UNIVERSIDADE FEDERAL DE SANTA CATARINA - CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Análise Comparativa de Grandes Modelos de Linguagem na Avaliação de Sentimentos Sobre o IBOVESPA em Redes Sociais

João Vitor Beltramini

Florianópolis
2024

Universidade Federal de Santa Catarina
Departamento de Informática e Estatística

Análise Comparativa de Grandes Modelos de Linguagem na Avaliação de Sentimentos Sobre o IBOVESPA em Redes Sociais

Trabalho de Conclusão de Curso de Graduação em Sistemas de Informação, do Departamento de Informática e Estatística, do Centro Tecnológico da Universidade Federal de Santa Catarina, requisito parcial à obtenção do título de Bacharel em Sistemas de Informação.

Autor: João Vitor Beltramini

Orientador: Prof. Dr. Elder Rizzon Santos

Florianópolis
2024

RESUMO

Com o crescente aumento de investidores brasileiros no mercado de ações e a expansão de discussões financeiras em plataformas como o X (anteriormente Twitter), compreender o sentimento dos investidores tornou-se essencial para processos de tomada de decisão. Este estudo realiza uma análise comparativa de grandes modelos de linguagem (LLMs) aplicados à tarefa de análise de sentimentos sobre o IBOVESPA em redes sociais. A pesquisa examinou a eficácia de dois LLMs distintos: o FinBERT-PT-BR, uma adaptação do modelo BERT para textos financeiros em português, e o GPT-3.5 Turbo, na classificação de sentimentos em publicações relacionadas à BOVESPA. Um levantamento dos principais trabalhos relacionados indicou que a maior parte dos estudos está focada no mercado de ações americano e na análise de sentimentos em língua inglesa, utilizando métodos como aprendizado supervisionado (Naive Bayes, SVM e Random Forest) ou técnicas baseadas em dicionários, como o SentimentAnalysis para R. Trabalhos como o de Bollen, Mao e Zeng (2011) alcançaram uma precisão de 86,7% na análise de sentimentos, enquanto outros, como Pagolu et al. (2016), reportaram acurácia de 70,18%. No entanto, nenhum dos trabalhos revisados utilizou LLMs como o BERT ou GPT, destacando o diferencial do presente estudo. Foram coletados dados de postagens no X ao longo de um período de 30 dias, pré-processados para remover conteúdos irrelevantes e, em seguida, analisados com ambos os modelos. Com base na análise inicial dos dados, foram constatadas algumas diferenças na maneira em que os modelos classificam os sentimentos. Dessa forma, 20% dos textos foram classificados manualmente para calcular as métricas de ambos os modelos. As métricas de desempenho indicam que o GPT-3.5 Turbo atinge um *F1-score* ponderado de 0,863 em comparação com 0,652 do FinBERT-PT-BR, superando-o em todas as principais métricas (precisão, recall e F1-score). O GPT mostrou-se mais eficiente ao classificar os textos como positivos e neutros, enquanto o FinBERT-PT-BR conseguiu classificar a maioria dos textos negativos corretamente, mas mostrou problemas na interpretação de textos subjetivos, ironia e linguagem coloquial. Esses resultados contribuem com o desenvolvimento de índices baseados em sentimentos que podem aprimorar a avaliação do sentimento dos investidores, influenciando potencialmente a tomada de decisões financeiras no mercado de ações brasileiro.

ABSTRACT

With the increasing number of Brazilian investors in the stock market and the expansion of financial discussions on platforms like X (formerly Twitter), understanding investor sentiment has become essential for decision-making processes. This study conducts a comparative analysis of large language models (LLMs) applied to the task of sentiment analysis regarding the IBOVESPA on social networks. The research evaluated the effectiveness of two distinct LLMs: FinBERT-PT-BR, an adaptation of the BERT model for financial texts in Portuguese, and GPT-3.5 Turbo, in classifying sentiments expressed in posts related to BOVESPA.

A review of the main related works indicated that most studies focus on the American stock market and sentiment analysis in the English language, employing methods such as supervised learning (Naive Bayes, SVM, and Random Forest) or dictionary-based techniques, like SentimentAnalysis for R. Studies such as Bollen, Mao, and Zeng (2011) achieved an accuracy of 86.7% in sentiment analysis, while others, like Pagolu et al. (2016), reported an accuracy of 70.18%. However, none of the reviewed studies utilized LLMs such as BERT or GPT, highlighting the unique contribution of this study.

Data was collected from posts on X over a 30-day period, pre-processed to remove irrelevant content, and subsequently analyzed with both models. Based on the initial analysis of the data, some differences were observed in how the models classify sentiments. Accordingly, 20% of the texts were manually classified to calculate the metrics for both models. The performance metrics indicate that GPT-3.5 Turbo achieves a weighted F1-score of 0.863 compared to 0.652 for FinBERT-PT-BR, outperforming it in all key metrics (precision, recall, and F1-score). GPT proved to be more efficient in classifying texts as positive and neutral, while FinBERT-PT-BR correctly classified most negative texts but showed difficulties in interpreting subjective texts, irony, and colloquial language. These results contribute to the development of sentiment-based indices that can enhance the assessment of investor sentiment, potentially influencing financial decision-making in the Brazilian stock market.

SUMÁRIO

1. INTRODUÇÃO

Segundo a B3 (Brasil, Bolsa, Balcão – a bolsa de valores oficial do Brasil), a BOVESPA (Bolsa de Valores de São Paulo) movimentou, no ano de 2021, mais de R$ 7 trilhões com a compra e venda de ações e derivativos, com um aumento de 9% em relação ao ano anterior. Esse valor corresponde, segundo dados disponíveis pelo IBGE, à 80% do PIB do Brasil em 2021 (R$ 8,7 trilhões). Ainda, de acordo com a B3, em janeiro de 2022 cerca de 4,2 milhões de brasileiros possuíam conta em alguma corretora de renda variável, representando um aumento de mais de 1,5 milhão de investidores em comparação com o mesmo período em 2021.

Os investidores individuais (pessoa física) se organizam em grandes comunidades através de várias redes sociais, como o X, Investing, TradersClub e demais plataformas especializadas no mercado financeiro. Segundo a Folha de S. Paulo (2019), nestas plataformas, como o X, grandes investidores institucionais (membros de bancos ou fundos de investimentos), gestores, analistas e criadores de conteúdo interagem entre si e com diversos tipos de investidores menores e entusiastas, onde compartilham os pensamentos acerca do mercado e debatem sobre diversos ativos do mercado financeiro.

Segundo um relatório publicado pela ANBIMA (Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais), em dezembro de 2021, perfis focados em mercado financeiro somavam cerca de 91,5 milhões de seguidores e 406 mil publicações entre as redes sociais mais populares, como Instagram, Youtube, Facebook e o X, sendo esta a rede mais popular entre os seguidores e somando 6 a cada 10 publicações em todas as redes. Entre os assuntos mais abordados estão moedas, ações e criptomoedas.

Devido a grande quantidade de informações que impactam no preço de uma ação, vários métodos qualitativos e quantitativos são utilizados para auxiliar investidores na tomada de decisão. Há fundos específicos em métodos quantitativos, que contam com doutores em física e matemática no seu quadro de sócios e utilizam métodos de *machine learning* e *data science* para tomar decisões. Um desses fundos é o Medallion, liderado pelo matemático americano Jim Simons, que administra uma quantia de US$ 75 bilhões segundo a Forbes (2020) e é um dos fundos mais rentáveis da história.

