Tendências e Lacunas em Processamento de Linguagem Natural: Uma Revisão Bibliográfica com Foco em Classificação de Texto

Thales Henrique Silva (Aluno)
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
thaleshenrique@ufmg.br

Pedro Olmo Stancioli Vaz De Melo (Orientador)
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
olmo@dcc.ufmg.br

***Resumo*—Este trabalho de revisão analisa artigos acadêmicos sobre Processamento de Linguagem Natural/Natural Language Processing (NLP), com foco na classificação de texto e com o objetivo de identificar as principais tendências de pesquisa na área. A revisão abrange principalmente um período de tempo mais recente para selecionar trabalhos mais relevantes. Os artigos são analisados quanto aos métodos, algoritmos e técnicas empregados, bem como aos resultados e conclusões alcançados. Os resultados da revisão são apresentados e discutidos em relação ao estado atual da pesquisa em NLP e às possíveis direções futuras.**

***Index Terms*—Text Classification; Natural Language Processing (NLP); Large Language Models (LLMs); ChatGPT.**

I. Introdução

O Processamento de Linguagem Natural (Natural Language Processing - NLP) é uma área interdisciplinar da computação que se concentra no desenvolvimento de métodos, algoritmos e técnicas capazes de permitir que os computadores compreendam e manipulem a linguagem humana de forma eficaz. A importância do NLP tem crescido exponencialmente com o avanço das tecnologias de inteligência artificial, especialmente com o surgimento dos Grandes Modelos de Linguagem (Large Language Models - LLMs). Estes modelos, como o GPT-3 por Brown et al. [1], o BERT por Devlin et al. [2], o T5 por Raffel et al. [3], e as suas respectivas variantes, têm demonstrado capacidades impressionantes em diversas tarefas, desde a geração e compreensão de texto até a tradução automática e a análise de sentimentos.

Os LLMs têm revolucionado o campo do NLP ao proporcionar melhorias significativas na precisão e na versatilidade dos sistemas de processamento de linguagem. No entanto, essa rápida evolução tecnológica e a proliferação de novas pesquisas têm gerado uma quantidade vasta de publicações, o que pode tornar difícil para os pesquisadores acompanhar as tendências e avaliar as inovações mais recentes. Nesse cenário, uma revisão bibliográfica abrangente torna-se essencial para mapear o estado atual da pesquisa e identificar tendências emergentes, lacunas e áreas promissoras para futuras investigações.

Esta revisão se concentra especificamente na tarefa de classificação de texto, uma das aplicações mais estudadas e amplamente implementadas em NLP. A classificação de texto envolve a atribuição de rótulos a entradas textuais com base em categorias discretas predefinidas. Esta tarefa não apenas representa um desafio significativo em termos de modelagem e algoritmos, mas também é fundamental para uma série de aplicações práticas, como filtragem de spam, análise de sentimentos, categorização de notícias, dentre outras. A escolha de focar na classificação de texto é motivada por seu papel central na área de NLP e pela quantidade substancial de pesquisa dedicada a essa tarefa, tornando-a um parâmetro valioso para avaliar o progresso e o impacto dos LLMs.

Dada a vastidão e a profundidade da literatura existente, é imperativo realizar uma revisão detalhada para proporcionar uma visão clara e estruturada das principais metodologias, técnicas e resultados recentes na classificação de texto. Esta revisão não só buscará sintetizar os avanços mais recentes, mas também identificará lacunas na pesquisa atual e sugerirá direções para futuras investigações. Além disso, será abordado o impacto dos LLMs na tarefa de classificação de texto, analisando como essas tecnologias têm influenciado e transformado as práticas e resultados na área.

Além do mais, com o rápido avanço dos LLMs e da inteligência artificial generativa, surgem desafios éticos significativos que precisam ser abordados. A capacidade cada vez maior de tais modelos para gerar texto convincente e realista levanta questões sobre a responsabilidade no uso e a potencial criação de desinformação. A facilidade com que esses sistemas podem ser utilizados para criar conteúdos falsos ou enganosos pode impactar negativamente a opinião pública e a integridade das informações. Ainda por cima, a questão da privacidade dos dados utilizados para treinar esses modelos também é uma preocupação relevante, pois o uso de grandes volumes de dados pode incluir informações sensíveis ou pessoais sem o devido consentimento. O viés presente nos dados de treinamento é outro desafio crítico, já que modelos de linguagem podem perpetuar ou até amplificar preconceitos existentes, afetando a equidade e a justiça na tomada de decisões automatizadas. Assim, é essencial que esta revisão de literatura explore os desafios éticos envolvidos, como desinformação, privacidade dos dados e viés. Ao examinar as abordagens e diretrizes atuais, a revisão contribuirá para identificar lacunas e propor recomendações para um uso mais responsável e transparente dessas tecnologias.

Portanto, este trabalho visa fornecer uma visão abrangente das principais tendências e desenvolvimentos em NLP, com um foco específico na classificação de texto, para ajudar

pesquisadores e profissionais a navegar pelas complexidades do campo e identificar novas oportunidades para avanço e inovação.

II. Trabalhos Relacionados

Dado que a classificação de texto é uma das principais tarefas em NLP, existe uma grande variedade de revisões bibliográficas explorando o estado da literatura. Por exemplo, Kowsari et al. [4] fornecem uma visão geral das técnicas de classificação de texto, abrangendo desde a extração de características e redução de dimensionalidade até a seleção de classificadores e métodos de avaliação. Os autores detalham técnicas como TF-IDF, Word2Vec e GloVe para a representação de texto, além de métodos de classificação como Naive Bayes, SVM e algoritmos de aprendizado profundo. O artigo também discute a importância da limpeza de texto, as técnicas de redução de dimensionalidade, e métodos de avaliação como F-beta Score e AUC. Além disso, explora-se as limitações e desafios dessas técnicas e suas aplicações práticas em áreas como recuperação de informação, sistemas de recomendação e saúde pública.

