Opportunities vs. Risks: Exploring Automatic Annotation of Financial Polarity Biases via Large Language Models

Viviane Romero[1], Gabriel Assis[1], Jonnathan Carvalho[2], Paulo Mann[3], Aline Paes[1]

[1] Universidade Federal Fluminense, Brazil,
{vivianeromero,assisgabriel}@id.uff.br,alinepaes@ic.uff.br
[2] Instituto Federal Fluminense, Brazil
joncarv@iff.edu.br
[3] Universidade Federal do Rio de Janeiro, Brazil
paulomann@ic.ufrj.br

Abstract. The financial market encompasses investors with distinct risk profiles — conservative, moderate and aggressive — each reflecting different attitudes toward gains and losses. These profiles influence the interpretation of financial news, particularly regarding the comparison between the labels "positive"/"negative" and the categories "opportunity"/"risk". Although these pairs of terms may initially appear equivalent, their practical application reveals notable inconsistencies. This paper employs large language models (LLMs) to annotate financial news, investigating whether such models capture the biases associated with each investor profile. We analyze the correlation between "opportunity" and "positive" and between "risk" and "negative" in the labels generated by the models. Furthermore, we examine whether, in the absence of explicit instructions regarding risk preference, the LLMs implicitly adopt a default bias when performing sentiment analysis on financial texts. Our findings provide insights into how risk profiles influence model behavior and suggest directions for improving both the personalization and accuracy of polarity detection in financial news analysis.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: financial polarity, large language models, prompt engineering

1. INTRODUÇÃO

O mercado financeiro é fortemente influenciado pelas notícias diárias relacionadas a empresas, políticas públicas e questões ambientais, impactando diretamente preços de ativos e decisões dos investidores [Anbaee Farimani et al. 2022]. Dada a relevância dessas informações e o alto volume publicado diariamente, torna-se fundamental o uso de abordagens automatizadas para classificar rapidamente esses conteúdos, destacando-se especialmente a análise de sentimentos [Saad and Saberi 2017]. Contudo, notícias financeiras frequentemente envolvem aspectos específicos que indicam oportunidades ou riscos concretos para investidores ou corporações [Chen et al. 2023].

Nesse contexto, diretrizes ambientais, sociais e de governança (ESG) têm ganhado importância na análise financeira, refletindo mudanças estruturais na economia global, impulsionadas pela adoção crescente desses princípios por investidores e empresas [UNPRI 2006]. A relevância dessas diretrizes se relaciona diretamente à concretização dos Objetivos de Desenvolvimento Sustentável (ODS) da ONU [United-Nations 2015], influenciando significativamente as percepções e decisões de diferentes perfis de investidores sobre setores e empresas. Com efeito, notícias relacionadas a práticas de ESG

Os autores agradecem ao financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), do Instituto Nacional de Inteligência Artificial (INCT IAIA), da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/002930/2024, SEI-260003/000614/2023, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código Financeiro 001.

podem influenciar diretamente a percepção dos investidores sobre uma empresa ou setor, tornando essencial a incorporação dessa ótica na análise de sentimentos aplicada ao contexto financeiro.

Os investidores, por outro lado, podem ser vislumbrados de acordo com seus perfis de investimento, aderentes ao nível de sua tolerância ao risco: conservadores, moderados ou arrojados [Pompian 2016; Mak and Ip 2017]. Embora categorias de sentimento como "positivo" e "negativo" sejam semanticamente próximas a "oportunidade" e "risco", elas podem diferir dependendo do perfil do investidor. Uma mesma notícia pode ser interpretada de forma distinta por investidores com diferentes tolerâncias ao risco, revelando a complexidade da classificação automática de textos financeiros.

