas para as métricas m-SBERT e v-SBERT, indicando que, por mais que um resumo com alta similaridade entre as frases seja um indicativo de redundância e comprometa a concisão do texto, é necessário manter certo grau de similaridade

para garantir que o texto seja coerente, com transições fluidas entre as frases. A variância, nesse caso, indica que é altamente desejável ter em um diagnóstico pares de frases com diferentes graus de similaridade.

A correlação negativa com a média ponderada w-SBERT, sugere ainda que alta similaridade entre frases consecutivas é um indicativo de redundância local que pode afetar a qualidade do resumo, criando um fluxo artificial e repetitivo, enquanto certa similaridade média geral entre as frases pode ser positiva para a coerência de um texto, por meio de transições mais fluidas e reforçando as ideias principais.

4.2.3 Consistência

A consistência está diretamente relacionada à capacidade dos modelos em resumir os comentários dos clientes mantendo as informações mais relevantes sem distorções no sentido original dos comentários. Para diagnósticos em experiência do cliente, este é um critério fundamental, e as métricas que apresentaram maiores correlações positivas com este critério foram o recall do BERTscore e da sobreposição semântica KCs e entropia. Para este critério, nenhuma das métricas apresentou correlações negativas fortes (coeficientes maiores ou iguais a 0.7).

Um resumo com alto RBERT e KCs(recall) tende a ser mais consistente, pois mantém o sentido original da maioria das informações importantes dos comentários dos clientes. Além disso, resumos com alta entropia tendem a ser mais consistentes, pois provavelmente mantêm a diversidade de informações e ideias dos comentários, sem simplificar excessivamente ou omitir aspectos relevantes. O resumo gerado pelo gemini-1.5-flash, apesar de ter tido a menor média de concisão, foi o mais bem avaliado em termos de consistência das informações (média de 4.29 pontos, na avaliação humana), além de apresentar a maior entropia entre todos os resumos (6.28) e maior RBERT (0.69).

4.2.4 Relevância

A relevância de um resumo está relacionada à consistência, porém os critérios possuem diferenças importantes. Considerar um resumo como relevante ou não depende se o diagnóstico

traz as informações consideradas mais importantes para as áreas de Experiência do Cliente, não apenas sua capacidade de reproduzir e sintetizar o conteúdo dos comentários sem introduzir informações falsas e distorcidas (como é o caso da consistência).

Em termos de correlações, observamos que tanto RBERT e KCs(recall) apresentaram correlações positivas, indicando que resumos que cobrem o maior número de palavras-chave (no caso do KCs) e mantêm o sentido original dos comentários (no caso do RBERT) foram considerados mais relevantes pelos avaliadores humanos. O resumo gerado pelo copilot foi considerado o segundo mais relevante (média de 4.2 pontos), e apresenta o segundo maior RBERT e o maior KCs(recall).

4.2.5 Qualidade Geral

As métricas mais consistentes para avaliação da qualidade dos resumos gerados para diagnósticos em experiência do cliente foram RBERT, KCs(recall), entropia e KCr(precision). A precisão da sobreposição exata de palavras-chave foi a métrica com correlação negativa mais forte, enquanto as demais apresentaram correlações positivas, contribuindo para melhorar a qualidade geral dos diagnósticos.

Avaliar a qualidade geral de um texto gerado especificamente para nosso contexto de negócio é uma tarefa que exige uma abordagem multifacetada, pois cada métrica automática consegue dar conta de apenas alguns aspectos necessários para considerar um texto como tendo boa qualidade. Isto implica na necessidade estratégica de definir quais aspectos são mais importantes para cada contexto, para realizar os ajustes necessários no framework.

5.DEFINIÇÃO DO FRAMEWORK

Após as análises de correlação e performance das métricas em avaliar a qualidade dos diagnósticos automáticos gerados por LLMs, definimos o framework NEOscore, uma métrica unificada e flexível baseada na média ponderada das métricas utilizadas para avaliação da qualidade dos resumos.

