Pontifícia Universidade Católica
do Rio de Janeiro

Bruno Frederico Maciel Gutierrez

Geração de descrições de produtos a partir de avaliações de usuários usando um LLM

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática da PUC-Rio.

Orientador : Prof. Hélio Côrtes Vieira Lopes
Coorientador: Dr. Fernando Alberto Correia dos Santos Júnior

Rio de Janeiro
abril de 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Bruno Frederico Maciel Gutierrez

Geração de descrições de produtos a partir de avaliações de usuários usando um LLM

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Prof. Hélio Côrtes Vieira Lopes
Orientador
Departamento de Informática – PUC-Rio

Dr. Fernando Alberto Correia dos Santos Júnior
Coorientador
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Prof. Bruno Feijó
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Prof. Marcos Kalinowski
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Dr. Jonatas dos Santos Grosman
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Rio de Janeiro, 12 de abril de 2024

Bruno Frederico Maciel Gutierrez

Graduou-se em Engenharia de Produção pela PUC-Rio. Fez mestrado no Departamento de Informática da PUC-Rio, especializando-se na área de Ciência de Dados.

Ficha Catalográfica

Gutierrez, Bruno Frederico Maciel

Geração de descrições de produtos a partir de avaliações de usuários usando um LLM / Bruno Frederico Maciel Gutierrez; orientador: Hélio Côrtes Vieira Lopes; coorientador: Fernando Alberto Correia dos Santos Júnior. – 2024.

96 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2024.

Inclui bibliografia

1. keywordpre – Teses. 2. keywordpre – Teses. 3. Geração de texto. 4. Mineração de informação. 5. Inteligência artificial generativa. 6. Large Language Model. 7. Aprendizado de máquina. 8. Comércio eletrônico. I. Côrtes Vieira Lopes, Hélio. II. Alberto Correia dos Santos Júnior, Fernando. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Agradecimentos

Agradeço ao meu orientador, Hélio Lopes, pela oportunidade e confiança durante o desenvolvimento dessa pesquisa. Um agradecimento muito especial ao Fernando e Jonatas pelo apoio e orientação ao longo de todo o processo. Sem dúvida nenhuma foram nas nossas trocas onde eu senti mais ter aprendido nessa experiência.

Agradeço à minha família, pelos ensinamentos e por seus sacrifícios. Mais do que palavras, eles me ensinam com o exemplo, e é o maior dos privilégios ter uma família assim. Agradeço também aos meus amigos, que tanto somam a vida.

À PUC-Rio, pelo ambiente tão acolhedor e pela qualidade acadêmica que tanto possibilita. Um agradecimento carinhoso a todos que participaram voluntariamente do trabalho. Sem dúvida, essa etapa foi fundamental para a sua conclusão.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Resumo

Gutierrez, Bruno Frederico Maciel; Côrtes Vieira Lopes, Hélio; Alberto Correia dos Santos Júnior, Fernando. **Geração de descrições de produtos a partir de avaliações de usuários usando um LLM**. Rio de Janeiro, 2024. 96p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

No contexto de comércio eletrônico, descrições de produtos exercem grande influência na experiência de compra. Descrições bem feitas devem idealmente informar um potencial consumidor sobre detalhes relevantes do produto, esclarecendo potenciais dúvidas e facilitando a compra. Gerar boas descrições, entretanto, é uma atividade custosa, que tradicionalmente exige esforço humano. Ao mesmo tempo, existe uma grande quantidade de produtos sendo lançados a cada dia. Nesse contexto, este trabalho apresenta uma nova metodologia para a geração automatizada de descrições de produtos, usando as avaliações deixadas por usuários como fonte de informações. O método proposto é composto por três etapas: (i) a extração de sentenças adequadas para uma descrição a partir das avaliações (ii) a seleção de sentenças dentre as candidatas (iii) a geração da descrição de produto a partir das sentenças selecionadas usando um *Large Language Model* (LLM) de forma *zero-shot*. Avaliamos a qualidade das descrições geradas pelo nosso método comparando-as com descrições de produto reais postadas pelos próprios anunciantes. Nessa avaliação, contamos com a colaboração de 30 avaliadores, e verificamos que nossas descrições são preferidas mais vezes do que as descrições originais, sendo consideradas mais informativas, legíveis e relevantes. Além disso, nessa mesma avaliação replicamos um método da literatura recente e executamos um teste estatístico comparando seus resultados com o nosso método, e dessa comparação verificamos que nosso método gera descrições mais informativas e preferidas no geral.

Palavras-chave

Geração de texto; Mineração de informação; Inteligência artificial generativa; Large Language Model; Aprendizado de máquina; Comércio eletrônico.

Abstract

Gutierrez, Bruno Frederico Maciel; Côrtes Vieira Lopes, Hélio (Advisor); Alberto Correia dos Santos Júnior, Fernando (Co-Advisor). **Product description generation from user reviews using a LLM**. Rio de Janeiro, 2024. 96p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In the context of e-commerce, product descriptions have a great influence on the shopping experience. Well-made descriptions should ideally inform a potential consumer about relevant product details, clarifying potential doubts and facilitating the purchase. Generating good descriptions, however, is a costly activity, which traditionally requires human effort. At the same time, there are a large number of products being launched every day. In this context, this work presents a new methodology for the automated generation of product descriptions, using reviews left by users as a source of information. The proposed method consists of three steps: (i) the extraction of suitable sentences for a description from the reviews (ii) the selection of sentences among the candidates (iii) the generation of the product description from the selected sentences using a *Large Language Model* (LLM) in a *zero-shot* way. We evaluate the quality of descriptions generated by our method by comparing them to real product descriptions posted by sellers themselves. In this evaluation, we had the collaboration of 30 evaluators, and we verified that our descriptions are preferred more often than the original descriptions, being considered more informative, readable and relevant. Furthermore, in this same evaluation we replicated a method from recent literature and performed a statistical test comparing its results with our method, and from this comparison we verified that our method generates more informative and preferred descriptions overall.

Keywords

Text generation; Data mining; Generative artificial intelligence; Large Language Model; Machine Learning; e-commerce.

Sumário

Lista de Figuras

Lista de Tabelas

1
Introdução

A geração de descrições de produtos de forma automática é uma prática que tem ganhado cada vez mais destaque no contexto do comércio eletrônico e varejo online. Na era digital, onde a competição é acirrada e a demanda por conteúdo é constante, a geração manual de descrições de produto costuma ser uma atividade custosa e em muitos casos proibitiva. Isso resulta, muitas vezes, em descrições de produto que oferecem pouco valor ao consumidor, contrastando, todavia, com o papel fundamental que as descrições de produtos desempenham na experiência de compra do consumidor.

Ao fornecer informações detalhadas sobre as características, funcionalidades e especificações dos produtos, as descrições permitem que os consumidores melhor compreendam o que estão adquirindo, inclusive respondendo dúvidas ao abordar questões comuns sobre o uso, tamanho, compatibilidade, entre outros aspectos importantes do produto. Assim, de forma geral, descrições de produto podem ajudar o consumidor a tomar decisões de compra mais informadas e alinhadas com suas necessidades e preferências. Nesse sentido, a automação na geração de descrições de produtos surge como uma solução inovadora para atender às necessidades tanto dos consumidores, interessados em descrições informativas sobre os aspectos que os interessam apresentadas na forma de um texto conciso e legível, quanto dos anunciantes, interessados em expor seus produtos em um ambiente digital de forma rápida e dinâmica.

Com os avanços de modelos de inteligência artificial, principalmente na área de processamento de linguagem natural, novas técnicas tem sido introduzidas com o objetivo de criar descrições detalhadas e atrativas para uma ampla gama de produtos de maneira eficiente e escalável. Essa prática não apenas atende às demandas do ambiente digital atual, mas também oferece oportunidades significativas para as empresas se destacarem e prosperarem em um cenário de negócios em constante evolução.

De fato, já encontramos no cenário atual muitas empresas de comércio eletrônico adotando ferramentas de geração automática de descrições de produtos. Um exemplo é a Amazon, que adotou recentemente uma ferramenta

desse tipo para auxiliar seus anunciantes[1].

Dentre as técnicas existentes, temos uma grande diversidade em relação a quais dados são utilizados para gerar as descrições de forma automática, combinando múltiplas informações, como atributos fornecidos pelo anunciantes (Wang et al., 2017), o título do produto (Zhan et al., 2021), e até slogans publicitários (Zhang et al., 2022) . No nosso trabalho, seguimos a linha proposta por Novgorodov et al. (2019) e utilizamos as avaliações deixadas por usuários do produto, entendendo que essa é uma fonte fonte rica de informações. Por conter uma perspectiva única da interação entre o usuário e produto, as avaliações contém experiências que podem enriquecer significativamente a compreensão de um produto, fornecendo percepções reais e autênticas que vão além das informações fornecidas pelo fabricante ou pelo varejista. Ao considerar produtos com abundâncias de comentários, teremos então um potencial enorme de encontrar informações relevantes sobre o produto e que sejam uma boa fonte para gerar uma descrição.

Afim de corroborar a riqueza de informações presente nas avaliações, notamos algumas novas ferramentas adotadas em plataformas de comércio eletrônico no sentido de reunir as informações contidas nas avaliações. Um exemplo de aplicação pode ser notado no site do Mercado Livre, que apresenta para produtos com muitos comentários um resumo das avaliações com o título "Resumo com base em opiniões de compradores". Afim de ilustrar essa aplicação, mostramos um resumo gerado para um tênis Converse, em que foi gerado "O tênis é amplamente elogiado por sua beleza e conforto, sendo uma escolha adorada por crianças e adultos. A qualidade do produto é frequentemente mencionada, assim como a satisfação dos que adquiriram o produto para presentear. Destacamse também o design e a cor como pontos positivos." [2]

Somado a isso, pretendemos potencializar o processo de geração de descrições utilizando avanços recentes na área de geração de texto em linguagem natural, especificamente os avanços introduzidos com os *Large Language Models* (LLM, do inglês, Grandes Modelos de Linguagem). Com esse novo paradigma de modelo, propomos no nosso trabalho um novo método, que combina a riqueza de informações contidas nas avaliações de usuários com a capacidade

[1] A descrição dessa ferramenta está disponível em: <`https://www.aboutamazon.com/news/small-business/amazon-sellers-generative-ai-tool`>, acessado pela última vez em 28/03/2024.

[2] Esse caso pode ser encontrado na seção de opiniões do produto no final da página aberta com o link `https://produto.mercadolivre.com.br/MLB-4194254180-tnis-all-star-cano-alto-2543-botinha-adulto-e-infantil-_JM#position=6&search_layout=grid&type=item&tracking_id=c8c5d388-51d2-4dc9-a62d-3370ade4030a`, acessado pela última vez em 28/03/2024. Ressaltamos que o erro de português na última frase fazia parte da descrição.

de síntese e articulação de um modelo de linguagem. Assim, definimos como a nossa questão principal de pesquisa como: *É possível enriquecer a geração descrições de produtos a partir de avaliações de usuário com o uso de um LLM?*

Em seguida, quebramos essa questão principal em duas subquestões mais específicas:

- *SQP1: Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?*
- *SQP2: Como se comparam as descrições geradas com as descrições originais?*

Para responder essas questões, executamos os seguintes passos. Primeiro revisamos a literatura no contexto de geração automatizada de descrições de produtos para obter compreender o estado da arte nesse tema. Em seguida, utilizamos o conhecimento obtido para desenhar um método que combine um LLM com as informações deixadas por usuários na forma de avaliações de produto. Assim, propomos um método composto por uma série de subetapas, em que nossa entrada é uma coleção de avaliações sobre um produto e nossa saída é uma descrição desse produto, que esperamos ser legível e informativa. Para definir cada uma das subetapas do nosso método, combinamos referências da literatura recente com uma série de experimentos, afim de encontrar boas soluções. Em seguida, avaliamos o nosso método, inclusive comparando as descrições geradas pelo nosso método com as postadas pelos anunciantes.

Baseados nos nossos resultamos, observamos que as nosso método é capaz de gerar descrições que são preferidas em relações as descrições originais em muitas dimensões, como legibilidade e informatividade. Além disso, replicamos um método (Novgorodov et al., 2019) encontrado na literatura, e o utilizamos como referencial para melhor entender os resultados do nosso método. Os resultados dessa comparação novamente indicaram a preferência dos avaliadores pelas descrições geradas pelo nosso método, indicando que o uso de um LLM dentro do processo de geração de descrição é muito benéfico. Dentre nossas principais contribuições, mostramos como o uso de LLMs possibilita uma linha de pesquisa interessante no contexto de geração de descrições de produtos, e propomos um novo formato de avaliação inteiramente reprodutível comparando as descrições de produto geradas com as descrições originais postadas pelos anunciantes. Além disso, nossa principal contribuição é sem dúvida o método em si.

O restante desse documento foi estruturado da seguinte forma. No capítulo 2 revisitamos alguns conceitos importantes tratados no trabalho. No Capítulo 3 conduzimos uma revisão do literatura sobre a geração de descrições

de produtos. Já no Capítulo 4 discutimos os dois conjuntos de dados utilizados nesse trabalho. No Capítulo 5 apresentamos o nosso método, oferecendo uma visão geral de cada uma de suas etapas. Já no Capítulo 6 discutimos em mais detalhes algumas subetapas do nosso método e os experimentos realizados para defini-las. Em seguida, no Capítulo 7 avaliamos as descrições geradas pelo nosso método, buscando avaliar como o LLM utilizado contribui para o processo. Por último, no capítulo 8 concluímos nosso trabalho destacando as principais contribuições e limitações do nosso método, e apontamos possíveis trabalhos futuros.

2
Fundamentação

Este capítulo apresenta dois dos principais conceitos que fundamentam nosso trabalho, e que são intrinsecamente relacionados, o de modelos pré-treinados e *Large Language Models* (Grandes Modelos de Linguagem). Nele, propomos então uma breve análise histórica, em que comentamos da evolução do uso de modelos pré-treinados, inicialmente no campo de visão computacional, para sua posterior popularização no campo de processamento de linguagem natural (PLN), culminando finalmente nos *Large Language Models* (LLM).

2.1
De modelos pré treinados até *Large Language Models*

Modelos pré-treinados são modelos de aprendizado de máquina (ML, do inglês *Machine Learning*) que foram treinado em um grande conjunto de dados e que podem ser ajustados para uma tarefa específica, como reconhecimento de imagem, processamento de linguagem natural ou reconhecimento de fala. A ideia por trás desses modelos é a de *transfer learning* (aprendizagem por transferência), em que o conhecimento adquirido no treinamento em uma tarefa, ou conjunto de dados, é transferido para uma outra tarefa diferente, mas relacionada.

Esses modelos são frequentemente usados como ponto de partida para o desenvolvimento de soluções complexas a partir do seu ajuste fino, que consiste na sua adaptação a tarefas em domínios específicos. Nesse sentido, permitem que desenvolvedores e pesquisadores aproveitem o conhecimento aprendido codificado na forma de parâmetros do modelo, ao mesmo tempo que podem ajustar o modelo para suas necessidades específicas com conjuntos de dados menores. Assim, esses modelos se tornaram ferramentas essenciais em diversas áreas, possibilitando novas soluções eficientes para atividades complexas mas sem exigir treinamento extenso do zero.

Inicialmente, a popularização dos modelos pré-treinados foi amplamente relacionada com o surgimentos de grandes conjuntos de dados anotados, como ImageNet (Deng et al., 2009), que é um conjunto voltado para tarefas de classificação de imagens. A partir desses grandes conjunto de dados, os pesquisados começaram a experimentar estratégias de pré-treinamento para

ajudar a inicializar redes neurais já com representações significativas dos dados de entrada, permitindo uma convergência mais rápida durante o ajuste fino subsequente em tarefas específicas com conjuntos de dados menores.

Nesse contexto, um dos primeiros avanços que impulsionou os modelos pré-treinados para o centro das atenções foi no campo de visão computacional, com o lançamento da arquitetura AlexNet (Krizhevsky et al., 2012). Treinado no conjunto de dados ImageNet, o AlexNet demonstrou desempenho significativamente melhorado na classificação de imagens em comparação com métodos anteriores. Este sucesso despertou o interesse em técnicas de pré-treinamento e lançou as bases para o desenvolvimento de modelos mais sofisticados.

Nos anos seguintes, a comunidade de visão computacional testemunhou uma proliferação de modelos pré-treinados em vários domínios. Modelos como VGG (Simonyan e Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), e ResNet (He et al., 2016), por exemplo, foram desenvolvidos para tarefas de reconhecimento de imagens. Contudo, ainda que o campo da visão computacional tenha se aproveitado inicialmente de grandes conjuntos de dados anotados para um aprendizado supervisionado, no campo do PLN foram avanços desenvolvidos com o uso do pré-treinamento semi-supervisionado que impulsionaram a área (Han et al., 2021).

Na medida que a anotação em larga escala de dados textuais é mais complexa que a anotação de imagens, devido as infinitas variações semânticas e nuances da linguagem, os primeiros avanços no contexto de modelos pré-treinados de PLN foram no sentido de representação das palavras, como no caso do *word2vec* (Mikolov et al., 2013). Nesse caso, grandes conjunto de textos não anotados eram utilizados de forma não supervisionada, afim de capturar a relação semântica entre as palavras próximas uma das outras. As representações da palavras geradas por esses modelos eram então utilizadas como entrada para outros modelos de arquiteturas de aprendizado profundo, como redes neurais recorrentes (RNN) e redes de memória de curto e longo prazo (LSTM). Essas abordagens iniciais foram importantes, na medida que demonstraram o potencial para gerar texto coerente e contextualmente relevante, embora com limitações na captura de dependências de longo alcance e na compreensão de estruturas linguísticas complexas.

Os avanços na geração de textos em linguagem natural foram acelerados então com a introdução da arquitetura *Transformer* (Vaswani et al., 2017), que são particularmente adequadas para lidar com dados sequenciais longos, como texto. Com esse novo paradigma, os modelos pré-treinados para tarefas de PLN entraram em um novo estágio, na medida que foi possível treinar modelos de linguagem mais profundos em comparação com os convencionais,

e assim aprender padrões e estruturas intrínsecas da linguagem humana.

Nessa linha, se baseando na arquitetura *Transformer*, o ano de 2018 marcou um divisor de águas com o lançamento do modelo GPT (Generative Pre-trained Transformer) (Radford et al., 2018) e do modelo BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). Pré-treinados em grandes quantidades de dados de texto, ambos os modelos demonstraram capacidades notáveis na compreensão e geração de linguagem natural, mostrando o potencial de modelos pré-treinados para tarefas relacionadas à linguagem. Além disso, diferente dos modelos iniciais pré-treinados no nível da representação de palavra que geravam representações para serem utilizadas como entrada para outros modelos, modelos baseados em *Trasnformer*, como os dois citados anteriormente, podem ser usados como a espinha dorsal de modelos em várias outras tarefas.

Desde então, o campo dos modelos pré-treinados continuou a avançar rapidamente. Pesquisadores e organizações lançaram modelos cada vez maiores e complexos, desenvolvendo novas variações de arquiteturas baseadas no paradigma *Trasnformer* e contendo uma quantidade cada vez maior de parâmetros, como o caso do modelo GPT-3 (Brown et al., 2020), que contém 175 bilhões de parâmetros. Na medida que esse modelos também são treinados em corpora de texto cada vez maiores, muitas vezes compreendendo bilhões ou até trilhões de palavras, para aprender padrões e estruturas intrínsecas da linguagem humana, esses modelos ampliaram os limites do desempenho em tarefas como compreensão, tradução e geração de idiomas, assim como permitiram uma nova gama de atividades. Assim, essa nova classe de grandes modelos de linguagem pré-treinados ficaram conhecidos como *Large Language Model*.

3
Trabalhos Relacionados

Este capítulo apresenta os trabalhos existentes na literatura que mais se relacionam com o nosso. Propomos uma revisão extensiva sobre a literatura relacionada ao tema específico de geração de descrições de produto de forma automática. Primeiro destacamos que essa é uma área relativamente recente, com o trabalho mais antigo aqui discutido sendo de 2017. Contudo, é uma área dinâmica, na medida que verificamos uma quantidade de trabalhos crescentes no tema, e que tem se beneficiado muito dos avanços recentes na área de geração de linguagem natural (GLN).

Assim, nos propomos a cobrir os trabalhos mais importantes publicados na área, de forma que vários foram lidos, mas apenas abordaremos em detalhes 8 trabalhos. Realizamos uma revisão histórica, em que ordenamos os trabalhos por ano de publicação. Para cada um dos trabalhos comentamos alguns do seus objetivos, ressaltamos o que os autores trouxeram de novidades e discutimos como funciona a sua solução. Além disso, focamos também nos métodos empregado pelos autores para avaliarem seus resultados, apontando as métricas e referencias que foram utilizadas em casos de comparações. Outro foco importante que buscamos explorar na hora de revisar os trabalhos é apontar algumas de suas limitações.

Como essa revisão influenciou fortemente o nosso trabalho, na medida que proporcionou ideias e inspirações para o nosso método, concluímos essa capitulo apresentado como esses trabalhos se relacionam com o nosso, no sentido de apresentar semelhanças e diferenças.

3.1
Geração de descrições de produtos

Começamos discutindo o trabalho de Wang et al. (2017), que propõe gerar descrições de produtos acuradas e fluentes a partir do uso de *templates* estatísticos com base em atributos do produto. Para isso, desenvolvem um sistema que aprende *templates* adequados no nível da sentença para cada atributo de um produto, isto é, aprendem um formato de sentença que é apropriado para cada atributo possível. Em seguida, experimentam para cada um dos atributos do produto os *templates* identificados, e geram as descrições a partir de todas

combinações possíveis de *templates*. Treinam então um modelo supervisionado, *SVM-rank* (Joachims, 2002), para ranquear as descrições candidatas usando a métrica *BLUE* entre a descrição candidata e a de referência como rótulo de cada descrição. Além de métricas quantitativas, como do *BLEU* e *top-k recall*, propõem avaliações qualitativas com anotadores humanos, e avaliam tanto as descrições humanas quanto as originais em múltiplas dimensões, como fluência e completude. Como algumas das limitação, ressaltamos que a abordagem proposta pelos autores é restrita aos produtos e atributos observados na base de treino, da qual os *templates* foram extraídos. Além do mais, pelo mesmo motivo, a estrutura do discurso das descrições geradas é bastante limitada.

Já Novgorodov et al. (2019) propõe uma abordagem extrativa a partir de uma nova fonte de informações para gerar as descrições: as avaliações deixadas por usuários. Conforme observado pelos autores, as avaliações deixadas por usuários são potencialmente uma fonte de informações muito rica e diversa, e que diferem dos atributos informados pelos anunciantes, utilizadas no caso trabalho anterior. Mais especificamente, os autores conduzem uma análise sobre como avaliações podem ser usadas para gerar descrições de produtos, examinando quais sentenças são adequadas para pertencer a uma descrição. Com esse objetivo, Novgorodov et al. (2019) propõe um *dataset* a partir da classificação de sentenças extraídas de avaliações e treinam um classificador para identificar sentenças adequadas a pertencerem a uma descrição de produto. Em posse dessas sentenças candidatas, os autores propõem uma abordagem de sumarização extrativa. Para isso, primeiro testam alguns métodos para ranquear quais sentenças são as mais interessantes e que de fato serão selecionadas. Com base nesse ranqueamento, os autores concatenam uma quantidade pré-definida de sentenças, experimentados descrições compostas por diferentes quantidades de sentenças, 3, 5 e 7. Como avaliação, utilizam uma escala *Likert* de 5 pontos, avaliando as descrições geradas em múltiplas dimensões. Contudo, ressaltamos que as descrições geradas são restritas, na medida que combinam uma quantidade de informação limitada a um número pequeno de sentenças. Ao mesmo tempo, pela natureza extrativa e devido a concatenação de sentenças, as descrições ficam sujeitas a problemas de legibilidade, algo observado pelos autores em suas avaliações.

Após a publicação de Novgorodov et al. (2019), os mesmos autores do trabalho o estendem com Novgorodov et al. (2020), adicionando um teste A/B em uma plataforma real de comércio eletrônico com as descrições geradas e mostrando resultados muito positivos. Selecionaram aleatoriamente 100 produtos e adicionaram em sua respectiva página web a descrição gerada sob o título de "*Descriptions from our costumers*" (tradução, descrições de nossos

clientes). Para cada produto, comparam o tráfego gerado nos dias anteriores com os 30 dias posteriores à adição da descrição. Conforme observado, o aumento do tráfego quando comparado com o grupo de controle (também composto por 100 produtos com os mesmos níveis de popularidade), foi estatisticamente significante, chegando a 23.8% de aumento do tráfego de produtos populares (produtos dentre o top 0.5% de acordo visualizações diárias da página do produto) da categorias *Fashion* contra 2.4% de aumento do respectivo grupo de controle.