Este estudo propõe um método automático para investigar empiricamente o alinhamento entre a classificação tradicional de polaridade de sentimentos e a categorização por "oportunidades" e "riscos" no contexto financeiro e de ESG. Para isso, são abordadas as seguintes questões de pesquisa:

(1) *Existe um alinhamento entre as classificações de "oportunidade" e "risco" feitas pelo mercado e os sentimentos "positivo" e "negativo", considerando diferentes perfis de investidores?*
(2) *Qual o impacto de incorporar vieses característicos dos diferentes perfis de investidores no processo de anotação feito por modelos de linguagem de grande escala (LLMs)?*
(3) *Na ausência de um perfil especificado, o modelo assume implicitamente algum perfil padrão?*

Para responder a essas questões, utilizamos LLMs como anotadores automáticos [Zhang et al. 2023], aproveitando sua capacidade de personalizar respostas com base em *prompts* e contextos específicos [Tan et al. 2024]. Este estudo avalia a capacidade de os modelos capturarem os vieses característicos dos perfis conservador, moderado e arrojado, analisando se há uma correlação consistente entre as classificações de "oportunidade" e "positivo" e entre "risco" e "negativo". Adicionalmente, exploramos qual perfil é adotado implicitamente pelo modelo na ausência de instruções explícitas, destacando especialmente o foco em notícias com aspectos ESG devido à sua relevância crescente.

Os resultados indicam que, embora haja um alinhamento geral entre as classificações, existem discrepâncias importantes dependendo do perfil de investidor, confirmando que o modelo captura parcialmente os vieses associados a cada perfil. Esses achados ressaltam tanto o potencial quanto a complexidade da aplicação de LLMs na análise automática de sentimentos no domínio financeiro.

2. TRABALHOS RELACIONADOS

Os trabalhos relacionados abrangem três áreas relevantes para este estudo: anotação automatizada com LLMs, análise de sentimentos no domínio financeiro e classificação de oportunidades e riscos.

Na anotação automatizada, LLMs como GPT-4 [OpenAI-Team 2024] e Gemini [Gemini-Team 2024] têm se mostrado eficazes para superar limitações de recursos e conhecimento especializado exigidos pela anotação manual [Wang et al. 2024]. Estudos recentes indicam que esses modelos podem oferecer eficiência e precisão comparáveis às anotações humanas, destacando seu potencial em tarefas complexas e especializadas [Yadav et al. 2024; Ahmed et al. 2024; Pangakis and Wolken 2024; Wu et al. 2024].

Na análise de sentimentos financeiros, diversas abordagens têm sido propostas para prever movimentos de mercado e auxiliar decisões estratégicas [Khadjeh Nassirtoussi et al. 2014; Heston and Sinha 2017]. Inicialmente limitadas por léxicos genéricos, inovações como dicionários específicos de domínio [Loughran 2011] e modelos de linguagem neural, como BERT, trouxeram avanços significativos [Day and Lee 2016; Man et al. 2019; Hiew et al. 2022]. Adicionalmente, análises de aspectos específicos, como risco e oportunidade, têm refinado ainda mais a interpretação de textos financeiros [Yang et al. 2018].

A categorização explícita de textos financeiros em "risco" e "oportunidade" também recebeu atenção significativa. Trabalhos recentes como o FinSentGPT [Ardekani et al. 2024], utilizando versões

adaptadas de LLMs, demonstraram eficácia em captar nuances contextuais importantes para investidores. A influência dos perfis comportamentais dos investidores e suas percepções de risco também foi amplamente investigada [Bashar Yaser Almansour and Almansour 2023; Dickason and Ferreira 2018; Nguyen et al. 2017], revelando relações complexas entre tolerância ao risco, decisões financeiras e fatores psicológicos.

O presente estudo diferencia-se ao integrar essas abordagens, investigando especificamente como LLMs interpretam sentimentos positivos e negativos a partir da perspectiva dos perfis de investidores em textos previamente rotulados como risco ou oportunidade. Destaca-se ainda pela introdução de personas específicas (conservador, moderado e arrojado) e uma persona genérica nos *prompts*, permitindo uma análise aprofundada dos vieses implícitos nas classificações automáticas de sentimentos no domínio financeiro.