Para John McCarthy (2007), a inteligência artificial pode ser considerada como "a ciência e a engenharia de fabricar máquinas inteligentes, especialmente programas de computador inteligentes". Aliada a esta definição, McCarthy cita que utilizar uma IA é basicamente utilizar computadores para compreender a inteligência humana. De acordo com o autor, inteligência é a habilidade de conquistar objetivos e atingir metas.

Seguindo uma abordagem para categorizar os tipos de inteligência humana, Stuart Russell e Peter Norvig (1995) propõem quatro definições possíveis de IA: sistemas que pensam como pessoas, sistemas que agem como pessoas, sistemas que pensam racionalmente e sistemas que agem racionalmente.

De acordo com Mike Loukides (2010), "*a ciência de dados permite a criação de produtos de dados*", e não apenas a utilização dos dados para análise. Ou seja, mais do que apenas analisar os dados, a ciência de dados consiste em gerar valor para um produto ou aplicação através dos próprios dados e gerar novos dados como resultado.

Segundo Keith D. Foote (2019), no começo do século XX, o professor e pesquisador Saussure definiu uma linguagem como um sistema onde o significado é criado nas relações e diferenças entre as partes da linguagem. Ou seja, um sistema de linguagem compartilhada faz com que a comunicação seja possível. Com o avanço da pesquisa em computação de Alan Turing e do modelo de Hodgkin-Huxley, que explica como o cérebro utiliza neurônios para formar uma rede elétrica, pesquisas em inteligência artificial e processamento de linguagem natural tornaram-se cada vez mais comuns.

O processamento de Linguagem Natural (PLN), como um aspecto de inteligência artificial, auxilia computadores a entender, interpretar e utilizar a língua humana. Nesse sentido, um modelo de PLN consegue auxiliar em tarefas como categorização de conteúdo, extração de contexto, resumo de textos, descoberta e modelagem de tópicos e análise de sentimentos, por exemplo.

A análise de sentimentos, ou mineração de opinião, consiste em um subcampo de PLN cujo objetivo é analisar fragmentos textuais e avaliar a emoção, opinião ou sentimento do autor em relação a algo e detectar se um texto possui conotação positiva ou negativa (Sobkowicz, Kaschesky, & Bouchard, 2012).

Portanto, o presente trabalho tem como objetivo coletar publicações do X em língua portuguesa sobre o IBOVESPA e comparar o desempenho de duas LLMs – FinBERT-PT-BR e GPT – na classificação de sentimentos sobre a Bolsa de Valores de São Paulo, diferenciando-se dos trabalhos relacionados e das abordagens tradicionais de PLN para análise de sentimentos.

1.2. OBJETIVOS

1.2.1 Objetivo geral

O objetivo geral deste trabalho é realizar uma análise comparativa da eficácia de Grandes Modelos de Linguagem (LLMs) na análise de sentimentos sobre o IBOVESPA, com base em dados coletados da rede social X. Os modelos selecionados para a comparação incluem o FinBERT-PT-BR, uma versão pré-treinada do BERT (*Bidirectional Encoder Representations from Transformers*) com textos financeiros em português brasileiro e o GPT (*Generative Pre-trained Transformer*), referências em processamento de linguagem natural.

1.2.2 Objetivos Específicos

- O1. Analisar o estado da arte em processamento de linguagem natural e análise de sentimento;
- O2. Comparar a performance do FinBERT-PT-BR e GPT na tarefa de análise de sentimentos em relação ao IBOVESPA, utilizando publicações da rede social X;
- O3. Analisar a eficácia dos modelos na identificação de sentimentos por meio de métricas de desempenho;
- O4. Analisar os resultados obtidos dos modelos, identificando padrões e discrepâncias no desempenho entre eles;
- O5. Propor melhorias e sugestões para futuras aplicações de análise de sentimentos sobre o mercado financeiro em português brasileiro, considerando as limitações encontradas e as possíveis implicações dos resultados obtidos;
- O6. Publicar um *dataset* rotulado de publicações do X sobre o IBOVESPA para contribuir com o avanço de pesquisas e o desenvolvimento de novos modelos de linguagem em português.

1.3. METODOLOGIA DE PESQUISA

Para o desenvolvimento deste trabalho, foi realizado um estudo dos conceitos e teorias principais sobre o mercado financeiro, processamento de linguagem natural (PLN) e análise de sentimentos. Esta etapa inicial teve como objetivo criar uma base sólida de fundamentação teórica, analisando diferentes abordagens e aplicabilidades de PLN no contexto financeiro.

Após consolidada a fundamentação teórica, foi realizada uma pesquisa de estado da arte para identificar e selecionar artigos relevantes. Esse levantamento permitiu extrair informações essenciais sobre os modelos e métodos mais recentes utilizados na análise de sentimentos em redes sociais.

Na sequência, foram configurados e comparados os modelos de análise de sentimentos GPT e FinBERT-PT-BR. As publicações foram coletadas por meio da API oficial do X, e, em seguida, os modelos foram ajustados para interpretar os sentimentos nas postagens coletadas. Com os resultados das classificações dos modelos, foi realizada uma análise exploratória dos dados e a rotulação manual de uma amostra do

conjunto de publicações para calcular métricas de desempenho dos modelos, a fim de avaliar a precisão e a sensibilidade na classificação dos sentimentos.

Por fim, os resultados obtidos foram analisados, o que permitiu identificar padrões de comportamento dos modelos ao classificar os sentimentos sobre o IBOVESPA, relacioná-los com as flutuações do mercado e comparar a performance do GPT e do FinBERT-PT-BR, consolidando as conclusões sobre a eficácia dos modelos na análise de sentimentos em um contexto financeiro no ambiente de redes sociais.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Análise de Sentimentos

A análise de sentimentos, situada no domínio do Processamento de Linguagem Natural (PLN) e da Inteligência Artificial, concentra-se na computação e análise de opiniões, sentimentos e emoções expressas em textos. Essa área explora como as atitudes e percepções são manifestadas linguisticamente, identificando se um texto transmite uma perspectiva positiva, negativa ou neutra (Liu, 2012).

A importância da análise de sentimentos tem crescido em paralelo com o aumento do conteúdo gerado por usuários em plataformas digitais, como redes sociais, blogs e fóruns de discussão. A habilidade de processar esses dados textuais para extrair *insights* emocionais e de opinião é crucial para diversas aplicações, desde a análise de tendências de mercado até o monitoramento da saúde mental pública (Pang e Lee, 2008).

Um dos campos de aplicação mais comuns da análise de sentimentos é o monitoramento e a análise da opinião pública em redes sociais. Empresas e organizações utilizam essas análises para entender melhor a percepção do consumidor, permitindo uma resposta mais eficaz a crises ou oportunidades de mercado (Liu, 2012; Pak e Paroubek, 2010).

No contexto financeiro, a análise de sentimentos de notícias e postagens nas redes sociais tem sido explorada como uma ferramenta potencial para prever as tendências do mercado de ações. Embora essa aplicação esteja ainda em estágios iniciais de pesquisa e desenvolvimento, ela representa um novo horizonte na interseção entre dados financeiros e opinião pública como exemplificado pelo trabalho de Bollen, Mao, e Zeng (2011).