Assim como no trabalho Novgorodov et al. (2019), Elad et al. (2019) propõe uma abordagem extrativa de sumarização, porém com o objetivo de gerar descrições de produtos personalizadas. Fazem isso predizendo a personalidade do usuário baseados em sua atividade em um site de comércio eletrônico, seguido de um algoritmo extrativo que seleciona quais sentenças devem ser usadas como parte da descrição. Contudo, usam como fonte de informação as próprias descrições originais para gerar descrições personalizadas e menores, com apenas três sentenças. Mais especificamente, os autores adotam o modelo *Big Five* de personalidades (John et al., 1999) e propõem identificar a personalidade do usuário, baseado em variáveis relacionadas ao seu comportamento na plataforma de comércio eletrônico, como número de compras, padrões temporais, categorias dos produtos comprados, preços, dentre outros. Em seguida, conduzem uma análise sobre as características linguísticas das descrições de produtos compradas por diferentes traços de personalidades, e baseados nesse estudo formulam um algoritmo que usa sumarização extrativa para selecionar sentenças devem compor uma descrição de três sentenças. No trabalho, os autores propõem avaliar os resultados pedindo a usuários que tiveram suas personalidades inferidas que comparem a descrição gerada com mais duas alternativas e selecionem a mais atraente. Dentre as alternativas, uma é construída utilizando os aspectos opostos da personalidade predita, enquanto outra é construída de forma neutra. Propõem uma avaliação usando então uma escala Likert de 5 pontos, e obtém resultados indicando a preferência dos usuários por descrições personalizadas.

Após avanços significativos na área de redes neurais e sequence-to-sequence Learning, Chen et al. (2019) estendem o *framework Transformer* (Vaswani et al., 2017) para a atividade de geração de descrição de produtos. Nesse sentido, geram descrições informativas e, assim como Elad et al. (2019), personalizadas, porém baseadas em categorias de usuários e não na personalidade. Desenvolvem então um modelo generativo batizado de KOBE. Diferentemente de Wang et al. (2017) e Novgorodov et al. (2019), combinam aspectos do produto com a categoria do usuário, segmentados em 24 catego-

rias com base em seus interesses, afim de gerar descrições personalizadas. Além disso, incorporam também uma base de conhecimento externa como entrada para o modelo. Por utilizar um modelo generativo, não se restringem na estrutura do discurso. Na etapa de avaliação, comparam o modelo como uma série de alternativas para verificar se o condicionamento do modelo aos atributos utilizados contribui para melhores resultados, concluindo que sim. Eles conduzem também uma avaliação qualitativa dos resultados em múltiplas dimensões, comparando com as descrições geradas pelo modelo comparativo *Transformer* que utiliza apenas o título do produto como entrada. Contudo, por novamente não haver nenhuma comparação com as descrições originais, não fica clara a real qualidade dos dados gerados. Além disso, o modelo proposto é limitado ao conjunto de aspectos do produto e categorias de usuário definidos pelos autores, de forma que seria necessário a reconstrução do *dataset* e o retreinamento do modelo conforme novos aspectos e categorias surgissem.

Em outra abordagem também com o *framework Transformer*, Zhan et al. (2021) propõem um novo modelo com foco no aspectos que interessam aos usuários. Nesse sentido, assim como Novgorodov et al. (2019), também usam as avaliações, mas não como fonte de informação e sim com o objetivo de incorporar na geração de texto os aspectos úteis aos usuários. Dessa forma, adicionam à arquitetura *Transformer* um módulo de destilação posterior (Hinton et al., 2015), introduzindo um novo modelo chamado Adaptative Posterior Distillation Transformer (APDT). Com esse módulo, os autores transferem à geração das descrições as preferências dos consumidores aprendidas com base nas avaliações, e usam como entrada do modelo o título e atributos do produto. Na etapa de avaliação, os autores comparam o modelo com diversos alternativas em métricas automáticas (*BLEU* (Papineni et al., 2002), *ROUGE-L* (Lin, 2004)) e com avaliações humanas, superando o modelo KOBE proposto por Chen et al. (2019) em todas essas métricas.

Todavia, novamente após avanços na área de geração de linguagem natural, dessa vez com o sucesso dos modelos generativos pré-treinados, Nguyen et al. (2021) propõem aplicar o modelo *Generative Pre-trained Transformer* (GPT-2) (Radford et al., 2019) na atividade de geração de descrições de produto. Com esse objetivo, os autores adaptam o modelo ao domínio de descrições de produto, utilizando a técnica de pré-treinamento adaptativo, e depois fazem mais uma etapa de ajuste fino do modelo condicionando-o a gerar uma descrição de produto ao receber a sua categoria, seu título, marca e atributos. Afim de cobrir vários aspectos do produto e mitigar problemas de dependências de longa distância, os autores propõem gerar múltiplas pequenas descrições de produtos, cada uma cobrindo um aspecto particular do produto, e obtém a

descrição final ao combinar as múltiplas descrições geradas.

Ao comparar os resultados com a alternativa *Transformer* em avaliações automáticas e manuais, Nguyen et al. (2021) verificam como o modelo pré-treinado precisa de muitos menos dados, além de generalizar muito melhor quando utilizado em outras categorias de produto. No entanto, como limitações, destacamos que os autores utilizam um conjunto muito restrito de atributos na etapa de condicionar o modelo. Além disso, propõem um modelo que precisa ser pré-treinado e depois ajustado, e que apresenta uma piora significativa na avaliação humana no contexto de generalização, ainda que melhor que a alternativa comparada.

Por último, apresentamos o método proposto por Zhang et al. (2022), que propõem um sistema para geração automática de descrições de produtos que ora utiliza uma rede *Trasnformer-pointer*, que é uma rede formada a partir da combinação de uma rede *Pointer-generator* (See et al., 2017) com um rede *Transformer* (Vaswani et al., 2017), e ora utiliza um modelo de linguagem pré-treinado, no caso de escassez de dados para produtos novos. Com base no título, pares de atributos e valores e um *slogan* publicitário, os autores querem gerar uma descrição de produto que apresente suas características e atraia interesse dos usuário, informando-os rapidamente sobre seus aspectos mais importantes do produto. Pretendem utilizar a descrição gerada de várias formas dentro do contexto da plataforma de comércio eletrônico, incluindo uma plataforma de recomendação de produtos e uma plataforma de transmissões ao vivo. Treinam seus modelo em um *dataset* proprietário, composto por informações extraídas de fontes diversas, como o título do produto, lista de atributos e trechos obtidos a partir de avaliações de usuários, e suas correspondentes descrições de produtos escritas por especialistas.

Para avaliar seus resultados Zhang et al. (2022) propõem comparar suas descrições geradas com dois modelos alternativos, sendo um deles o método baseado em *templates* proposto por Wang et al. (2017). Para isso, usam métricas quantitativas, como SacreBleu (Post, 2018), ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) e Meteor (Lavie et al., 2004), e também avaliações qualitativas, em que pedem para os anotadores ranquearem as descrições, e verificam que o modelo pré-treinado obtém os melhores resultados. No sistema proposto, após a geração de texto, o conteúdo gerado é filtrado, inclusive com intervenção humana, para garantir que apenas descrições de alta qualidade e informações corretas possam por ventura ser introduzidas em aplicações reais na plataforma. Assim como no trabalho de Novgorodov et al. (2020), relatam a implementação do sistema proposto na plataforma de comércio eletrônico[1],

[1] `https://global.jd.com`

também indicando resultados positivos, como a geração de 2,53 milhões de descrições de produto, e um aumento de 4.22% na taxa de cliques e de 3.61% na taxa de conversão dos produtos em um espaço de 6 meses, quando comparado com os resultados obtidos com técnicas utilizadas anteriormente na plataforma.

3.2 Comparação com o nosso método

Os trabalhos apresentados na seção anterior, foram aqueles que consideramos mais relevantes e nos quais baseamos nossa solução. Discutiremos nessa seção como o nosso trabalho se relaciona com esses trabalhos. Destacamos que os trabalhos que mais influenciaram o nosso método foram Novgorodov et al. (2019) e Nguyen et al. (2021), na medida que um nos inspirou a usar as avaliações de produto como fonte de informação e o outro nos inspirou a usar um LLM nesse processo.

Com base no trabalho de Novgorodov et al. (2019), adotamos a ideia de utilizar as avaliações deixadas por usuários como fonte de informação dos produtos. Para isso, adaptamos muitas de suas etapas, principalmente as etapas relacionadas a seleção e ranqueamento de sentenças extraídas de avaliações. Nos diferenciamos, contudo, na medida que adotamos uma abordagem abstrativa para combinar múltiplas sentenças e o título do produto em uma descrição de produto usando um LLM. Assim pretendemos mitigar os problemas de legibilidade identificados pelos autores, ao mesmo tempo que conseguimos utilizar mais sentenças para gerar descrições ainda mais informativas. Além disso, buscamos replicar na medida do possível o método proposto pelos autores, enriquecendo a nossa avaliação com a comparação das descrições geradas pelo seu método, o que nos permite, em certa medida, investigar como um LLM contribui para a atividade de geração de descrições de produto.

A ideia de usar um LLM, porém, foi influenciada pelo trabalho de Nguyen et al. (2021), que demonstrou o sucesso do modelo pré-treinado GPT-2 na comparação com a alternativa *Transformer*. Além disso, outra semelhança é que também utilizamos o título do produto como entrada do modelo. Uma diferença, porém, é que damos um uso diferentes para as avaliações, uma vez que os autores as utilizam para incorporar na geração de texto a preocupação com os aspectos úteis aos usuários, enquanto nós as utilizamos como fonte de informação parar gerar as descrições. Mais uma diferença relevante está relacionada a forma como usamos o LLM, uma vez que os autores optam por adaptar o modelo pré-treinado enquanto nós optamos por usá-lo de forma *zero-shot*, apenas calibrando o prompt. Assim, destacamos que a forma como usamos o LLM não é restrita a um *dataset*.

Destacamos ainda uma semelhança relevante com o trabalho de Wang et al. (2017) em relação a avaliação qualitativa dos resultados. Na nossa avaliação, também comparamos as descrições geradas pelo nosso método com as descrições originais postadas pelos anunciantes, diferente dos demais trabalhos que quando comparam suas descrições o fazem com descrições geradas por outros métodos.

Outra semelhança é com o trabalho de Chen et al. (2019), no sentido de adotar a ideia de utilizar conhecimento externo no processo de geração de descrição. Enquanto os autores fazem isso propondo utilizar uma base de dados, estabelecendo um mecanismo de busca de informações em uma base externa, no nosso trabalho potencializamos essa ideia ao utilizar um modelo pré-treinado, que foi exposto a uma quantidade de dados muito maior. Assim, esperamos que o nosso método seja capaz de utilizar o conhecimento prévio adquirido para gerar descrições ainda mais informativas.

Por último, destacamos as experiências positivas ao executarem seus métodos em aplicações reais nos trabalhos de Novgorodov et al. (2019) e Zhang et al. (2022) como exemplos motivadores na área.

4
Datasets

Neste capítulo apresentamos os conjuntos de dados utilizados no trabalho. Foram dois, o primeiro, o *Amazon Review Data* (Ni et al., 2019) é um *dataset* público[1] que serviu amplamente como base para o desenvolvimento da metodologia aqui proposta, fornecendo as avaliações deixadas por usuários e as descrições originais de produto usadas para comparação. Já o segundo *dataset* foi publicado por Novgorodov et al. (2019), que também está em domínio público[2] e foi utilizado especificamente para as etapas de classificação de sentenças adequadas para pertencer a uma descrição de produto (essa etapa é discutida em detalhes na metodologia, Capítulo 5). Nas seções desse capítulo apresentaremos uma breve análise exploratória sobre cada um dos *datasets*.

4.1
Amazon Review Data

O *Amazon Review Data* é um *dataset* que contém uma grande coleção de avaliações. Cada avaliação é um texto escrito livremente pelo próprio usuário relatando a sua experiência com o produto ou a compra, que vem acompanhado possivelmente de outras informações, como título da avaliação, nota de 0 a 5, imagens, identificação do produto e usuário e data da avaliação. A coleção contém 233 milhões de avaliações relativas a 15,5 milhões de produtos distribuídos em mais de 20 categorias. Elas foram coletadas pela Amazon em sua loja online durante o período entre maio de 1996 e outubro de 2018. Além disso, o *dataset* incluí metadados relativos a cada produto avaliado, como descrição do produto, informações da categoria, preço, marca, imagem, etc.

Uma vez que propomos a comparação com o método dos autores (Novgorodov et al., 2019), que será discutido no capítulo 7, optamos por trabalhar com a mesma categoria de produto. Dessa forma, seguimos com a categoria *"Clothing, Shoes and Jewlery"*, com um total de 32 milhões de avaliações e 2,7 milhões de produtos. Contudo, como o nosso trabalho não tem como objetivo

[1] *Dataset Amazon*: *dataset* público acessível no link <`https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/`>, último acesso em 11/03/2024.

[2] *Dataset Description Generation from avaliações*: *dataset* público acessível no link <`https://tdk.cs.technion.ac.il/research-files/description_generation_from_avaliaÃğÃµes.zip`>, último acesso em 11/03/2024.

uma análise exaustiva sobre essa coleção, reduzimos, à priore, esse recorte selecionando aleatoriamente uma coleção de 3 milhões de avaliações e 2 milhões de descrições de produtos.

Como será apresentado no capítulo 5, no nosso método propomos utilizar como unidade de análise as sentenças de um texto que descrevem avaliações, e não as avaliações em si. Nesse sentido, descrevemos a seguir os passos tomados afim de construir um novo *dataset*, desta vez centrado em sentenças.

4.1.1 Desagregação das avaliações em sentenças

Primeiramente, estabelecemos alguns critérios de exclusão de produtos e de avaliações baseado no tamanho do texto e ausência de informações. Descartamos produtos que não tivessem um texto de descrição, título ou pelo menos alguma imagem do produto. Essas informações serão necessárias para comparar as descrições geradas pelos nosso método com as descrições originais na etapa de avaliação, como será apresentado no Capítulo 7. Com isso, dos dois milhões de produtos, aproximadamente metade tinha alguma informação ausente, restando 1.051.958 produtos. Também descartamos avaliações que não tinham textos (3 mil) – exemplo, avaliações em que usuários atribuíram título e nota, mas não digitaram um texto – e também removemos duplicadas (560 mil). Ao fim dessa etapa, restamos portanto com 2.436.748 de avaliações de usuários.

Como produtos e avaliações são dados de tabelas diferentes, em seguida, agregamos essas duas informações vinculando avaliações e produtos a partir do id da cada produto (presente nas duas tabelas). Neste processo, foram descartados as avaliações não vinculados a produtos, restando 2.050.234 avaliações referentes a um total de 13.251 mil produtos.

Em seguida, quebramos os textos das avaliações em sentenças, utilizando a implementação da biblioteca NLTK[3]. Como resultado, obtivemos um conjunto de 5.997.555 sentenças com as informações de 13.251 produtos aos quais se referem. Para ilustrar o resultado dessa etapa, exibimos alguns dados descritivos do nosso *dataset* centrado em sentenças na Tabela 4.1. Conforme pode ser observado, temos uma média de 154,7 avaliações por produto, com cada avaliação contendo 2,9 sentenças. Cada sentença, por sua vez, contém 11,8 palavras.

Exploramos também a distribuição do número de sentença por avaliações, apresentada na Figura 4.1. Podemos observar como a grande maioria das

[3]Natural Language Toolkit: acessível no link <`https://www.nltk.org`>, último acesso em 12/03/2024.

Tabela 4.1: Características do *dataset* desagregado em Sentenças

	Média	Desvio Padrão	Mediana	Máximo
Avaliações por produto	154,7	479,9	27	9820
Sentenças por avaliação	2,9	2,8	2	144
Sentenças por produto	452,6	1264,7	95	21660
Palavras por sentença	11,8	8,8	10	459

avaliações possui 1 ou 2 sentenças, e como as avaliações não costumam passar de 5 sentenças, compreendendo 89% das avaliações.

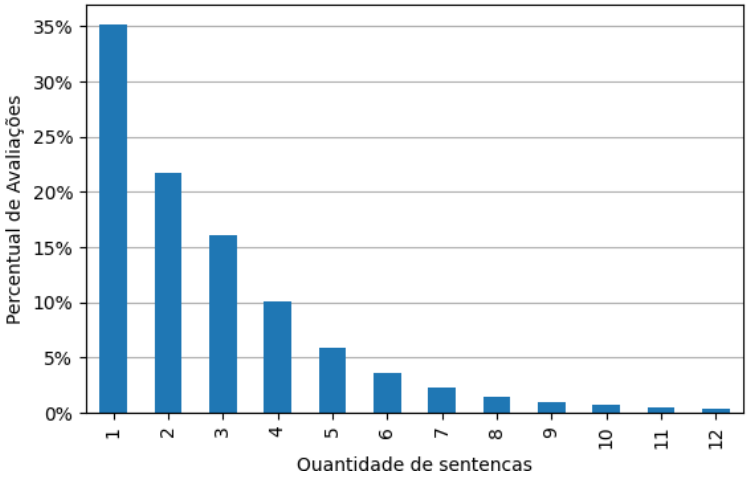


Figura 4.1: Distribuição da quantidade de sentenças por avaliação

Já na quantidade de palavras por sentenças, observamos uma distribuição bastante diferente (Figura 4.2), sendo o valor mais frequente o de 4 palavras em uma sentença. Verificamos que os 50% centrais das sentenças possuem entre 5 e 16 palavras.

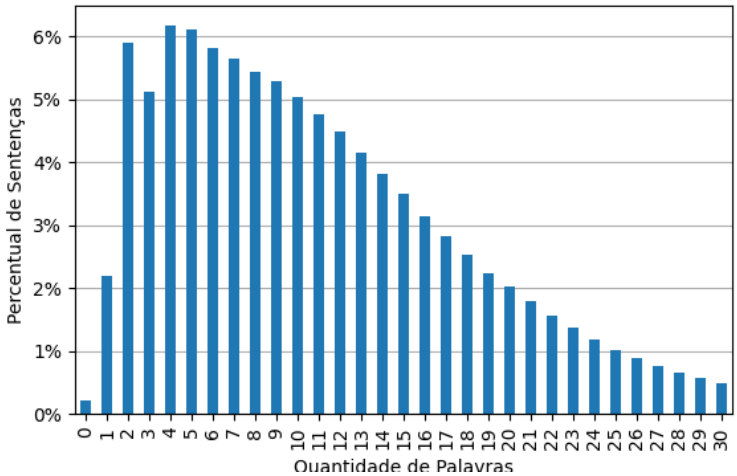


Figura 4.2: Distribuição da quantidade de palavras por sentença

Em relação às descrições originais dos produtos, referenciamos a análise conduzida na Seção 6.2, na qual destacamos algumas das características dos dados e definimos que tipo de descrições queremos gerar.

4.2 *Dataset* proposto por Novgorodov et al. (2019)

O segundo *dataset*, que vamos nos referir como *dataset* de sentenças adequadas, foi publicado no trabalho de Novgorodov et al. (2019), e contém sentenças extraídas de avaliações obtidas de uma grande plataforma de *e-commerece* estadunidense, referentes a duas categorias diferentes de produtos: *Fashion* (*"clothing, shoes and jewelry"*) e *Motors* (*"automotive parts and accessories"*). Ainda que não seja explícito o nome da plataforma de comércio eletrônico das quais as sentenças foram extraídas, suspeitamos fortemente se tratar de um subconjunto do *dataset* da Amazon citado acima, pelo fato do nome das categorias serem os mesmos e pela limite da janela temporal das avaliações (julho de 2018).

O *dataset* contém sentenças manualmente classificadas como sendo adequadas ou não a pertencer a uma descrição de produto. Para definir esse conceito, os autores conduziram uma análise pré-liminar sobre uma amostra de sentenças em que identificaram 10 motivos que tornam uma sentença inadequada para pertencer a uma revisão de produto. Já uma sentença adequada seria uma sentença que poderia ser inserida na sua forma original em uma descrição de produto, tendo como base a definição de descrição de produto proposta pelos próprios autores: "uma apresentação textual do que é o produto, como pode ser utilizado e por que vale a pena comprá-lo". Reforçam que o objetivo da descrição é fornecer aos clientes detalhes sobre os recursos e benefícios do produto para os incentivar a comprar.

Conforme mencionado, o *dataset* proposto por Novgorodov et al. (2019) se divide em duas categorias: *Fashion* e *Motors*, cada uma contendo 24.034 e 24.527 sentenças respectivamente. Além do rótulo, no caso da classe inadequada, a sentença é acompanhada de um motivo justificando a classificação, que podem ser um dos 10 apontados na Tabela 4.2, onde apresentamos as sentenças no seu formato original em inglês[4].

[4]Apresentamos aqui traduções livres de cada sentença: "Foi o início mais fácil de todos os tempos." "Caso contrário, permanece ocioso." "10 dólares por 3 pares é um ótimo negócio." "Esta camisa é ótima." "A configuração extremamente fácil permite que você obtenha o código do seu veículo rapidamente." "O único problema é o material bastante fino." "Provavelmente bom também para pneus de bicicleta." "O chapéu é exatamente como descrito." "Como outros aqui disseram, esta lata de gás tem um bico giratório longo." "Ótimo para Honda 2003 2.0L." "Produto fantástico, brilhante como m*rda"

Tabela 4.2: Lista de motivos que justificam a inadequação de sentenças acompanhados do seu percentual em cada conjunto de dados. Apresentamos as sentenças no seu formato original em inglês.

Motivo	*Fashion* (%)	*Motors* (%)	Exemplo
Subjtetivo	52,5	52,4	*It was the easiest jumpstart ever.*
Faltando contexto	16,9	16,8	*Otherwise it remains idle.*
Referência a um aspecto do anúncio	8,4	6,7	*10 bucks for 3 pairs is a great deal.*
Não informativo	7,9	6,4	*This shirt is great.*
Linguagem e ortografia ruins	4,9	5,2	*Extremely easy setup lets you pull your vehicle's code fast.*
Sentença Negativa	3,9	4,2	*Only issue is the pretty thin material.*
Expressa dúvida	2,4	2,3	*Probably good also for bicycle tires.*
Referência a descrição	1,8	1,7	*The hat is exactly as described.*
Outro	1,5	1,2	*Like others here have said, this gas can has along rotating nozzle.*
Muito específico/detalhado	0,6	2,5	*Great for Honda 2003 2.0L.*
Linguagem ofensiva	0,1	0,2	*Fantastic product, bright as sh*t.*

Uma vez que optamos por gerar descrições de produto da categoria *Fashion*, conforme discutido na Seção anterior, analisamos mais detalhadamente as sentenças do *dataset* nessa categoria. Apresentamos na Figura 4.3 a distribuição de sentenças pelo número de palavras, adicionando que uma sentença contém em média 10,1 palavras (6,3 de desvio padrão), e destacando que 75% dessas sentenças possuem 13 palavras ou menos.

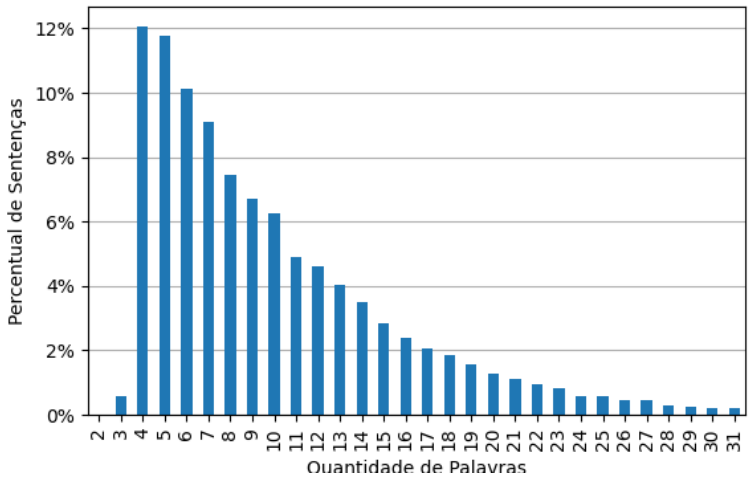


Figura 4.3: Distribuição da quantidade de palavras por sentenças da categoria *Fashion* no *dataset* de sentenças adequadas.