3. METODOLOGIA

Este estudo investiga a relação entre os rótulos "oportunidade" e "risco", obtidos a partir de uma base de notícias relacionadas a empresas, comparando-os com classificações feitas pelas personas que representam diferentes perfis de investidores, utilizando os rótulos "positivo" e "negativo". O objetivo central é analisar como investidores com perfis distintos percebem essas notícias e determinar se suas interpretações e sentimentos estão alinhados às avaliações feitas pelas próprias empresas, especialmente sob a perspectiva ESG, ou se apresentam diferenças significativas. O código do presente estudo pode ser encontrado no GitHub[1].

Conjunto de dados. Utilizamos os dados da *shared task* "Multi-Lingual ESG Impact Type Identification" [Chen et al. 2023], proposta no *workshop* FinNLP-2023, cujo objetivo é classificar notícias financeiras segundo o impacto ESG ("*Opportunity*", "*Risk*" ou "*Cannot Distinguish*"). Este trabalho utilizou especificamente o conjunto de dados ML-ESG-2, obtido após solicitação formal aos organizadores, concentrando-se apenas nas notícias em inglês. O conjunto de dados contém 1.024 instâncias, originalmente dividido em 808 instâncias para treinamento e 216 para teste, com uma média de 63 palavras por notícia. O rótulo "*Cannot Distinguish*" não foi considerado por ausência de instâncias anotadas para as notícias em inglês.

3.1 Elaboração de *prompts*

A etapa de elaboração de *prompts* no processo de anotação automática desempenha um papel fundamental no esforço de assegurar a qualidade e a precisão das anotações geradas [Dong et al. 2024]. Para desenvolver instruções claras, objetivas e bem estruturadas para os modelos, foram incorporados conceitos de engenharia de *prompt*. O uso de personas é uma destas técnicas, que consiste em atribuir uma identidade específica ao modelo [White et al. 2023], moldando seu comportamento por meio da adaptação do tom, estilo e perspectiva das respostas.

Os rótulos "Oportunidade" e "Risco" já fornecidos foram anotados com base na visão da empresa, avaliando como as notícias coletadas impactam diretamente a operação e a estratégia da empresa no aspecto ESG. Com o uso de personas, é possível expandir essa análise para incorporar o ponto de vista de investidores de diferentes perfis, utilizando os rótulos "Positivo" e "Negativo", permitindo explorar como diferentes perfis de investidores interpretam as mesmas notícias sob a ótica tradicional de análise de sentimentos. Utilizamos quatro perfis de investidores, cada um resultando em um *prompt*: genérico, conservador, moderado e arrojado [Pompian 2016]. O investidor conservador tende a perceber maior risco em cenários incertos, priorizando estabilidade e segurança. Para ele, notícias que introduzem volatilidade ou custos adicionais são frequentemente vistas como negativas. O investidor moderado procura avaliar riscos e oportunidades de forma equilibrada, ponderando se os benefícios potenciais

[1] `https://github.com/AIDA-BR/polarity_ratings`

superam os desafios no médio e longo prazo. Por sua vez, o investidor arrojado foca em ganhos futuros e demonstra mais tolerância a riscos. Ele pode interpretar notícias anotadas como risco de forma positiva, desde que apontem para oportunidades de liderança ou inovação para a empresa. Adicionalmente, elaboramos um *prompt* genérico, sem perfil de investidor, utilizado como *baseline* para comparar as respostas do modelo sem associação de persona com aquelas geradas a partir das personas definidas. Também permitiu identificar tendências gerais nas respostas do modelo, como, por exemplo, se algum dos perfis gerou anotações similares às do perfil genérico, entre outras análises que serviram como referência para apurar os resultados personalizados. Além deste *prompt*, também criamos um *prompt* para a análise das classes originais de "Risco" e "Oportunidade".