Minaee et al. [5] também fizeram uma contribuição semelhante, porém voltada especificamente para os principais modelos baseados em arquiteturas de aprendizado profundo. Os autores analisaram cerca de 150 modelos aplicados à classificação de texto. Eles discutiram as contribuições técnicas, semelhanças e pontos fortes desses modelos, abrangendo tarefas como análise de sentimentos, categorização de notícias, question answering e inferência de linguagem natural. Eles também providenciaram um resumo de mais de 40 datasets populares utilizados em classificação de texto e uma análise quantitativa do desempenho de diferentes modelos em benchmarks. Por fim, os autores ressaltaram os desafios restantes e futuras direções de pesquisa na área.

Revisões desse tipo são úteis como ferramentas de consulta. Elas permitem uma compreensão abrangente dos métodos mais eficazes e das abordagens emergentes, facilitando a comparação entre diferentes técnicas e a identificação de melhores práticas. Além disso, elas ajudam a destacar as lacunas existentes na literatura, fornecendo uma base sólida para o desenvolvimento de novas estratégias.

Com a popularização dos LLMs, surge um novo paradigma para abordar as tarefas relacionadas à classificação. Não é mais estritamente necessário treinar um modelo do zero para obter um bom classificador. Aplicações como o ChatGPT, da OpenAI, permitem utilizar instruções ou prompts diretamente e sem a necessidade de realizar um fine-tuning ou ajuste dos pesos com os dados de interesse. Isso revolucionou a área de NLP, e a comunidade acadêmica tem investigado o seu potencial a fundo.

Nesse sentido, inúmeros trabalhos têm sido publicados comparando as capacidades (bem como as limitações) dos LLMs da família GPT com classificadores mais tradicionais ou até mesmo avaliadores humanos. Isso indica um interesse significativo por parte dos pesquisadores, e essas contribuições serão discutidas em detalhes nas próxima seção.

Uma revisão focada em classificação de texto usando arquiteturas do tipo transformer foi realizada por Fields et al. [6]. Os autores exploraram aplicações desde análise de sentimentos até sistemas de perguntas e respostas, e propuseram uma nova taxonomia. A revisão avaliou a precisão dos modelos em 358 datasets, desafiando a ideia de que LLMs são sempre superiores, e discutiu a evolução histórica dos transformers, questões de custo e acessibilidade, e novas implicações sociais e éticas, como viés e direitos autorais. O estudo enfatizou a necessidade de uma compreensão detalhada do desempenho dos modelos e uma abordagem holística para sua implementação.

Outra revisão relevante, por Chang et al. [7], forneceu uma análise abrangente dos métodos de avaliação dos LLMs, abordando três dimensões principais: o que avaliar, onde avaliar e como avaliar. O estudo examinou a aplicação dos LLMs em várias tarefas, como processamento de linguagem natural, raciocínio, uso médico, ética, educação e ciências sociais e naturais. Os autores analisaram métodos e benchmarks de avaliação, discutiram casos de sucesso e falhas dos LLMs em diferentes tarefas, e destacaram desafios futuros, incluindo questões de custo e acesso. A pesquisa enfatizou a importância de tratar a avaliação como uma disciplina essencial para o desenvolvimento de modelos mais eficazes e seguros.

Por fim, é importante lembrar que tais inovações em NLP também estão associadas a controvérsias e desafios éticos. Trabalhos como os de Ray [8], Wu et al. [9], e Stahl e Eke [10], aprofundaram os riscos que o ChatGPT e tecnologias afins impõem à comunidade acadêmica e à sociedade como um todo. Ameaças à privacidade, falta de acurácia, viés, plágio, dentre outros, são alguns dos perigos presentes que requerem atenção e medidas visando garantir um uso mais seguro.

III. Contribuições

A. Metodologia

Para conduzir esta revisão, foi realizado um levantamento exaustivo utilizando mecanismos de busca acadêmicos, com foco em palavras-chave relacionadas à área, como ”classification”, ”LLM”(Large Language Models), ”ChatGPT”, ”few-shot learning”, ”zero-shot learning”, dentre outras. O primeiro passo consistiu em filtrar os artigos com base no número de citações, pois esse critério ajuda a identificar trabalhos com maior impacto e relevância na comunidade acadêmica. Além disso, foi considerada a recência das publicações, priorizando estudos mais recentes devido ao rápido avanço tecnológico no campo de NLP. Essa abordagem garantiu a inclusão de pesquisas que refletissem as tendências e inovações mais atuais.

Os artigos foram então analisados com atenção especial para aqueles que mencionavam explicitamente LLMs, particularmente o ChatGPT, uma vez que esses modelos representam uma inovação significativa no processamento de linguagem natural. A análise envolveu a classificação dos artigos com base nas metodologias empregadas e nos resultados obtidos. Cada artigo foi avaliado quanto à sua abordagem metodológica, como o uso de técnicas de aprendizado de máquina, estratégias

de fine-tuning, e experimentos específicos com diferentes datasets. Os resultados foram comparados para identificar padrões emergentes e abordagens promissoras na classificação de texto.

Após a análise, as principais informações foram extraídas e organizadas em uma planilha, permitindo uma visão estruturada dos dados. Essa organização facilitou a redação deste trabalho e a identificação dos tópicos de pesquisa vigentes em NLP, com um foco específico na classificação de texto. O processo de seleção e análise dos artigos foi conduzido com o objetivo de fornecer uma visão abrangente das metodologias atuais e das tendências de pesquisa, oferecendo uma base sólida para entender o estado atual da área e identificar direções futuras para pesquisas.