A análise de sentimentos enfrenta desafios significativos, principalmente devido à natureza subjetiva e complexa da linguagem humana. A ambiguidade, o uso de ironia e a dependência de contextos culturais e situacionais tornam a tarefa de interpretação automática de textos particularmente desafiadora (Liu, 2012; Pang e Lee, 2008).

A análise de sentimentos pode ser realizada usando uma variedade de técnicas e algoritmos, geralmente classificados em duas categorias principais: métodos baseados em aprendizado de máquina e métodos baseados em léxico.

2.1.1 Aprendizado de Máquina em Análise de Sentimento

No campo da análise de sentimentos, o aprendizado de máquina (Machine Learning - ML) desempenha um papel crucial ao fornecer sistemas capazes de aprender automaticamente e melhorar a partir da experiência. Algoritmos de ML são treinados para classificar textos (como comentários ou *posts*) em categorias sentimentais como positivas, negativas ou neutras. Pang e Lee (2008) oferecem uma visão detalhada das técnicas de ML aplicadas à análise de sentimentos, destacando sua eficácia em diversas configurações.

2.1.2 Árvores de Decisão

As árvores de decisão são um dos métodos mais intuitivos e amplamente utilizados em aprendizado de máquina para classificação e regressão. Uma árvore de decisão é construída a partir de um processo de decisões sequenciais, onde cada nó representa uma característica do conjunto de dados e cada ramificação representa uma regra de decisão, levando a uma conclusão ou previsão no nó folha. Breiman et al. (1984) fornecem uma fundamentação teórica abrangente para árvores de decisão. As árvores de decisão funcionam da seguinte forma:

1. **Seleção de atributos**: No nó de cada árvore, o algoritmo seleciona o atributo que melhor divide o conjunto de dados de acordo com um critério específico, como ganho de informação ou índice de Gini.

2. **Divisão do nó**: Baseando-se no atributo selecionado, o conjunto de dados é dividido em subconjuntos. Cada subconjunto corresponde a um ramo saindo do nó.

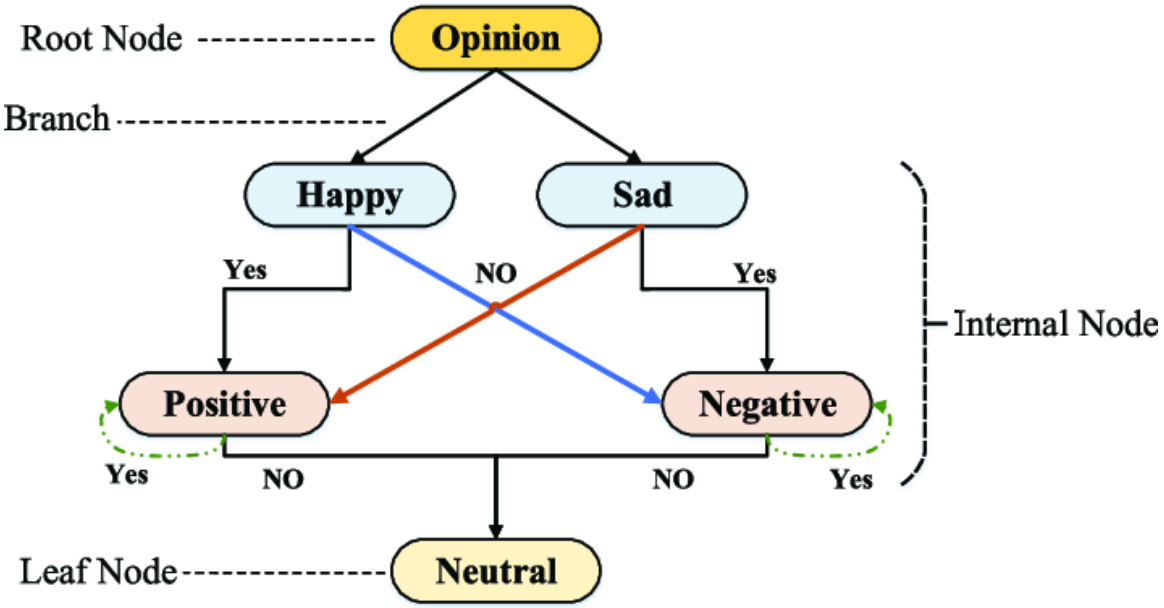
3. **Construção da árvore**: Esse processo é repetido recursivamente para cada ramo, criando uma estrutura de árvore até que um critério de parada seja atingido, como

um número mínimo de registros em um nó folha ou uma profundidade máxima da árvore.

4. **Poda da árvore**: Para evitar o overfitting, as árvores podem ser podadas, removendo seções da árvore que fornecem pouco poder de previsão.

A Figura 1, apresentada abaixo, contém um exemplo de uma árvore de decisão para análise de sentimentos. O processo se inicia no nó raiz, "*Opinion*", onde uma opinião é avaliada. A partir desse ponto, os sentimentos são direcionados para os nós intermediários, "*Happy*" e "*Sad*", que funcionam como pontos de decisão baseados na identificação de felicidade ou tristeza. O fluxo é guiado pelas respostas "*Yes*" ou "*No*", determinando o direcionamento para os nós subsequentes. Caso o sentimento não seja classificado como esperado em um nó específico, a análise pode ser redirecionada, conforme ilustrado pelas conexões cruzadas, como no caso de "Happy" levando a "Negative" ou "Sad" levando a "Positive". Por fim, o processo resulta nos nós folha, que representam as categorias de sentimento.

Figura 1 - Exemplo de árvore de decisão para compra de carro



Fonte: Al Qudah *et al* (2020).

As árvores de decisão podem ser usadas para classificar textos com base em características como frequência de palavras, presença de palavras-chave específicas ou combinações de termos. Por exemplo, um nó pode dividir textos com base na presença

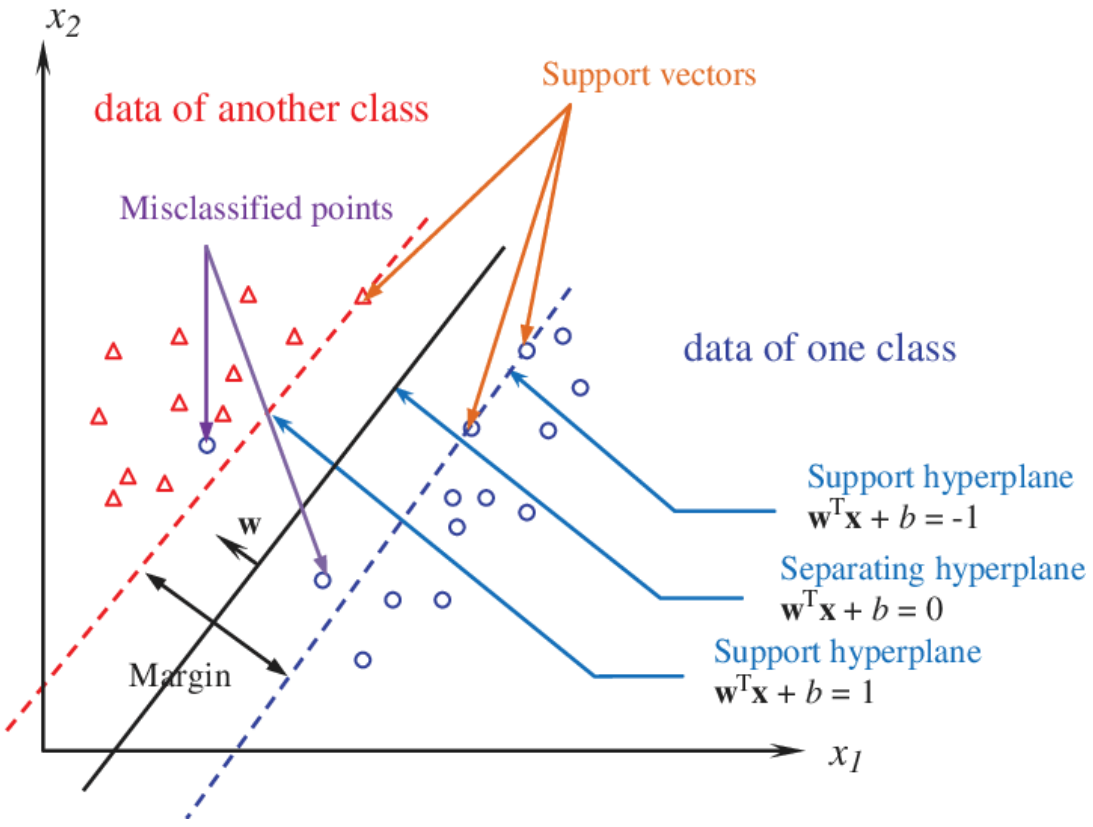
de palavras positivas ou negativas. Quinlan (1986) descreve o uso de árvores de decisão em várias aplicações de classificação, incluindo o contexto de PLN.