Os pesos considerados para cada métrica foram definidos de acordo com os coeficientes de correlação de Pearson, de modo que métricas mais fortemente correlacionadas com os critérios de avaliação humana recebem pesos maiores (tanto positivos quanto negativos) na composição do NEOscore (tabela 2).

Métrica	Coerência	Concisão	Consistência	Relevância	Qualidade Geral
FBERT	-0,754	0,000	0,000	0,000	0,000
PBERT	-0,872	0,000	0,000	0,000	0,000
RBERT	0,000	0,000	0,000	0,000	0,925
Entropia	0,000	0,000	0,819	0,902	0,924
KCr(precision)	-0,641	0,659	0,000	-0,696	-0,734
KCs(precision)	0,000	0,977	0,000	0,000	0,000
KCs(recall)	0,000	0,000	0,753	0,000	0,826
m-SBERT	0,758	-0,674	0,000	0,000	0,000
v-SBERT	0,825	0,000	0,000	0,000	0,000
w-SBERT	0,000	0,000	0,716	0,829	0,708

Tabela 2. Pesos considerados para o cálculo do NEOscore.

Deste modo, definimos o framework NEOscore como a média ponderada de cada critério de qualidade, com os pesos k_i definidos pelos coeficientes de correlação dos indicadores mais correlacionados ao critério de qualidade, conforme formulação abaixo.:

$$NEOscore = (k_1 * Consistência + k_2 * Concisão + k_3 * Coerência + k_4 * Relevância + k_5 * Qualidade Geral) / (k_1 + k_2 + k_3 + k_4 + k_5) \quad (8)$$

A Tabela 3 apresenta as médias obtidas por cada modelo na pesquisa de avaliação humana e os respectivos valores de NEOscore. Marcamos em azul o resumo com a maior nota e em vermelho o resumo com menor nota em cada critério. Interessante observar que os valores de NEOscore são coerentes com as médias observadas para cada diagnóstico, em especial para os gerados pelos modelos gpt-4.0 e gemini-1.5-flash.

Modelo	Coerência	Concisão	Consistência	Relevância	Qualidade Média	NEOScore
Copilot	3,714	4,142	3,928	4,142	3,982	4,253
GPT-3.5-turbo	4,142	4,071	4,000	3,785	4,000	4,181
GPT-4.0	3,928	4,000	3,714	3,214	3,714	3,259
Gemini-1.5-flash	4,142	3,428	4,285	4,500	4,089	4,354
Gemini-1.5-pro	4,285	3,928	3,571	4,000	3,946	3,605
Humano						3,494

Tabela 3. Comparação entre os resultados da pesquisa de avaliação humana e NEOScore.

O primeiro obteve a menor nota para coerência, relevância e média geral, e a menor pontuação NEOScore. Já o resumo gerado pelo gemini-1.5-flash obteve as maiores pontuações nos critérios de consistência, relevância e média geral, apresentando também a maior pontuação de NEOScore, demonstrando que o framework avaliou corretamente os diagnósticos, em concordância com o julgamento humano.

O limiar de qualidade do NEOScore foi definido a partir da pontuação obtida pelo diagnóstico humano, tendo a qualidade necessária para ser considerado satisfatório para a empresa. Deste modo, resumos com NEOScore abaixo de 3,5 serão descartados (qualidade insatisfatória).

6. TESTES DE VALIDAÇÃO

Para garantir a validade e robustez do framework estabelecido, performamos uma série de testes para analisar quão bem o NEOScore é capaz de diferenciar um diagnóstico satisfatório de um diagnóstico insatisfatório. Os testes foram os seguintes:

Teste 1. Resumo composto apenas por alucinações e imprecisões factuais: introduzimos um resumo com diversas distorções para avaliar se o framework é robusto em classificá-lo como insatisfatório.