B. Taxonomia da Classificação de Texto e Principais Arquiteturas

De acordo com Fields e Chovanec [6], os problemas de classificação de texto podem, em boa parte, serem subdivididos segundo a taxonomia abaixo[1]:

Tabela I: Tipos de Tarefas de Classificação

Tipo	**Definição**
Análise de sentimento	Classificar com base na polaridade ou emoção contida no texto (por exemplo positivo, negativo ou neutro)
Classificação de notícias	Associar categorias a artigos de notícias baseado no assunto
Rotulação de tópicos	Identificar o assunto de um texto
Resposta a perguntas	Fornecer respostas à consultas
Inferência de linguagem natural	Determinar qual é a relação lógica entre sentenças
Classificação de atos de diálogo	Identificação da intenção contida em um texto (solicitar, questionar ou informar, por exemplo)
Reconhecimento de entidades nomeadas	Classificar entidades nomeadas como pessoas, organizações e locais
Análise sintática	Analisar a estrutura gramatical de um texto visando entender as relações entre palavras e frases

Por outro lado, Fields e Chovanec [6] também ressaltam que recentemente várias aplicações de classificação têm surgido. Dessa forma, os autores desenvolveram outra taxonomia mais abrangente:

[1]Ambas as tabelas foram adaptadas de Fields e Chovanec [6]

Tabela II: Tipos de Tarefas de Classificação de Texto

Tipo	**Definição**
Análise de sentimento fina/detalhada [11]	Considera as nuances e diferentes níveis de intensidade dos sentimentos
Análise de sentimento baseada em aspecto [12]	Identifica opiniões relacionadas a aspectos específicos
Detecção de linguagem ofensiva [13]	Reconhecer a ocorrência de conteúdo inapropriado
Reconhecimento de intenção [14]	Treinar através de conversas para identificar o que o usuário deseja
Classificador de Documentos [15]	Utilizar técnicas novas para textos maiores
Classificação de notícias falsas [16]	Identificação de padrões que indiquem notícias falsas/verdadeiras
Classificação interlingual [17]	Trabalha com textos contendo mais de um idioma
Detecção de posicionamento [18]	Identificar o ponto de vista de um ator em relação a um texto (favorável, contrário ou neutro)
Detecção de emoção/saúde mental [19]	Avaliar o estado mental de alguém com base no texto
Classificação de sentença [20]	Associar uma categoria à sentença inteira
Múltiplos rótulos [21]	Texto pode receber mais de uma classe simultânea
Multimodal	Utilizar vários formatos de dados, como texto, vídeo, áudio, etc.

Percebe-se que esse problema pode se manifestar em diferentes domínios e atender a diversas aplicações.

Do ponto de vista histórico, a classificação de texto é uma das tarefas mais fundamentais em NLP, tendo sido estudada pela literatura desde os anos 1950-60. Inicialmente, as técnicas aplicadas eram rudimentares e dependiam da criação de regras manuais, o que prejudicava a eficiência e escalabilidade. Com a evolução do campo de inteligência artificial, algoritmos de aprendizado de máquina passaram a ser aplicados, incluindo regressão logística, Naive Bayes Multinomial (NVM), k-vizinhos mais próximos (KNN), Máquina de Vetores de Suporte (Support Vector Machine), Árvore de Decisão (Decision Tree), Floresta Aleatória (Random Forest) e Adaboost.

A partir dos anos 2010, arquiteturas baseadas em redes neurais profundas (Deep Learning), como RNNs e CNNs, foram empregadas com sucesso em problemas de classificação. Isso foi possível em parte graças à popularização de GPUs e

à disponibilização de grandes volumes de dados.

Mais recentemente, o desenvolvimento dos LLMs representou uma verdadeira revolução em NLP. Tais modelos são extremamente sofisticados e capazes de realizar uma variedade de tarefas linguísticas, em particular a classificação, sem a necessidade de serem treinados novamente, bastando fornecer instruções adequadas (prompting).

De acordo com Sun et al. [22], pode-se dividir os LLMs em 3 tipos principais do ponto de vista da arquitetura:

1) Arquiteturas encoder: modelos como BERT (Bidirectional Encoder Representations from Transformers) por Devlin et al. [2] se encaixam nessa categoria. Essa classe de modelo utiliza camadas de transformers exclusivamente no lado do encoder para criar representações de alta qualidade do texto. Ele é treinado em tarefas de modelagem de linguagem mascarada e de previsão de sentenças adjacentes, o que lhe permite capturar relações contextuais profundas. Após esse treinamento, é possível realizar um ajuste (fine-tuning) em uma tarefa downstream específica.
2) Arquiteturas decoder: modelos da família GPT (Generative Pre-trained Transformer) por Radford et al. [23] e Brown et al. [1], projetados para a geração de texto, focando em prever a próxima palavra em uma sequência com base no contexto anterior. Utilizam as camadas de tranformers apenas no lado do decoder.
3) Arquiteturas encoder-decoder: modelos como o T5 por Raffel et al. [3], os quais combinam o encoder e o decoder e transformam todas as tarefas em um formato de texto-para-texto.

C. Few-Shot e Zero-Shot Learning com LLMs: Avanços, Comparações e Experimentos

Uma das principais vantagens dos LLMs em relação aos algoritmos tradicionais de aprendizado de máquina e aprendizado profundo é que eles não precisam passar por um novo treinamento para cada tarefa específica, dado que eles já foram expostos a uma vasta quantidade de dados. Essa flexibilidade tem o potencial de democratizar as aplicações de NLP, antes reservadas a organizações com acesso a bons recursos computacionais.

Nesse sentido, duas abordagens se destacam: few-shot learning e zero-shot learning. No primeiro caso, o modelo recebe algumas instâncias de exemplo, enquanto que no outro, apenas a descrição da tarefa é dada. Vários artigos explorando tais metodologias foram identificados, sendo esta uma sub-área de pesquisa dinâmica e promissora. Nos próximos parágrafos, seus principais resultados e conclusões serão detalhados.