Os *prompts* foram escritos em inglês, uma vez que o conjunto de dados estava nesse idioma. Os *prompts* estão apresentados a seguir. No primeiro e segundo *prompt*, troca-se entre o par *risk/opportunity* ou *positive/negative* a depender do tipo de classificação. No *prompt* por persona (segundo *prompt*), substitua `{persona}` por *conservative*, *moderate* ou *aggressive*, e "*a*" ou "*an*", conforme o caso.

(1) ***Prompt* Genérico**: *Act as an investor in the financial market; you are interested in staying informed about the market when making investment choices. Classify the sentiment of the following headline as either* `{"Positive"/"Opportunity"}` *or* `{"Negative"/"Risk"}`.

(2) ***Prompt* por Persona**: *Act as an investor with* `{a/an}` `{persona}` *risk tolerance; you are interested in staying informed about the market when making investment choices. Classify the sentiment of the following headline as either* `{"Positive"/"Opportunity"}` *or* `{"Negative"/"Risk"}`.

3.2 Processamento das Anotações

Para cada iteração de anotação, para extrair um rótulo único — já que as LLMs podem produzir textos diferentes do formato exato esperado — procura-se a ocorrência de "*positive*", "*negative*" (ou "*opportunity*" e "*risk*" se for o caso) em cada resposta, retornando o rótulo correspondente ou "*undetermined*", caso nenhum seja encontrado. Por fim, as respostas são consolidadas em um único conjunto que preserva todas as instâncias originais e adiciona colunas com os rótulos atribuídos por cada modelo. As marcações que são diferentes dos rótulos esperados ("*undetermined*") são excluídas da contagem, garantindo que toda instância receba ao final um rótulo válido.

Ao fim, para avaliar a concordância nas anotações realizadas por perfil de investidor, é consolidado um arquivo em que cada perfil de investidor é representado por uma coluna. Essa coluna contém os rótulos mais frequentes atribuídos a cada instância da base de dados para o respectivo perfil, permitindo uma análise comparativa das classificações predominantes entre os diferentes perfis.

4. RESULTADOS EXPERIMENTAIS

4.1 Configuração dos Experimentos

Foram selecionados os modelos Qwen-3-32B [Yang et al. 2025], Gemma-3-27B [Gemma Team 2025] e Gemini-2.5-flash [Google DeepMind Team 2025]. Optamos por esses modelos devido ao seu equilíbrio entre custo e desempenho[2], adequando-se às nossas restrições. Os parâmetros de configuração do Gemini incluem parâmetros de segurança para assegurar que o modelo não produza conteúdo prejudicial ou inapropriado. Para todas as categorias de tais parâmetros, optou-se pela configuração "BLOCK ONLY HIGH" para que apenas conteúdo considerado de alto risco nessas categorias seja bloqueado, sem restringir excessivamente as saídas. Os demais parâmetros, definidos para todos os modelos são, temperatura de 0, para garantir previsibilidade nas respostas ; top_p de 0,95, o que permitiu a seleção das palavras mais prováveis até atingir 95% da distribuição acumulada, equilibrando a geração

[2] https://artificialanalysis.ai/#intelligence-vs-price

Tabela I. Desempenho (macro) da anotação automática para cada modelo e perfil de *prompt*, comparado com a anotação original.

Modelo	*Prompt*	Acurácia	Precisão	Sensibilidade	F1
Qwen-3-32B	genérico	0,899	0,767	**0,909**	0,813
	conservador	0,889	0,753	0,887	0,796
	moderado	**0,908**	**0,779**	0,881	**0,817**
	arrojado	0,894	0,751	0,807	0,774
Gemma-3-27B	genérico	0,848	0,716	0,897	0,754
	conservador	0,724	0,621	0,843	0,639
	moderado	0,866	0,733	0,908	0,776
	arrojado	0,862	0,717	0,855	0,756
Gemini-2.5-flash	genérico	0,894	0,757	0,873	0,797
	conservador	0,866	0,726	0,874	0,766
	moderado	0,894	0,759	0,890	0,802
	arrojado	0,894	0,755	0,857	0,792

de respostas criativas com a precisão; e *max_ tokens* de 20, para garantir saídas dentro das categorias esperadas, como "*Risk*", "*Opportunity*", "*Positive*", e "*Negative*".