Teste 2. Resumo pouco informativo: texto muito breve, sem informações relevantes

Teste 3. Resumo com inserção de uma informação imprecisa: texto coerente, que cobre vários pontos abordados pelos clientes, mas que possui uma informação falsa

Teste 4. Resumo altamente redundante

Teste 5. Resumo que foge do tema canais digitais

Teste 6. Resumo gerado a partir de 500 comentários sobre canais digitais: aumentar o volume de comentários para identificar o impacto no score geral do modelo

Teste 7. Resumo gerado com prompt direto, pouco instrutivo, para avaliar o impacto do prompt nos resultados

Teste 8. Resumo gerado com prompt detalhado, fornecendo mais contexto para os modelos.

Um bom framework precisa ser capaz de identificar imprecisões factuais, fuga do tema, alucinações e textos pouco informativos, e classificá-los corretamente como insatisfatórios. Por outro lado, o framework também precisa classificar corretamente bons resumos gerados a partir de um grande volume de comentários e com diferentes prompts de comando, permitindo maior flexibilidade nas requisições sem comprometer a qualidade geral do diagnóstico.

De acordo com os resultados, o sistema de pontuação NEOScore conseguiu classificar adequadamente todos os diagnósticos utilizados para teste (tabela 4). A força da classificação é a diferença entre a nota obtida pelo resumo e o limiar estabelecido (3,5 pontos), podendo ser positiva (força de aceitação) ou negativa (força de rejeição).

	Teste	NEOScore	Classificação	Força da classificação
Qualidade insuficiente	Teste 1	2,638	Insatisfatório	-0,862
	Teste 2	1,806	Insatisfatório	-1,694
	Teste 3	2,9	Insatisfatório	-0,6
	Teste 4	1,982	Insatisfatório	-1,518
	Teste 5	2,101	Insatisfatório	-1,399
Qualidade suficiente	Teste 6	4,627	Satisfatório	1,127
	Teste 7	3,95	Satisfatório	0,45
	Teste 8	4,68	Satisfatório	1,18

Tabela 4. Resultados dos testes realizados para validar o framework.

Os resumos utilizados para os testes 1 e 3 são particularmente interessantes, pois tiveram as maiores pontuações de NEOscore dentre os textos com qualidade insuficiente (pois apresentam características altamente desejáveis para um bom diagnóstico, sendo coerentes, concisos, fluidos etc.), porém são compostos por imprecisões factuais e informações falsas que podem distorcer o diagnóstico sobre a percepção dos clientes. A presença dessas alucinações faz com que a pontuação dos resumos fique abaixo do limiar de qualidade exigido. Este caso reforça a capacidade do framework em avaliar os resumos, penalizando corretamente os aspectos que comprometem a qualidade geral do texto.

Os valores de NEOscore foram muito baixos para os testes 2 e 4, evidenciando que o framework penalizou corretamente a redundância excessiva e o resumo com poucas informações relevantes. Além disso, o resumo com fuga do tema (teste 5) também recebeu uma pontuação baixa, conforme o esperado.

Os testes 6, 7 e 8 são testes de stress, avaliando quão bem o NEOscore se adapta a alterações de prompt e volume de comentários de origem. No dia a dia da empresa, o volume de comentários recebidos é muito maior que os 15 comentários selecionados para as análises. Para validar se as métricas e os limiares definidos estão de acordo com as demandas reais, o diagnóstico gerado para realizar o teste 6 utilizou 500 comentários como referência e obteve um bom NEOscore (4.627 pontos) e foi classificado adequadamente como satisfatório.

Os testes realizados com diferentes prompts também tiveram boas pontuações, com o prompt mais detalhado obtendo o maior valor de NEOscore, sendo um indício de que a qualidade do resumo pode estar diretamente relacionada ao prompt utilizado durante a requisição. Os resumos e prompts utilizados durante os testes podem ser consultados no apêndice C ao final deste artigo.

7.DISSCUSSÃO

Avaliar objetivamente a qualidade dos textos gerados por modelos de linguagem é desafiador principalmente porque não existe uma métrica unificada que consiga abranger os diversos aspectos do texto, tais como precisão factual, coerência, coesão, consistência e relevância. Isto

faz com que sejam necessárias várias métricas diferentes, o que pode levar a interpretações complexas. Por mais que as métricas estejam em constante evolução, ainda hoje o padrão ouro para avaliação da qualidade de textos é o julgamento humano. Entretanto, quanto mais intenso for o uso destes modelos em contextos de negócio, como é o caso da Neoenergia, a avaliação humana pode ser uma tarefa onerosa, o que ressalta a importância de estudos que contribuam para o refinamento de métricas automáticas.