Por exemplo, Wang et al. [24] realizaram uma série de experimentos para comparar o desempenho do Llama2, GPT-3.5, GPT-4 e algoritmos de aprendizado de máquina tradicionais e de aprendizado profundo num contexto de zero-shot learning. Os quatro datasets utilizados consistiram em tweets relacionados ao COVID-19, textos econômicos, textos sobre e-commerce, e mensagens de SMS para detecção de spam. Os autores concluíram que os LLMs são classificadores zero-shot eficazes, embora em algumas circumstâncias o desempenho deles tenha ficado abaixo de métodos baseados em aprendizado profundo. Ademais, os autores ressaltaram que uma futura pesquisa poderia incluir refinamento nos prompts ou implementação de agentes críticos para avaliar as saídas dos LLMs.

Por sua vez, Loukas et al. [25] exploraram a aplicação do GPT-3.5 e GPT-4 no domínio das finanças por meio do dataset Banking77, o qual consiste em consultas típicas de online banking. Foi observado que os LLMs obtiveram um desempenho superior ao de modelos não-generativos, mesmo com menos exemplos. Em suma, os resultados da classificação few-shot foram diretos e eficazes, demonstrando o potencial dos LLMs de forma prática.

Kuzman et al. [26] exploraram a abordagem zero-shot. A tarefa consistia em identificar o gênero de um texto de acordo com a sua função e forma. Isso não é algo trivial, visto que as anotações geradas manualmente por seres humanos frequentemente apresentavam discordância. Os autores compararam o ChatGPT com um modelo fine-tuned da arquitetura RoBERTa em datasets nas línguas inglesa e eslovena. Os experimentos realizados demonstraram que o ChatGPT obteve o desempenho superior, principalmente no caso do inglês. Isso foi surpreendente na visão dos autores, pois o modelo rival passou por um fine-tuning com dados manualmente anotados e específicos do problema. Em suma, esse artigo é mais um exemplo de como um LLM pode ter o potencial de auxiliar tarefas que, tradicionalmente, necessitavam de um esforço manual intenso.

De maneira semelhante, Savelka e Ashley [27] avaliaram o uso dos modelos GPT em uma tarefa de classificação semântica composta por documentos jurídicos, algo que normalmente requer anotadores humanos. Os experimentos foram realizados em um contexto zero-shot, e apesar dos resultados não terem sido tão satisfatórios em comparação a LLMs ajustados com várias instâncias, os autores concluíram que essa abordagem zero-shot é promissora e pode contribuir positivamente para os fluxos de trabalho no campo do direito. Este artigo, assim como os outros descritos anteriormente, ilustram que os LLMs podem ter aplicações significativas em diversos ramos do conhecimento.

Continuando, uma contribuição interessante feita por Gilardi et al. [28] foi a comparação entre o desempenho do ChatGPT e anotadores humanos. Os autores utilizaram quatro datasets consistindo de tweets e notícias previamente rotulados. Observou-se que, tanto a acurácia quanto o "intercoder agreement"(medida de concordância e confiabilidade das anotações) do ChatGPT foram superiores ao de anotadores da plataforma online MTurk. Assim, esses resultados indicam que, pelo menos em certos contextos, o uso de um LLM pode ser mais adequado que um ser humano. Tais resultados são particularmente relevantes considerando-se que boa parte da literatura se limita a analisar as capacidades dos LLMs, mas a questão de compará-las com as habilidades do ser humano nem sempre é investigada. Isso tende a se tornar mais importante conforme os modelos de inteligência artificial

avançam a passos largos.

Prosseguindo, uma das sub-tarefas da classificação de texto é a detecção de discurso de ódio e/ou conteúdo tóxico. Tal problemática tem se tornado cada vez mais relevante com o crescimento das mídias sociais e a disseminação rápida e ampla de informações. Nesse contexto, as técnicas de NLP podem auxiliar na identificação automática de tais conteúdos inapropriados, contribuindo para um ambiente virtual mais seguro.

Vários trabalhos da literatura averiguaram as capacidades dos LLMs, especialmente o ChatGPT, em identificar linguagem tóxica. Por exemplo, Li e Fan [29] compararam o desempenho de anotadores da plataforma MTurk com o ChatGPT. De forma geral, os autores observaram que o modelo alcançou bons resultados em métricas quantitativas, embora houvesse uma certa discordância com os avaliadores humanos. Além disso, a importância de se desenvolver prompts adequados foi um aspecto enfatizado pelos autores, já que diferentes instruções podem afetar as saídas.

Oliveira et al. [30], por sua vez, conduziram um estudo semelhante. Os autores utilizaram o dataset ToLD-Br, proposto por Leite et al. [31], para comparar a eficácia do ChatGPT e de modelos baseline do tipo BERT. Assim como no artigo anterior, os experimentos demonstraram que o ChatGPT apresenta uma grande proficiência no problema. Por outro lado, a escolha do prompt novamente foi um fator capaz de influenciar os resultados. Isso ressalta a necessidade de se desenvolver instruções precisas e avaliar o seu impacto na geração das respostas, principalmente em tarefas subjetivas como a detecção de conteúdo tóxico.

Em contraste aos trabalhos que identificaram um desempenho impressionante do ChatGPT e afins, Kocoń et al. [32] realizaram uma série de experimentos extensos para avaliar esse LLM em 25 tarefas de NLP (a maioria das quais se reduzia àlgum tipo de classificação). Os autores analisaram cerca de 48 mil respostas do ChatGPT e compararam com o estado-da-arte. Os resultados quantitativos demonstraram que, apesar deste modelo ser capaz de resolver de forma razoável a maioria das tarefas, ele obteve um desempenho inferior ao das melhores soluções. Assim, ele foi descrito pelos autores como um *Jack of all trades, master of none*. Outra questão investigada se refere à possibilidade dos datasets utilizados para a avaliação terem sido empregados no próprio treinamento do ChatGPT, o que certamente poderia afetar os resultados. Os autores estimaram o quão provável era cada conjunto de dados ter sido disponibilizado para o ChatGPT durante a sua etapa de fine-tuning. As observações indicam que ele teve um desempenho melhor nos dados aos quais ele provavelmente foi exposto.