Para medir o desempenho da classificação automática comparado à anotação original, adotamos as métricas de acurácia, precisão, sensibilidade e F1, computadas sobre o conjunto de teste. Para mensurar o grau de concordância da anotação automática entre os quatro perfis de investidor (genérico, conservador, moderado e arrojado), calculamos o Alfa de Krippendorff [Krippendorff 2011], com interpretação que varia de excelente a baixa concordância.

4.2 Classificação de Oportunidade e Risco

A Tabela I apresenta a comparação dos modelos na tarefa original de classificação entre oportunidade e risco. O desempenho foi medido utilizando os quatro perfis distintos, que simulam as perspectivas de um investidor conservador, moderado ou arrojado de acordo com o *prompt* 2, e o investidor genérico com o *prompt* 1 (com as classes *risk-opportunity*) apresentados na seção 3.1. Os melhores resultados por modelo/*prompt* estão sublinhados e os melhores resultados em geral estão em negrito. A Tabela I evidencia três tendências principais: (i) o desempenho varia mais entre as personas do que entre as famílias de LLM de tamanho semelhante; (ii) a persona moderada atinge, de forma consistente, o maior F1 dentro de cada modelo; e (iii) o par Qwen-3-32B e persona moderada obtém o melhor F1 macro geral (0,817).

A análise comparativa revela que a engenharia de *prompt*, por meio da atribuição de personas, atua como um mecanismo importante que influencia o desempenho de LLMs. A suscetibilidade a esse enquadramento semântico, no entanto, varia drasticamente entre arquiteturas. O modelo Gemma-3-27B demonstrou ser o mais sensível, cujo desempenho sofreu um colapso de 13,7 pontos percentuais em F1-score sob a persona conservadora ($F_1 = 0,639$), resultado de uma queda abrupta na precisão que sinaliza um viés de classificação para o rótulo positivo. Em nítido contraste, o Gemini-2.5-flash exibiu uma robustez intrínseca superior, mantendo seus resultados em um intervalo estreito ($0,766 \leq F_1 \leq 0,802$), ainda que também tenha se beneficiado da persona moderada para atingir seu pico de desempenho. O Qwen-3-32B, por sua vez, não apenas validou essa tendência, mas estabeleceu o teto de desempenho ($F_1 = 0,817$) utilizando a perspectiva moderada.

Como consequência destes resultados, podemos responder à pergunta de pesquisa "*Qual o impacto de incorporar vieses característicos dos diferentes perfis de investidores no processo de anotação feito por LLMs?*". A escolha da persona não é um mero ajuste, mas um fator que modula o comportamento e desempenho do modelo. A perspectiva moderada consistentemente alinhou as predições de forma mais fidedigna aos rótulos de referência, enquanto perfis extremos, como conservador e agressivo, introduziram vieses significativos que pioram o desempenho. Além disso, a seleção do modelo subjacente continua a ser um fator determinante. Embora o Qwen-3-32B ofereça o desempenho máximo, o desempenho competitivo e estável do Gemini-2.5-flash o posiciona como uma alternativa viável e eficiente em cenários com restrições de custo ou latência.