Para mais detalhes sobre a construção do *dataset*, recomendamos a consulta ao trabalho original (Novgorodov et al., 2019), que contém também uma análise detalhada da diferença entre sentenças adequadas e inadequadas em termos de unigramas e bigramas. O trabalho apresenta também dados sobre as sentenças positivas, como a sua distribuição em termos de quantidade de palavras, posição das sentenças nas avaliações originais, e tamanho das avaliações originais.

4.3 Comparação entre os dois *datasets*

Conforme discutido na Seção anterior, suspeitamos fortemente que o segundo *dataset* tenha sua origem no primeiro. De qualquer forma, mesmo que se trate de um subconjunto do mesmo *dataset*, propomos aqui uma pequena análise entre as semelhanças dos dois *datasets* mencionados acima na categoria *Fashion*.

Figura 4.4: Nuvem de Palavras do *dataset* de sentenças adequadas na categoria *Fashion*

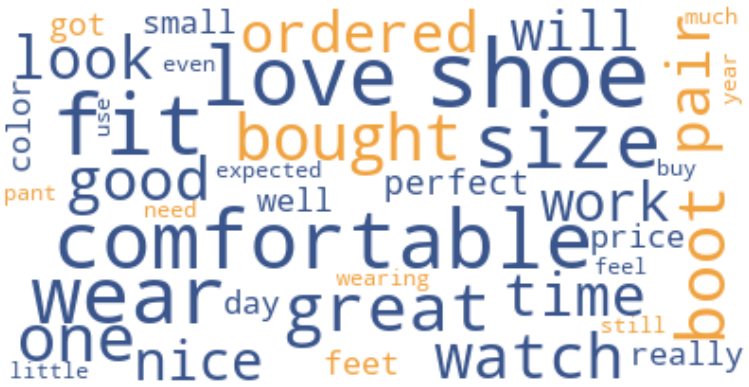


Figura 4.5: Nuvem de Palavras do *dataset* Amazon na categoria *Fashion*

Destacamos primeiro a nuvem de palavra de cada *dataset* nas Figuras 4.4 e 4.5, formada pelas 40 palavras mais frequentes. Colorimos as palavras comuns

as duas nuvens de azul e deixamos as demais em laranja. Verificamos que ambas nuvens são formadas por vários adjetivos, muito deles em comum, como *comfortable, well, nice, great* (confortável, bem, bom, ótimo) e *small* (pequeno). Além disso ambas as nuvens contém verbos como *wear, work* (vestir, trabalhar) e *fit* (ajustar), e substantivos indicando produtos como *shoe* (sapato) e *watch* (relógio).

Além da análise dos unigramas, exibimos também a Tabela 4.3 com os 20 bigramas mais frequentes de cada *dataset*. Destacamos que os *datasets* possuem 8 bigramas em comum (em negrito), com a maior parte incluindo a palavra "fit". Já no caso dos demais bigramas verificamos uma forte semelhança de tema e vocabulário, se tratando no geral de bigramas relacionados a experiência positiva do usuário.

Tabela 4.3: Comparação entre os 20 bigramas mais frequentes de cada *dataset*.

dataset	Bigramas mais frequentes
Amazon	*"another pair", "arch support", "first pair",* ***"fit great"****, "fit perfect",* ***"fit perfectly"****,* ***"fit well"****,* ***"good quality"****,* ***"great fit"****,* ***"great price"****, "half size", "highly recommend",* ***"looks great"****, "love shoes", "perfect fit", "second pair", "true size",* ***"well made"****, "would recommed"* e *"year old".*
Novgorodov et al. (2019)	*"easy assemble", "easy put", "easy use", "fit expected",* ***"fit great"****,* ***"fit perfectly"****,* ***"fit well"****, "fits well", "good fit", "good price",* ***"good quality"****,* ***"great fit"****,* ***"great, price"****, "great product", "great quality", "high quality", "light weight",* ***"looks great"****, "put together"* e ***"well "made"****.*

5
Metodologia

Neste capítulo apresentamos a metodologia desenvolvida no trabalho. Assim como Novgorodov et al. (2019), consideramos as avaliações de produto uma fonte rica de informações, e as utilizamos para gerar uma descrição.

Ao avaliar um produto, o usuário está explicitamente contando sua perspectiva enquanto consumidor, deliberadamente compartilhando uma experiência que possivelmente pode interessar outros usuários. Essas experiências podem enriquecer significativamente a compreensão de um produto, fornecendo percepções reais e autênticas que vão além das informações fornecidas pelo fabricante ou pelo varejista. Além disso, destacamos que as avaliações dos usuários oferecem uma perspectiva genuína sobre as características, qualidades e usos de um produto. Enquanto as descrições fornecidas pelo fabricante podem ser tendenciosas ou exageradas para promover as vendas, as opiniões dos consumidores refletem experiências reais e sem interesses econômicos por trás.

Ao considerar produtos com abundâncias de comentários, teremos então um potencial enorme de encontrar informações relevantes sobre o produto e que sejam uma boa fonte para gerar uma descrição. Assim, entendemos que ao usar informações de avaliações podemos ajudar os potenciais compradores a entenderem melhor como o produto se comporta na prática, suas vantagens e desvantagens, e se atende às suas expectativas e necessidades específicas.

Por esse motivo, nosso método de geração de descrições é baseado na seleção de sentenças relevantes extraídas de avaliações de usuários. Somado a isso, propomos combinar as informações contidas nas sentenças com uma etapa de geração de texto, e para isso precisamos articular esse conteúdo em uma descrição de produto. Com esse objetivo, propomos usar um LLM, contando com a sua capacidade de sumarização. Mais especificamente, esperamos que o modelo seja capaz de selecionar, agrupar e transmitir parte desse conteúdo em uma descrição que seja informativa, concisa e legível.

De forma geral, dividimos nosso método em três macro etapas, apresentadas no diagrama apresentado na Figura 5.1.

A base de nosso método é o conjunto de múltiplas avaliações de um produto. Entendemos que para gerar uma descrição de qualidade precisamos

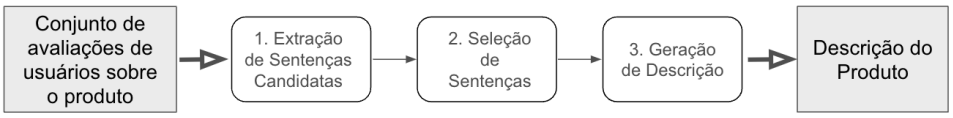


Figura 5.1: Visão geral do método proposto

de várias sentenças contendo uma quantidade razoável de informação sobre o produto, e isso implica em múltiplas avaliações refletindo diferentes pontos de vista e cobrindo diferentes características do produto. Contudo, entendemos também que nem toda informação contida nas avaliações é interessante para gerar uma descrição de produto, como observado por Novgorodov et al. (2019). De fato, por se tratar de um experiência pessoal, em muitos casos a avaliação pode trazer questões que sejam muito individuais, ou então específicas demais para interessar outros usuários. Assim, o primeiro passo do nosso método é separar as avaliações em sentenças, e separar as sentenças que podem ser interessantes para serem utilizadas no nosso método, chamadas de sentenças candidatas.

Em seguida, o segundo passo do nosso método é selecionar, dentro de conjunto de sentenças candidatas, quais de fato serão utilizadas. Para isso, nessa etapa nos propomos a selecionar as sentenças mais importantes dentre as candidatas, assim como garantir que essas sentenças sejam diversas entre si. Na terceira e última etapa, utilizamos então um LLM para gerar as descrições de produtos a partir das sentenças selecionadas na etapa anterior. Para fazer isso, utilizamos o modelo de forma *zero-shot*, fornecendo as sentenças selecionadas como parte do *prompt* junto com uma instrução de gerar uma descrição de produto. Com isso, a saída do nosso método é uma descrição de produto, que reúne as perspectivas únicas dos usuários deixadas nas avaliações em uma descrição de produto.

Conforme discutido, nos baseamos parcialmente no método proposto por Novgorodov et al. (2019), replicando algumas de suas soluções propostas e adaptando alguns outros pontos. Além disso, somamos uma nova etapa de geração de descrição, que envolve um LLM. Nesse sentido, optamos por separar a apresentação do método proposto, discutida nesse capítulo, dos experimentos e estudos conduzidos para definir algumas das etapas específicas da solução proposta, que serão apresentados no capítulo 6.

Dividimos esse capítulo abordando cada uma das etapas exibidas no diagrama. Entramos em mais detalhes sobre a extração de sentenças candidatas na Seção 5.1, enquanto mais detalhes sobre a seleção de sentenças são abordados na Seção 5.2. Na Seção 5.3 discutimos a questão da geração de descrição, comentando sobre a escolha de modelo e como utilizá-lo. Por último, apresen-

tamos uma visão em mais detalhes do método proposto, discutindo também quais configurações foram ajustadas no capítulo seguinte.

5.1 Primeira Etapa: Extração de Sentenças Candidatas

A extração de sentenças candidatas é uma etapa que se divide em duas subetapas, recebendo como entrada as avaliações quebradas em sentenças relativas a um determinado produto e gerando um conjunto de sentenças candidatas para serem utilizadas no processo de geração de uma descrição de produto.

Baseamos nossa solução na abordagem proposta por Novgorodov et al. (2019). Assim, a primeira subetapa é a aplicação de 3 filtros propostos pelos autores para identificar sentenças que não são interessantes para uma descrição de produto. Nesse trabalho, os autores conduzem uma análise sobre as diferenças linguísticas entre descrições e avaliações de usuários, notando serem textos que apresentam características contrastantes. Enquanto as avaliações refletem opiniões subjetivas sobre uma experiência individual, espera-se que descrições sejam mais objetivas, explicando o que o produto é e porque comprá-lo. Na sua análise, os autores usam uma técnica chamada divergência de Kullback-Leibler, que é uma medida de distância não simétrica entre duas distribuições (Berger e Lafferty, 2017), para destacar os termos que contribuem para a divergência entre a linguagem das avaliações contra a linguagem das descrições, e vice-versa.

Com base nessas diferenças, os autores propuseram os seguintes filtros, apresentados abaixo. Além dos filtros, também comentamos para cada um deles o percentual de sentenças filtradas ao aplicá-los no *dataset* de sentenças discutido na Seção 4.1.1, que é o que usamos para gerar as descrições.

1 **Sentenças pequenas**. A primeira regra proposta na seleção de sentenças é o filtro de sentenças pequenas (até 3 palavras). Conforme observado, sentenças pequenas são no geral pouco informativas. No *dataset* trabalhado, 13% das sentenças foram identificadas como possuindo 3 palavras ou menos. Exemplos de sentenças que se adéquam a essa regra são "*Great choice!*", "*Perfect*" e "*work as expected*".

2 **Sentenças pessoais**. A segunda regra proposta é a de filtragem de sentenças pessoais. Nesse sentido, os autores propõem a identificação de sentenças pessoais pela presença de unigramas como pronomes de primeira pessoa ou pronomes pessoas de terceira pessoa. No geral, 57% das sentenças foram classificadas como pessoais. Exemplos de sentenças

são "*I absolutely love it*", "*it suits all my needs*" e "*My brother can't stop using.*"

3 **sentenças relacionadas ao anúncio**. A terceira regra proposta é a filtragem de sentenças relacionadas ao anúncio do produto, em vez do produto em si. Essa filtragem também foi realizada pela identificação de unigramas de um conjunto que inclui sentenças que se referem a aspectos como entrega do produto, preço ou atendimento do vendedor. Com isso, propõe-se eliminar sentenças que se referem a outros aspectos da experiência de compra e não ao produto em si. No caso, 5% das sentenças foram identificadas como relacionadas ao anúncio. Exemplos são "*Great price for value*" e '*This is definitely a bargain*'.

No total, ao aplicar simultaneamente as 3 regras no *dataset* de sentenças foram filtradas 71% das sentenças. Após o filtro inicial, o passo seguinte na extração de sentenças candidatas é a classificação das sentenças filtradas. Nessa etapa, utilizamos um classificador para distinguir sentenças como sendo adequadas ou não a pertencer a uma descrição de produto, que é um conceito proposto pelos autores. Sentenças adequadas são, na visão dos autores, sentenças que descrevem o que o produto é, como pode ser utilizado, ou porque deve ser comprado, sendo sentenças que na sua forma original poderiam ser incluídas em uma descrição de produto. Já sentenças inadequadas são sentenças que não se encaixam nessa definição, sendo que os autores identificam 10 motivos que desqualificam sentenças, conforme discutido em 4.2. Ao usar esse classificador, portanto, estamos selecionando sentenças ricas em informações sobre o produto em questão.

Todavia, como não foram disponibilizados os modelos treinados pelos autores, mas sim o conjunto de dados (discutido em detalhes em 4.2), tivemos que treinar alguns classificadores próprios. Assim, optamos por utilizar a mesma métrica que a usada pelos autores, a área debaixo da curva ROC (AUC) (Bradley, 1997) e definimos 5 modelos candidatos para serem treinado na atividade binária de classificar uma sentença como sendo adequada ou não para uma descrição de produto. Destacamos que um dos modelos obteve uma performance maior que a encontrada pelos autores, considerada então satisfatória. Tratamos da discussão sobre os classificadores experimentados, resultados encontrados e uma discussão sobre generalização e métrica na Seção 6.1.

5.2 Segunda Etapa: Seleção de Sentenças

Uma vez classificadas, temos então um conjunto de sentenças adequadas para compor uma descrição, e a segunda etapa do nosso método é selecionar quais dessas sentenças serão utilizadas. Para fazer isso, dividimos essa etapa em três subetapas, e a saída dessa etapa é justamente um subconjunto das sentenças mais relevantes e diversas.

De fato, ainda que todas sentenças sejam adequadas para pertencer a uma descrição, as sentenças extraídas são variadas, cobrindo diferentes aspectos do produto e trazendo diferentes informações. Para ilustrar a ideia, exibimos algumas das sentenças extraídas de produto na Tabela 5.1, na sua forma original em inglês[1], acompanhada de uma foto do respectivo produto na Figura 5.2.

Tabela 5.1: Sentenças Extraídas de avaliações de um tênis Converse

	Sentenças
1	*Laced through metal loops for a durable sneaker.*
2	*Brilliant colors and finely made.*
3	*Canvas material is very durable.*
4	*Great all around shoe for athletics and motorcycle riding.*
5	*The shoe fabric has a rich texture and feel.*
6	*Good for everyday wear, not athletics.*
7	*Wash and tumble dry makes cleaning a breeze.*
8	*Shoes fit like a glove.*
9	*The pair fits like a glove!*

Conforme pode ser observado, as sentenças cobrem uma ampla gama de conteúdos, falando sobre a lavagem do tênis na sentença 8 ou tratando de usos específicos do produto na sentença 4, ao falar que é um ótimo tênis para "andar de moto" (*motorcycle riding*). Além disso, é interessante notar que as sentenças podem conter opiniões contrastantes, como a 4 e a 6. Por último, observamos também sentenças com mensagens bem semelhantes, como a 8 e 9. Por isso, seguimos a linha proposta por Novgorodov et al. (2019) para selecionar as sentenças que usaremos para gerar as descrições, baseando nossa escolha no ranqueamento e diversificação das sentenças. Para isso, a primeira decisão tomada foi no sentido de qual representação vetorial utilizar, de modo que optamos por utilizar forma de representar as sentenças diferente da que os autores utilizaram.

[1] Apresentamos aqui traduções livres de cada sentença:"Atado através de presilhas de metal para um tênis durável., "Cores brilhantes e finamente feitas", "O material da lona é muito durável.", "Ótimo sapato versátil para atletismo e motociclismo.", "O tecido do sapato tem uma textura e toque ricos", "Bom para uso diário, não para atletismo.", "Ótimo sapato versátil para atletismo e passeios de motocicleta.", "Lavar e secar torna a limpeza muito fácil.", "Os sapatos cabem como uma luva.", "O par cai como uma luva!"

Figura 5.2: Exemplo de produto: um tênis da Converse.

No seu trabalho, (Novgorodov et al., 2019) experimentam 3 técnicas diferentes de representação das sentenças e obtiveram os melhores resultados treinando um modelo *word2vec* Mikolov et al. (2013) em um *dataset* de 10 milhões de avaliações de usuários específicas ao domínio do produto. Contudo, frente a rápidos avanços na área de representação textual, e pensando na generalização do nosso método, buscamos um outra forma de representar os textos que não fosse específica a um domínio apenas.

Adotamos um modelo *Sentence-Transformer*, descrito inicialmente em Reimers e Gurevych (2019), para representar vetorialmente as sentenças. No trabalho, os autores propõem uma variação da arquitetura pré-treinada BERT (Devlin et al., 2018) para derivar representações vetoriais das sentenças que possam ser comparadas usando similaridade de cosseno de forma de forma muito mais eficiente, e superam outros modelos do estado da arte de representação de sentenças.

Mais especificamente, adotamos o modelo *open-source all-mpnet-base-v2* [2]. Baseado no MPNET Song et al. (2020), esse modelo foi desenvolvido a partir do seu ajuste fino em um *dataset* de 1 bilhão de pares de sentenças com um objetivo de aprendizado contrastivo: dada uma sentença, o modelo deve prever qual, de um conjunto de outras sentenças amostradas aleatoriamente, é a sentença originalmente emparelhada com ela no conjunto de dados.

Para escolher esse modelo, nos baseamos nos resultados do *benchmark "Massive Text Embedding Benchmark"*, publicado por Muennighoff et al. (2022). Esse *benchmark* compreende um total de 8 atividade relacionadas a

[2]Mais informações em: <`https://huggingface.co/sentence-transformers/all-mpnet-base-v2`>

representação de texto, cobrindo um total de 58 *datasets* e avaliando um total de 33 modelos. Umas das atividades compreendida é a de *"Semantic Textual Similarity"* (STS), em que o objetivo é determinar a similaridade entre um par de sentenças, sendo o modelo desenvolvido a partir do ajuste fino do MPNET é um dos que obtém maior performance nos 10 *datasets* avaliados. Somado a isso, após avaliar extensamente os modelos no *benchmark*, os autores recomendam utilizar o modelo MPNET ajustado quando deseja-se combinar velocidade e performance, justificando nossa escolha.

Após selecionada a forma de representação vetorial das sentenças, seguimos a abordagem em Novgorodov et al. (2019) e replicamos uma subetapa de ranqueamento das sentenças mais importantes. Aqui a ideia é que algumas sentenças são mais interessantes que outras, sejam pelo aspecto do produto que abordam ou pela riqueza de detalhes que trazem, conforme exemplificado na Figura 5.2. Para o ranqueamento, adotamos o método apontado pelo autores como o o que teve melhor performance dentre os experimentados, isto é, o LexRank Erkan e Radev (2004). No caso, o LexRank é um método que atribui um *score* de importância para cada sentença com base no conjunto como um todo, usando passeios aleatórios e auto-centralidade vetorial. A partir disso, temos portanto uma lista ordenada das sentenças mais importantes.

Como terceira subetapa da seleção de sentenças, seguimos a abordagem dos autores de diversificação das sentenças, utilizando a similaridade de cosseno para identificar sentenças similares. Assim, adotamos um valor de similaridade máxima, que é o mesmo que o reportado pelos autores ao conduzir uma análise sobre as descrições originais. Nessa análise, os autores mediram o valor de similaridade no 90º percentil de um subconjunto de descrições de produtos postadas por anunciantes e curadas por especialistas no domínio, e definiram então que esse seria o valor máximo admitido para que duas sentenças não fossem consideraras similares. O valor reportado pelos autores, e que nós igualmente adotamos como limiar, foi de 0,73. Com base nessa valor, seguimos uma abordagem gulosa, em que vamos adicionando as sentenças pela ordem definida na subetapa anterior caso ela não seja similar à nenhuma das adicionadas anteriormente.

5.3 Terceira Etapa: Geração de Descrição de Produto com LLM

A ultima etapa do nosso método, portanto, é a geração de uma descrição de produto a partir das sentenças selecionadas na etapa anterior, que são sentenças adequadas para pertencerem a uma descrição de produto, selecionadas com base na sua relevância e diversas entre si. Com esse objetivo, propomos a

geração de descrições de produto a partir da sumarização das sentenças selecionadas usando um LLM.

Sendo o estado da arte em uma série de tarefas, os modelo de linguagem pré-treinados revolucionaram uma série de atividades na área de linguagem natural e potencializaram novas aplicações. Uma das atividades revolucionada foi a de sumarização, surgindo aplicações em áreas diversas, como na área médica (Kieuvongngam et al., 2020), de notícias (Goyal et al., 2022) e sumarização de opiniões (Bhaskar et al., 2022), dentre alguns exemplos. Nesse sentido, no nosso método usamos um LLM de forma *zero-shot*, instruindo o modelo a gerar uma descrição de produto a partir das sentenças selecionados.

Na sua terceira iteração com o modelo GPT (*"Generative Pre-training Transformer"*), iniciada em 2018 com o trabalho (Radford et al., 2018), a empresa OpenAI desenvolveu o modelo GPT-3, que é um dos maiores LLM desenvolvido até o momento, com um total de 175 bilhões de parâmetros, e performa bem em uma ampla gama de atividades, principalmente no contexto de *zero-shot* ou *few-shot* (Brown et al., 2020). No nosso trabalho, optamos por utilizar o modelo *"gpt-3.5-turbo-0613"*[3], versão mais recente dos modelos da família GPT no momento de desenvolvimento desse trabalho. Com um janela de contexto de 4.096 tokens, o modelo foi treinado com dados de até setembro de 2021.

Por ser um modelo proprietário da empresa, o uso do modelo impõe um custo por uso, que é feito via a API da OpenAI. Contudo, ressaltamos que essa é uma forma rápida e prática de utilizar um LLM, uma vez que não exige custos de infraestrutura e bastante simples em termos de interface. Por isso, julgamos essa escolha adequada para o nosso método. Em relação a hiper-parâmetros, optamos por utilizar os valores padrão.

Ressaltamos que após o desenvolvimento desse trabalho, uma nova versão do modelo foi lançada, GPT-4 Achiam et al. (2023). Contudo, existe uma diferença significativa de custo no uso dos modelos. Enquanto o modelo *"gpt-3.5-turbo-0613"* tem um custo de 2 dólares por milhão de *tokens* gerados, o do modelo gpt-4 é de 60 dólares por milhão de *tokens* gerados[4], tornando sua aplicação em muitos casos proibitiva.

No nosso trabalho, optamos por usar o modelo de forma *zero-shot*, por considerar essa uma abordagem com muitos benefícios. A primeira é em relação a custo, uma vez que qualquer abordagem de ajuste fino pressupõe uma etapa de treinamento. O segundo benefício é o de não exigir uma coleção de dados

[3]Mais informações em: <`https://platform.openai.com/docs/models/gpt-3-5-turbo`>

[4]Para mais informações sobre os preço do modelo, referimos o site da OpenAI <`https://openai.com/pricing`>

anotados. E o terceiro, e principal, é de tornar a nossa abordagem generalizável para outros domínios: uma vez que não estamos treinando o modelo gerador de nenhuma forma, temos todo motivo para pressupor que ele também seria capaz de gerar descrições para outros tipos de produtos além do domínio experimentado. Dito isso, um desafio imediato que surge é a definição do *prompt* para geração de descrições.

Propomos pensar no *prompt* como a combinação de uma instrução e de um conteúdo. A instrução é a função que atribuímos ao modelo, no nosso caso caso gerar uma descrição de produto. Já o conteúdo é o texto fornecido como contexto para o modelo. No nosso caso, o conteúdo inclui justamente as sentenças selecionadas para servirem de base para a descrição. Além disso, consideramos interessante incluir também o título do produto, no sentido de fornecer mais informações para o modelo pré-treinado e deixar claro a que as sentenças se referem. Feita essa importante separação, nos debruçamos na questão do *prompt* examinando primeiro a questão da instrução.

Ao pensar em qual instrução fornecer ao modelo, nos deparamos com a questão básica de de que tipo de descrições queremos gerar. Conforme discutido anteriormente, baseamos nosso método na riqueza de informações contidas nas avaliações, de modo que queremos gerar descrições que reflitam essa riqueza. Contudo, entendemos também que descrições mais informativas serão mais longas, exigindo portando um custo maior do leitor, de forma que o tamanho das sentenças é uma questão relevante. Assim, para examinar essa questão, analisamos as descrições originais postadas pelos anunciantes no *dataset* da Seção 4.1. Nesse estudo (que será explicado em mais detalhes na Seção 6.2), constatamos a diversidade das descrições no que se refere ao seu tamanho, e optamos então por definir um tamanho máximo para as descrições que respeitasse o observado nas descrições postadas pelos anunciantes. Como consequência desse estudo, incluímos como um pré-requisito da instrução a necessidade de impor um limite de máximo de palavras nas descrições geradas.