As métricas escolhidas para este estudo demonstraram boas correlações com os critérios de avaliação humana, sendo um avanço importante para a confiabilidade das métricas em questão na aplicação prática de frameworks e rotinas automáticas de avaliação de qualidade, garantindo a eficiência na geração de diagnósticos sem comprometer a integridade das informações.

Além disso, os modelos apresentaram bons resultados tanto em termos de métricas automáticas quanto em termos de avaliação humana, evidenciando que são boas ferramentas para as tarefas de sumarização abstrativa de textos em contextos reais de negócio, como é o caso dos diagnósticos para Experiência do Cliente. Em alguns casos, conseguiram notas maiores que o diagnóstico humano utilizado como baseline.

Os testes de validação também evidenciam que o NEOscore é uma boa medida da qualidade dos resumos, pois o framework foi capaz de classificar corretamente todos os resumos utilizados para teste.

7.1 FORÇAS DO ESTUDO

- Abordagem multifacetada, utilizando várias métricas
- Validação por meio de protocolo de avaliação humana
- Framework flexível de acordo com as necessidades de negócio
- Diagnóstico humano como baseline
- Proposta de avaliação objetiva dos textos gerados por IA a partir de uma única métrica combinada (NEOscore)

7.2 LIMITAÇÕES DO ESTUDO

- Os resultados podem não ser generalizáveis para outros domínios, por focar em critérios particularmente importantes para diagnósticos em CX. No entanto, a estratégia por trás do framework pode ser adaptada de acordo com as necessidades de outras áreas de negócio.
- Métricas calculadas com base em textos em português: modelos para representação de embeddings em inglês são geralmente mais abrangentes para identificação de similaridade semântica, pois foram treinados em corpus textuais maiores.
- Métricas amplamente utilizadas como ROUGE, BLEU e METEOR não foram utilizadas para compor o NEOscore.
- Os LLMs foram utilizados em abordagens zero-shot, ou seja, sem treinamento específico para a tarefa de geração de diagnósticos.

7.3 IMPLICAÇÕES PRÁTICAS PARA A NEOENERGIA

O desenvolvimento e adoção do framework tem implicações práticas significativas para a empresa, pois além de gerar diagnósticos sobre a experiência dos clientes com rapidez e eficiência, contribui para aumentar a confiabilidade ao repassar as informações aos executivos. Além disso, conseguimos aprimorar o fluxo para geração destes resumos, alterando etapas de pré-processamento, melhorando as instruções dos prompts utilizados nas requisições etc.

Com esta abordagem, conseguimos manter sempre o foco em alcançar valores mais altos de NEOscore e, consequentemente, gerar diagnósticos melhores, aproveitando todo o potencial dos LLMs para esta tarefa sem comprometer a qualidade e confiabilidade dos resumos. Neste estudo de caso nos concentramos na aplicação de geração de diagnósticos com base em comentários de clientes, mas o framework pode ser testado e adaptado para outras aplicações, tais como reescrita de documentos, higienização de textos para comunicação com o cliente, avaliação da qualidade das interações em *chatbots*, dentre outras. Este é apenas um ponto de partida para novos desenvolvimentos.

8. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste estudo de caso apresentamos o framework NEOscore, um sistema de pontuação multifacetado para avaliação da qualidade dos textos gerados por diferentes modelos de linguagem, com foco na geração de diagnósticos para as áreas de Experiência do Cliente. Os critérios de qualidade foram definidos de acordo com o domínio específico de negócio, em que a consistência, relevância, concisão e coerência dos resumos são cruciais para um resumo efetivo, o que pode não ser diretamente aplicável a outros domínios, porém demonstrou ser efetivo para as aplicações em Experiência do Cliente na Neoenergia.