4.3 Análise de Concordâncias

Agora que estabelecemos o desempenho das LLMs na atribuição dos rótulos "oportunidade" e "risco" originais, investigamos a sua concordância com os sentimentos "positivo" e "negativo". Para tanto, empregamos o primeiro e o segundo *prompt* descrito na seção 3.1 — com os rótulos positivo/negativo —, que instrui o modelo a classificar cada notícia como positiva ou negativa de acordo com a correspondência positivo-oportunidade e risco-negativo. A avaliação de concordância é conduzida em dois níveis, utilizando o Alfa de Krippendorff: (1) *Concordância inter-persona*, que mede o grau de consistência entre as quatro personas de investidor (genérico, conservador, moderado e arrojado) quando rotulam as notícias como positivas ou negativas; e (2) *Concordância inter-taxonomia*, que quantifica o alinhamento entre os rótulos oportunidade/risco e positivo/negativo atribuídos às mesmas instâncias.

Tabela II. Alfa de Krippendorff para alinhamento entre perfis de investidor e modelos de LLM.

Inter-persona	**Alfa**	**Inter-taxonomia (pos-neg)**	**Alfa**	**Inter-taxonomia (risco-opor)**	Alfa
Genérico	0.872	Gemma	0.866	Gemma	0.804
Conservador	0.761	Gemini	0.927	Gemini	0.961
Moderado	0.893	Qwen	0.983	Qwen	0.905
Arrojado	0.860	—	—	—	—

Concordância inter-persona

Para verificar a concordância entre diferentes perfis e as classes "positivo" ou "negativo", consideramos os resultados de classificação para cada notícia para cada uma das três LLMs experimentadas por persona. Isso resultou em uma tabela para cada persona, e portanto, um valor de Alfa de Krippendorff para cada tabela. Com efeito, ao considerar apenas o *prompt* para a persona genérica (*prompt* 1), obtivemos o resultado de Alfa de Krippendorff de 0.872, de acordo com os resultados apresentados na Tabela II. Segundo a interpretação do Alfa de Krippendorff, valores de $0.8 < \alpha < 1.0$ apresentam alta confiabilidade entre os avaliadores. No entanto, para a persona do tipo conservador, o valor de alfa ficou abaixo, que ainda é considerado uma boa confiabilidade, mas que indica uma discordância maior entre as LLMs quando orientadas de acordo com o perfil de conservador. Esse resultado também está alinhado com a classificação dos rótulos segundo a tarefa original na Tabela I que apresenta desempenho de F_1 pior sob o perfil conservador.

A partir desta análise, percebe-se que a influência das personas é sutil quando as LLMs rotulam entre os sentimentos positivo ou negativo. Considerando a pergunta de pesquisa "*Na ausência de um perfil especificado, o modelo assume implicitamente algum perfil padrão?*", não fica evidente, de acordo com os resultados experimentais apontados, que o perfil genérico adota um viés em particular. No entanto, como mostrado na Tabela I, o desempenho medido de acordo com a métrica de classificação F_1 pode ser significativamente diferente para o investidor genérico, por exemplo, a diferença entre os perfis genérico e conservador para Gemma-3-27B.

Concordância inter-taxonomia

Neste cenário, usamos novamente o primeiro e o segundo *prompt* da seção 3.1 para rotular o dado como "positivo" ou "negativo", para em seguida, analisarmos a correlação deste rótulo com os rótulos originais de oportunidade e risco de cada notícia para cada modelo de LLM diferente. Esses resultados são explorados para responder à pergunta de pesquisa *"Existe um alinhamento entre as classificações de 'oportunidade' e 'risco' feitas pelo mercado e os sentimentos 'positivo' e 'negativo', considerando diferentes perfis de investidores?"*.

A hipótese considera a possibilidade de uma correlação entre os rótulos, mas também reconhece que, dependendo do perfil do investidor, a interpretação de notícias classificadas como oportunidade ou risco poderia divergir. No domínio financeiro, o conceito de risco não é intrinsecamente negativo, assim como o de oportunidade não é necessariamente positivo, sendo ambos dependentes do contexto e da tolerância ao risco de cada perfil de investidor.