Estabelecido esse pré-requisito, começamos a busca por uma instrução para o modelo. Entendemos que essa é uma parte fundamental do nosso processo de geração de descrição, ao mesmo tempo que é um grande desafio. Por haver uma infinidade de instruções disponíveis, e pela avaliação de textos no contexto de LLMs também ser um grande desafio, que será discutido na Seção 7, optamos por adotar um processo de melhoria contínua na etapa de definição da instrução, entendendo também que essa não é uma questão definitiva e que sempre pode ser revisitada.

Nesse processo de escolha da instrução, conduzimos alguns experimentos. Nesse sentido, o primeiro passo foi definir um conjunto de instruções candida-

tas, e em um segundo momento exploramos a questão de incorporar a limitação de palavras discutida anteriormente, buscando variações dessas instruções que respeitassem o limite máximo definido.

Como resultado desse experimento, chegamos em mais um conjunto de candidatos, e conduzimos então um segundo experimento no qual avaliamos qualitativamente as descrições geradas por cada instrução, chegando então a uma instrução final. Explicamos em detalhes os experimentos conduzidos para selecionar a instrução na Seção 6.3, assim como algumas das limitações do LLM que foram observadas nesse processo.

Por último, abordamos a questão de quantas sentenças incluir no processo de geração da descrição. Essa etapa tem um efeito direto na definição do conteúdo, ou seja, vamos definir quantas sentenças incluir. Conforme discutido anteriormente, a ideia do nosso método é potencializar a informatividade das descrições a partir do uso das informações presentes nas avaliações. Assim, em certa medida, ao usar mais sentenças esperamos obter descrições mais informativas. Para confirmar essa questão, realizamos mais um experimento, em que geramos descrições de produtos variando a quantidade de sentenças fornecidas, e em seguida comparamos essas descrições entre si com avaliadores humanos. Dessa comparação, escolhemos então o cenário que gerou as descrições preferidas pelos anotadores, fixando portanto a quantidade de sentenças a ser utilizada. Novamente ampliamos a discussão e explicamos em detalhes o experimento conduzido na Seção 6.3.

5.4 Visão do método em mais detalhes

Uma vez enumeradas os processos que compõem cada uma das etapa da metodologia, podemos construir uma visão mais detalhada do nosso processo. Para isso, apresentamos uma visão geral do método proposto com cada uma de suas subetapas detalhadas na Figura 5.3. Contudo, conforme comentado nas seções anteriores, algumas das subetapas do nosso método precisam ainda ser configuradas via experimentos e avaliações. Assim, incluímos na visão detalhada do nosso método as questões que precisam ser investigas para melhor definir cada uma das subetapas em aberto. Deixamos essas questões, assim com as subetapas que dela dependem em negrito.

Conforme pode ser observado, na primeira etapa do nosso método, de extração de sentenças candidatas trataremos ainda da definição de um classificador que distingue as sentenças adequadas das não adequadas. Já na terceira etapa do método, de geração da descrição usando o modelo GPT3-5 turbo, trataremos de 3 questões, ambas relacionadas diretamente a

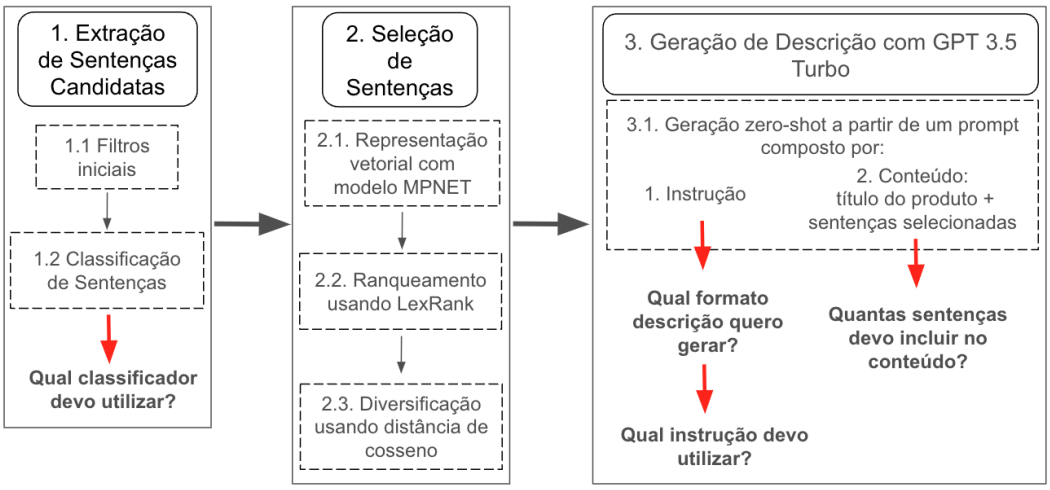


Figura 5.3: Visão detalhada do método proposto com subetapas a serem definidas.

construção do *prompt*. A primeira das questões tem impacto na definição da instrução do modelo, no sentido de definir qual o formato de descrições que queremos gerar. Já a segunda questão é a definição da instrução em si. Por último, abordamos a questão da quantidade de sentenças a ser passada para o modelo, definindo portanto o conteúdo do *prompt*. Investigamos a definição das subetapas deixadas em aberto no capítulo 6

6
Definição das subetapas com experimentos

Neste capítulo investigamos as subetapas deixadas em aberto na metodologia afim de definir a nossa solução. Para isso, conduzimos uma série de experimentos afim de buscar soluções para cada uma.

O primeiro experimento discutido se refere a etapa inicial de extração de sentenças candidatas. Discutimos aqui o treinamento de um classificador para ser utilizado nessa primeira etapa, de seleção de sentenças adequadas a pertencerem a uma descrição de produto. Experimentamos um total de 5 classificadores, variando desde modelos tradicionais até fazer o ajuste fino de um LLM. Destacamos que nosso melhor modelo obtém uma performance melhor do que a reportada pelos autores, e discutimos a generalização dos modelos para outros domínios além dos dados rotulados (outras categorias de produtos), parecendo obter resultados positivos. Outro ponto abordado é uma discussão sobre a métrica e como a questão de falsos positivos e falsos negativos se aplica no nosso método. Entramos em mais detalhes sobre a classificação de sentenças na Seção 6.1.

Em seguida, tratamos da terceira etapa do nosso método, de geração de descrições de produto com um LLM. Nessa etapa, fazemos uma distinção entre instrução e conteúdo, em que no nosso caso a instrução é a função que atribuímos ao modelo (gerar descrições) e o conteúdo é o título do produto junto com as sentenças selecionadas nos quais o modelo deve se basear. Assim, primeiro abordamos a questão da instrução.

Começamos abordando a questão do formato das descrições que queremos gerar, o que tem um impacto direto na busca por uma instrução para o modelo. Para explorar esse questão, apresentamos um estudo na Seção 6.2 sobre as descrições originais postadas pelos anunciante do *dataset* Amazon (descrito na Seção 4.1). Nesse estudo, verificamos primeiro que as descrições possuem alta variabilidade em relação a quantidade de palavras. Optamos então por controlar a quantidade de palavras geradas em nossas descrições, e escolhemos adotar um limite baseado na distribuição observada. Definimos portanto essa questão como um pré-requisito para a instrução.

Depois, na Seção 6.3, apresentamos o experimento conduzido para de fato definir a instrução do modelo, em que propomos um processo de melhora

contínua a partir de um conjunto de instruções candidatas. Nesse processo, incorporamos a questão do limite de palavras discutida anteriormente, adicionando às instruções candidatas a característica de impor um máximo de palavras. Interessantemente, verificamos que o LLM desrespeita esse limite, e conduzimos um experimento para encontrar uma variação da instrução que contorne esse problema. Por último, para definir quais dos candidatos utilizar, conduzimos uma avaliação qualitativa das descrições geradas por cada instrução, comparando pares entre si. Após apresentar os resultados, terminamos essa seção discutimos algumas das limitações do LLM observadas no que tange a geração de descrições.

Por último, exploramos a questão do conteúdo, investigando quantas sentenças fornecer para o modelo para gerar nossas descrições. Conforme discutido, o conteúdo passado para o nosso modelo é composto pelo título do produto e por uma determinada quantidade de sentenças, de modo que nesse experimento pretendemos definir qual deve ser essa quantidade. Com esse objetivo, conduzimos um experimento com 10 avaliadores humanos para comparar 3 cenários em que usamos quantidades de sentenças diferentes. Apresentamos na Seção 6.4 esse estudo e seus resultado.

6.1 Treinamento de um Classificador

Conforme discutido na Seção 5.1, a extração de sentenças candidatas utiliza um classificador binário com o objetivo de identificar as sentenças adequadas para pertencerem a uma descrição de produto. Tratamos aqui então do treinamento desse classificador. Para isso, usamos o *dataset* disponibilizado por Novgorodov et al. (2019), discutido em mais detalhes na Seção 4.2, e adotamos a mesma métrica utilizada pelos autores, a área debaixo da curva ROC (AUC)(Bradley, 1997)[1].

Na etapa de treinamento dos classificadores, dividimos o *dataset* com 80% dos dados para treinamento e 20% para teste. Da parte destinada a treinamento, separamos 20% dos dados para validação. Em relação aos classificadores, foram experimentados os seguintes modelos:

– ***Naive Bayes* com TF-IDF** - experimentamos usar o *Naive Bayes* Rish et al. (2001), modelo amplamente utilizado em classificação de texto. Como *features*, utilizamos a representação textual da sentenças produzida pelo TF-IDF (Salton e Buckley, 1988). Em relação aos parâmetros

[1] A métrica AUC mede toda a área bidimensional abaixo da curva ROC inteira, que é a curva que mostra a taxa de verdadeiro positivo (recall) pela taxa de falso positivo em diferentes limiares de classificação. Assim, a AUC oferece uma medida agregada de desempenho em todos os limites de classificação possíveis.

ajustados, usamos a técnica de validação cruzada para escolher a melhor representação n-grama do TF-IDF.

- ***RandomForest Classifier* com TF-IDF** - outro modelo tradicional de classificação de texto (Breiman, 2001), também usando a representação obtida via TF-IDF para representar as sentenças. Além da forma de representação n-grama do TF-IDF, calibramos também o número máximo de *features*, mínimo de folhas sorteadas e número de árvores do algoritmo.
- ***XGBoost* com TF-IDF** - experimentamos outro modelo, *XGBoost* Chen e Guestrin (2016), agora também utilizando a representação obtida via TF-IDF para representar as sentenças. Além da representação n-grama, calibramos a profundidade máxima, peso mínimo do filho, e os parâmetros gamma e alpha.
- ***XGBoost* com ajuste fino do modelo *Masked and Permuted Pre-training for Language Understanding* (MPNET)** - Novamente experimentamos o *XGBoost*, porém optamos por utilizar o modelo domínio público *all-mpnet-base-v2*, desenvolvido a partir do ajuste fino do modelo MPNET (Song et al., 2020) em um *dataset* de 1 bilhão de pares de sentenças com um objetivo de aprendizado contrastivo para representar as sentenças. Calibramos os mesmos parâmetros mencionados no caso anterior.
- **Ada** - por último, experimentamos usar um LLM para realizar a classificação das sentenças, após os resultados demonstrados em Chae e Davidson (2023). Mais especificamente, condicionamos o modelo ada da OpenAI, uma das menores variante do modelo GPT-3, com 350 milhões de parâmetros. Notadamente, esse não é um modelo de domínio público, e é necessário pagar para utilizar o modelo via API da OpenAI. O modelo Ada é recomendado para classificação[2] por possuir menor custo para ajuste fino e uso na comparação com seus pares[3], ao mesmo tempo que mantém um performance próxima dos modelos com mais parâmetros.

Os classificadores foram treinados separadamente em cada um dos domínio de *Fashion* e *Motors* do *dataset* comentado. Apresentamos os resultados obtidos com o conjunto de teste na Tabela 6.1. Como podemos observar, o modelo que obteve melhor resultado foi o ada, obtendo uma performance bastante

[2]A OpenAI apresenta a seguinte mensagem: *Para classificação, recomendamos experimenta um dos modelo mais rápidos e baratos, como o 'ada"*

[3]Conforme pode ser observado em: <`https://platform.openai.com/docs/deprecations`>

superior nas comparações com os demais. Destacamos que esse performance foi superior a melhor reportada por Novgorodov et al. (2019), que foi de 0,924.

Tabela 6.1: AUC dos classificadores na atividade de identificação de sentenças adequadas a pertencerem a uma descrição de produto.

Modelo	Categoria	
	Fashion	*Motors*
Naive Bayes com TF-IDF	0,908	0,910
RandomForest com TF-IDF	0,886	0,897
XGBoost com TF-IDF	0,867	0,890
XGBoost com MPNET	0,886	0,907
Ada	**0,943**	**0,948**

Além de reportar os resultados, comparamos brevemente os nossos resultados com os resultados apresentados por Novgorodov et al. (2019), nos casos em que experimentos modelos de classificação em comum, mais especificamente o *RandomForest* e *XGBoost*. Enquanto os autores, que experimentaram usar os unigramas, bigramas e trigramas das palavras obtiveram uma performance de 0,779 e 0,831 na categoria *Fashion* para o *Naive Bayes* e *XGBoost*, nós utilizamos a representação das sentenças usando o método TF-IDF, que melhorou a classificação, obtendo valores de 0,886 e 0,867 no comparativo. Além disso, ao experimentar uma outra forma de representação das sentenças baseada no ajuste fino do modelo MPNET, verificamos que o *XGBoost* teve uma melhora de performance.

Frente a esses resultados, adotamos então o modelo Ada como sendo o classificador utilizado em nosso método. Contudo, discutimos adiante algumas questões relacionadas a métrica utilizada e outras possibilidades.

6.1.1 Generalização para outras categorias de produto

Afim de explorar a capacidade de generalização dos modelos para outros domínios, experimentamos avaliar novamente os classificadores mas agora usando o conjunto de teste do outro domínio. Ou seja, selecionamos o modelo treinado no conjunto *Fashion* e medimos a sua performance no conjunto de teste *Motors*, e vice-versa. Os resultados podem ser verificados na tabela 6.2. Indicamos na coluna o conjunto de dados no qual o modelo foi treinado, ou seja, na coluna *Fashion* indicamos o score do modelo treinado nessa categoria e avaliado no conjunto de teste da categoria *Motors*. Já a segunda coluna se refere ao modelo treinado na categoria *Motors* e avaliado no conjunto de teste da categoria *Fashion*.

No geral, observamos que os modelos pareceram generalizar bem para outros domínios. Comparando os resultados entre as duas tabelas, verificamos

Tabela 6.2: AUC dos classificadores no caso de generalização, em que treinamos o modelo em um categoria (nome da coluna) e o testamos na outra.

Modelo	Categoria	
	Fashion	*Motors*
Naive Bayes com TF-IDF	0,870	0,867
RandomForest com TF-IDF	0,862	0,854
XGBoost com TF-IDF	0,850	0,838
XGBoost com MPNET	0,842	0,859
Ada	**0,917**	**0,902**

que o modelo Ada treinado no *dataset Motors* e avaliado no *Fashion* teve uma performance razoavelmente próxima do treinado no próprio domínio, com uma diferença de 0,041, enquanto o treinado no *Fashion* e avaliado no *Motors* teve uma diferença de 0,031. O mesmo é verdade para os demais modelos, em que observamos uma piora da performance limitada.

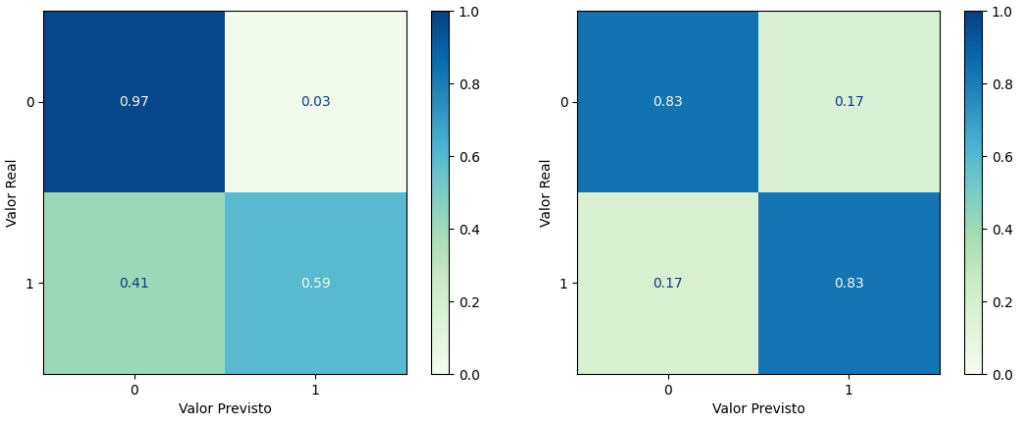
Isso parece indicar que, em certa medida, a natureza das sentenças de produtos de categorias diferentes é parecida. Como essa etapa, no contexto do método que propomos, é a única que depende de um modelo supervisionado, e mesmo assim os modelos possuem uma capacidade razoável de generalização. Verificamos, com isso, que nosso método seria aplicável em outros domínios além dos quais possuímos dados anotados.

6.1.2 Discussão sobre métrica utilizada e limitações

Ainda que tenhamos optado utilizar a mesma métrica que os autores para seleção do classificador (AUC-ROC), vale a discussão sobre as possíveis diferenças mascaradas por trás de pontuações aparentemente semelhantes. Exibimos nas Figuras 6.1(a) e 6.1(b) a matriz de confusão normalizada dos dois classificadores com melhores resultados na etapa anterior, o Naive Bayes e o ada.

Conforme pode ser observado, existe uma diferença muito grande entre os dois modelos no que tange falsos positivos e falsos negativos. Enquanto o Ada classifica quase perfeitamente os casos negativos, ele o faz a custa de uma quantidade grande de falsos negativos. Já o *Naive Bayes* tem uma performance muito equilibrada nos dois casos, ou seja, ele classifica mais sentenças como positivas, a um custo de mais falsos positivos.

Na nossa metodologia, temos ainda uma etapa de geração de texto a partir das sentenças selecionadas, conforme discutido na Seção 6.4. Nessa etapa um LLM deve selecionar quais informações das sentenças utilizar, de modo que ocorre mais uma etapa de filtro de sentenças. Assim, fica a questão se falsos positivos são de fato muito prejudiciais ao nosso método. Por outro



6.1(a): Classificador Ada

6.1(b): Classificador *Naive Bayes*

Figura 6.1: Matriz de Confusão normalizada dos dois classificadores com melhores resultados.

lado, sabemos que um método que classifica muitas instâncias positivas como falsas acaba sendo prejudicial, no sentido que exige mais avaliações deixadas por usuários para ser capaz de gerar uma descrição. Deixamos essa discussão como um trabalho futuro.

6.2 Definição do formato das descrições a serem geradas

Examinamos aqui uma das questões diretamente relacionada com a etapa final da nossa proposta de geração da descrição de produto. Ao utilizar um LLM estamos usando um modelo gerador de texto em que as possibilidades de saída são infinitas, variando tanto no seu conteúdo, quanto na sua forma. Dessa forma, para pensar em como utilizar o modelo precisamos primeiro pensar em que tipo de texto queremos gerar. Assim, nos debruçamos nessa seção na questão da forma do texto gerado, e mais especificamente no seu tamanho.

Quando pensamos no tamanho das descrições de produtos, existe um *tradeoff* a ser considerado. Ao gerar uma descrição muito longa, possivelmente estaremos adicionando mais informação sobre o produto, porém impomos um custo maior para o leitor. Com isso em mente, e uma vez que gostaríamos de gerar descrições que fossem mais informativas que as descrições originais, dada a constatação da riqueza de informações presente nas avaliações deixadas pelos usuários, fica a questão de qual tamanho nossas descrições devem possuir.

Para responder essa questão, optamos por basear nossa decisão no tamanho de descrições reais. Nesse sentido, conduzimos uma análise descritiva sobre as descrições de produto originais contidas no *dataset* descrito na Seção 4.1, que são descrições reais postadas pelos anunciantes. Com essa finalidade, selecionamos 1 milhão de descrições originais de produtos da

categoria *"Clothing, Shoes and Jewlery"*, que é o domínio que pretendemos gerar as nossas descrições. Afim de evitar qualquer tipo de viés, evitamos qualquer tipo de filtro das descrições selecionadas, entendendo que boas descrições podem apresentar tamanhos e formatos diversos. Examinamos então a quantidade de sentenças e palavras dessas descrições.

Exibimos primeiro na Figura 6.2, o percentual de descrições por quantidade de sentenças, e verificamos que a maior parte das descrições é composta por apenas uma sentença (28%). Por outro lado, verificamos também que 50% das descrições possuem até 3 sentenças, e 75% possuem até 5.

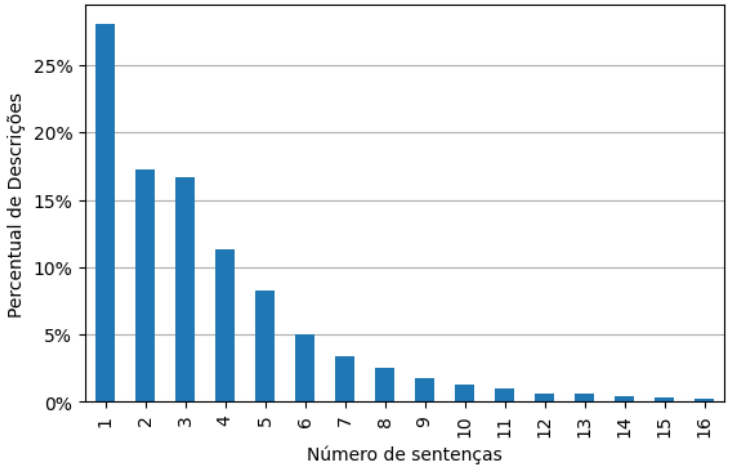


Figura 6.2: Percentual de descrições por número de sentenças das descrições originais postadas pelos anunciantes.

Por outro lado, examinando agora as descrições a nível de palavras, exibimos na Figura 6.3 o histograma da quantidade de palavras por percentual de sentenças. Verificamos como a distribuição do número de palavras se concentra no lado esquerdo, com 50% das descrições contendo menos de 45 palavras e 75% contendo menos de 75 palavras. Além das distribuições, exibimos na tabela 6.3 algumas estatísticas das sentenças e palavras por descrição, chamando atenção para o alto desvio padrão da quantidade de palavras (50 palavras).

Tabela 6.3: Características das descrições originais postadas pelos anunciantes.

	Média	Desvio Padrão	Mediana	Máximo
Sentenças por descrição	3.7	3.3	1	64
Palavras por descrição	56.4	50.1	44	1172

Ao refletir sobre o nosso método e pensar de que forma podemos nos basear nas descrições originais, ressaltamos dois pontos da nossa metodologia que consideramos como pontos fortes: o primeiro é a riqueza das informações contidas nas sentenças das avaliações, e o segundo é a capacidade do modelo

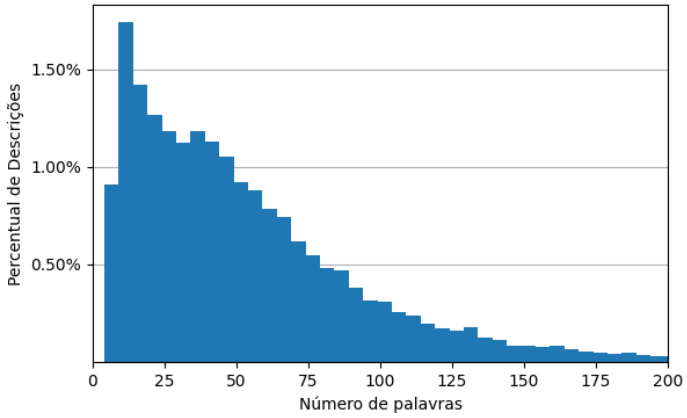


Figura 6.3: Percentual de descrições por número de palavras das descrições originais postadas pelos anunciantes.

gerador em articular essas informações em um texto conciso e articulado. Com isso em mente, consideramos que a forma adequada de nos basear nas descrições originais seria adotando um tamanho máximo para as descrições conforme observado na distribuição, e permitindo então bastante liberdade ao modelo na hora de gerar as descrições.