2.1.3 Support Vector Machines

Outro método de aprendizagem supervisionada amplamente utilizado em classificação e regressão são as Máquinas de Vetores de Suporte (SVM, do inglês Support Vector Machines). Conforme descrito por Cortes e Vapnik (1995), o SVM é um método que busca a melhor fronteira de decisão linear - o hiperplano que melhor separa diferentes classes de dados.

O SVM funciona identificando o hiperplano que maximiza a margem entre as duas classes no espaço de características. A margem é definida como a distância entre o hiperplano de decisão e os pontos mais próximos de cada classe, conhecidos como vetores de suporte. A escolha do hiperplano é crítica e é feita de modo a garantir a melhor generalização para dados não vistos. Schölkopf e Smola (2002) detalham a teoria por trás do SVM, incluindo conceitos como margens e *kernels*. A Figura 2, apresentada abaixo, exemplifica o funcionamento de uma classificação com SVM.

Figura 2 - Explicação visual de uma classificação com SVM



Fonte: Jwo et al. (2021).

Funções *kernel* permitem que o modelo opere em espaços de características não lineares. *Kernels* comuns incluem polinomial, radial (RBF) e linear. A escolha do *kernel* apropriado depende da natureza dos dados e da tarefa específica. Schölkopf, Burges e Smola (1999) fornecem uma visão abrangente sobre a teoria e aplicação de *kernels* em SVM.

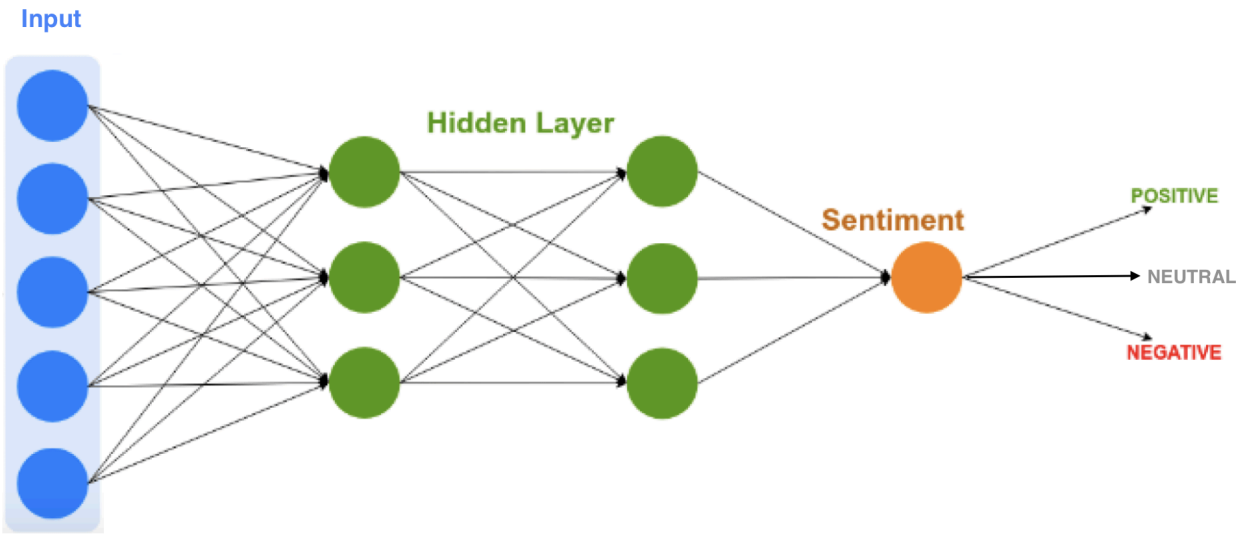
Na análise de sentimentos, o SVM é particularmente útil devido à sua eficácia em lidar com espaços de alta dimensão e sua robustez a overfitting, especialmente em conjuntos de dados grandes. Joachims (2002) explora o uso de SVM em classificação de texto, demonstrando sua aplicabilidade em categorizar opiniões em textos como positivas ou negativas.

2.1.4 Redes Neurais

As redes neurais, inspiradas pelo funcionamento do cérebro humano, são um pilar fundamental no campo do aprendizado de máquina e do processamento de linguagem natural (PLN). Elas são particularmente eficazes na análise de sentimentos devido à sua capacidade de aprender representações complexas e capturar nuances contextuais em dados textuais. LeCun, Bengio e Hinton (2015) oferecem uma visão abrangente das redes neurais profundas e suas aplicações em diversas áreas, incluindo PLN.

Conforme a representação visual de Dang *et al* (2021), apresentada abaixo na Figura 3, redes neurais consistem em camadas de neurônios, onde cada neurônio em uma camada está conectado a vários neurônios na camada seguinte. Essas conexões têm pesos que são ajustados durante o treinamento. Ao fim, é retornado o resultado mais provável. Em análise de sentimentos, as redes neurais são treinadas para classificar textos com base em características aprendidas dos dados.

Figura 3 - Estrutura de uma rede neural



Fonte: Dang *et al* (2021).

As redes neurais são aplicadas para extrair e aprender características relevantes dos textos, como a presença de palavras específicas, frases ou construções gramaticais associadas a sentimentos positivos ou negativos. Socher et al. (2013) exploram o uso de Redes Neurais Profundas para análise de sentimentos, demonstrando como essas redes podem capturar a estrutura semântica e sintática de frases.

2.1.5 Métodos Baseados em Léxico

Métodos baseados em léxico são uma abordagem fundamental na análise de sentimentos, onde o sentimento de um texto é determinado com base na presença de palavras que têm polaridades sentimentais pré-definidas. Esses métodos dependem de léxicos de sentimentos, que são listas de palavras e frases com anotações de sentimentos positivos, negativos ou neutros. Turney e Littman (2003) fornecem uma visão inicial sobre como os léxicos de sentimentos podem ser utilizados para a classificação de sentimentos em textos.