Tais análises sobre um possível vazamento de informação, como aquela realizada por Kocoń et al. [32], nem sempre estão presentes nos trabalhos que avaliam empiricamente as capacidades de um LLM. Isto é, os pesquisadores frequentemente se limitam a reportar os resultados do modelo em dados públicos da internet, os quais poderiam ter sido aplicados no próprio treinamento ou fine-tuning. Por outro lado, essa questão tende a se tornar um tópico de pesquisa cada vez mais importante conforme os LLMs continuam a avançar. O ato de avaliar as suas habilidades de forma justa não é algo trivial. Isso foi investigado a fundo em um survey por Chang et al. [7]. Os autores ressaltaram a necessidade dos protocolos e benchmarks estarem em constante evolução, permitindo identificar de forma eficaz as capacidades e limitações dos LLMs. Uma metodologia mais robusta poderia, portanto, contribuir para o avanço da área de NLP, já que a comunidade teria uma noção mais precisa e correta sobre os LLMs.

Numa direção mais prática, Zhu et al. [33] desenvolveram o PromptBench, uma biblioteca open-source em Python voltada para a avaliação dos LLMs. Ela conta com suporte para várias tarefas, abordagens e análises. A importância de ferramentas desse tipo não pode ser subestimada, considerando que, como discutido acima, o ato de avaliar os LLMs está se tornando um tópico de pesquisa com inúmeras sutilezas e complexidades. Assim, é razoável pensar que, num futuro próximo, haverá um interesse crescente em utilizar frameworks para avaliar LLMs.

Uma tendência clara é que boa parte dos trabalhos explorando as capacidades dos LLMs num contexto de classificação utiliza os modelos GPT. Isso é esperado considerando a sua ampla popularidade, documentação extensa e adoção em larga escala pela comunidade acadêmica. Entretanto, é preciso ter em mente que existem outros LLMs competindo com as aplicações da OpenAI. Dessa forma, uma observação é que foram identificados relativamente poucos artigos que comparam de maneira sistemática as capacidades do ChatGPT e afins com LLMs competidores na tarefa de classificação. Isso pode ser visto como uma lacuna na literatura, dado que há uma oportunidade para estudos comparativos mais abrangentes. Explorar tais diferenças poderia enriquecer o entendimento das capacidades específicas de cada modelo e elucidar suas vantagens e limitações em diferentes cenários.

Nesse sentido, Mardiansyah e Surya [34] abordaram tal questão na tarefa de classificação de spam.[2] Os autores compararam o ChatGPT-4 com o Google Gemini através do dataset SpamAssassin. Os resultados indicam que o ChatGPT-4 apresenta um desempenho equilibrado, com altos valores de precisão e recall. Por sua vez, o Google Gemini foi mais eficaz na captura de spam, tendo um alto recall porém com uma tendência maior em classificar incorretamente e-mails legítimos como spam. Isso sugere que o ChatGPT-4 seria mais adequado fazendo o papel de um detector de spam genérico, enquanto que o Google Gemini seria útil quando o risco de deixar escapar um e-mail de spam é mais significativo que o risco de cometer um falso positivo. Embora tais resultados sejam interessantes, há de se considerar as limitações desse estudo. Testar em mais datasets, por exemplo, poderia servir para fortalecer ou refutar a tese de que os dois modelos de fato seriam úteis em cenários diversos. Ainda por cima, ressalta-se que, mesmo com essas diferenças observadas, não há como prever exatamente o comportamento que os modelos teriam

[2]Este artigo é uma pré-impressão e ainda não foi revisado por pares em alguma revista.

numa aplicação downstream em ambiente de produção.

Outra análise comparativa por Dao [35] testou os modelos ChatGPT, BingChat e o antigo Google Bard no dataset VNHSGE. Os dados consistem em 2500 perguntas de múltipla escolha sobre uma variedade de tópicos lecionados no ensino médio. Apesar da tarefa não configurar uma classificação de texto exatamente, que é foco principal dessa revisão, os resultados e a metodologia continuam sendo relevantes para a discussão. Os autores utilizaram uma abordagem zero-shot e observaram que o BingChat foi superior ao Google Bard e este, ao ChatGPT(3.5). Porém, é preciso ter cautela e não considerar tal conclusão como absoluta. Mais experimentos em outros datasets seriam necessários para afirmar com confiança que existem diferenças substanciais na capacidade de raciocínio dos modelos. Outra questão válida seria investigar o impacto de uma abordagem few-shot learning ou de técnicas de engenharia de prompting. Ademais, os três LLMs tiveram um desempenho melhor de que estudantes vietnamitas nos exames de língua inglesa. Na visão do autor, isso demonstra que tais ferramentas podem servir de auxílio na educação.

Prosseguindo, Zhong et al. [36] conduziram uma comparação do ChatGPT com quatro arquiteturas da família BERT em tarefas de compreensão linguística. O ChatGPT obteve um desempenho comparável ao dos modelos rivais nos problemas de classificação (análise de sentimento) e respostas a perguntas (QA). Por outro lado, os resultados do ChatGPT não foram bons nas tarefas relacionadas a similaridade e paráfrase, mas superaram todos os modelos do tipo BERT na inferência linguística. Além disso, os autores investigaram o impacto de estratégias de engenharia de prompting, incluindo few-shot learning e Chain-of-thought. Assim como em outros trabalhos, foi observado que tais abordagens melhoraram o desempenho do modelo. Por outro lado, o ChatGPT se mostrou sensível no cenário 1-shot. Isso foi atribuído a uma menor correlação entre o exemplo amostrado e os dados de teste, já que uma instância pouca relacionada iria prejudicar a saída. Em suma, tais análises serviram para avaliar de forma quantitativa as capacidades e limitações do ChatGPT.