Consideramos então que uma forma prática de controlar o texto gerado seria controlando a quantidade máxima de palavras geradas. Da distribuição exibida, definimos o nosso limite de palavras a ser gerado com base no 95º percentil dos dados observados, que é de 150 palavras. Assim, estabelecemos esse limite como um pré-requisito da etapa de geração, isto é, as descrições de produto geradas devem ter no máximo 150 palavras.

6.3 Construção da Instrução

Examinamos agora uma subetapa fundamental da etapa final do nosso método de geração da descrição de produto. Conforme discutido anteriormente, existe uma quantidade infinita de instruções possíveis, sendo que cada instrução pode gerar um texto diferente. Assim, consideramos essa uma etapa desafiadora.

Uma vez definido o modelo utilizado como o GPT 3-5 turbo, conforme discutido na Seção 5.3, para definir qual instrução utilizar para gerar as descrições de produto, optamos por adotar uma abordagem de aprimoramento contínuo.

6.3.1 Definindo as instruções candidatas iniciais

Inicialmente, transformamos a definição de descrição de produto proposta pelos autores Novgorodov et al. (2019) em um instrução. Com esse objetivo, buscamos seguir as orientações propostas pela OpenAI[4]. no que tange a construção do prompt. Após uma série de experimentações empíricas, chegamos ao seguinte resultado:

– **Instrução Base**: *"You will be provided with a product title and sentences extracted from avaliações. Your task is to write a product description. We refer to a product description as textual presentation of what the product is, how it can be used, and why it is worth purchasing. The purpose of a product description is to provide customers with details about the features and benefits of the product so they are compelled to buy. However, a product description doesn't have to necessarily contain all this information, as it should be based only on the title and extracted sentences. That being said, we expect a good description to be informative, readable, objective, and relevant to the product."*[5].

Uma vez definida uma instrução base para o modelo, propomos ainda três variações mais enxutas desse mesma instrução. Com essas variações, queremos explorar como variam os resultados. Essas variações podem ser vistas a seguir, e foram consideradas instruções candidatas para serem utilizadas no nosso método.

– **Variação 1**: *"You will be provided with a product title and sentences extracted from reviews. Your task is to generate a product description based only on the title and extracted sentences. The product description should only contain objective, relevant and positive information from the reviews"*[6].

[4]As orientações de *prompting* da própria OpenAI estão disponíveis em sua página web, acessível em: <`https://platform.openai.com/docs/guides/prompt-engineering`>

[5]**Tradução da instrução base**: "Você receberá o título do produto e frases extraídas das avaliações. Sua tarefa é escrever uma descrição do produto. Referimo-nos à descrição do produto como uma apresentação textual do que é o produto, como pode ser usado e por que vale a pena comprá-lo. O objetivo da descrição de um produto é fornecer aos clientes detalhes sobre os recursos e benefícios do produto, para que eles sejam obrigados a comprar. Porém, a descrição de um produto não precisa necessariamente conter todas essas informações, pois deve ser baseada apenas no título e nas frases extraídas. Dito isto, esperamos que uma boa descrição seja informativa, legível, objetiva e relevante para o produto".

[6]**Tradução da variação 1**: "Você receberá o título do produto e frases extraídas dos comentários. Sua tarefa é gerar uma descrição do produto baseada apenas no título e nas frases extraídas. A descrição do produto deve conter apenas informações objetivas, relevantes e positivas provenientes das avaliações".

– **Variação 2**: *"Write an objective and informative product description based only on the product title and received sentences extracted from reviews"*[7].

– **Variação 3**: *"Write a product description based only on the following title and sentences extracted from reviews"*[8].

6.3.2 Limitando a quantidade de palavras

Conforme detalhado na Seção 6.2, estabelecemos um tamanho desejado para as descrições a serem geradas, baseados na distribuição observada das descrições de produto originais. Estamos interessados então em instruções que respeitem a condição de gerarem descrições com uma quantidade de palavras inferior ao limite de 150 palavras definido em 6.2.

Com isso em mente, experimentamos gerar descrições de produtos para um subconjunto de produtos para obter uma ideia inicial dos tamanhos das descrições geradas. Para tal, geramos para cada instrução 100 descrições a partir dos mesmos 100 produtos selecionados aleatoriamente, passando em cada instrução 20 sentenças selecionadas pela metodologia discutida nas Seções anteriores. Computamos então algumas estatísticas do tamanho das descrições geradas em termos de número de palavras utilizadas, exibidas na Tabela 6.4.

Tabela 6.4: Quantidade de Palavras Geradas por cada instrução.

instrução	Média de Palavras	Desvio Padrão	Máximo
Instrução Base	231	60	410
Variação 1	196	54	360
Variação 2	192	47	325
Variação 3	198	51	362

Conforme pode ser observado, as descrições geradas pelo modelo são significativamente mais longas que as descrições que pretendemos gerar, no sentido de que em todos os casos a média de palavras foi muito superior a 150 palavras. Por isso, pensamos então em formas de limitar o tamanho de textos gerados.

Para isso, exploramos combinar os instruções iniciais com terminações que limitassem o tamanho dos textos gerados, de forma que geramos 4 terminações para serem adicionadas ao final de cada instrução. Cada terminação funciona como um sufixo concatenado à instrução. Especificamos em cada terminação a quantidade limite de palavras a serem geradas no texto, trabalhando

[7]**Tradução da variação 2**: "Escreva uma descrição objetiva e informativa do produto com base apenas no título do produto e nas frases recebidas extraídas dos comentários".

[8]**Tradução da variação 3**: "Escreva uma descrição do produto com base apenas no seguinte título e frases extraídas de comentários".

com o valor definido anteriormente de 150 palavras. Apresentemos essas terminações a seguir, e geramos um total de 16 instruções a partir de todas as combinações possíveis de instrução inicial e terminação.

- **Terminação 1:** *"Write a text of no more than 150 words"*[9].
- **Terminação 2:** *"The product description cannot contain more than 150 words"*[10].
- **Terminação 3:** *"The product description must be limited to 150 words"*[11].
- **Terminação 4:** *"Use 150 words at maximum to write the product description"*[12].

Para cada par (instrução e terminação), geramos novamente as 100 descrições dos produtos selecionados, e novamente observamos a quantidade de palavras geradas. Constatamos então que para nenhum dos 16 pares de instrução inicial e terminação experimentados o número limite de palavras foi respeitado no caso das 100 descrições, isto é, em todos casos pelo menos uma descrição teve mais de 150 palavras. Contudo, ao comparar com as descrições geradas sem nenhum limite, observamos que a quantidade de palavras média diminuiu. Para todos pares experimentados a média de palavras ficou entre 130 e 150 palavras, bem abaixo do verificado sem as terminações.

Percebendo portanto que o modelo não respeita o limite exato determinado na instrução, mas é sim afetado pelo limite de palavras imposto, experimentamos reduzir a quantidade máxima de palavras esperando dessa forma gerar descrições ainda menores, de modo que para os 100 produtos do nosso conjunto sejam gerados textos com menos de 150 palavras. Com esse intuito, propomos repetir a geração de descrições de produto para cada instrução só que passando um quantidade limite ainda menor. Para simplificar o experimento, em vez de de gerar descrições para cada uma das 16 combinações de instrução formuladas, optamos por escolher uma das 4 terminações experimentadas. Adotamos então o critério de escolher a terminação que menos vezes ultrapassou o limite estipulado na instrução de 150 palavras. Exibimos essa informação na tabela 6.5, e escolhemos a primeira terminação.

Feito isso, repetimos agora a geração das 100 descrições de produtos, usando as 4 variações de instrução com a terminação definida anteriormente,

[9]Tradução da terminação 1: "Escreva um texto com no máximo 150 palavras".

[10]Tradução da terminação 2: "A descrição do produto não pode conter mais de 150 palavras".

[11]Tradução da terminação 3: "A descrição do produto deve ser limitada a 150 palavras".

[12]Tradução da terminação 4: "Use no máximo 150 palavras para escrever a descrição do produto".

Tabela 6.5: Percentual de descrições geradas por cada terminação com mais de 150 palavras. Cada terminação foi combinada com as 4 instruções candidatas, gerando um total de 400 descrições.

Terminação	Percentual
Terminação 1	7,5
Terminação 2	13,8
Terminação 3	14,8
Terminação 4	21,0

porém reduzindo de forma iterativa o limite de palavras máximo estipulado na instrução de 25 em 25. Definimos como condição de parada para cada instrução que todas as descrições geradas tivessem uma quantidade de palavras inferior ao limite real estabelecido de 150 palavras. Ou seja, primeiro repetimos as instruções mas impondo no máximo 125 palavras, e em seguida para as instruções que não respeitaram o limite repetimos a geração mas agora com limite de 100 palavras, e assim em diante. Dessa forma, chegamos finalmente às três instruções:

- **Instrução 1**: "You will be provided with a product title and sentences extracted from avaliações. Your task is to write a product description. We refer to a product description as textual presentation of what the product is, how it can be used, and why it is worth purchasing. The purpose of a product description is to provide customers with details about the features and benefits of the product so they are compelled to buy. However, a product description doesn't have to necessarily contain all this information, as it should be based only on the title and extracted sentences. That being said, we expect a good description to be informative, readable, objective, and relevant to the product.The product description cannot contain more than 75 words"[13].
- **Instrução 2**: "Write an objective and informative product description based only on the product title and received sentences extracted from reviews. The product description cannot contain more than 75 words"[14].

[13]Tradução da Instrução 1:"Você receberá o título do produto e frases extraídas das avaliações. Sua tarefa é escrever uma descrição do produto. Referimo-nos à descrição do produto como uma apresentação textual do que é o produto, como pode ser usado e por que vale a pena comprá-lo. O objetivo da descrição de um produto é fornecer aos clientes detalhes sobre os recursos e benefícios do produto, para que eles sejam obrigados a comprar. Porém, a descrição de um produto não precisa necessariamente conter todas essas informações, pois deve ser baseada apenas no título e nas frases extraídas. Dito isto, esperamos que uma boa descrição seja informativa, legível, objetiva e relevante para o produto. A descrição do produto não pode conter mais de 75 palavras".

[14]Tradução da Instrução 2:"'Escreva uma descrição objetiva e informativa do produto com base apenas no título do produto e nas frases recebidas extraídas das avaliações. A descrição do produto não pode conter mais de 75 palavras".

– **Instrução 3**: "Write a product description based only on the following title and sentences extracted from reviews. The product description cannot contain more than 125 words"[15].

Dentre as 4 variações experimentadas, ressaltamos que para a variação 2 não foi atingida a condição de parada para nenhum dos valores experimentados, gerando descrições de produto com mais de 150 palavras mesmo quando limitamos a descrição à 25 palavras. Por esse motivo, ela foi descartada.

6.3.3 Avaliando as descrições geradas

Após as etapas descritas, terminamos então com 3 instruções candidatas, e optamos por conduzir uma avaliação qualitativa das descrições geradas por cada instrução. Estruturamos essa avaliação da seguinte forma: para um conjunto de 80 produtos aleatórios, geramos descrições de produto usando cada uma das 3 instruções. Em seguida, geramos 120 pares combinando uniformemente descrições geradas por cada instrução, ou seja, formamos 40 pares combinado descrições geradas pela instrução 1 e 3, 40 pares combinando descrições geradas pela instrução 1 e 4 e 40 pares combinando descrições geradas pela instrução 3 e 4. Em seguida, apresentamos para um avaliador, no caso o autor principal do trabalho, uma descrição original de um produto seguida por um dos pares de descrições em ordem aleatória. Nesse cenário, o anotador foi questionado com a seguinte pergunta:

"Qual descrição você escolheria se tivesse que substituir a original?"

Como resposta, o anotador poderia indicar uma das duas descrições apresentadas, ou então indicar um empate em caso de nenhuma preferência clara. Desse experimento, obtivemos então os seguintes dados, apresentados na tabela 6.6. Conforme observado, a instrução 3 foi escolhida como pior instrução em mais da metade das comparações. Por esse motivo, ele foi descartado.

Tabela 6.6: Resultado em percentual da comparação direta entre pares de descrições

Instrução	Como Melhor (%)	Como Pior (%)	Empate (%)
Instrução 1	35	19	46
Instrução 2	45	9	46
Instrução 3	10	62	28

Já em relação aos instruções 1 e 2, verificamos que na comparação direta entre os dois a instrução 2 foi preferida 9 vezes, a instrução 1 foi preferida 5

[15]Tradução da Instrução 3:"Escreva uma descrição do produto com base apenas no seguinte título e frases extraídas dos comentários. A descrição do produto não pode conter mais de 125 palavras".

vezes e 26 vezes ocorreram empates. Além disso, observamos que a instrução 2 possui quase metade da quantidade de tokens da primeira (69 tokens de diferença), e portanto é mais barato [16]. Por esse motivo, optamos por seguir com a instrução 2.

6.3.4 Limitações identificadas

Uma vez selecionado a melhor instrução dentre as candidatas, no espírito de melhoria contínua da nossa solução, observamos algumas limitações identificadas na avaliação qualitativas das descrições. Foram duas as questões abordadas e achamos interessante reportá-las, ainda que não tenhamos encontrado soluções. Apresentamos esses casos na Tabela 6.7, em que selecionamos dentre as descrições geradas pela instrução que obteve melhores resultados algumas que apresentaram as limitações discutidas.

A primeira limitação abordada foi a forma que os as descrições de produto eram iniciadas, que consideramos muito engessadas. Dos 120 pares de descrições de produto geradas, observamos que em todas as descrições repetiam o título na exata forma como foi passado, e o faziam sempre na primeira sentença. Isso pode ser observado na Tabela 6.7. Ainda que não seja um problema por si só, consideramos essa uma questão curiosa, e investigamos fazer ajustes na instrução selecionada como a melhor. Dentre os ajustes, experimentamos incluir de diversas formas na instrução para o modelo que o título não deveria ser repetido, e essa instrução foi completamente ignorada.

A segunda limitação verificada foi a de que algumas das descrições geradas incluíram de alguma forma na sua última sentença declarações de marketing, isto é, declarações puramente apelativas para a compra do produto. Por não adicionaram informações sobre o produto, consideramos isso um problema. Para dimensionar essa questão em mais detalhes, avaliamos 100 descrições de produto em busca das declarações que nos referimos e as encontramos em 24 ocasiões. Conforme pode ser visto na Tabela 6.7, tivemos muitas descrições afirmando que determinado produto é um *"must-have"*, como exemplificado na descrição 3 e 4, afirmações elogiando a escolha do produto, como na descrição 2, ou adjetivos apelativos, como *"ultimate comfort"* na primeira descrição.

Para mitigar esse problema, experimentamos variações tentando controlar a terminação das descrições, como ao pedir que as descrições terminem apresentando informações objetivas do produto, ou que não terminassem com

[16] Apenas para ilustração, se fossem geradas 10 mil descrições usando cada instrução, teríamos uma diferença de custo de 1,36 dólares

Tabela 6.7: Descrições de produto geradas que ilustram as limitações discutidas.

	Exemplos de descrições geradas
1	*The ECCO Men's Yucatan Sandal is a high-quality and comfortable sport sandal, perfect for hot summer days and all-day wear. With excellent arch support and great traction, these sandals provide the perfect fit and feel. Made with superior workmanship and sturdy materials, they ensure durability and long-lasting comfort. These lightweight sandals are also suitable for all-terrain use and provide fantastic support for those with flat feet. Experience ultimate comfort with the ECCO Yucatan Sandal.*
2	*Experience the classic Birkenstock fit and feel with the Women's Milano Backstrap Sandal. The contoured cork mid-sole molds to your foot, providing amazing support for your entire foot. Made with quality materials, these sandals are well-made and comfortable. With their good arch support, they can be worn with anything from jeans to dresses. Whether you're going on a walking vacation or simply looking for a comfortable sandal for everyday wear, the Milano is the perfect choice.*
3	*Introducing the Rhode Island Novelty Child Pink Blinking Tiara Cowboy Hat! This adorable cowgirl hat is perfect for playing dress up or as a prop for a cowgirl-themed birthday party. Made with high-quality materials, this well-made hat features a blinking tiara on the front that lights up in different colors. The flashing lights are super bright and the battery life is long-lasting. This cute and sassy hat is a must-have for all the little princess cowgirls out there!*
4	*The Timex Cavatina Expansion Band Watch is a sleek and stylish timepiece designed for everyday wear. With a feminine design and a black leather strap, this watch is comfortable to wear on your hand. The genuine leather strap adds a touch of quality, and the easy-to-read face makes it convenient to check the time. Whether you're at work or dressing up for a special occasion, this watch is perfect for any occasion. Reliable and attractive, it's a must-have accessory for every woman.*

declaração de marketing. Contudo, também não obtivemos resultados iniciais muito promissores, e preferimos deixar essa questão como um possível trabalho futuro.

6.4 Quantidade de sentenças a ser passada para o modelo

Por último, uma vez definido a instrução para o modelo generativo, o passo final para a geração de descrições de produto foi definir seu conteúdo. Conforme discutimos, o conteúdo passado para o modelo vai ser composto pelas sentenças junto do título do produto, de forma que precisamos definir o número de sentenças para a ser utilizado. Para abordar essa questão, discutimos primeiro um *tradeoff* que esperamos existir no que tange a qualidade das descrições geradas e o custo de gerá-las.

No nosso método, conforme discutido em mais detalhes no Capítulo 5, estamos interessados em potencializar a informatividade das descrições ao utilizar sentenças extraídas de avaliações. Dessa forma, uma questão que surge naturalmente é a partir de quantas sentenças devo gerar minha descrição. Por um lado, ao usar mais sentenças, estou possibilitando que uma maior

quantidade de informações seja condensada em única descrição.

De fato, ao usar um LLM esperamos gerar descrições que sejam mais informativas do que uma abordagem extrativa de reunir as sentenças, esperando que informações contidas em múltiplas sentenças sejam condensadas em um texto conciso e legível. Assim, temos a expectativa de que um número maior de sentenças, e que portanto reúna mais conteúdo, gere descrições mais informativas. Contudo, pela natureza do modelo, não conseguimos prever qual será o efeito de incluir mais sentenças para serem sumarizadas, entendendo que pode ser possível que problemas de alucinação, que será discutido na avaliação (Capítulo 7), apareçam com mais frequência conforme aumente a quantidade de sentenças. Além disso, como desejamos gerar descrições com uma quantidade limitada de palavras, conforme discutido na Seção 6.2, fica o questionamento se ao adicionar mais sentenças elas de fato serão utilizadas.

Nesse contexto, optamos por desenvolver 3 cenários distintos em que fornecemos diferentes quantidades de sentenças para o modelo gerado. Com isso, a ideia é explorar a capacidade de sumarização da metodologia, isto é, em que medida ele é capaz de condensar informações contidas em uma série de sentenças não articuladas em uma única descrição. Assim, primeiro buscou-se definir um cenário base, com um mínimo de sentenças, para depois se estender essa quantidade e avaliar se houve melhorias na qualidade da descrição. Para se definir o número de sentenças no cenário de menor quantidade, foi feita a seguinte a análise.

No trabalho de referência (Novgorodov et al., 2019) foram explorados 3 cenários de descrições extrativas, usando 3, 5 e 7 sentenças. No geral, ainda que não tenha sido estatisticamente significativo, os autores observaram que o melhor resultado geral foi usando 5 sentenças. Por esse motivo, e afim de explorar possíveis comparações com o trabalho de referência, definimos esse como o primeiro cenário a ser testado na nossa metodologia. Consideramos esse portanto o cenário base para geração de descrições.

Contudo, como estamos interessados em potencializar as capacidades de sumarização de LLM para geração de descrições, propomos aumentar nos dois outros cenários a quantidade de sentenças. Nesse sentido, notamos o custo de usar uma quantidade maior de sentenças, na medida que isso exige uma quantidade maior de avaliações de usuários para um determinado produto e portanto impedindo a aplicação da solução proposta para produtos recentes ou com pouco engajamento. Além disso, há também um custo relacionado a quantidade de tokens, isto é, quanto mais sentenças mais tokens. Por isso, definimos para esses dois cenários um total de 10 e 20 sentenças.

Para avaliarmos se de fato vale a pena utilizar mais sentenças e assim

definirmos a quantidade de sentenças a ser passada para o modelo, propomos então um avaliação qualitativa com anotadores humanos, em que geramos descrições para 100 produtos usando os três cenários de quantidade de sentenças definidos anteriormente.

O objetivo dessa avaliação é o de verificar se existe alguma diferença significativa na qualidade geral das descrições geradas ao passar diferentes quantidades de sentenças para o modelo, e, se existe, qual é a quantidade que gera as melhores descrições. Assim, não nos interessamos em comparar as descrições em múltiplos aspectos (como legibilidade e fluência), entendendo que isso poderia enviesar a escolha do método, e sim em selecionar a metodologia que gera as melhores descrições.

Outro ponto relevante a ser notado é que, por mais que tenhamos adotado uma definição de descrição de produto, não definimos o que é uma boa descrição, deixando a cargo do anotador fazer uma análise com base em suas preferências pessoais. Por avaliarmos se tratar de um conceito intrinsecamente subjetivo, consideramos que qualquer tipo de orientação ao anotador por definição geraria um viés na escolha do método. Esse viés, por sua vez, poderia gerar descrições piores de forma geral, na medida que a definição proposta poderia falhar em capturar aspectos importantes de uma boa descrição.

Assim, optamos por basear nossa avaliação no ranqueamento entre as três descrições proposta. Assim como nos trabalhos de Liu e Lapata (2019), Puduppully et al. (2019) e Amplayo et al. (2021), utilizamos a técnica *Best-Worst Scalling* (Louviere et al., 2015). Conforme demonstrado por Kiritchenko e Mohammad (2017), e citando Gehrmann et al. (2023), *Best-Worst Scalling* é uma técnica eficiente e confiável para a coleta de anotações de sobre a comparação de texto gerado.

No nosso cenário, pedimos para os anotadores selecionarem dentre as 3 descrições apresentadas qual era a melhor ("Qual é a melhor descrição dentre as 3 apresentadas?") e qual era a pior ("Qual é a pior descrição dentre as 3 apresentadas?"). O score de cada metodologia foi calculado como o percentual de vezes em que foi escolhida como a melhor menos o percentual de vezes que foi escolhido como a pior, variando portanto entre 1 e -1 (Orme, 2009). Cada anotador teve que indicar para um total de 10 produtos a pior e melhor descrição.

Foram utilizados nessa etapa 10 anotadores voluntários. Por se tratar de uma atividade complexa, foi definido um perfil alvo para os anotadores. Primeiramente, por se tratar de textos em inglês, os anotadores precisam serem capazes de ler e compreender textos na língua inglesa. Em segundo lugar, consideramos que a experiência com descrições de produtos como fundamental,

justamente por entendermos que uma boa descrição é um conceito amplo e aberto. Por isso, convidamos pessoas com habitualidade com compras online.

Os resultados do experimento podem ser observados na Tabela 6.8. Verificamos que quanto mais sentenças melhor avaliado foi o método, com o método de 20 sentenças ficando em primeiro e o de 10 sentenças em segundo. Nessa avaliação, o método com 20 sentenças foi amplamente preferido, sendo escolhido como melhor em 45% das vezes e pior em 17%. Já o método de 5 sentenças foi amplamente escolhido como o pior, em 47% das vezes.

Tabela 6.8: Resultados *Best-worst Scalling* variando a quantidade de sentenças insluídas no conteúdo do *prompt*.

Qtd. de Sentenças	Como melhor (%)	Como pior (%)	Pontuação
20 sentenças	45	17	0.28
10 sentenças	32	36	-0.04
5 sentenças	23	47	-0.24

Esse resultado confirma nossa expectativa de que o uso de mais sentenças gera descrições melhores. Por conter mais sentenças, esperávamos que essas descrições fossem mais informativas, na medida que LLM tem acesso a uma quantidade maior de informação. Seguimos no nosso método com o cenário de 20 sentenças.

6.5 Visão final do método

Para concluir, apresentamos então como ficou configurada nossa solução após a definição de cada subetapa com o aprofundamento realizado em cada uma das seções anteriores. Ilustramos essa configuração final na Figura 6.4, apresentando como ficou definida cada subetapa.