A construção de um léxico de sentimentos pode ser feita manualmente, por especialistas, ou automaticamente, através de algoritmos de PLN. O SentiWordNet, por exemplo, é um léxico amplamente utilizado, construído automaticamente a partir do WordNet, um banco de dados léxico para a língua inglesa. Baccianella, Esuli e Sebastiani (2010) descrevem a metodologia de construção do SentiWordNet.

Na análise de sentimentos, os métodos baseados em léxico envolvem a varredura do texto para identificar e agregar as polaridades das palavras presentes. A pontuação geral de sentimentos pode ser calculada somando as polaridades de todas as palavras identificadas ou por meio de métodos mais sofisticados que levam em conta a intensidade e a negação. Taboada et al. (2011) exploram diferentes estratégias de agregação de polaridades em análises de sentimentos baseadas em léxico.

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial que se concentra na interação entre computadores e linguagem humana, mais especificamente na forma como programar computadores para processar e analisar grandes quantidades de dados de linguagem natural. O objetivo do PLN é permitir que os computadores compreendam, interpretem e manipulem a linguagem humana (Manning e Schütze, 1999).

Uma das principais aplicações do PLN é na análise de textos, permitindo a extração de informações e *insights* a partir de grandes volumes de dados textuais. Isso inclui tarefas como reconhecimento de fala, tradução automática, extração de entidades nomeadas, classificação de texto e análise de sentimentos. Estes avanços têm sido impulsionados tanto pelo aumento da disponibilidade de grandes conjuntos de dados quanto pelo desenvolvimento de modelos de aprendizado de máquina mais sofisticados (Jurafsky e Martin, 2014).

O PLN utiliza várias técnicas para processar e entender a linguagem, incluindo tokenização, análise sintática e representações de texto.

2.2.1 Tokenização

A tokenização é o processo de dividir o texto em tokens, que são as unidades mínimas para análise. Estes podem ser palavras, frases ou até mesmo caracteres. A tokenização é crucial para muitas tarefas de PLN, pois permite que o texto seja transformado de forma que os algoritmos possam processá-lo. Além disso, a tokenização pode envolver a remoção de pontuações e a separação de palavras compostas (Jurafsky & Martin, 2014).

Após realizar a tokenização da frase *"A WEG é uma grande empresa. Os investidores estão otimistas!",* por exemplo, obtemos a seguinte lista de tokens: 'A', 'WEG', 'é', 'uma', 'grande', 'empresa', '.', 'Os', 'investidores', 'estão', 'otimistas', '!'.

2.2.2 Análise sintática

A análise sintática, ou *parsing*, refere-se ao processo de analisar a estrutura gramatical de uma sentença. Esse processo envolve a identificação de componentes como sujeito, verbo e objeto e como eles se relacionam, o que é essencial para entender o significado de uma sentença. Existem diversos métodos de *parsing*, incluindo *parsing* baseado em regras e estatístico. O *parsing* estatístico, por exemplo, utiliza grandes coleções de texto anotado para aprender como analisar estruturas sintáticas (Manning & Schütze, 1999).

2.2.3 Representações de texto

Os métodos de representação de texto consistem, basicamente, em técnicas para converter texto em formatos numéricos que modelos de PLN conseguem processar, capturando aspectos semânticos e contextuais das palavras. Técnicas de representação de texto incluem *Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency)* e *Word Embeddings.*

O *BoW* é uma representação simples onde cada documento é representado pela frequência de palavras nele. Esta técnica, contudo, ignora a ordem das palavras e o contexto (Manning & Schütze, 1999). O BoW é um método simples de implementar e funciona bem para várias tarefas básicas de PLN, como identificação de *spam* e análise de sentimentos. Porém, devido à sua simplicidade, possui algumas desvantagens, como problemas com palavras raras em um conjunto de documentos, falta de contexto semântico entre as palavras e a ordem das palavras não é levada em conta (Brownlee, 2019). Tomando como exemplo as seguintes frases: "o gato comeu o rato" e "o rato comeu o queijo", o vocabulário seria *{o, gato, comeu, rato, queijo}*. A representação BoW para cada frase seria a seguinte, onde cada número corresponde à contagem da palavra no documento:

- "o gato comeu o rato": [2, 1, 1, 1, 0]
- "o rato comeu o queijo": [2, 0, 1, 1, 1]

A representação TF-IDF pode ser considerada uma melhoria em relação ao BoW, uma vez que este pondera a frequência das palavras pela sua raridade em um conjunto de documentos (*corpus*), o que ajuda a destacar palavras importantes que são específicas de um documento (Salton & McGill, 1986). Algumas de suas aplicações incluem ranquear a relevância de um documento com base em uma consulta de pesquisa (Jabri et al., 2018), identificação de palavras-chaves (Lee et al., 2009) e classificação e

agrupamento de textos (Bafna et al., 2016). Supondo um *corpus* com 100 documentos, em que a palavra "economia" aparece em 5 deles, o IDF (*Inverse Document Frequency*) para "economia" seria calculado como log(100/5). Se em um documento específico "economia" aparece 3 vezes e o documento tem 100 palavras, o TF (*Term Frequency*) seria 3/100. Assim, a pontuação TF-IDF para "economia" nesse documento seria (3/100) × log(100/5).

Métodos de *Word Embedding* criam representações vetoriais de palavras que capturam informações contextuais e semânticas, tornando possível identificar relações e semelhanças entre elas. Palavras com significados semelhantes tendem a ter representações vetoriais próximas (Mikolov et al., 2013; Pennington et al., 2014). Ao contrário de abordagens mais antigas, como Bag of Words ou TF-IDF, que se concentram na frequência das palavras, os embeddings capturam a essência semântica e contextual das palavras.

O modelo Word2Vec, desenvolvido por Mikolov et al. (2013), é um dos mais influentes nesta área. Utilizando redes neurais, o Word2Vec aprende representações vetoriais de palavras de grandes conjuntos de dados textuais. Ele opera em dois modos principais: Skip-Gram e CBOW (Continuous Bag of Words). No modelo Skip-Gram, o objetivo é prever o contexto a partir de uma palavra, enquanto no CBOW, o objetivo é prever uma palavra a partir de seu contexto.

Outro modelo importante é o GloVe (*Global Vectors for Word Representation*), criado por Pennington et al. (2014). O GloVe se baseia na co-ocorrência de palavras em grandes *corpora* para produzir um espaço vetorial global, onde a distância entre dois vetores de palavras reflete a proximidade semântica delas.

Esses embeddings têm uma ampla gama de aplicações, desde a melhoria da análise de sentimentos até a tradução automática e reconhecimento de entidades nomeadas. Eles são capazes de capturar nuances semânticas de palavras em diferentes contextos, o que os torna ferramentas poderosas para tarefas complexas de PLN. No entanto, também existem desafios associados aos Word Embeddings. Eles requerem uma quantidade significativa de dados para treinamento e podem ser computacionalmente intensivos. Além disso, lidar com a polissemia – palavras com múltiplos significados – permanece um desafio (Mikolov et al., 2013; Pennington et al., 2014).