D. Abordagens Inovadoras em Classificação com LLMs: Métodos Indiretos e Estratégias de Prompting

Uma tendência presente em boa parte dos trabalhos revisados é que eles empregam o LLM para realizar a classificação explicitamente. Isto é, o modelo recebe um prompt e as instâncias para serem categorizadas, o que é a abordagem mais óbvia e natural. Porém, Shi et al. [37] propuseram um paradigma inovador para tirar proveito das capacidades do ChatGPT de maneira indireta. Primeiramente, o modelo é instruído a refinar o texto original e corrigir possíveis erros ortográficos ou gramaticais. Em seguida, um novo prompt é aplicado para extrair um grafo de conhecimento, o qual modela relacionamentos entre as entidades do texto. Por exemplo, se a entrada original era *bank of france leaves intervention rate unchanged at pct official*, então um possível resultado após o refinamento seria *The Bank of France has decided to maintain its intervention rate at the current percentage, according to an official statement*, enquanto que o grafo resultante seria uma tupla no formato *('Bank of France', 'maintain', 'intervention rate')*. Na sequência, um novo grafo textual é construído, sendo possível incrementá-lo com conhecimento externo via ponderação TF-IDF, e isso será a entrada para uma rede neural de grafos, a qual realiza a classificação. Uma vantagem desse método é que, devido ao fato de o modelo ser linear, ele é inerentemente interpretável, em contraste à maioria dos algoritmos de aprendizado de máquina e inteligência artificial. Os resultados demonstraram que tal abordagem apresenta um bom equilíbrio entre desempenho e interpretabilidade. Pesquisas futuras poderiam desenvolver técnicas similares para outras tarefas e avaliar o potencial do ChatGPT de uma forma não tão direta.

Outra proposta inovadora, desenvolvida por Sun et al. [22], chama-se CARP: Clue And Reasoning Prompting. Os autores observaram que, embora os LLMs tenham obtido um sucesso impressionante nas tarefas de classificação, o desempenho é prejudicado devido à dificuldade associada aos aspectos linguísticos complexos e à restrição no número de tokens disponíveis. Visando superar tais obstáculos, essa metodologia primeiramente instrui o modelo a encontrar pistas no texto que possam auxiliar a classificação (palavras chaves, frases, informação contextual, etc.). Então, o LLM é induzido a realizar um processo de raciocínio mais aprofundado para tomar a decisão, levando em consideração a evidência presente. Isso se assemelha ao paradigma Chain-of-thought por Wei et al. [38], uma técnica que aprimora o desempenho em tarefas de aritmética e raciocínio simbólico ao instruir o modelo a descrever seus passos intermediários.

Um dos prompts da metodologia CARP é:

> This is an overall sentiment classifier for opinion snippets.
> First, list CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) for determining the overall sentiment of the input. Next, deduce a diagnostic reasoning process from clues and the input to determine the overall sentiment.
> Finally, determine the sentiment of input as Positive or Negative considering clues, the reasoning process and the input. INPUT: <text >
> CLUES:

Veja que essa é uma configuração zero-shot, dado que o modelo não recebe instâncias ilustrativas. Também é possível incluir demonstrações no contexto few-shot, e os autores propuseram algumas estratégias para fazer a amostragem de exemplos a partir do conjunto de treinamento. A solução mais óbvia e simples é amostrar aleatoriamente itens do conjunto. Todavia, talvez essas instâncias não sejam muito relacionadas semanticamente aos dados de teste, o que poderia prejudicar o desempenho. Outra abordagem proposta emprega uma busca KNN para selecionar os dados relevantes e com maior similaridade. Isso se mostrou mais eficaz que a amostragem

aleatória.

Em suma, a metodologia CARP foi capaz de atingir resultados estado-da-arte em 4 dos 5 benchmarks utilizados. Futuramente, outras tarefas em NLP, além da classificação de texto, poderiam tirar proveito de tal abordagem.

Outra abordagem interessante, que também explora a engenharia de prompting de uma maneira inovadora, foi desenvolvida por Wu et al. [39]. O problema investigado pelos autores é a classificação de textos curtos ou Short Text Classification (STC), cuja complexidade se dá pela concisão das entradas. Textos pequenos naturalmente impõem uma dificuldade para o classificador, já que existe menos informação sintática e semântica para ser aproveitada. Visando empregar os LLMs de forma eficaz, os autores propuseram a metodologia *Quartet Logic: A Four-Step Reasoning* (QLFR), a qual se assemelha a uma sequência de prompts usando um raciocínio Chain-of-thought. Na primeira etapa, o modelo é instruído a identificar os conceitos principais do texto original. Então, essa saída é concatenada ao prompt anterior, e na segunda etapa, o modelo deve recuperar conhecimento ou informação de senso comum. O resultado é concatenado novamente, e, na terceira etapa, o modelo irá reescrever o texto para aprimorar a sua legibilidade e clareza. Por fim, na quarta e última etapa, esse texto refinado servirá de entrada para o modelo realizar a classificação com base nas categorias pré-definidas. Perceba que tal metodologia apresenta uma grande semelhança com as propostas de Shi et al. [37] e Sun et al. [22], discutidas acima, no sentido de que o LLM é induzido a descobrir os fatos e relações do texto original antes da classificação propriamente dita ocorrer.

Outra contribuição relevante por Wu et al. [39] foi o método *CoT-Driven Multi-task learning* (QLFR-CML), uma extensão do QLFR que incrementa modelos menores por meio de uma transferência de conhecimento dos LLMs. Os experimentos realizados em seis datasets de benchmark demonstram a eficácia das abordagens propostas. Em particular, o QLFR atingiu resultados estado-da-arte em todos esses datasets.

E. Aspectos Éticos

Apesar do enorme potencial benéfico dos grandes modelos de linguagem, é necessário ressaltar que eles estão associados a uma série de controvérsias e debates éticos, os quais tendem a se intensificar conforme o campo de NLP continua a avançar rapidamente. Assim, é importante ressaltar tais questões e avaliar criticamente o impacto dessas novas tecnologias.