No caso da extração de sentenças candidatas, selecionamos o modelo Ada para classificar as sentenças, que obteve a maior pontuação na métrica proposta (AUC), conforme indicado em subetapa 1.2 na Figura. Já na etapa de geração de descrição, definimos como formato das descrições apenas a condição de ter menos que 150 palavras. Em relação a instrução, definimos alguns candidatos e selecionamos um após conduzir uma experimento para incluir o limite de palavras na instrução e conduzir uma avaliação qualitativa. Podemos observar esses dois ajustes na instrução final, na subetapa 3.1 na Figura.

Por último, conduzimos mais uma avaliação qualitativa com múltiplos avaliadores e verificamos que, dentre os 3 cenários testados, o cenário com mais sentenças gerou as descrições preferidas, fixando então em 20 o total de sentenças a serem fornecidas junto com o título como conteúdo para o modelo. Exibimos isso no conteúdo que compõe a subetapa 3.1 na Figura.

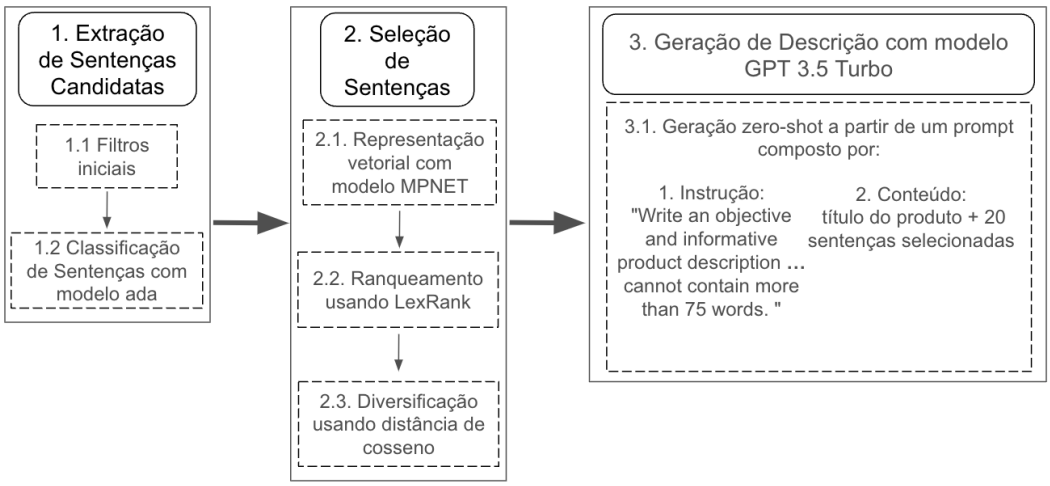


Figura 6.4: Solução proposta após definição de cada subetapa.

7
Avaliação

Com os recentes avanços em modelos de redes neurais profundas na área de linguagem, novos modelos tem alcançado progresso significativo em atividades desafiadoras, como na geração de textos longos. Contudo, com o progresso recente, surge também a necessidade de novas métricas e avaliações. Conforme constatado no trabalho de Celikyilmaz et al. (2020), a avaliação de tarefas envolvendo a geração de textos longos é uma atividade desafiadora. Métricas tradicionais baseadas na interseção de palavras apresentam em muitos casos baixa correlação com avaliações humanas, conforme apontado nos trabalhos de Goyal et al. (2022), Bhaskar et al. (2022) e Zhu et al. (2020), principalmente em tarefas que permitem uma diversidades de textos e múltiplos resultados para uma mesma entrada.

Discutindo ainda mais o tema, Gehrmann et al. (2023) observam que com o surgimento de LLMs capazes de continuar sequencias de texto com uma alta acurácia, métricas baseadas em um resumo de referência, como BLEU (Papineni et al., 2002) e ROUGE (Lin, 2004), tem apresentado baixa performance, no sentido de que frequentemente o texto gerado pelo modelo é julgado como superior ao de referência. Nesse contexto, para avaliar a qualidade de textos gerados é muito comum o uso de avaliações humanas.

Nesse sentido, com o objetivo de avaliar nosso método, conduzimos tanto avaliações quantitativas como avaliações qualitativas. Em relação as avaliações quantitativas, analisamos primeiro de que formas as sentenças selecionadas das avaliações dos consumidores e inseridas no *prompt* influenciam a LLM na geração de descrições. Para tal, exploramos primeiro métricas que se baseiam na interseção de termos entre as descrições geradas e o *prompt*. Em seguida, buscamos explorar a questão da consistência factual da descrições geradas, utilizando duas métricas baseadas em classificadores propostas recentemente na literatura. Contextualizamos mais essa discussão e apresentamos os resultados obtidos nessas métricas na Seção 7.1.

Em relação a avaliações qualitativas, conduzimos um experimento para avaliar a qualidade das descrições em algumas dimensões: legibilidade, objetividade, informatividade, relevância para o produto e preferência no geral. Fizemos isso comparando as descrições geradas com as originais (a descri-

ção fornecida pelo próprio anunciante), utilizando uma escala Likert (Amidei et al., 2019) de 7 pontos, em que cada item da escala é uma comparação entre as duas descrições, considerando cada uma das dimensões mencionadas. Com essa avaliação, buscamos responder a subquestão de pesquisa *SQP2 (Como se comparam as descrições geradas com as descrições originais?)*.

Além disso, a fim de estabelecer um comparativo com a literatura recente, reproduzimos o método proposto por Novgorodov et al. (2019) e também o comparamos com as descrições originais nas mesmas dimensões. Ao buscar esse comparativo, estamos buscando também responder a nossa subquestão de pesquisa *SQP1 (Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?)*, na medida que esse é um trabalho que também utiliza as avaliações de usuários como fonte de informação, mas sem utilizar um LLM.

Na condução do experimento, contamos com a colaboração voluntária de 30 avaliadores, que analisaram pares de descrições referentes a um total de 150 produtos. Além de apresentar os resultados obtido com a escala Likert, realizamos também um teste estatístico comparando os resultados do nosso método com o método de referência. Discutimos esse experimento em mais detalhes e apresentamos seu resultado na Seção 7.2

7.1 Avaliação de consistência

Uma questão natural que surge na nossa metodologia é a preocupação com relação a consistência das informações apresentadas na descrição produzida. Em nosso método, conforme mencionado na Seção 5.3, selecionamos uma quantidade de sentenças contendo informações relevantes sobre o produto em questão e as incluímos como parte do *prompt* para serem utilizadas pelo modelo para gerar a descrição de produto. Nesse sentido, ao fazer isso, estamos propondo como uma etapa do nosso processo a sumarização das informações contidas nas sentenças. Contudo, pela natureza complexa do modelo gerador, não sabemos se essas informações de fato são apresentadas na descrição, e caso sejam, se são apresentadas corretamente.

Assim, consideramos também essa questão interessante para melhor entender os efeitos de utilizar um LLM, e entender se estamos conseguindo sintetizar as informações, como é o nosso objetivo explícito na nossa subquestão de pesquisa *SQP1 (Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?)*. Começamos a explorar um pouco mais essa questão contextualizando o problema.

Modelos baseados em redes neurais profundas para a sumarização abs-

trata podem gerar com alta frequência sumários com inconsistências de fatos, como observados nos estudos de Falke et al. (2019), Zhu et al. (2020), Pagnoni et al. (2021). Possíveis inconsistências são diversas, podendo ocorrer casos como inversão semântica (uma palavra com significado de negação pode inverter complementarmente o sentido da frase), uso incorreto de uma entidade (como trocar o sujeito de uma ação), imprecisão de valores numéricos e alucinações (introdução de uma entidade não presente no documento original). Para uma revisão sobre o assunto de consistência factual no contexto de sumarização abstrata, referimos o trabalho de Huang et al. (2021).

No nosso caso específico, nos preocupamos em avaliar em que medida as informações apresentadas na descrição gerada possuem embasamento nas sentenças passadas para o modelo gerador. Notadamente, nem toda informação presente nas sentenças deverá aparecer na descrição, e devemos considerar também que, por ser um modelo treinado em uma quantidade enorme de texto, o modelo gerador pode possuir conhecimento externo sobre o tipo de produto que está sendo descrito e suas características. Além disso, em muitos casos é possível que o modelo possua conhecimento externo do produto especificamente, ativado a partir do momento que passamos o seu título. Para ilustrar essa ideia, pedimos ao LLM que gerasse um descrição para o tênis apresentado na Figura 5.2, usando como contexto apenas o título do produto.

> *"The Converse Chuck Taylor All Star High Top is a classic and timeless sneaker that has been loved by generations. With its iconic design and high-top silhouette, this shoe offers style and versatility. It features a durable canvas upper, a rubber toe cap, and a rubber outsole for traction. Perfect for casual wear or as a statement piece, the Converse Chuck Taylor All Star High Top will never go out of style"*[1].

Como pode ser observado na Figura 5.2, as características apresentadas na descrição, como silhueta de cano alto, parte superior de lona e biqueira de borracha, são verdadeiras. Assim, destacamos não ser possível afirmar a priori que uma informação que esteja presente na descrição e não esteja nas sentenças seja incorreta. Essa é justamente a diferença entre consistência factual e verificação de fatos, atividades bem parecidas mas ligeiramente diferentes, como bem discutido em Laban et al. (2022). Todavia, para melhor entendermos

[1]Tradução: "O Converse Chuck Taylor All Star High Top é um tênis clássico e atemporal que é amado por gerações. Com seu design icônico e silhueta de cano alto, este sapato oferece estilo e versatilidade. Possui uma parte superior de lona durável, uma biqueira de borracha e uma sola de borracha para tração. Perfeito para uso casual ou como peça de referência, o Converse Chuck Taylor All Star High Top nunca sai de moda."

o processo de geração do texto, é interessante sabermos o quanto da informação presente na descrição é consistente com informações apresentadas na sentenças, e o quanto é inconsistente.

Por esse motivo, exploramos nessa seção algumas métricas quantitativas relacionadas a consistência dos textos gerados. Mais especificamente, optamos por explorar a questão da consistência factual, que busca responder se um sumário gerado está consistente com o texto original (texto de referência). No nosso caso, consideramos como texto de referência o título mais as sentenças selecionadas das avaliações, isto é, o conteúdo do *prompt* passado para o modelo gerador, conforme discutido na Seção 5.3

Dentre as métricas utilizadas, primeiro utilizamos uma métrica tradicional de sumarização para avaliar como as sentenças influenciam o processo de geração de texto. Em seguida propomos duas métricas baseadas em modelos semi-supervisionados desenhados para detectar consistência factual entre sumários e documentos. Para cada uma das métricas, calculamos os resultados para um total de 333 descrições de produtos.

Além disso, afim de enriquecer essa discussão e obter uma melhor compreensão dos resultados obtidos em cada métrica, propomos também algumas descrições alternativas, e apresentamos seus resultados. Dessa forma, nas subseções seguinte, primeiro, apresentamos as descrições comparativas na Seção 7.1.1 e comentamos alguns dos motivos para sua inclusão, assim como algumas expectativas. Em relação às métricas, começamos explorando uma métrica tradicional de sumarização, ROUGE (Lin, 2004), na Seção 7.1.2. Embora um alto valor ROUGE não necessariamente indique consistência factual, conforme discutido em Zhu et al. (2020), consideramos uma métrica pertinente por ser baseado na ocorrência de palavras, possuindo uma boa interpretabilidade.

Em seguida, na Seção 7.1.3, exploramos uma das métricas baseadas em modelos voltadas a classificação da consistência dos textos gerados, mais especificamente a abordagem proposta por Kryściński et al. (2019). Experimentamos o modelo proposto pelos autores, *FactCC*, para calcular se uma dada sentença está factualmente consistente com um documento de referência. Por último, na Seção 7.1.4, exploramos mais uma métrica, proposta por Laban et al. (2022), de detecção de inconsistência, experimentando o modelo *SummaConv* desenvolvido pelos autores.

7.1.1 Descrições alternativas

Além de exibir os resultados obtido em cada métrica para o modelo, exibimos também os resultados obtidos ao comparar algumas descrições alternativas com o mesmo texto de referência discutido anteriormente. Com essas descrições alternativas, buscamos tanto validar as métricas selecionadas quanto melhor compreender os resultados obtidos pelo nosso método a partir da comparação de resultados. Descrevemos a seguir as descrições alternativas utilizadas nessa avaliação comparativa:

- **Descrições aleatórias**. Afim de obter uma limite inferior para as métricas selecionadas, selecionamos para a comparação com o texto de referência uma descrição de produto que fosse pouco relacionada com as sentenças selecionadas. Assim, selecionamos para cada produto avaliado uma descrição de um produto aleatório de uma outra categoria bastante diferente da nossa categoria (*"Clothing, Shoes and Jewlery"*), mais especificamente da categoria *"Eletronics"* do *dataset* discutido na Seção 4.1.
- **Descrições originais**. Da mesma forma que no caso anterior, estamos interessado em estabelecer mais um referencial inferior, porém agora um pouco mais alto. Por isso, selecionamos as descrições originais dos produtos trabalhados. Ainda que as descrições em si não tenham relação com as sentenças selecionadas, por se tratarem do mesmo produto, esperamos o uso de palavras do mesmo domínio, e possivelmente até mesmo informações similares. Contudo, dado as origens diferentes dos textos, e levando em conta também o estudo realizado por Novgorodov et al. (2019) sobre como apenas 45% das sentenças de descrições originais de produto são adequadas para pertencer a uma descrição, esperamos resultados bastante baixos nas métricas utilizadas.
- **Descrições baseadas no título**. Uma vez que estamos utilizando um modelo pré-treinado para gerar as descrições, consideramos interessante comparar em que medida as sentenças selecionadas influenciam o processo de geração de texto. Assim, propomos como um terceira alternativa a descrição de produto gerada apenas com o título, isto é, sem as sentenças extraídas das avaliação. Nesse caso, adaptamos a instrução definida na Seção 6.3, retirando a menção das sentenças e pedindo que o LLM gerasse uma descrição de produto com base apenas no título. Dessa forma, poderemos avaliar em que medida as informações geradas com base ape-

nas no título divergem das informações contidas nas sentenças, e também quanto das sentenças é utilizado.

– **Descrição extrativa**. Por último, propomos uma descrição gerada replicando o método proposto pelo autores Novgorodov et al. (2019), no formato de 5 sentenças, que foi a configuração identificada nesse trabalho como sendo a com melhor resultado na avaliação com avaliadores humanos. Pela natureza do método dos autores, obteremos uma noção da performance de uma abordagem extrativa em que um pequeno pedaço do texto foi selecionado como sumário e que portanto é perfeitamente consistente com o documento de origem. Além disso, justamente por ser um método perfeitamente consistente, adicionamos esse método também como uma validação das métricas e como limite superior.

Assim como no nosso método, geramos cada uma das descrições alternativas para os mesmos 333 produtos, e calculamos as métricas com base nessas descrições. Durante a discussão dos resultados, nos referimos a cada uma das classes de descrições usando o seu título definido acima.

7.1.2 Avaliando a influência do conteúdo com ROUGE

Afim de explorar como a geração de descrição de produtos é afetada pelas informações passadas como conteúdo do *prompt*, avaliamos agora as descrições de produto usando a métrica ROUGE. Ainda que não seja a métrica ideal para avaliar possíveis inconsistências do modelo, conforme observado por Zhu et al. (2020), uma vez que não propõe uma análise semântica do texto, consideramos o ROUGE interessante por ser uma métrica baseada na ocorrência de palavras.

Tradicionalmente utilizado para comparação de um resumo candidato com um resumo de referência, aqui, no nosso caso, propomos a comparação da descrição gerada com o próprio documento de referência, isto é, o conteúdo que compõe o *prompt*, que são as sentenças selecionadas e o título do produto. Assim, pretendemos estabelecer se o texto gerado é de fato influenciado pelas sentenças que foram passadas na instrução e, se sim, em que medida. Consideramos essa questão fundamental antes de propor qualquer avaliação de consistência.

Em relação à métrica ROUGE, as seguintes variações foram utilizadas: (i) o **ROUGE-1**, calculado com base na sobreposição de unigramas (palavras); e (ii) o **ROUGE-2**, calculado com base na sobreposição de bigramas (sequências de duas palavras).

Sobre o uso do ROUGE no comparativo, reforçamos algumas expectativas. No comparativo com as descrições aleatórias, estamos intuitivamente

propondo um limite inferior, para termos uma noção da semelhança mínima de palavras entre uma descrição bastante diferente do texto de referência, formado pelas sentenças adequadas a pertencerem a uma descrição de um produto completamente diferente. Já na comparação com as descrições originais esperamos obter uma ideia da semelhança de palavras entre uma possível descrição e o texto de referência, dado que ambos os textos se referem a um mesmo produto mesmo que a partir de diferentes perspectivas.

Quando comparamos com as descrições baseadas no título, queremos ter uma noção da similaridade entre o texto gerado pelo modelo em um cenário em que a LLM não sofre nenhuma influência das sentenças passadas. Nossa expectativa nesse caso é que *score* obtido seja inferior ao obtido pelo nosso método, e pela diferença de valores poderemos ter uma noção de quanto o modelo generativo é influenciado pelas sentenças que compõem o conteúdo do *prompt*.

Por fim, na comparação com a descrições geradas de forma extrativa, sabemos a priori que todas as palavras contidas no sumário estão contidas no documento original, e também que é um método que se propõe a cobrir uma pequena parte das sentenças selecionadas. Dessa forma, é um método que obtém uma precisão perfeita e baixo *recall*. Reforçamos que com essa comparação estamos interessados apenas em ter uma noção do que seria uma abordagem extrativa que resumisse o texto significativamente, sem estarmos interessados fazer qualquer comparação entre a qualidade dos métodos.

Os *scores* obtidos para cada um dos 5 cenário podem ser observados na Tabela 7.1. Além da média obtida para os 333 produtos, apresentamos também o desvio padrão. Vamos analisar primeiro os dados do ROUGE-1, mais especificamente a precisão. Começamos destacando a similaridade grande que existe no vocabulário de descrições de um mesmo produto - mesmo no caso das descrições originais 40,7% das palavras utilizadas na descrição aparecem em alguma das sentenças. Contudo, no caso descrições de eletrônicos esse percentual é de apenas 15,6%, sugerindo haver uma diferença significativa no vocabulário ao se alterar o domínio do produto.

Interessantemente, as descrições observadas com base no título tiveram uma precisão de 53%, acima das descrições originais. Isso indica que o modelo gerador é capaz de se adequar bem ao vocabulário do produto. Já no caso do no nosso, 72% das palavras presentes na descrição podem ser encontrados no documento de referência, indicando uma alta similaridade de palavras. Ao comparar nossa método com as descrições geradas com base no título, observamos uma diferença significativa, de 19,1%. A partir dessa diferença observamos que o nosso método está sendo bastante influenciado pela sentenças, que é um

requisito mínimo para consistência.

Tabela 7.1: ROUGE entre cada classe de descrição e conteúdo do *prompt*.

Método	ROUGE-1			ROUGE-2		
	Precision	Recall	F1	Precision	Recall	F1
Extrativo	**1.00** ± 0.00	0.24 ± 0.04	0.38 ± 0.06	**1.00** ± 0.00	**0.23** ± 0.04	**0.37** ± 0.06
Proposto	0.72 ± 0.07	**0.35** ± 0.06	**0.47** ± 0.06	0.31 ± 0.08	0.15 ± 0.04	0.20 ± 0.05
Título	0.53 ± 0.07	0.23 ± 0.04	0.32 ± 0.04	0.15 ± 0.05	0.07 ± 0.02	0.09 ± 0.03
Original	0.41 ± 0.15	0.12 ± 0.08	0.17 ± 0.09	0.07 ± 0.09	0.02 ± 0.02	0.03 ± 0.02
Aleatório	0.16 ± 0.10	0.07 ± 0.07	0.08 ± 0.07	0.01 ± 0.02	0.00 ± 0.01	0.00 ± 0.01

Por outro lado, observamos que o *recall* é limitado em todos os casos. Nesse sentido, recordamos que o *recall* é proporcional ao tamanho do texto: um resumo que contenha apenas metade das palavras do texto original poderá alcançar no máximo um *recall* de 50%. É isso, por exemplo, que justifica o baixo *recall* das descrições geradas pelo método extrativo.

Analisando cada cenário, observamos primeiro que as descrições originais possuem um *recall* baixo, de apenas 12%, indicando que grande parte da informação contida nas sentenças não é coberta. Isso reforça, em certa medida como as sentenças extraídas de uma avaliação são uma fonte rica de informação, na medida que contém perspectivas únicas não cobertas pelos anunciantes.

Já no caso das descrições baseadas no título, há um aumento do *recall*, atingindo 23%, justificado em parte pela repetição do título do produto (em pelo menos 96,0% das descrições geradas por esse método o título é repetido de alguma forma no texto). Ao comparar com as descrições do nosso método, verificamos um aumento de 12,6% do *recall*, atingindo 35,5%. Quando comparamos os tamanhos, verificamos que na média as descrições geradas pela nossa metodologia contém 49,8% da quantidade de palavras que o texto de referência, enquanto no caso das descrições geradas com base apenas no título, esse percentual é de 44,0%. Sob essa perspectiva, verificamos ser um aumento considerável do *recall*, na medida que as novas palavras que estão sendo adicionadas contribuem significativamente no aumento da cobertura do texto de referência.

Além disso, notamos que o *recall* do nosso método foi maior que o obtido pelo método extrativo. Ainda que isso seja esperado, dado que esse método é limitado a 5 sentenças, esse resultado confirma que uso maior de palavras se reflete em uma maior cobertura do conteúdo das sentenças. Isso indica, em certa medida, que o nosso método parece cobrir uma quantidade de informação considerável do texto de referência.

Analisando agora os resultados obtido com o ROUGE-2, verificamos que com exceção do método extrativo, a precisão e *recall* em todos cenários naturalmente caem bastante. Por outro lado, é interessante notar que enquanto para as descrições originais a quantidade de bigramas em comum é muito baixa

(7%), no nosso método ainda existe uma quantidade relevante de bigramas em comum com o documento de origem. No caso, a precisão dos bigramas, mesmo sendo um método abstrato, é de 31,0%, ou seja, grande parte dos bigramas encontrados nas descrições geradas pelo método podem ser encontrados no documento de origem. Isso é o dobro da observada no texto gerado apenas com o título (15%), ressaltando novamente a influência que as sentenças exercem na geração de texto.

7.1.3 Avaliando a consistência factual com o modelo FactCC

Avaliada a questão de similaridade de palavras, a primeira métrica de consistência factual é baseada em um modelo que checa a consistência factual (FactCC, do inglês *fact consistancy checking*), e foi proposto no trabalho de Kryściński et al. (2019). O FactCC é um classificador que tem como objetivo atestar a consistência de uma sentença em relação a um documento original. Conforme os autores observam, o problema de consistência factual é intimamente relacionado com o problema de inferência de linguagem natural (*natural language inference*, NLI), que foca em classificar a implicação lógica entre pares de sentenças, com a diferença de que a consistência factual requer incorporar o contexto inteiro do documento de referência. Com isso em mente, os autores propõem um abordagem de comparação sentença-documento. No nosso contexto de comparação entre resumo e texto original, executamos a verificação individual de cada sentença contida na descrição com o corpo inteiro do texto de referência. A métrica, no nosso estudo, surge como o percentual de sentenças classificadas como factual.

Comentando um pouco sobre o trabalho, e devido a ausência de *datasets* supervisionados na área de consistência factual, os autores constroem um novo *dataset*, gerando dados de treino sintéticos a partir do *dataset* de sumarização CNN/DailyMail Nallapati et al. (2016). Para isso, sorteiam sentenças dos seus respectivos documentos de origem e conduzem uma série de transformações de texto pré-definidas. Entre essas transformações, incluem o parafraseamento, a mudança de entidades e valores numéricos, a mudança de pronomes, a negação de sentenças e a injeção de ruídos. As sentenças que sofreram mudanças semanticamente invariante foram rotuladas como positivas, e as variantes negativas, de forma que os autores conseguem construir um *dataset* de treino a baixo custo. Com o novo *dataset*, os autores propõem um classificador, chamado FactCC, implementado a partir do *finetuning* do BERT (Devlin et al., 2018) na atividade de classificação binária para rotular as sentenças em consistentes ou inconsistentes.

Para avaliação, os autores apresentam um conjunto de validação e teste, a partir de resumos gerados por modelos que são o estado da arte para a tarefa de sumarização. Para cada resumo, os autores quebram os textos em sentenças e avaliam manualmente cada par sentença-documento, rotulando cada par como consistente ou não. Geram assim conjuntos com 931 e 503 exemplos respectivamente, e propõem validar e avaliar o modelo nesse conjunto. Como resultado, alcançam uma acurácia balanceada por classe de 74,15% no conjunto de teste, e um pontuação F1 de 0,5106. Assim, superam muito modelos também baseados no BERT mas treinados em *datasets* de inferência textual.