2.3 Modelos de linguagem

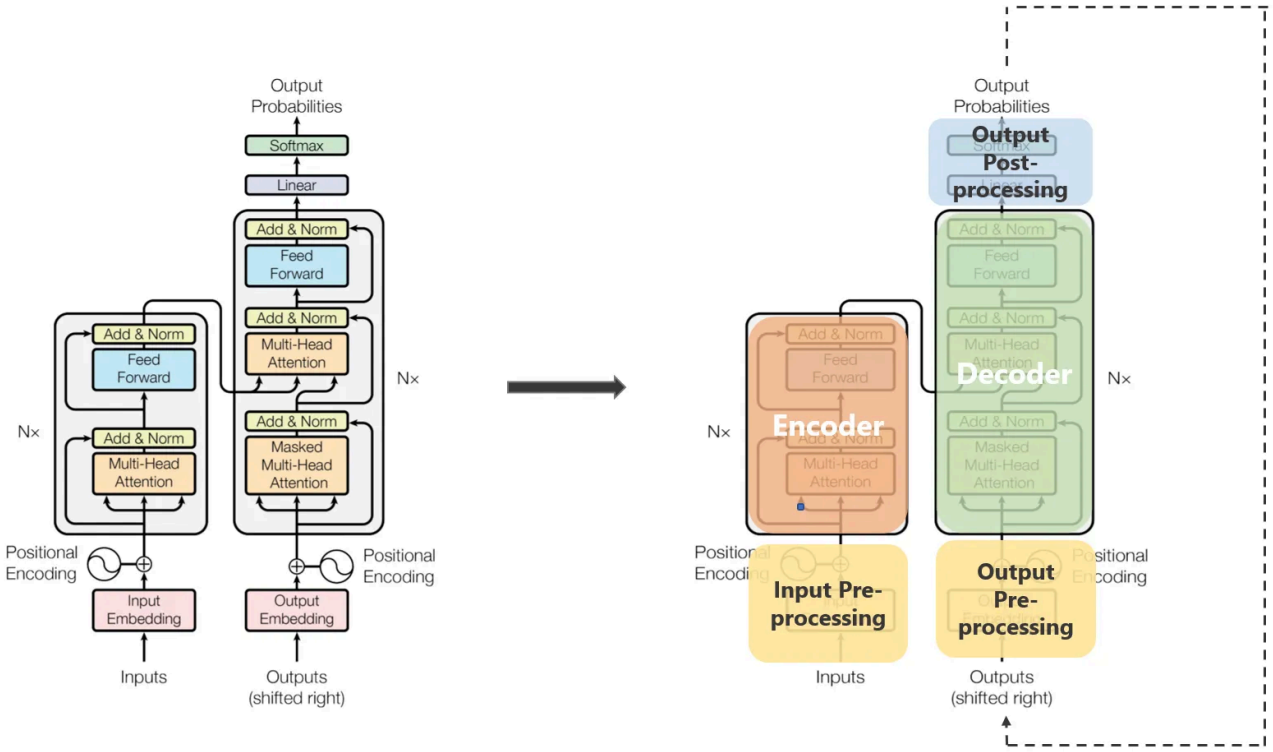
Modelos de linguagem são uma peça central no Processamento de Linguagem Natural (PLN). Eles são sistemas de aprendizado de máquina treinados para entender, interpretar e gerar linguagem humana. Os modelos de linguagem podem ser baseados em diferentes abordagens, como estatística, redes neurais ou *transformers*.

Os modelos estatísticos de linguagem, como os desenvolvidos por Shannon (1948), são baseados na probabilidade de sequências de palavras. Eles utilizam a ideia de que a probabilidade de uma palavra em um texto depende apenas de um número limitado de palavras anteriores. Esses modelos são frequentemente utilizados em tarefas como correção automática e previsão de texto.

Com o advento das redes neurais, os modelos de linguagem evoluíram para serem mais eficientes e precisos. O modelo neural mais comum é o *Recurrent Neural Network* (RNN), que é particularmente bom para lidar com sequências de dados, como texto. O Long Short-Term Memory (LSTM), uma variante do RNN, é especialmente notável por sua habilidade em capturar dependências de longo prazo no texto (Hochreiter & Schmidhuber, 1997).

Uma inovação recente nos modelos de linguagem é o uso de *transformers*, uma arquitetura introduzida por Vaswani et al. (2017). Os *transformers* são baseados em mecanismos de atenção que permitem que o modelo pese a importância relativa de diferentes partes do texto. O BERT (Devlin et al., 2018) é um exemplo proeminente de modelo de linguagem baseado em *transformers*. A Figura 4, apresentada abaixo, contém uma representação visual da arquitetura de um *transformer*.

Figura 4 - Arquitetura de um *transformer*



Fonte: Arquitetura à esquerda por Vaswani et al. (2017), adaptação à direita por Kulshrestha (2020).

Os *Transformers* baseiam-se em uma estrutura de Encoder-Decoder. Um encoder transforma a sequência de entrada em uma série de representações, com cada encoder contendo duas subcamadas: uma para autoatenção múltipla (*Multi-Head Attention*) e outra para uma rede *feed-forward* posicional. Já um decoder produz a sequência de saída, com cada decoder contendo três subcamadas: uma para autoatenção múltipla mascarada, uma para atenção entre o encoder e o decoder, e uma rede *feed-forward*. Essas subcamadas são essenciais para o processamento paralelo e a captação da dinâmica sequencial nos dados.

A arquitetura utiliza o conceito de *Multi-Head Attention* para permitir que o modelo foque simultaneamente em diferentes partes da entrada. Isso é possível por meio de conjuntos de vetores de consulta, chave e valor. A dimensão dos vetores é ajustada para manter a eficiência computacional, e em vez de uma média, os vetores são concatenados e dimensionados antes de passar para a próxima subcamada.

Visto que os Transformers não têm natureza sequencial como os RNNs (*Recurrent Neural Networks*), eles utilizam codificações posicionais para incorporar informações sobre a posição das palavras na sequência. As saídas do decoder são retroalimentadas para a próxima etapa, com a adição de codificação posicional. A autoatenção no decoder é mascarada para evitar que as posições atuais atendam a posições futuras (Kulshrestha, 2020).

Modelos de linguagem têm uma ampla gama de aplicações, incluindo tradução automática, geração de texto, resumo automático e sistemas de resposta a perguntas nos mais diversos campos, desde arte até direito, biologia e medicina. Eles são fundamentais para o desenvolvimento de assistentes virtuais e outras interfaces de linguagem natural (Kaddour et al., 2023).

Apesar dos avanços, os modelos de linguagem ainda enfrentam desafios, especialmente em entender a ambiguidade, a ironia e o contexto cultural. O desenvolvimento contínuo de modelos mais sofisticados e a integração de conhecimento e lógica do mundo real são áreas de pesquisa ativa (Kaddour et al., 2023).

2.3.1 BERT

BERT, que significa Bidirectional Encoder Representations from Transformers, é um modelo de linguagem baseado em *transformers* desenvolvido por Devlin et al. (2018). É notável por sua capacidade de capturar contexto bidirecional em textos, proporcionando um avanço significativo no campo do Processamento de Linguagem Natural (PLN).

A arquitetura do BERT é baseada em um modelo de *transformer*, uma abordagem que se afasta das redes neurais recorrentes tradicionais. Os *transformers* usam mecanismos de atenção que permitem que o modelo pondere a importância relativa de diferentes palavras em uma sentença. Isso permite que o BERT compreenda melhor o contexto em que cada palavra está inserida, diferentemente de modelos anteriores que não capturavam contexto bidirecional de forma tão eficaz (Devlin et al., 2018).