O valor obtido pelo Alfa para o modelo Gemini foi 0.927 e 0.961 quando consideramos os *prompts* com os rótulos positivo/negativo e oportunidade/risco respectivamente. Isso indica uma alta confiabilidade entre os rótulos "oportunidade" e "positivo", bem como entre "risco" e "negativo" de acordo com o modelo Gemma para os quatro diferentes perfis de investidor. No entanto, o modelo Qwen apresentou uma queda de aproximadamente 8 p.p entre o positivo-negativo e risco-oportunidade. Embora existam diferenças, os valores ainda refletem alta confiabilidade de concordância entre os avaliadores — neste caso, os diferentes perfis de investimento.

Em conjunto, esses resultados confirmam que existe, de modo geral, um forte alinhamento entre as classificações de oportunidade/risco do mercado e os sentimentos positivo/negativo. Em especial, os perfis moderado e arrojado e LLMs como Gemini e Qwen revelam maiores valores para a concordância. Por outro lado, ainda existe nuances que demarcam a diferença entre as duas taxonomias de positivo/negativo e oportunidade/risco, como demonstrado pela queda no valor de Alfa de Krippendorff, em particular para Gemma e Qwen na Tabela II.

5. CONCLUSÃO

Os resultados mostram um alinhamento consistente entre as classificações "oportunidade" e "positivo", e entre "risco" e "negativo", independentemente das diferentes perspectivas de empresas e perfis de investidores. Essa correlação foi confirmada quantitativamente, por meio do Alfa de Krippendorff de resultados semelhantes entre os pares oportunidade/risco e positivo/negativo com valores chegando a 0.927 e 0.961 respectivamente para a LLM Gemini.

Em relação à capacidade dos LLMs de incorporar os vieses de diferentes perfis de investidores, os efeitos foram limitados, devido ao alto nível de concordância entre as personas. Contudo, nas discordâncias, observou-se que o perfil conservador foi mais propenso a classificar notícias com "negativo", enquanto os perfis moderado e arrojado tenderam a ver mais oportunidades em situações de risco. Um resultado inesperado foi a maior tolerância ao risco do perfil moderado em comparação ao arrojado, indicando que o modelo capta parcialmente os vieses, mas com inconsistências. Por fim, a persona genérica (sem perfil definido) não apresentou alinhamento consistente com nenhum dos perfis analisados, o que pode indicar uma falta de direcionamento claro do modelo na ausência de instruções específicas. Esse comportamento levanta preocupações quanto à confiabilidade e à consistência das análises geradas por LLMs sem um direcionamento explícito, sobretudo em contextos financeiros, nos quais decisões sensíveis dependem de interpretações claras, consistentes e bem fundamentadas.

Trabalhos futuros podem conduzir uma anotação manual dos rótulos de "oportunidade" e "risco" com diferentes perfis de investidores e comparar os achados com a anotação automática. Outras bases de dados com foco mais amplo do que ESG também podem ser uma contribuição relevante.

REFERENCES

Ahmed, T., Devanbu, P., Treude, C., and Pradel, M. Can llms replace manual annotation of software engineering artifacts?, 2024.

Anbaee Farimani, S., Vafaei Jahan, M., Milani Fard, A., and Tabbakh, S. R. K. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowledge-Based Systems* vol. 247, pp. 108742, 2022.

Ardekani, A. M., Bertz, J., Bryce, C., Dowling, M., and Long, S. C. Finsentgpt: A universal financial sentiment engine? *International Review of Financial Analysis* vol. 94, pp. 103291, 2024.

Bashar Yaser Almansour, S. E. and Almansour, A. Y. Behavioral finance factors and investment decisions: A mediating role of risk perception. *Cogent Economics & Finance* 11 (2): 2239032, 2023.