Ray [8] explorou as aplicações e os dilemas atuais relacionados ao ChatGPT. O autor identificou uma série de desafios que esta ferramenta impõe, os quais são resumidos a seguir. Por exemplo, ele, como qualquer outro modelo de inteligência artificial, pode reproduzir o viés contido nos seus dados de treinamento, possivelmente gerando saídas discriminatórias. Além disso, considerando que boa parte do conteúdo fornecido para treinar os LLMs está em inglês, o risco de vieses culturais e linguísticos se torna maior.

Outra questão se refere à atribuição de propriedade intelectual. Os LLMs podem auxiliar na produção científica, contribuindo para a criação de hipóteses e na escrita de trabalhos. Isso pode servir para aumentar a produtividade. Porém, surgem questionamentos sobre transparência e integridade acadêmica, principalmente quando o uso de tais ferramentas não é devidamente documentado pelo pesquisador.

Prevenir o uso malicioso dos LLMs, como a desinformação, spam e deepfakes, é outro ponto crucial. É sabido que tais modelos podem ser vulneráveis a ataques adversários e serem induzidos a gerar conteúdo prejudicial. Portanto, é fundamental por parte dos desenvolvedores estabelecer salvaguardas para reduzir esses riscos. A comunidade acadêmica pode continuar investigando o quão robustos os LLMs são, identificando vulnerabilidades a serem corrigidas.

Outros aspectos relevantes identificados por Ray [8] estão relacionados à transparência e à explicabilidade. Os LLMs são ferramentas extremamente complexas e difíceis de interpretar, o que é agravado pelo fato de boa parte deles serem código-fechado, e os detalhes de como foram treinados exatamente não são conhecidos pelo público em geral. Conforme o seu uso em aplicações comerciais é ampliado, torna-se necessário ser capaz de explicar as decisões tomadas. Ademais, a responsabilidade por eventuais consequências negativas advindas de algum modelo deve ser delimitada. Isso é particularmente relevante em contextos críticos, como a área da medicina, por exemplo.

Ainda por cima, cabe ressaltar o impacto da inteligência artificial generativa nas áreas do conhecimento que envolvem criatividade, Embora os modelos de linguagem possam contribuir para o aumento de produtividade, existe o risco de eles tornarem os trabalhadores humanos obsoletos ou dispensáveis.

Há também a possibilidade dos LLMs contribuírem para uma redução geral do pensamento crítico da população. A conveniência que eles proporcionam e a sua capacidade de automatizar certas tarefas cognitivas pode levar a uma dependência por parte de seus usuários, especialmente estudantes.

Outro aspecto relevante são as limitações inerentes ao ChatGPT. É bem documentado que o modelo pode gerar informações incorretas ou duvidosas e se basear em conhecimento desatualizado. Além disso, ele pode ter problemas ao lidar com: solicitações inadequadas; consultas multilíngues; linguagem não literal ou figurativa; consultas ambíguas; dentre outras dificuldades.

Devido à sua complexidade inerente e natureza interdisciplinar, as questões elencadas acima não podem ser resolvidas facilmente. É preciso um esforço coletivo dos vários atores e *stakeholders* envolvidos, devendo cada um tomar as devidas precauções. Os desenvolvedores dos LLMs, por exemplo, devem estar ciente dos eventuais impactos prejudiciais de suas tecnologias, minimizando esses riscos na medida do possível. As autoridades públicas e governamentais, por sua vez, devem gradualmente desenvolver enquadramentos jurídicos e regulações para fomentar um uso ético e responsável de tais modelos, o que é uma tarefa árdua considerando a rápida evolução tecnológica e as limitações do processo político. Por sua vez, a comunidade acadêmica de forma geral pode contribuir para o avanço seguro e responsável da área ana-

lisando criticamente os impactos dessas novas tecnologias, identificando pontos de vulnerabilidade a ser aprimorados.

IV. Conclusão

Por meio desta revisão bibliográfica, foi possível delimitar certas tendências recentes da literatura em NLP.

O grande potencial dos LLMs, principalmente o ChatGPT, já foi demonstrado em vários estudos, e tais ferramentas podem servir como uma alternativa eficaz a modelos tradicionais que precisam ser treinados do zero. Dependendo das particularidades da tarefa, observa-se que o ChatGPT pode atingir um desempenho estado-da-arte ou algo próximo disso. Há também a possibilidade de utilizar tal ferramenta para automatizar processos manuais e que normalmente necessitariam de avaliadores humanos. Isso pode contribuir para um aumento da produtividade em várias áreas do conhecimento.

Entretanto, uma questão de pesquisa ainda aberta é se tais resultados impressionantes podem, ainda que parcialmente, ser atribuídos ao fato de que os datasets utilizados nos experimentos são públicos e, logo, poderiam ter sido acessados pelo modelo durante a etapa de treinamento. Isso possivelmente iria significar que, numa aplicação real e com dados inéditos, o desempenho não seria tão bom.

Nesse contexto, análises como a de Kocoń et al. [32] são especialmente interessantes. A tentativa de quantificar a probabilidade do modelo ter sido exposto aos dados anteriormente é algo inovador, considerando que muitos outros trabalhos semelhantes se limitam a reportar os resultados de experimentos sem aprofundar no assunto do vazamento de informação. Essa questão se torna mais relevante conforme o progresso dos LLMs continua e eles são treinados com dados públicos mais recentes. Uma abordagem útil seria testar o modelo em datasets privados, o que alguns pesquisadores fizeram ocasionalmente. Tal metodologia é mais custosa e nem sempre viável.