O BERT é pré-treinado em dois conjuntos de dados de texto não rotulados: o *BookCorpus* (conjunto de dados que consiste no texto de cerca de 7.000 livros autopublicados) e a versão em inglês da Wikipedia. Esse pré-treinamento é realizado através de duas tarefas: previsão de palavras mascaradas e previsão de próximas sentenças. Após o pré-treinamento, é possível realizar um ajuste fino (*fine-tune)* do BERT com dados adicionais rotulados para tarefas específicas de PLN, como classificação de

texto, análise de sentimentos e reconhecimento de entidades nomeadas (Devlin et al., 2018).

Desde o seu lançamento, o BERT tem sido utilizado em uma variedade de aplicações de PLN, como classificação de notícias (Kuncahyo et al., 2021) e análise de sentimentos (Hoang et al., 2019). O BERT estabeleceu novos padrões de desempenho em várias tarefas de benchmark de PLN. Sua introdução levou a um rápido desenvolvimento de modelos baseados em transformadores, incluindo variantes como RoBERTa, DistilBERT, e ALBERT, que buscam otimizar e expandir as capacidades do modelo original (Schütz et al., 2021).

2.3.2 FinBERT-PT-BR

Segundo Santos, Bianchi e Costa (2023), autores do modelo, o FinBERT-PT-BR é um modelo de análise de sentimentos desenvolvido para textos em português no contexto do mercado financeiro, com o objetivo de capturar nuances específicas do vocabulário financeiro em português.

O FinBERT-PT-BR foi treinado a partir de um *fine-tune* do modelo BERTimbau, uma variante do BERT ajustada para o português, com 1,4 milhões de textos financeiros coletados de veículos de comunicação especializados, incluindo Valor Econômico, Exame e InfoMoney, abrangendo notícias de 2006 a 2022. O modelo alcançou uma perplexidade de 1,24, uma métrica satisfatória para este domínio específico (Santos, Bianchi e Costa, 2023). A perplexidade (PPL) é uma das métricas mais comuns para avaliação de modelos de linguagem. Ela é definida como a exponenciação da média do *log-likelihood* negativo de uma sequência, calculada com base na exponencial natural (base *e*). Dessa forma, a perplexidade mede o quão bem o modelo prevê uma sequência de tokens, sendo que valores menores indicam melhor desempenho preditivo (Hugging Face, 2024).

2.3.3 GPT

Segundo Radford et al. (2018), O GPT, abreviação de *Generative Pretrained Transformer*, representa uma evolução significativa no campo do Processamento de Linguagem Natural. Desenvolvido pela OpenAI, o GPT e suas versões subsequentes, como GPT-2 e GPT-3, marcaram uma mudança na maneira como os modelos de linguagem são construídos e utilizados. O GPT original, introduzido por Radford et al. (2018), já demonstrava uma capacidade notável de gerar texto coerente, preparando o caminho para avanços futuros.

Em 2019, Radford, Wu, Child, et al., expandiram essa ideia com o GPT-2, um modelo mais robusto com 1,5 bilhão de parâmetros. A maior inovação veio com o GPT-3, descrito em detalhes por Brown, Mann, Ryder, et al. (2020). Com 175 bilhões de parâmetros, o GPT-3 não apenas aprimorou a capacidade de geração de texto, mas também demonstrou habilidades de aprendizado com poucos exemplos (f*ew-shot learning*), tornando-o um modelo de linguagem extremamente versátil e poderoso.

Os modelos GPT têm sido empregados em uma variedade de aplicações, desde a geração automática de texto até sistemas sofisticados de chatbot e tradução de linguagem. A versatilidade do GPT-3, em particular, abre caminho para novas possibilidades em interfaces de linguagem natural e aplicações de PLN. No entanto, essa capacidade também traz desafios, especialmente em termos de viés e uso ético dos modelos gerativos, como discutido por Bommasani, Hudson, Adeli, et al. (2021).

Enquanto os modelos GPT demonstram avanços impressionantes em termos de capacidade de geração de linguagem e aprendizado de máquina, eles também enfrentam críticas. Questões relacionadas ao viés intrínseco nos dados de treinamento, a possibilidade de uso mal-intencionado e a falta de transparência nos modelos de linguagem de grande escala são áreas de preocupação ativa na comunidade de PLN, conforme apontado por Bommasani e colaboradores.

2.4 Mercado de ações

O mercado de ações é um importante componente do sistema financeiro, tanto para empresas quanto para investidores. De acordo com Graham e Dodd (2014), as ações representam uma fração do capital social das companhias, conferindo aos seus proprietários direitos políticos e econômicos, como o recebimento de dividendos e a possibilidade de participar nas decisões estratégicas da empresa.

É importante destacar a existência de diferentes tipos de ações, como as ordinárias e preferenciais. As ações ordinárias proporcionam aos seus detentores direito de voto nas assembleias da empresa e, consequentemente, uma participação mais ativa nas decisões estratégicas da organização. Já as ações preferenciais conferem prioridade no recebimento de dividendos e reembolso do capital em caso de liquidação da empresa (ASSAF NETO, 2019).

Para Hull (2018), o mercado de ações é uma forma de investimento com possibilidade de altos retornos financeiros, embora também envolva riscos consideráveis. Investidores que desejam se expor ao mercado de ações devem estar dispostos a lidar com a volatilidade e incerteza presentes nesse mercado. Para tanto, é necessário que realizem uma análise criteriosa das informações financeiras da empresa e do contexto macroeconômico em que ela está inserida.

Segundo Damodaran (2017), a avaliação de ações é uma tarefa complexa e envolve a análise de diversos fatores, como o fluxo de caixa futuro esperado da empresa, sua estratégia de crescimento e os riscos envolvidos no negócio. Desse modo, é fundamental que investidores e gestores financeiros estejam atentos às particularidades do mercado de ações e utilizem ferramentas de análise adequadas para tomadas de decisão mais precisas e conscientes.

2.5 X

O X, anteriormente conhecido como *Twitter,* é uma plataforma de mídia social conhecida por seu formato único de microblogging, onde os usuários podem publicar mensagens curtas, conhecidas como "*posts*" (anteriormente como *tweets)*, limitadas a 280 caracteres. Desde seu lançamento em 2006 por Jack Dorsey, Noah Glass, Biz Stone e Evan Williams, o X evoluiu para se tornar uma ferramenta global de comunicação e informação, sendo amplamente utilizado para disseminação de notícias, discussões políticas, marketing, entretenimento e interações sociais. A plataforma se destaca por sua capacidade de fornecer atualizações em tempo real e facilitar discussões públicas, tornando-se uma fonte valiosa para análise de tendências e opiniões públicas. O X também tem um papel significativo na mobilização social e no jornalismo cidadão, como discutido por Kwak et al. (2010) em seu estudo sobre as características únicas do X como um meio de comunicação.

3. TRABALHOS RELACIONADOS

Com o intuito de identificar trabalhos relacionados ao projeto, foram realizadas buscas por estudos que utilizassem métodos de processamento de linguagem natural e análise de sentimentos extraídos de redes sociais e sites de notícias relacionados ao mercado financeiro. Nesta seção, serão analisados, discutidos e comparados os três estudos mais relevantes encontrados em relação ao projeto atual. Foram selecionados trabalhos em três diferentes esferas: análise de sentimento de publicações do X em

inglês; análise de sentimento de publicações do X em português; e uma dissertação sobre análise de sentimentos por meio de notícias. Também serão discutidos, brevemente, outros trabalhos relacionados. Por fim, o capítulo será concluído com uma avaliação sobre os tópicos relevantes e as metodologias utilizadas no trabalhos relacionados, e como estas podem auxiliar no desenvolvimento do projeto.