Outros estudos contribuiriam expressivamente para a área de geração textual aplicada a negócios, testando diferentes técnicas de prompts e métricas que ficaram de fora de nosso estudo, como ROUGE, BLEU, METEOR e métricas de legibilidade, como índice de Gunning-Fog e índice de Flesch-Kincaid. Além disso, o framework proposto se beneficiaria de estudos aplicados a outras áreas da empresa e até mesmo a negócios diferentes do setor elétrico, para avaliar a generalização do NEOscore.

9.SOBRE O CONEXÃO DIGITAL

Este estudo foi suportado pelo Conexão Digital, um projeto de Transformação Digital de P&D ANEEL, de caráter de Pesquisa e Desenvolvimento, que busca transformar digitalmente o atual modelo de relacionamento dos clientes com as distribuidoras de energia para proporcionar experiências mais ágeis, personalizadas a partir da inovação e redesenho das estratégias e processos da Neoenergia.

REFERÊNCIAS

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics

Eddine, M. K., et al. (2022). FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1305-1318)

He, T., et al. (2022). On the blind spots of model-based evaluation metrics for text generation. arXiv preprint arXiv:2212.10020

Nguyen, H., et al. (2024). A comparative study of quality evaluation methods for text summarization. arXiv preprint arXiv:2407.00747.

Peruchini, M., et al. (2024). Between artificial intelligence and customer experience: A literature review on the intersection. Discov Artif Intell, 4(4). <https://doi.org/10.1007/s44163-024-001058>

Rane, N. (2023). Role and challenges of ChatGPT and similar generative artificial intelligence in business management. SSRN Electronic Journal. <http://doi.org/10.2139/ssrn.4603227>

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084

Verma, R. K., & Kumari, N. (2023). Generative AI as a tool for enhancing customer relationship management automation and personalization techniques. International Journal of Responsible Artificial Intelligence, 13(9), 1-8

Von Däniken, P., et al. (2022). On the effectiveness of automated metrics for text generation systems. arXiv preprint arXiv:2210.13025

Zhang, T., et al. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675

APÊNDICES

A. DATASETS E DIAGNÓSTICOS GERADOS

Comentários Originais

- *Acesso a 2 via no site um pouco burocrático. Tarifas caras*
- *Mais lojas físicas com mais serviços disponíveis e melhorar os erros no site e app*
- *O atendimento digital sempre resolve rápido às solicitações, facilita muito.*
- *O atendimento pelo WhatsApp é muito demorado, enviam um longo questionário e que antes de conseguir responder já encerram o atendimento fazendo a gente ter que esperar novamente, tive que aguardar umas duas horas para atendimento e quando consegui e estava respondendo o questionário, simplesmente encerram o atendimento e tive que fazer novamente.*
- *Foi bom o acesso no app foi tranquilo Só não consegui parcelar a fatura*
- *Deveria ter no app, a opção pra informar pagamento de faturas.*
- *Gosta muito do atendimento no WhatsApp*
- *Foi rápido aprovação hoje estou tendo problemas em entrar no aplicativo*
- *Avisar via WhatsApp quando houver contas em atraso*
- *Acho que a agência virtual pode melhorar, com melhor acesso e menos burocracia.*
- *Atendimento pelo WhatsApp é show!!!*
- *Atendimento mesmo que de forma digital, foi nota máxima. Parabéns a todos*
- *Ao tentar acessar o aplicativo, preenchi todos os dados e ao tentar finalizar a solicitação não digitava o número de celular, por mais que tentasse*
- *Simplificar a solicitação de serviços pelo site e agência virtual*
- *Boa tarde. Pelo app o atendimento ainda anda, mas quando e direcionado para o site, nada funciona*

Resumo gerado por humano (baseline)

“Os clientes expressaram opiniões mistas sobre os canais digitais da empresa. Alguns elogiaram o atendimento pelo WhatsApp e digital, no geral, reconhecendo que os canais facilitam a experiência e resolução de problemas de forma fácil e rápida. No entanto, alguns clientes reclamam da burocracia e erros para acessar o app (problemas de login), e para acessar a segunda via pelo site. Além disso, os clientes sugerem aviso por WhatsApp de contas em atraso, bem como a opção de parcelamento de faturas no aplicativo. É imprescindível corrigir os erros de acesso aos canais digitais, bem como diminuir a burocracia para solicitação de alguns serviços, aumentando ainda mais a praticidade para o cliente.”