Outra área de interesse da comunidade consiste em avaliar o efeito das estratégias de engenharia de prompting. É amplamente conhecido que a escolha das instruções fornecidas tem um forte impacto nas saídas do LLM. Vários trabalhos investigaram tal questão, e o consenso geral parece ser que prompts didáticos, informativos e contextuais são mais eficazes que requisições diretas e imediatas. Metodologias como CARP por Sun et al. [22], Chain-of-thought por Wei et al. [38] e QLFR por Wu et al. [39] se enquadram nesse paradigma. Todas elas têm em comum o fato de elicitarem do modelo um raciocínio mais explícito e focado nos passos intermediários. Entretanto, a desvantagem é que tais métodos consomem mais tokens e dificultam a realização de experimentos em larga-escala devido ao custo associado às requisições das APIs. Futuros estudos poderão aprofundar tais metodologias e avaliar quantitativamente a sua eficácia em mais datasets e benchmarks.

Também foi observado que, de maneira geral, a abordagem few-shot learning é mais eficaz que a zero-shot. Por outro lado, a ideia de fornecer exemplos gera a dúvida de como exatamente eles devem ser escolhidos. Pode-se optar por uma amostragem aleatória, que é a ideia mais simples e direta. Entretanto, é razoável pensar que, especialmente quando os datasets apresentam uma distribuição desbalanceada de classes, a amostragem aleatória iria enviesar o contexto em favor de certas categorias, prejudicando o desempenho. Dessa forma, a metodologia por Sun et al. [22], a qual seleciona os exemplos mais similares, se mostra interessante. Pesquisas futuras poderiam propor métodos alternativos de obter as instâncias mais semelhantes, visando melhorar a abordagem few-shot.

Outra questão que a comunidade poderá explorar se refere a análises comparativas entre os diferentes LLMs. O ChatGPT é um dos grandes modelos mais investigados pela literatura. Porém, graças ao forte interesse comercial que a área de NLP adquiriu nos últimos anos, alternativas estão sendo desenvolvidas, como o Google Gemini (anteriormente conhecido como Bard). Comparar de forma quantitativa os LLMs, assim como no estudo de Mardiansyah e Surya [34], pode trazer insights relevantes sobre as capacidades e limitações de cada um. Observa-se que isso tem recebido relativamente pouca atenção, pelo menos no contexto restrito à classificação, e, logo, existe um potencial para futuros estudos.

Por fim, é preciso ter em mente os desafios éticos que essas novas tecnologias impõem à sociedade. O desenvolvimento e a aplicação de modelos de NLP levantam preocupações sobre privacidade, uso de dados pessoais, e a potencial perpetuação de preconceitos e discriminações presentes nos dados de treinamento. Além disso, há o risco de criação e disseminação de informações falsas ou enganosas, o que pode impactar negativamente a confiança pública e o discurso político e social. As empresas e pesquisadores têm a responsabilidade de implementar práticas de desenvolvimento éticas, garantindo a transparência e a explicabilidade dos modelos, bem como a mitigação de vieses e a proteção da privacidade dos usuários. Nesse contexto, a criação de regulamentações e diretrizes de governança irá desempenhar um papel crucial na orientação e fiscalização do uso dessas tecnologias, assegurando que o seu impacto seja positivo e justo para toda a sociedade.

Referências

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[4] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[5] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.

[6] John Fields, Kevin Chovanec, and Praveen Madiraju. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 2024.

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[8] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[9] Xiaodong Wu, Ran Duan, and Jianbing Ni. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of Information and Intelligence*, 2(2):102–115, 2024.

[10] Bernd Carsten Stahl and Damian Eke. The ethics of chatgpt–exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700, 2024.

[11] Manish Munikar, Sushil Shakya, and Aakash Shrestha. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE, 2019.

[12] Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299, 2019.

[13] AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. Leveraging affective bidirectional transformers for offensive language detection. *arXiv preprint arXiv:2006.01266*, 2020.

[14] Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip S Yu. Are pretrained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. *arXiv preprint arXiv:2106.04564*, 2021.

[15] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.

[16] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062, 2019.

[17] Cindy Wang and Michele Banko. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, 2021.

[18] Bowen Zhang, Daijun Ding, and Liwen Jing. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*, 2022.

[19] Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. Transformer-based language models for mental health issues: a survey. *Pattern Recognition Letters*, 167:204–211, 2023.

[20] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*, 2019.

[21] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171, 2020.

[22] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[24] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*, 2023.

[25] Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. Breaking the bank with chatgpt: Few-shot text classification for finance. *arXiv preprint arXiv:2308.14634*, 2023.

[26] Taja Kuzman, Igor Mozetic, and Nikola Ljubešic. Chatgpt: beginning of an end of manual linguistic data annotation. *Use Case of Automatic Genre Identification. ArXiv abs/2303.03953*, 2023.

[27] Jaromir Savelka and Kevin D Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6, 2023.

[28] Fabrizio Gilardi, Meysam Alizadeh, and Ma¨el Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[29] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36, 2024.

[30] Amanda S Oliveira, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas, and Eduardo JS Luz. How good is chatgpt for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103. SBC, 2023.

[31] João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China, December 2020. Association for Computational Linguistics.

[32] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023.

[33] Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*, 2023.

[34] Ketut Mardiansyah and Wayan Surya. Comparative analysis of chatgpt-4 and google gemini for spam detection on the spamassassin public mail corpus. 2024.

[35] Xuan-Quy Dao. Performance comparison of large language models on vnhsge english dataset: Openai chatgpt, microsoft bing chat, and google bard. *arXiv preprint arXiv:2307.02288*, 2023.

[36] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*, 2023.

[37] Yucheng Shi, Hehuan Ma, Wenliang Zhong, Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 515–520. IEEE, 2023.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[39] Hui Wu, Yuanben Zhang, Zhonghe Han, Yingyan Hou, Lei Wang, Siye Liu, Qihang Gong, and Yunping Ge. Quartet logic: A four-step reasoning (qlfr) framework for advancing short text classification. *arXiv preprint arXiv:2401.03158*, 2024.