Contudo, como principal limitação, destacamos que o modelo foi treinado em um *dataset* artificial, conforme detalhado acima. Dessa forma, é importante entender que o modelo é limitado ao conjunto de transformações textuais pré-estabelecidas pelos autores.

Por ser um abordagem de comparação sentença-documento, para executar o modelo na nossa avaliação, o primeiro passo foi quebrar as descrições geradas em sentenças. Em seguida, executamos o modelo para cada par sentença-documento. Calculamos então algumas estatísticas sobre as probabilidades retornadas pelo modelo para cada classe, apresentado na Tabela 7.2. Nessa Tabela, exibimos na penúltima coluna o percentual de sentenças classificadas como positivas, usando como limiar de decisão o valor de 0,5 para a classificação binária, e na última coluna o percentual de descrições em que todas as sentenças foram classificadas como consistentes.

Tabela 7.2: Consistência factual entre descrições e conteúdo do *prompt* calculada pelo modelo FactCC.

Método	Média	% de sentenças consistentes	% de descrições com todas sentenças consistentes
Extrativo	0.999 ± 0.023	99,9	99,7
Proposto	0.917 ± 0.254	91,8	65,2
Título	0.903 ± 0.269	90,5	64,9
Original	0.627 ± 0.442	62,6	33,0
Aleatório	0.464 ± 0.449	45,0	13,2

Conforme pode ser observado pelos valores médios, existe uma diferença bem grande de sentenças consistentes quando comparamos as descrições aleatórias com as descrições originais, e uma diferença maior ainda quando comparamos as descrições humanas com as descrições geradas pelo GPT 3-5 Turbo (que são as descrições geradas pelo método proposto e baseadas no título).

Contudo, a diferença no *score* médio quando examinamos os dois conjuntos de descrições geradas pelo modelo pré-treinado é pequena - o que surpreende quando lembramos que um dos conjuntos não teve acesso as sentenças selecionadas das avaliações, e sim apenas ao título.

Refletindo sobre esses resultados, consideramos uma possibilidade o cenário em que, por possuir amplo conhecimento externo, as descrições baseadas no título sejam coerentes com as sentenças selecionadas das avaliações, no sentido de não contradize-las. As sentenças geradas nesse caso seriam então na verdade sentenças neutras, enquanto o modelo FactCC realiza uma classificação binária. Pela forma como o *dataset* foi construído, em que instâncias negativas de treino foram artificialmente geradas a partir de transformações de texto, seria possível que sentenças neutras não fossem consideradas inconsistentes.

De qualquer forma, verificamos que a inclusão de mais palavras compartilhadas com o documento original, verificada quando utilizamos o ROUGE (Seção 7.1.2), não parece ter gerado mais inconsistências. Isso indica que pelo menos as transformações textuais de inversão semântica, propostas pelos autores na hora de gerar o *dataset* em que o modelo foi treinado, não estão ocorrendo, caso contrário seriam apontadas pelo classificador.

Por último, frente a esses resultado, questionamos a qualidade do modelo, isto é, se as sentenças estão sendo classificando corretamente como inconsistentes ou consistentes e deixamos como um possível trabalho futuro se aprofundar nessa questão. Assim, decidimos buscar mais uma forma de avaliar a consistência das descrições.

7.1.4 Avaliando a consistência factual com o modelo SummaC

Como última métrica, apresentamos o modelo chamado SummaC, proposto por Laban et al. (2022). Diferentemente do modelo FactCC discutido na Seção 7.1.3, esse modelo propõe retornar um valor do quão consistente o resumo como um todo é em relação ao documento de origem.

Para isso, os autores novamente ressignificam o uso de NLI na atividade de detecção de inconsistência, chamando atenção para a diferença de granularidade em cada atividade. Enquanto nos *datasets* de NLI a granularidade comparada é tradicionalmente no nível de sentenças, na atividade de detecção de inconsistência factual a comparação se dá no nível de documento.

Laban et al. (2022), propõem então um novo método chamado SummaC que habilita modelos de NLI para a atividade de detecção de inconsistências. Fazem isso segmentando tanto o resumo quanto o seu documento de origem em blocos menores, e executam para cada par o modelo de NLI, obtendo uma distribuição de probabilidade para as categorias implicação, contradição e neutralidade. Em seguida, constroem uma matriz com o resultado de cada par e realizam operações para obter um *score* entre 0 e 1 para o resumo como um todo, em que quanto mais próximo de 1 mais consistente o resumo em

relação ao documento de origem.

Os autores experimentam uma série de configurações, tanto em relação ao nível de granularidade dos blocos de texto quanto em relação uso das probabilidades e operações realizadas na matriz. Em relação ao nível de granularidade, obtém melhores resultados quebrando o resumo em sentenças e o documento de origem em sentenças ou pares de sentenças. Já em relação à matriz formada pela execução do modelo para cada par, obtém melhores resultados executando uma transformação na matriz e realizando uma operação de convolução, exigindo portanto o treinamento de parâmetros. Por último, verificam que o melhor resultado é obtido quando usando apenas a probabilidade de implicação, descartando portanto as probabilidades de contradição e neutralidade.

Além do modelo, os autores propõem também um novo *benchmark* ao padronizar 6 grandes *datasets* para a atividade de classificação binária. Para avaliar o modelo, definem para cada *dataset* um limiar para separar a classe consistente da inconsistente a partir de um conjunto de validação. Com o novo *benchmark*, comparam o novo método com uma série de alternativas e métodos do estado da arte, incluindo o classificador proposto no trabalho de Kryściński et al. (2019), comentado na Seção 7.1.3.

Obtém a maior performance geral na métrica proposta de acurácia balanceada, atingindo um valor de 74,4%, 5 pontos percentuais acima do modelo com segunda melhor performance, indicando uma melhoria com um intervalo de confiança de 99%. Além disso, na segunda métrica proposta pelos autores, ROC-AUC, o método alcança um *score* geral de 77,8%, novamente mais de 5 pontos percentuais acima do segundo melhor modelo e também indicando uma melhoria com 99% de confiança.

Como limitações, os autores destacam a baixa interpretabilidade do modelo com melhor resultado e que os 6 datasets que compõem o *benchmark* proposto possuem resumos do domínio de notícias, havendo portanto espaço para outros domínios menos frequentes no contexto de sumarização automática. Além disso, destacamos o fato do método retornar um valor entre 0 e 1, não sendo diretamente interpretável se o resumo é consistente ou não.

Comentaremos agora os resultados obtidos, apresentados na Tabela 7.3. Primeiramente, apenas como uma validação, destacamos que as descrições extrativas tiveram um resultado praticamente igual a 1, conforme esperado, enquanto todas as demais classes de descrições tiveram resultados médios abaixo de 0.5. Vamos agora comparar com os resultados das demais descrições alternativas, lembrando que estamos interessados principalmente na diferença entre eles.

Tabela 7.3: Consistência factual entre descrições e conteúdo do *prompt* calculada pelo modelo SummaConv.

Método	Média	Mínimo	Máximo
Extrativo	0.988 ± 1,1	0.928	1.00
Proposto	0.440 ± 0.140	0.244	0.944
Título	0.304 ± 0.057	0.227	0.743
Original	0.272 ± 0.69	0.20	0.817
Aleatório	0.246 ± 0.036	0.188	0.611

Primeiro, observamos como as descrições geradas com base no título estão apenas ligeiramente acima das descrições originais. Isso é relevante ao recordarmos dos resultados obtidos com a métrica ROUGE, em que constatamos um aumento da precisão de 12,0% nos unigramas na comparação entre as duas classes. Entretanto, esse aumento não resulta em um aumento significativo do *score* obtido pelo modelo, que é apenas de 0.032. Esse resultado é ainda mais interessante quando levamos em conta o *score* médio obtido com as descrições aleatórias, de 0.246, com uma diferença de apenas 0.026 para o caso das descrições originais, indicando então que as descrições geradas com base no título estão bem próximas do mínimo esperado ao se comparar descrições de produtos diferentes. Isso é interessante no sentido que valida a métrica, ao mostrar que um aumento da interseção de palavras com o texto de referência não indica um aumento do *score* calculado pelo modelo.

Com isso em mente, ao comparar os resultados entre as descrições do nosso método e as descrições geradas com base no título, verificamos um aumento bem mais significativo da consistência, dessa vez de 0.136. Com isso, concluímos que o aumento do ROUGE na comparação entre as duas classes, discutido anteriormente, parece ter sido acompanhando de ganhos de consistência. Deixamos a questão de melhor dimensionar o quanto de consistência representa esse aumento de *score* como uma questão a ser investigada em trabalhos futuros, possivelmente comparando o nosso método com mais descrições alternativas que representassem outros limites superior.

7.1.5 Conclusão e discussão

Para encerrar nossa análise sobre a consistência das descrições geradas, reforçamos alguns pontos e indicamos alguns pontos de discussão. Primeiramente, conforme discuto anteriormente, nos debruçamos em avaliar se as sentenças são utilizadas no processo de geração de sentenças, e de que forma. A partir da análise ROUGE, pudemos responder essa questão, obtendo uma noção entre a semelhança de vocabulário.

Em seguida, exploramos um modelo de consistência factual (FactCC),

que sugeriu um alto score de sentenças consistentes no caso de textos gerados pelo modelo pré-treinado, mesmo quando este não teve acesso ao documento de origem. Isso indica, em certa medida, que o modelo não é capaz de avaliar com profundidade textos gerados pela LLM utilizada. De qualquer forma, nesse caso, parece ser necessária uma investigação mais profunda.

Por último, experimentamos mais um modelo, o SummaC, que indicou uma diferença de consistência significativa entre as descrições geradas pelo nosso método e as geradas com base no título apenas. Isso indica, pelo menos parcialmente, que parte das palavras em comum com o texto de referência está sendo utilizada de forma consistente.

Abrindo a discussão, consideramos interessante explorar outras métricas de consistência factual, ressaltando ser uma área bastante desafiadora e agitada pelo aparecimento de novos modelos de sumarização abstrata, como LLMs, conforme discutido em Celikyilmaz et al. (2020). Além disso, outra linha interessante seria a de checar a veracidade dos fatos apresentados na descrição gerada. Pela natureza do modelo gerador, sabemos que conhecimento externo pode ser incluído na descrição, e ainda que isso possa enriquecer a descrição, seria interessante verificar esse conhecimento.

Por último, resgatamos a subquestão *SQP1 (Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?)*, indicando que o nosso método parece sim ser capaz em sintetizar as informações contidas nas avaliações. Contudo, afim de examinar mais profundamente essa questão, principalmente em relação a parte de legibilidade e informatividade, propomos avaliar qualitativamente as descrições.

7.2 Avaliando a qualidade das descrições

Agora com o objetivo de avaliar a qualidade das descrições geradas conduzimos um experimento com 30 avaliadores humanos em que propomos a comparação dos textos com as descrições originais de 150 produtos. Mais especificamente, utilizamos uma escala Likert de 7 pontos para comparar as descrições nas mesmas dimensões que as propostas no trabalho de Novgorodov et al. (2019), isto é, legibilidade, objetividade, informatividade, relevância para o produto e uma comparação geral de preferência. Com essa investigação, pretendemos responder a nossa subquestão de pesquisa, *SQP2 (Como se comparam as descrições geradas com as descrições originais?)*.

Além do nosso método, reproduzimos o método proposto por Novgorodov et al. (2019), gerando descrições extrativas formadas por 5 sentenças, que foram as que obtiveram melhor resultado nas avaliações dos autores, e também as

comparamos com as descrições originais. A partir da comparação desses dois métodos com um denominador comum, as descrições originais, propomos a comparação entre as duas metodologias. Na medida que esse é um método que também utiliza um as avaliações de usuários como fonte de informações para gerar as descrições, concatenando as sentenças selecionadas, a partir dessa comparação com o nosso método pretendemos melhor entender os efeitos de utilizar um LLM no processo de geração. Assim, essa comparação entre as duas metodologias esclarece a questão levantada pelo nossa subquestão de pesquisa, *SQP1 (Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?).*

Para facilitar a leitura, apresentamos na Subseção 7.2.1 o objetivo da avaliação, e discutimos as técnicas empregadas. Na Subseção 7.2.2 discutimos a metodologia da avaliação e apresentamos o seu formato. Por último, apresentamos os resultados na Subseção 7.2.2.

7.2.1 Objetivo

Para avaliar as descrições geradas por nossa metodologia, optamos por utilizar uma escala Likert (Amidei et al., 2019) para avaliar os textos gerados. Conforme estudo conduzido por Amidei et al. (2019) escalas de classificação, e mais especificamente escalas Likert são um método de avaliação amplamente utilizado no contexto de geração de texto natural. No caso, os autores verificaram que 135 trabalhos na área de GLN, 44% usaram escalas Likert. Citando alguns exemplos de trabalhos na própria área de geração de descrições de produto, podemos mencionar Novgorodov et al. (2019), Wang et al. (2017), Elad et al. (2019), Nguyen et al. (2021), e, assim como esses trabalhos, propomos uma avaliação que mensurasse os textos em múltiplas dimensões. No nosso trabalho, seguindo a recomendação de van der Lee et al. (2019), utilizamos uma escala Likert de 7 pontos.

Todavia, em nossa avaliação, optamos por seguir uma linha ligeiramente diferente da maioria dos trabalhos discutidos na Seção 3.1. Enquanto a maior parte deles optam por avaliar as descrições geradas isoladamente ou comparar com descrições geradas por métodos alternativos, como em Elad et al. (2019), Chen et al. (2019), optamos por seguir a mesma linha proposta por Wang et al. (2017) e preferimos comparar as descrições geradas por nosso método com as descrições originais postadas pelos anunciantes. Assim, ao estabelecer um referencia, poderemos obter uma compreensão mais tangível do que apenas um valor em uma escala de 1 a 5, por exemplo. Além disso, podemos obter também uma melhor compreensão sobre a substituição das descrições originais

pelas geradas por nossa metodologia, o que seria uma possível aplicação dos resultados da nossa metodologia.

Nessa mesma linha, e com o intuito de enriquecer o nosso trabalho, optamos também por comparar as descrições geradas por Novgorodov et al. (2019) com as descrições originais. Assim, como ambos os métodos se baseiam no uso de informações identificadas nas avaliações deixadas pelos usuários, poderemos avaliar como o uso de um LLM contribui para a atividade de geração de descrição de produtos. Do mesmo modo, poderemos avaliar em que medida conseguimos mitigar alguns dos problemas identificados pelo autores, sobretudo relacionados a legibilidade e informatividade - dimensões que obtiveram os piores resultados nas avaliações dos autores.

Contudo, na hora de estruturar uma avaliação com anotadores humanas no contexto de GLN, algumas ressalvas devem ser feitas. van der Lee et al. (2021) aponta a falta de questionários padronizados, assim como de uma nomenclatura padronizada para avaliar as diferentes qualidades de um texto. Como recomendação, apontam para o uso de diretrizes comum e compartilhadas. Na sua ausência, sugerem a definição explícita dos critérios mensurados, assim como o destaque de possíveis interseções entre eles. Nesse sentido, uma definição formal, assim como a apresentação de exemplos ajudam os participantes.

No seu trabalho, Novgorodov et al. (2019) propõem avaliar com anotadores as descrições geradas em 4 critérios, são eles: legibilidade, informatividade, objetividade e relevância do conteúdo para descrição do produto. Ainda que tenham sido apresentados exemplos para os anotadores ilustrando e desambiguando cada um dos conceitos, esses exemplos não foram disponibilizados pelos autores, dificultando sua reprodução. Do mesmo modo, não foram referenciadas ou propostas nenhuma definição dos 4 critérios. Nesse sentido, optamos por usar as mesmas dimensões que Novgorodov et al. (2019), mas agora trazendo definições para cada uma delas. Propomos então as seguintes definições, que foram apresentadas para os avaliadores em formato de afirmações[2]:

– **Legibilidade**: *"When comparing to the reference description, this description is more readable, being easier to read and understand."*

[2]Tradução das afirmações: "Quando comparada com a descrição de referência, esta descrição é mais legível, sendo mais fácil de ler e compreender.", "Quando comparada com a referência, esta descrição apresenta-se de forma mais objetiva, sendo mais sucinta e menos repetitiva.", "Quando comparada com a referência, esta descrição é mais informativa, pois apresenta mais informações e detalhes.", "Ao comparar com a referência, as informações apresentadas nesta descrição são mais relevantes para o produto, pois apresentam mais detalhes sobre características importantes.", "No geral, prefiro esta descrição do produto em comparação com a referência."

- **Objetividade**: *"When comparing to the reference, this description presents itself in a more objective way, being more succinct and less repetitive."*
- **Informatividade**: *"When comparing to the reference, this description is more informative, as it presents more information and details."*
- **Relevância para o produto**: *"When comparing to the reference, the information presented in this description is more relevant to the product, as it presents more details about important features."*
- **Preferência geral**: *"Overall, I prefer this product description when comparing to the reference."*

Discutimos a seguir como estruturamos nossa avaliação.

7.2.2 Formato da avaliação

Conforme discutido na seção anterior, optamos por avaliar as descrições geradas a partir da comparação com as descrições originais. Dessa forma, cada unidade de avaliação é composto por um par de descrições, sendo uma delas a descrição original e a segunda delas uma descrição alternativa que foi gerada ou por nossa metodologia ou pela metodologia proposta pelos autores (Novgorodov et al., 2019).

Ainda que tenhamos adotado a definição do que é uma descrição de produto proposta por Novgorodov et al. (2019), não nos propomos a definir o que é uma boa descrição. De fato, mesmo dentro do conceito proposto pelos autores, entendemos que descrições podem ter formatos muito diversos, tanto em relação a tamanho, nível de detalhe e estilo de escrita. Por isso consideramos importante deixar os avaliadores expressarem suas próprias preferência da forma menos controlada possível, de modo que não demos orientações nem exemplos do que seria uma boa descrição. Além disso, pelo mesmo motivo, consideramos importante uma quantidade grande de avaliadores.

A avaliação de cada par foi feita no seguinte formato. Primeiro apresentamos no topo da página a descrição original de um produto, acompanhada de seu título e imagem. Em seguida é apresentada abaixo uma descrição sob o título de descrição alternativa, junto de uma escala Likert de 7 pontos com as sentenças comparando as descrições, conforme apresentada na Seção 7.2.1. É solicitado então que o avaliador expresse o seu grau de concordância com cada afirmação, em uma escala que varia de *"Strongly disagree"* ("Discordo fortemente") para *"Storngly agree"* ("Concordo fortemente.").

Por último, exibimos uma escala de 7 pontos para o anotador marcar o grau de confiança em sua avaliação, variando de *"Very insecure."* ("Muito inseguro.") até *"Very confident."* ("Muito confiante"). Essa ultima pergunta era sinalizada como opcional. Exibimos um exemplo tela apresentada para o avaliador na Figura 7.1.

Figura 7.1: Tela apresentada ao avaliador.

Com base em uma primeiro teste piloto, consideramos como 10 um número razoável de comparações a serem avaliadas, resultando em uma duração de cerca de 25 minutos para completar a avaliação. Discutimos abaixo alguns pontos relacionados a distribuição de comparações para cada avaliador.

Um ponto de atenção na hora de desenhar o experimento é em relação ao uso de duas metodologias diferentes. No nosso experimento, estamos sempre comparando a descrição gerada com a descrição original, que são, conforme

discutido na Seção 6.2, bastante variadas. Além disso, não aferimos a qualidade da descrições originais nem propomos nenhum tipo de seleção, justamente por evitar definir o que é uma boa descrição. Assim, e recordando ainda a análise conduzida sobre sentenças extraídas de descrições de produto realizada por Novgorodov et al. (2019), em que os autores verificaram que 45% das sentenças não eram adequadas para pertencer a uma descrição (sendo o principal motivo anúncios de marketing com, 20% dos casos), é bastante razoável assumir que dentro do conjunto de descrições originais vão existir descrições de qualidade variadas.

Por esse motivo, para que a comparação fosse justa, consideramos fundamental que a comparação de cada método fosse com os mesmos produtos, para que fossem comparadas com as mesmas descrições originais. Seguida dessa decisão, optamos por apresentar para cada anotador em ordem aleatória a mesma quantidade de comparações usando o nosso método e o método de referência proposto pelo autores, isto é, cada anotador avaliou 10 pares de descrições, sendo 5 pares compostos pelo nosso método e pela descrição original e 5 pares compostos pelo método dos autores e pela descrição original. Além disso, impedimos que as duas comparações de um mesmo produto fossem sorteadas para um mesmo anotador, pois julgamos que dessa forma a primeira poderia enviesar a seguinte.

Conduzimos o experimento com um total de 30 anotadores, que foram voluntários a participar no trabalho. Julgamos ser uma quantidade suficiente de pessoas para que diferentes perspectivas fossem absorvidas na avaliação, dando portanto mais robustez aos resultados. Contudo, logicamente que um número maior de avaliadores enriqueceria ainda mais os resultados. Por se tratar de uma atividade complexa, foi definido um perfil alvo para os anotadores. Primeiramente, por se tratar de textos em inglês, os anotadores precisam serem capazes de ler e compreender textos na língua inglesa. Em segundo lugar, consideramos que a experiência com descrições de produtos como fundamental, justamente por entendermos que uma boa descrição é um conceito amplo e aberto. Por isso, selecionamos pessoas com o habito de fazer compras online. Além disso, outro ponto importante foi não repetir nenhum avaliador que já tivesse participado da etapa de avaliação conduzida para configurar a nossa solução, discutida na Seção 6.4, de modo que os anotadores não foram informados de qual era o método proposto pela nossa pesquisa.

7.2.3 Resultados

Antes de apresentar os resultados da avaliação, consideramos interessante considerar alguns cenários. Uma vez que avaliamos as descrições geradas sempre com base nas descrições originais postadas pelos anunciantes, vale uma discussão preliminar sobre como esse resultado pode influenciar na comparação indireta das duas metodologias que propomos avaliar, a nossa e a replicada dos autores Novgorodov et al. (2019). Assim, discutimos primeiro 3 cenários possíveis para os resultados que se aplicam a cada uma das dimensões das descrições avaliadas.

O primeiro cenário seria o caso em que ambos os métodos superam consistentemente as descrições originais. Nesse caso, não conseguiríamos comparar ambas as metodologias, de forma que o ponto de referência, no caso as descrições originais, estaria acima de ambos os métodos. Por outro lado, o segundo cenário, e também menos interessante, é o caso em que ambos os métodos não conseguem superar as descrições originais, de forma que novamente não conseguiríamos comparar as duas metodologias entre si. Por último, e mais interessante, é o caso em que uma metodologia apresenta resultados mais positivos em relação a descrição original do que a outra. Nesse caso, a descrição original pode sim ser utilizada como referencial e as duas metodologias podem ser de fato comparadas entre si.

Com isso em mente, apresentamos os resultados. Exibimos na Figura 7.2 o percentual de respostas para cada uma das afirmações da escala Likert. O primeiro ponto que destacamos da figura é que a soma do percentual de comparações em que os avaliadores concordaram com a afirmação proposta foi superior aos que discordaram em todas dimensões no caso nosso método. Já o mesmo não é verdade no caso do método extrativo, em que nas dimensões informatividade e relevância para o produto houve mais discordância que concordância.

Analisando agora os resultados de cada dimensão, destacamos primeiro como ambas as metodologias foram preferidas no geral, mas com margens bastantes diferentes. De fato, para mais de 62% dos produtos os avaliadores concordaram (ao usar concordaram, estamos nos referindo ao caso de *"Agree."* somado com *"Strongly Agree"*) que a descrição gerada pelo nosso método foi preferida em relação a original, enquanto 14% pelo menos discordou (igualmente, nesse caso estamos nos referindo a *"Disagree"* somado de *"Strongly disagree"*). Na comparação da metodologia extrativa, verificamos uma diferença significativa, uma vez que os avaliadores concordaram com a preferência para apenas 37% dos produtos, e discordaram em 36% dos casos.