Resumo gerado pelo Gemini-1.5-flash

“O feedback dos clientes indica uma experiência mista com os canais digitais da empresa. Enquanto o atendimento digital e o aplicativo são elogiados por sua rapidez e praticidade, há reclamações sobre a

burocracia no site, dificuldades na navegação e problemas com o aplicativo. A falta de opções de parcelamento da fatura no aplicativo e a necessidade de informar o pagamento de forma manual também são pontos a serem considerados. O atendimento via WhatsApp, apesar de positivo para alguns, apresenta problemas de demora e encerramento abrupto da conversa, prejudicando a experiência do cliente. É crucial investir em melhorias na usabilidade do site e aplicativo, simplificar processos e oferecer mais opções de pagamento e gerenciamento de contas. A comunicação proativa sobre contas em atraso via WhatsApp é um ponto positivo, mas a qualidade do atendimento precisa ser aprimorada para garantir uma experiência positiva para todos os clientes."

Resumo gerado pelo Gemini-1.5-pro

"A experiência dos clientes com os canais digitais da empresa apresenta uma dualidade. Enquanto o atendimento via WhatsApp e a resolução rápida de solicitações são elogiados, a experiência com o site e aplicativo é frustrante. Usuários relatam dificuldades em acessar a 2ª via de faturas, realizar pagamentos, navegar na agência virtual e finalizar solicitações, além de menções a erros e burocracia. Para uma experiência digital completa e satisfatória, é crucial investir em melhorias na usabilidade e funcionalidade do site e aplicativo, incluindo a opção de informar pagamentos e avisos de contas em atraso, além de otimizar o fluxo de atendimento no WhatsApp para evitar encerramentos prematuros."

Resumo gerado pelo GPT-3.5-turbo

"Caro time, o feedback dos clientes sobre os canais digitais tem sido misto. Embora o atendimento digital seja elogiado pela rapidez e eficácia, há reclamações sobre o atendimento pelo WhatsApp devido à demora e encerramento precoce das conversas. Além disso, há pedidos para melhorias no site e no aplicativo, como facilitar a emissão da 2ª via da fatura, parcelamento da mesma e notificações via WhatsApp sobre contas em atraso. Também foi mencionada a necessidade de mais acessibilidade e menos burocracia na agência virtual. Sugiro que priorizemos essas melhorias nos canais digitais para aprimorar a experiência do cliente."