Chen, C.-C., Tseng, Y.-M., Kang, J., Lhuissier, A., Seki, Y., Day, M.-Y., Tu, T.-T., and Chen, H.-H. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, C.-C. Chen, H.-H. Huang, H. Takamura, H.-H. Chen, H. Sakaji, and K. Izumi (Eds.). Association for Computational Linguistics, Bali, Indonesia, pp. 46–50, 2023.

Day, M.-Y. and Lee, C.-C. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. pp. 1127–1134, 2016.

Dickason, Z. and Ferreira, S. Establishing a link between risk tolerance, investor personality and behavioural finance in south africa. *Cogent Economics & Finance* 6 (1): 1519898, 2018.

Dong, X., Wang, S., Lin, D., Rajbahadur, G. K., Zhou, B., Liu, S., and Hassan, A. E. Promptexp: Multi-granularity prompt explanation of large language models, 2024.

Gemini-Team. Gemini: A family of highly capable multimodal models, 2024.

Gemma Team. Gemma 3: Open models from google deepmind. `https://huggingface.co/google/gemma-3-27b-it`, 2025. Modelo open-weight de 27 bilhões de parâmetros com capacidades multimodais.

Google DeepMind Team. Gemini 2.5 flash model card. `https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/gemini`, 2025. Modelo Gemini 2.5 Flash com foco em alta performance e baixa latência.

Heston, S. and Sinha, N. News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal* vol. 73, pp. 1–17, 06, 2017.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., and Xu, Y. Bert-based financial sentiment index and lstm-based stock return predictability, 2022.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. Text mining for market prediction: A systematic review. *Expert Systems with Applications* 41 (16): 7653–7670, 2014.

Krippendorff, K. Computing krippendorff's alpha-reliability, 2011.

Loughran, T. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* vol. 66, pp. 35 – 65, 02, 2011.

Mak, M. and Ip, W. An exploratory study of investment behaviour of investors. *International Journal of Engineering Business Management* vol. 9, pp. 184797901771152, 06, 2017.

Man, X., Luo, T., and Lin, J. Financial sentiment analysis(fsa): A survey. *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, 2019.

Nguyen, L., Gallery, G., and Newton, C. The joint influence of financial risk perception and risk tolerance on individual investment decision-making. *Accounting & Finance* vol. 59, 09, 2017.

OpenAI-Team. Gpt-4 technical report, 2024.

Pangakis, N. and Wolken, S. Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels, 2024.

Pompian, M. *Risk profiling through a behavioral finance lens*. CFA Institute Research Foundation, 2016.

Saad, S. and Saberi, B. Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology* vol. 7, pp. 1660, 10, 2017.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., and Liu, H. Large language models for data annotation and synthesis: A survey, 2024.

United-Nations. Transforming our world: The 2030 agenda for sustainable development, 2015. Accessed: 2025-02-02.

UNPRI. Principles for responsible investment, 2006. Accessed: 2025-02-02.

Wang, Y., Stevens, D., Shah, P., Jiang, W., Liu, M., Chen, X., Kuo, R., Li, N., Gong, B., Lee, D., Hu, J., Zhang, N., and Kamma, B. Model-in-the-loop (milo): Accelerating multimodal ai data annotation with llms, 2024.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Wu, J., Wang, X., and Jia, W. Enhancing text annotation through rationale-driven collaborative few-shot prompting, 2024.

Yadav, S., Choppa, T., and Schlechtweg, D. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4, 2024.

Yang, A., Liu, W., Zhu, B., et al. Qwen3 technical report. `https://github.com/QwenLM/Qwen`, 2025. Descrição técnica do modelo Qwen-3-32B, um grande modelo de linguagem com 32 bilhões de parâmetros.

Yang, S., Rosenfeld, J., and Makutonin, J. Financial aspect-based sentiment analysis using deep representations. *CoRR* vol. abs/1808.07931, 2018.

Zhang, R., Li, Y., Ma, Y., Zhou, M., and Zou, L. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali (Eds.). Association for Computational Linguistics, Singapore, pp. 13088–13103, 2023.