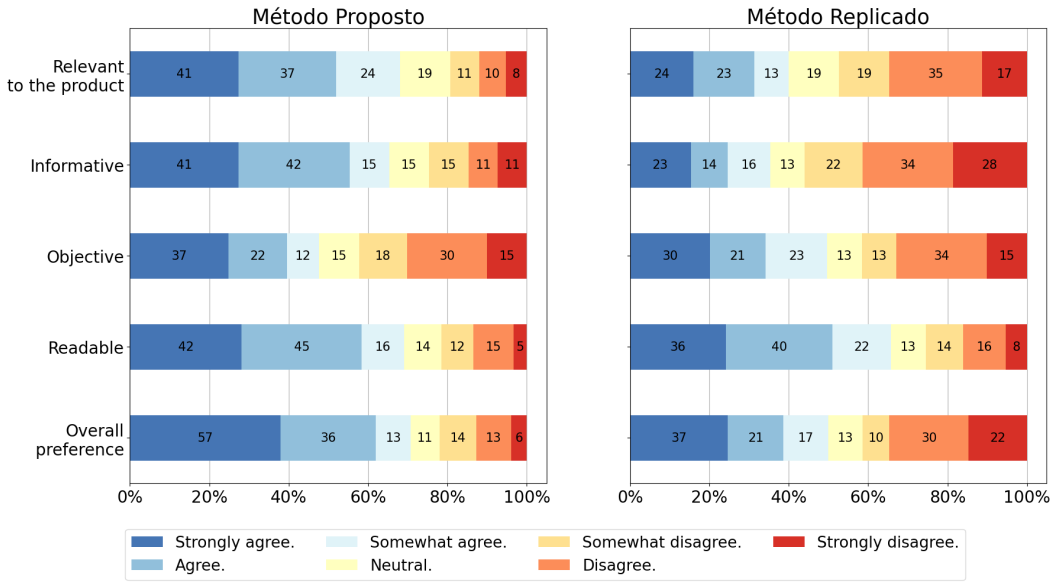


Figura 7.2: Resultados da avaliação qualitativa conduzida com 30 anotadores em 150 produtos.

Já no quesito legibilidade pareceu haver significativa preferência pelas duas metodologias na comparação com as originais, e de forma mais equilibrada. Para 59% das descrições os avaliadores concordaram com a maior legibilidade dos textos gerados pelo nosso método, enquanto apenas 14% discordou. No caso das descrições extrativas, os avaliadores também preferiram as descrições alternativas com ampla margem, concordando que 50% das descrições foram mais legíveis, e discordando em apenas 16% dos casos.

Quando analisamos a dimensão objetividade, ambas as metodologias foram preferidas em relação a original, mas com pouca margem nos dois casos. De fato, essa foi uma dimensão em que a preferência pela descrição alternativa foi menor nos dois casos, com os avaliadores concordando com a afirmação proposta no nosso método em 41% dos casos, e discordando em 30% dos casos. No caso da alternativa extrativa, os avaliadores concordaram com a afirmação em 34% das comparações e discordaram em 33%. Nessa dimensão, ressaltamos que novamente a preferência pela alternativa abstrata gerada pelo nosso método foi maior que a preferência pelo alternativa extrativa, ainda que essa tenha um número de sentenças pré-determinado de apenas 5 sentenças, sendo no geral menores às descrições que nós geramos (nossas descrições apresentaram uma média de 94 palavras, contra uma média de apenas 40 palavras no casos das descrições extrativas).

Ao analisar os resultados das dimensões informatividade e relevância para o produto, observamos resultados bem contrastantes para cada método. Em relação ao nosso método, novamente observamos uma tendência clara

de preferência, com os avaliadores concordando com as afirmações em 56% e 51% dos casos, respectivamente, enquanto a discordância foi apenas de 14% na dimensão informatividade e 12% na dimensão relevância para o produto. Por outro lado, na comparação do método extrativo verificamos justamente o oposto nessas duas dimensões, com os avaliadores preferindo as descrições originais. Na informatividade, 42% foram casos de discordância da afirmação, contra 24% de concordância, e na relevância para o produto foram 36% de discordâncias contra 29% de concordâncias.

Uma vez apresentados os resultados da avaliação, destacamos como, para a maioria das dimensões analisadas, nos encontramos no terceiro cenário considerado, em que a metodologia proposta nesse trabalho pareceu alcançar resultados mais positivos que a metodologia extrativa na comparação com a descrição original.

Por esse motivo, propomos então conduzir uma análise comparando as duas metodologias, em que por meio de testes estatísticos verificaremos se existe de fato uma diferença significante entre os métodos, e se sim, em que quesitos.

7.2.4 Comparação entre Metodologias

Afim de comparar as duas metodologias aqui avaliadas, conduzimos um teste de significância estatística comparando os resultados obtidos por cada método. Com esse teste, pretendemos verificar em que dimensões as diferenças observadas na são de fato estatisticamente significantes, e se podemos de fato fazer algum tipo de afirmação comparando as duas metodologias.

Para realizar os testes, primeiro tivemos que converter as respostas textuais em uma escala numérica de 1 a 7, em que o 1 representa o "Strongly disagree." ("Discordo fortemente".) e o 7 representa o "Strongly agree." ("Concordo fortemente"). Aplicamos o teste de *Mann-Whiteney U* (Nachar et al., 2008) para comparar a diferença das respostas entre cada método, lembrando que para cada questão proposta possuímos um total de 150 avaliações referentes a cada método, e adotamos um nível de significância de 0,05.

Uma vez que estamos interessado se o nosso método foi mais preferido do que o de referência nas dimensões propostas, consideramos adequados estabelecer a hipótese alternativa como sendo unilateral. Consideramos assim, apresentamos a hipótese nula e a hipótese alternativa:

– Hipótese nula: As descrições geradas por ambos os métodos foram igualmente preferidas na comparação com as descrições originais

– Hipótese alternativa: Na comparação com as descrições originais, as descrições geradas pelo nosso método foram mais preferidas do que as descrições geradas pelo método de referência.

Apresentamos os resultados da comparação na Tabela 7.4. De acordo com o teste, notamos como o nosso método foi mais preferido que o método de referência com significância estatística em 3 das 5 dimensões, sendo elas informatividade, relevância para o produto e preferência geral. Já nas outras dimensões observamos que o p-valor não foi pequeno suficiente para afirmarmos com 95% de confiança haver significância estatística.

Tabela 7.4: Resultados do teste *Mann-Whiteney U* para pares combinado. Destacamos em negritos os p-valores inferiores a 0,05, indicando a rejeição da hipótese nula.

Dimensão	Mediana			*p-valor*
	Nosso método	Referência	Diferença	
Legibilidade	6	6	0	0,1226
Objetividade	4	4	0	0,28006
Informatividade	6	3	3	**0,00000**
Relevância para o produto	6	4	2	**0,00000**
Preferencia geral	6	5	1	**0,00002**

Em posse desses resultados, respondemos a nossa subquestão de pesquisa *SQP1 (Como sintetizar as informações contidas em avaliações de usuário em uma descrição de produto legível e informativa?)*. Comparando ambos os métodos, lembramos que nós propomos usar um LLM na etapa de geração da descrição, ao contrário do método de referência que apenas concatena as sentenças selecionadas das avaliações de produto. Ao verificar que nosso método gera de fato descrições mais informativas sem prejudicar a legibilidade, confirmamos o sucesso da nossa abordagem.

8
Conclusões

Neste trabalho, contextualizamos o problema da geração da descrição de produtos de forma automática, apresentando os principais trabalhos na área, e propomos um novo método. Nosso método é inovador na medida que combina a riqueza de informações contida nas avaliações do produto deixadas por usuários com a capacidade de geração de texto de um LLM de forma *zero-shot*. Para fazer isso, propomos um método que extrai sentenças de avaliações, seleciona uma determinada quantidade dentre as extraídas e inclui as sentenças como conteúdo do *prompt* para o LLM, que é instruído então a gerar uma descrição com base nas sentenças.

Para configurar nosso método, realizamos uma série de experimentos afim de definir algumas de suas etapas, inclusive experimentando qual instrução utilizar no processo de geração de texto e também quantas sentenças incluir como conteúdo do *prompt*. Ao avaliar nosso método, investigamos primeiro a questão de como as descrições geradas se relacionam com as sentenças selecionadas, e nos debruçamos também sobre a questão da consistência factual, uma área bastante desafiadora frente aos novos avanços na área de geração de texto. Em seguida, conduzimos uma avaliação com 30 anotadores, analisando um total de 150 produtos, em que comparamos as descrições geradas pelo nosso método com as descrições originais postadas pelos anunciantes. Para isso, utilizamos um escala Likert de 7 pontos e examinamos as descrições nas dimensões de legibilidade, objetividade, informatividade, relevância e preferência geral. Além do nosso método, replicamos também uma proposta de geração de produtos da literatura recente que também utiliza sentenças extraídas de descrições de produto, esperando assim obter mais um referencial que nos permitisse melhor entender os efeitos de um LLM na atividade de geração de descrições de produto.

Ao analisar os resultados do nosso método na comparação direta com as descrições originais, observamos que nosso método foi amplamente preferido em quase todas dimensões, sendo a única dimensão que apresentou resultados menos contundentes a de objetividade. Em seguida, na comparação indireta com o método replicado utilizando as descrições originais em comum como ponto de referência, observamos que nosso método foi considerado mais inte-

ressante com forte significância estatística em 3 das dimensões avaliadas, sendo elas informatividade, relevância e preferência geral.

Assim, este capítulo apresenta as principais contribuições desta pequisa, assim como discute muitas de suas limitações. Por último, traçamos rumos para trabalhos futuros, indicando possíveis caminhos.

8.1 Principais Contribuições

Umas das contribuições do trabalho é no sentido de mostrar que o LLM utilizado é capaz de articular múltiplas sentenças em uma descrição de forma *zero-shot*. Ao compararmos o nosso método com um referencial replicado da literatura e que se baseia em sumarização extrativa, observamos que o uso de um LLM contribui para gerar descrições de produtos preferidas de forma geral.

No contexto de geração de texto *zero-shot*, contribuímos com a área ao apontar algumas limitações do modelo gerador de texto. Mais especificamente, exploramos a questão de limitar a quantidade de palavras geradas, mostrando que no geral o modelo não respeita com exatidão a instrução passada, mas é sim influenciado por ela. Além disso, contribuímos também ao reforçar a questão de como variações pequenas na instrução podem influenciar o resultado, ilustradas ao comparar instruções de um mesmo tema e obter diferenças significativas na qualidade das descrições geradas. Por último, foi interessante também notar os vícios de linguagem do modelo gerador no contexto de descrição de produto.

Outra contribuição importante é relacionada ao formato da avaliação proposta, principalmente na avaliação qualitativa. Ainda que tenhamos utilizado uma abordagem clássica de avaliação de texto, uma escala Likert, a utilizamos comparando diretamente as descrições geradas com as descrições originais postadas pelos anunciantes. Nesse novo formato, estabelecemos uma comparação com base em um referencial, diferentemente da maiorias dos trabalhos que propõe avaliar as descrições individualmente. Além disso, ao estabelecer esse referencial como sendo as descrições originais postadas pelos anunciantes, estamos explorando a substituição das descrições originais, o que seria uma aplicação bastante interessante para qualquer trabalho na área de geração de descrição de produtos.

Ainda sobre a avaliação, estruturamos nossa avaliação em um formato inteiramente reprodutível, de forma que diferentes métodos podem ser comparados entre si, o que inclusive fizemos ao replicar um método referência extraído da literatura. Dessa forma, qualquer trabalho futuro poderia replicar nossa avaliação e comparar os diferentes métodos. Nesse sentido, ao obter resultados positivos na comparação com as descrições originais sugerimos também

o nosso método como um referencial a ser comparado.

Por último, a principal contribuição do trabalho é sem dúvida o novo método proposto, que combina a riqueza das avaliações dos consumidores com a capacidade de sumarização de um LLM. Assim, reforçamos a ideia de que a as avaliações de produto contém um alto nível de informações do produto e mostramos uma nova forma de gerar uma descrição informativa, ao mesmo tempo que é articulada e legível.

8.2 Limitações

Pensando nas etapas que compõem o nosso método, um importante limitação do nosso trabalho é referente ao seu custo em duas etapas, na etapa de classificação de sentenças e na etapa de geração da descrição. De fato, em ambas as etapas estamos utilizando modelos proprietários da OpenAI, e o uso da sua API impõe um custo por uso. No caso do modelo classificador de sentenças, dada a média observada de palavras por sentença, estimamos que a classificação de 10 mil sentenças extraídas de avaliações de usuário teria um custo de 0,25 dólares, o que pode ser considerado alto pensando em uma aplicação em larga escala do método. Além disso, no caso da geração de texto em si, o custo seria consideravelmente maior, uma vez que as descrições geradas contém uma quantidade muito superior de palavras, e consequentemente tokens. Estimamos o custo em 7,56 dólares para gerar 10 mil descrições.

Além disso, um outro ponto fundamental relacionado ao uso de modelos proprietários da OpenAI é consequente dependência da empresa. Caso a empresa opte por descontinuar um dos modelos utilizados não seremos mais capazes de utilizá-lo. Isso ocorreu, por exemplo com o nosso modelo treinado para classificar as sentenças, de forma que teríamos que fazer mais um ajuste fino selecionando outro modelo disponibilizado, o que implicaria em mais custos.

Contudo, em relação a essas duas limitações, entendemos ser possível substituir os modelos da OpenAI por modelos de código aberto, inclusive por outro LLM. Nesse caso, porém, além de reconfigurar o método proposto, seria necessário arcar com os custos de armazenamento e execução do modelo.

Pensando ainda na etapa da geração de descrição, uma outra limitação é relacionada a natureza do LLM utilizado de forma *zero-shot*. Uma vez que não possuímos controle sobre o texto gerado, podem ser geradas descrições com inconsistências relativas ao produto, como sentenças apresentados atributos que o produto não possui ou detalhes técnicos imprecisos. Ainda que tenhamos investigado essa questão com avaliações de consistência factual, essa é uma

questão que merece especial atenção.

Pensando agora nas limitações apontadas nas avaliações de texto gerado, um ponto revelado foram os vícios de linguagem que o modelo possui quando gera uma descrição. No caso, apontamos dois deles, a forma engessada como o título do produto é repetido na primeira sentença e também a alta frequência com que declarações de marketing eram introduzidas no final da descrição. Conforme discutimos, esses vícios prejudicam a qualidade da descrição gerada, de forma que seria interessante pensar em mitigar esse problemas. Além disso, ao avaliar as descrições geradas em múltiplas dimensões, observamos como a objetividade foi a dimensão menos bem avaliada. Isso indica espaço para melhorias, de forma que seria interessante pensar em como gerar descrições com menos repetições e também mais sucintas.

Destacamos também uma limitação intrínseca de nossa metodologia, que é o fato de não sermos capazes de gerar descrições para produtos recém lançados ou com pouco engajamento do usuário. Esses são casos em que ainda não temos uma quantidade de avaliações suficiente para extrair sentenças adequadas para uma descrição de produto. Nesse sentido, essa é justamente a premissa fundamental para a aplicação do nosso método, isto é, precisamos de múltiplas avaliações de usuários para gerar uma descrição de produto. Por esse motivo, compartilhamos do entendimento elucidado no trabalho de Novgorodov et al. (2020) de que o nosso método, assim como o dos autores, pode melhorar a experiência de compra apenas de produtos já estabelecidos, impulsionando a sua venda ao propor informações valorizadas pelo cliente e não discutidas na descrição original. Contudo, já no caso de produtos que tiveram poucas avaliações, assim como no método proposto pelos autores, nosso método também sofre de um problema de inicialização, na medida que não é aplicável.

Também relacionada ao fato de usarmos sentenças extraídas de avaliações, apontamos uma outra limitação relevante que é a possibilidade de selecionar sentenças com informações falsas. De fato, é possível que sejam selecionadas pela nossa metodologia sentenças com informações imprecisas ou até mesmo incorretas, e que essas informações sejam refletidas nas descrições geradas. Assim, mesmo que consistentes com as informações das avaliações, seriam geradas descrições enganosas. Essa possibilidade levanta uma discussão muito importante sobre questões éticas relacionadas a aplicação da nossa metodologia, e, ainda que vislumbremos possíveis caminhos para mitigar esse problema de forma automática discutidos na Seção 8.3, entendemos que a aplicação do nosso método em um contexto real deve considerar esse cenário. Assim, uma possibilidade seria deixar explícito ao leitor que a descrição foi gerada automaticamente com base nas avaliações dos usuários, e que pode conter imprecisões.

Outra possibilidade seria a de apenas sugerir ao anunciante uma descrição alternativa, sendo ele responsável por aprovar ou não a sua substituição. Assim, ficaria a cargo do anunciante garantir a veracidade da descrição apresentada.

8.3 Trabalhos Futuros

Como trabalhos futuros de curto prazo, consideramos interessante explorar um pouco mais a etapa de extração de sentenças, principalmente no referente a etapa de classificação. Um ponto a ser explorada nesse contexto é entender quão generalista é o nosso método: conforme discutimos, a única etapa da nossa abordagem que é específica a um domínio é a na extração de sentenças, discutida em detalhes no capítulo 6, e mesmo nessa etapa verificamos que os classificadores treinados em um domínio parecem performar bem em outro. Por isso, acreditamos valer uma investigação mais profunda no tema, possivelmente com a avaliação de descrições de produtos de outros domínios geradas a partir do nosso método.

Outro ponto também interessante dentro do contexto de classificação é entender os efeitos de falsos positivos na etapa de geração da descrição, isto é, entender melhor se o modelo gerador é capaz de ignorar as sentenças fornecidas no *prompt* mas que não são interessantes e se basear apenas nas sentenças que possuem relevantes. Se esse fosse o caso, poderíamos então em pensar em alguma métrica que valorizasse mais a classificação correta das sentenças verdadeiramente positiva, mesmo que em detrimento de mais falsos positivos. Para investigar essa questão, poderíamos experimentar comparar as descrições geradas com mais falsos positivos com as geradas pelo nosso método no seu formato atual, e assim entender se há diferenças significativas.

Mais uma linha interessante a ser seguida a curto prazo é aprofundar na exploração da instrução do *prompt*. Em nossa abordagem, experimentamos instruir explicitamente a geração de uma descrição de produto, contudo, percebemos haver certos vícios do modelo ao gerar uma descrição. Assim, pode ser interessante mais uma rodada de experimentação de instruções. Nessa linha, poderíamos inclusive explorar instruções em que não seja solicitada uma descrição, e sim apenas um resumo das avaliações ou algo parecido. Novamente, poderíamos usar as descrições já geradas como referencial e estabelecer uma comparação direta.

Como trabalhos futuros de médio prazo, consideramos interessante experimentar outros LLMs, entendendo que essa é uma área bastante dinâmica e que produz novas descobertas com uma alta frequência. Entendemos que os LLMs tem cada um suas particularidades, e conforme surgem novos avanços

pode ser que surjam modelos ainda mais interessantes para o nosso método. Outro linha interessante de ser investiga é aprofundar a questão de quantas sentenças incluir como conteúdo do *prompt*. No nosso caso, verificamos que uma maior quantidade gerada gerou os melhores resultados, de forma que incluir ainda mais sentenças, no caso em que o produto possui muitas avaliações, parece uma ideia interessante. Nesse sentido, seria interessante entender a partir de que ponto já não faz muita diferença incluir sentenças adicionais. Além disso, podemos pensar no extremo oposto também, de entender qual o mínimo de sentenças que é suficiente para gerar descrições de qualidade.

Já para um trabalho de mais longo prazo, vislumbramos que seria interessante explorar combinar outras informações do produto com as sentenças selecionadas da avaliação. Um exemplo seriam detalhes técnicos fornecidos pelos fabricantes, como medidas ou materiais do qual o produto é feito, de forma a gerar descrições ainda mais detalhadas sobre o produto. Ainda nessa ideia, poderíamos pensar também em usar os detalhes técnicos como uma forma de verificar a veracidade das informações geradas, questão discutida na Seção 8.2. Outra ideia seria incorporar informações exibidas na imagem do produto, de forma que a descrição do produto dialogasse com a imagem e gerasse ainda mais interesse para o consumidor.

Referências Bibliográficas

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amidei, J., Piwek, P., e Willis, A. (2019). The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations. pages 397–402. Association for Computational Linguistics.

Amplayo, R. K., Angelidis, S., e Lapata, M. (2021). Aspect-controllable opinion summarization. *http://arxiv.org/abs/2109.03171*.

Berger, A. e Lafferty, J. (2017). Information retrieval as statistical translation. In *ACM SIGIR Forum*, volume 51, pages 219–226. ACM New York, NY, USA.

Bhaskar, A., Fabbri, A. R., e Durrett, G. (2022). Prompted opinion summarization with gpt-3.5. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 9282–9300.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I., e Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Celikyilmaz, A., Clark, E., e Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chae, Y. e Davidson, T. (2023). Large language models for text classification: From zero-shot learning to fine-tuning.

Chen, Q., Lin, J., Zhang, Y., Yang, H., Zhou, J., e Tang, J. (2019). Towards knowledge-based personalized product description generation in e-commerce. pages 3040–3050. ACM.

Chen, T. e Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., e Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Devlin, J., Chang, M. W., Lee, K., e Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

Elad, G., Guy, I., Novgorodov, S., Kimelfeld, B., e Radinsky, K. (2019). Learning to generate personalized product descriptions. pages 389–398. ACM.

Erkan, G. e Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Falke, T., Ribeiro, L. F., Utama, P. A., Dagan, I., e Gurevych, I. (2019). Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2214–2220.

Gehrmann, S., Clark, E., e Sellam, T. (2023). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Goyal, T., Li, J. J., e Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hinton, G., Vinyals, O., e Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, Y., Feng, X., Feng, X., e Qin, B. (2021). The factual inconsistency problem in abstractive text summarization: A survey.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

John, O. P., Srivastava, S., et al. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Kieuvongngam, V., Tan, B., e Niu, Y. (2020). Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*.

Kiritchenko, S. e Mohammad, S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. pages 465–470. Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kryściński, W., McCann, B., Xiong, C., e Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Laban, P., Schnabel, T., Bennett, P. N., e Hearst, M. A. (2022). Summac: Revisiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Lavie, A., Sagae, K., e Jayaraman, S. (2004). The significance of recall in automatic metrics for mt evaluation. In *Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6*, pages 134–143. Springer.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. pages 74–81.

Liu, Y. e Lapata, M. (2019). Hierarchical transformers for multi-document summarization. *http://arxiv.org/abs/1905.13164*.

Louviere, J. J., Flynn, T. N., e Marley, A. A. J. (2015). *Best-Worst Scaling*. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., e Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Muennighoff, N., Tazi, N., Magne, L., e Reimers, N. (2022). Mteb: Massive text embedding benchmark. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2006–2029.

Nachar, N. et al. (2008). The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

Nallapati, R., Zhou, B., dos Santos, C., Çaglar Gulçehre, e Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, pages 280–290.

Nguyen, M.-T., Nguyen, P.-T., Nguyen, V.-V., e Nguyen, Q.-M. (2021). Generating product description with generative pre-trained transformer 2. pages 1–7. IEEE.

Ni, J., Li, J., e McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 188–197.

Novgorodov, S., Guy, I., Elad, G., e Radinsky, K. (2019). Generating product descriptions from user reviews. In *The World Wide Web Conference*, pages 1354–1364. ACM.

Novgorodov, S., Guy, I., Elad, G., e Radinsky, K. (2020). Descriptions from the customers: comparative analysis of review-based product description

generation methods. *ACM Transactions on Internet Technology (TOIT)*, 20(4):1–31.

Orme, B. (2009). Maxdiff analysis: Simple counting, individual-level logit, and hb.

Pagnoni, A., Balachandran, V., e Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 4812–4829.

Papineni, K., Roukos, S., Ward, T., e Zhu, W.-J. (2002). pages 311–318.

Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Puduppully, R., Dong, L., e Lapata, M. (2019). Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6908–6915.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reimers, N. e Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Citeseer.

Salton, G. e Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513–523.

See, A., Liu, P. J., e Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, K., Tan, X., Qin, T., Lu, J., e Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., e Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

van der Lee, C., Gatt, A., van Miltenburg, E., e Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., e Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. pages 355–368. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., e Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J., Hou, Y., Liu, J., Cao, Y., e Lin, C.-Y. (2017). A statistical framework for product description generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 187–192.

Zhan, H., Zhang, H., Chen, H., Shen, L., Ding, Z., Bao, Y., Yan, W., e Lan, Y. (2021). Probing product description generation via posterior distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:14301–14309.

Zhang, X., Zou, Y., Zhang, H., Zhou, J., Diao, S., Chen, J., Ding, Z., He, Z., He, X., Xiao, Y., et al. (2022). Automatic product copywriting for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12423–12431.

Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., e Jiang, M. (2020). Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*.