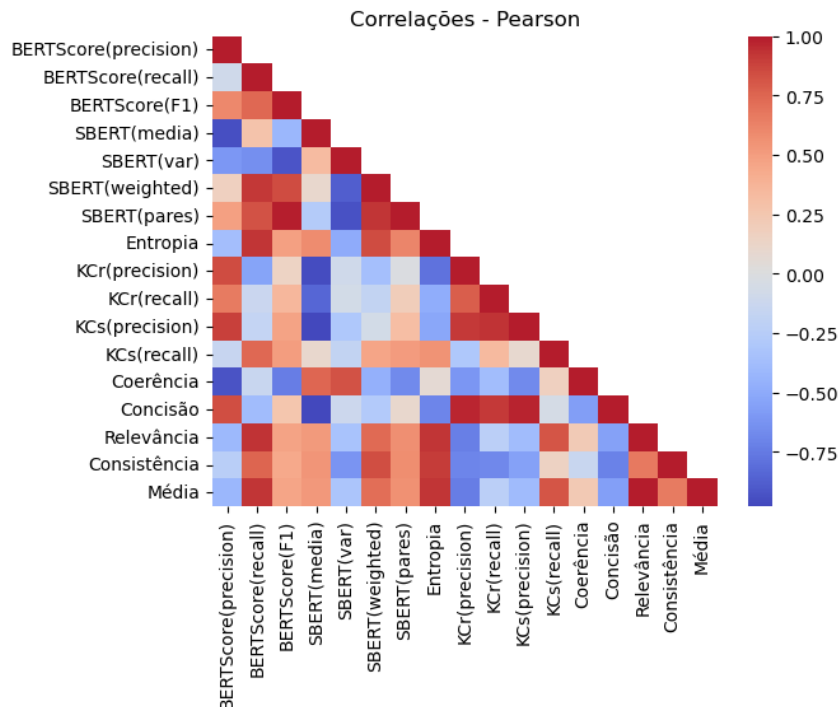
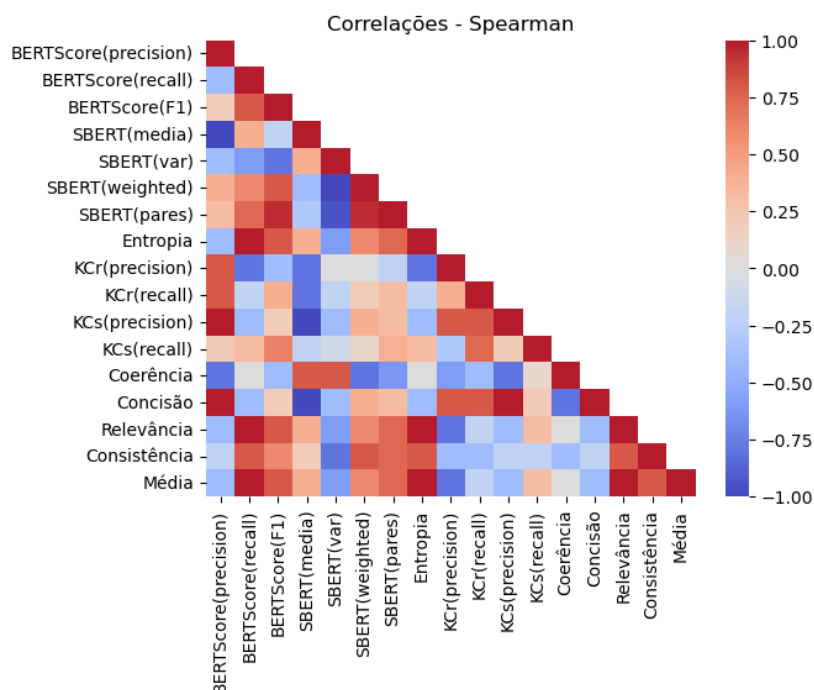
Resumo gerado pelo GPT-4.0

"Os clientes da empresa expressaram opiniões mistas sobre o uso dos canais digitais. Eles reconhecem a eficácia da resolução de solicitações nos meios digitais e apreciam o atendimento via whatsapp, mas tiveram problemas com a burocracia, erros no site e no aplicativo, e sentiram falta de algumas opções no aplicativo, como informar pagamento de faturas e parcelar a fatura. Além disso, o tempo de espera no atendimento pelo whatsapp foi descrito como demorado. Por fim, a experiência no site foi criticada devido à burocracia e problemas técnicos."

Resumo gerado pelo Microsoft Copilot

"Os clientes apreciam a rapidez do atendimento digital e a facilidade de uso do aplicativo, embora tenham enfrentado problemas ao tentar acessá-lo. Eles gostam do atendimento pelo WhatsApp, mas acham o processo demorado e burocrático. Além disso, eles sugerem melhorias no site e no aplicativo, como a simplificação do acesso à segunda via de contas e a inclusão de uma opção para informar pagamentos de faturas. Também expressaram o desejo de serem notificados via WhatsApp sobre contas em atraso. Além disso, os clientes gostariam de ver mais lojas físicas com mais serviços disponíveis. As tarifas foram consideradas altas por alguns clientes."

B. CORRELAÇÕES



C. RESUMOS USADOS PARA VALIDAÇÃO

Teste 1: Resumo composto por alucinações

"Os clientes demonstraram extrema satisfação com os canais digitais. O aplicativo é considerado fluido, completo e resolve todos os problemas dos clientes com extrema facilidade. O atendimento por whatsapp é frequentemente elogiado e o site e agência virtual também se destacam."

Teste 2: Resumo pouco informativo

"O feedback geral é extremamente negativo, sendo necessárias melhorias urgentes no acesso ao aplicativo, atendimento pelo whatsapp e funcionalidades do site."

Teste 3: Resumo bom, mas com imprecisões factuais

"Os clientes apresentaram comentários mistos. Exigem mais serviços nas lojas físicas, melhorias na agência virtual e site, além de maior facilidade para acessar o aplicativo. Também solicitam comunicação via whatsapp quando houver contas em atraso. Além disso, reclamam de tarifas caras de energia. O atendimento via whatsapp foi extremamente criticado"

Teste 4: resumo redundante

"O atendimento é bom, mas precisa melhorar o acesso ao aplicativo. Apesar de elogios ao atendimento, o aplicativo precisa de melhorias quanto ao login e funcionalidades."

Teste 5: Resumo com fuga do tema

"Os clientes estão insatisfeitos com o valor das tarifas e as formas de parcelamento disponíveis no site. Alguns elogiaram as equipes que realizaram a religação, mas criticaram a demora para finalizar os serviços."

Teste 6: Resumo gerado a partir de 500 comentários

"Os clientes têm opiniões diversas sobre a experiência com os canais digitais da empresa. Enquanto alguns elogiam a rapidez e praticidade no atendimento pelo WhatsApp e aplicativo, outros relatam dificuldades de acesso, problemas com a resolução de solicitações e falta de clareza nas informações fornecidas. É necessário melhorar a comunicação entre a central e as equipes de campo, assim como a qualidade e eficiência do atendimento digital para garantir uma experiência mais satisfatória para os clientes."

Teste 7: Resumo gerado a partir de prompt com poucas instruções

"Os clientes expressam opiniões mistas sobre o serviço. Eles acham o acesso à 2ª via no site e as tarifas caras, além de ressaltarem a necessidade de mais lojas físicas e correções de erros no site e no aplicativo. Observam também a demora no atendimento pelo WhatsApp, mencionando que, frequentemente, o serviço é encerrado antes que consigam completar o questionário. Apesar de tais problemas, o atendimento digital é elogiado por sua rapidez e eficiência, sendo mencionada a facilidade em acessar o aplicativo e a rápida

aprovação de pedidos. Os clientes sugerem a inclusão de notificações de contas em atraso via WhatsApp e o recurso de informar pagamentos de faturas pelo aplicativo. Além disso, indicam áreas de melhoria para o aplicativo, incluindo a correção de erros de entrada e a necessidade de tornar o acesso menos burocrático."

Prompt: "Resuma as informações dos comentários de clientes abaixo."

Teste 8: Resumo gerado a partir de prompt contextual

"Os clientes forneceram feedback variado sobre nossos canais digitais. Embora algumas pessoas achem o acesso a 2a via no site um pouco burocrático e identifiquem erros no site e aplicativo, muitos apreciam a rapidez com que nossos serviços digitais resolvem problemas. No entanto, foi apontada a demora e a limitação de nosso atendimento por WhatsApp, os clientes sugerem uma experiência mais eficiente e menos complicada. Outro ponto destacado é a necessidade de mais opções de serviços no aplicativo, como a capacidade de informar pagamentos de faturas e parcelar a fatura. Eles também sugerem melhorias na agência virtual para facilitar o acesso e diminuir a burocracia. Além disso, há um pedido para que eles sejam informados sobre contas em atraso via WhatsApp. Finalmente, alguns usuários encontram problemas ao tentar acessar nosso aplicativo, incluindo dificuldades para inserir o número de celular. Ainda assim, há um feedback positivo geral sobre o atendimento em nossos canais digitais, apesar das áreas de melhoria identificadas."

Prompt contextual: "você faz parte da equipe de experiência do cliente de uma grande empresa e precisa reportar o que os clientes dizem sobre os canais digitais da empresa. Faça isso em um parágrafo único."