Para selecionar os trabalhos, foram realizadas pesquisas por palavras-chave como "*sentiment analysis", "twitter", "stock market"* em plataformas como o *IEEE Xplore, Google Scholar* e o repositório de trabalhos acadêmicos da UFSC. A pesquisa sobre os trabalhos relacionados foi realizada entre setembro e dezembro de 2023, e foram escolhidos trabalhos publicados entre 2010 e 2023.

A escolha por três estudos principais foi motivada pela relevância, complementaridade e adequação ao contexto do projeto. O primeiro estudo, voltado para publicações em inglês, permite uma visão mais ampla e consolidada de metodologias utilizadas internacionalmente. O segundo estudo, focado em publicações em português, oferece uma análise alinhada com o escopo linguístico do projeto, refletindo os desafios específicos da língua portuguesa. Já o terceiro trabalho, uma dissertação que aborda análise de sentimentos por meio de notícias, amplia a perspectiva ao introduzir uma fonte distinta de dados, essencial para compreender diferentes aplicações no mercado financeiro.

A seleção desses trabalhos foi direcionada pela qualidade metodológica, abrangência dos dados e pela proximidade temática com o projeto atual. Outros trabalhos foram brevemente discutidos para garantir uma visão mais ampla do estado da arte, contudo, os estudos selecionados representam de forma eficaz as principais abordagens em análise de sentimentos e suas aplicações ao mercado financeiro.

3.1. Twitter mood predicts the stock market

A ideia principal do artigo de Bollen, Mao e Zeng (2011) baseia-se em investigar como o conceito de *behavioral economics*, cujo principal argumento é de que as emoções podem afetar profundamente a tomada de decisões e o comportamento individual, pode afetar sociedades e a tomada de decisões coletivas. Mais especificamente, o estudo investiga como as emoções coletadas através de publicações do X podem estar relacionadas a indicadores econômicos, como por exemplo o índice Dow Jones Industrial Average (DJIA), indicador da performance de 30 grandes companhias na bolsa de valores americana.

De acordo com a Teoria do Mercado Eficiente (Fama, E. F. 1965), o preço de ações é amplamente afetado por novas informações e notícias, as quais são imprevisíveis e, portanto, fazem com que não seja possível prever o preço de ações com mais de 50% de precisão. No entanto, novas pesquisas sugerem que alguns indicadores prévios podem ser extraídos de mídias sociais para prever mudanças em vários indicadores econômicos e comerciais.

Com base nessa premissa, os autores realizaram a coleta de 9.853.498 *posts* publicados entre 28 de fevereiro de 2008 até 19 de dezembro de 2008. Cada publicação contém informações importantes, como o conteúdo publicado, um identificador e a data e horário de publicação. Na fase de coleta dos *posts*, são removidas as *stop-words* (palavras de parada, como *"the", "and", "is", "in", "on", "a", "of"*, etc) e pontuações. Então, são agrupados *posts* publicados na mesma data e que contenham declarações explícitas de humor (*"i feel", "i am feeling", "i'm feeling", "i dont feel", "I'm", "Im", "I am",* e *"makes me"*). *Posts* que contêm as expressões "*http:*" e "*www.*" foram excluídos da análise para reduzir o ruído.

Os autores dividiram a análise dos dados em três etapas. Na primeira delas, utilizaram as ferramentas *OpinionFinder* (OF) e *Google Profile of Mood States* (GPOMS). O OpinionFinder é um software de análise de sentimentos disponível ao público que determina a subjetividade em nível de sentença, identificando a polaridade emocional (positiva ou negativa) das frases. Os autores o utilizaram para analisar o conteúdo emocional de grandes conjuntos de publicações, usando o léxico OF para determinar a proporção de postagens positivas e negativas em um dia específico. O léxico subjetivo do OF é baseado em trabalhos anteriores e inclui 2718 palavras positivas e 4912 negativas, e cada *post* é analisado para identificar termos negativos e positivos com base neste léxico. No entanto, o OF segue um modelo unidimensional de humor, distinguindo apenas entre sentimentos positivos e negativos.

Para abordar essa limitação, os autores desenvolveram a ferramenta GPOMS, que mede estados de humor em 6 dimensões diferentes: Calmo, Alerta, Certo, Vital, Gentil e Feliz. As dimensões e o léxico do GPOMS são baseados no instrumento psicométrico *Profile of Mood States* (POMS-Bi). Para adaptá-lo ao X, o léxico original de 72 termos do POMS foi expandido para 964 termos, permitindo que o GPOMS capture uma variedade maior de termos de humor nas publicações.

Para comparar as séries temporais do OF e GPOMS, elas são normalizadas para *z-scores* com base em uma média local e desvio padrão dentro de uma janela deslizante

de *k* dias. Essa normalização faz com que todas as séries temporais oscilem em torno de uma média zero e sejam expressas em uma escala de 1 desvio padrão.

Ambas as ferramentas foram testadas pelos autores em *posts* publicados entre 5 de outubro a 5 de dezembro de 2008 para avaliar sua capacidade de medir o humor público durante eventos como a eleição presidencial dos EUA (4 de novembro) e o Dia de Ação de Graças (27 de novembro). Os resultados indicaram que o OF capturou com precisão as reações emocionais positivas do público a ambos os eventos. Por outro lado, o GPOMS ofereceu uma visão mais detalhada do humor público, destacando emoções variáveis em torno do período eleitoral, com uma queda na Calma em 3 de novembro (dia anterior à eleição), níveis elevados de Vitalidade, Felicidade e Gentileza no dia da eleição, e um retorno ao padrão básico depois disso. O Dia de Ação de Graças mostrou apenas um pico de Felicidade no dia, segundo a medição GPOMS. Uma comparação entre as ferramentas revelou que a dimensão Feliz do GPOMS se assemelhava mais à tendência de humor do OF. A partir de uma análise quantitativa usando regressão múltipla, os autores encontraram correlações entre as tendências de humor do OF e dimensões específicas do humor GPOMS (Certeza, Vitalidade, Felicidade). No entanto, nem todas as dimensões do GPOMS se alinharam com o OF, sugerindo que o GPOMS capta um espectro mais amplo de nuances do humor público do que a abordagem unidimensional OF.

Bollen, Mao e Zeng, após concluírem que as variações no humor público podem ser capturados por uma série temporal de humor derivada do GPOMS e OpinionFinder em *posts*, investigaram se estas variações de humor correlacionam-se com mudanças no mercado de ações, particularmente os valores de fechamento do DJIA (*Dow Jones Industrial Average*). Usando a análise de causalidade de Granger, buscaram determinar se uma série temporal pode prever a outra, em vez de estabelecer uma relação direta de causa e efeito. O estudo foi conduzido pelos autores entre 28 de fevereiro e 3 de novembro de 2008, para remover a influência de eventos importantes como a eleição presidencial e o Dia de Ação de Graças. As séries temporais foram produzidas para 342.255 *posts* neste período, e o fechamento diário do DJIA foi coletado através do *Yahoo! Finance.* Os resultados indicaram que, entre várias dimensões de humor, apenas a dimensão de humor "Calmo" teve uma relação preditiva estatisticamente significativa com o DJIA para certos intervalos de tempo. No entanto, alguns eventos, como um anúncio de resgate bancário importante, causaram desvios. Isso mostra que, enquanto o

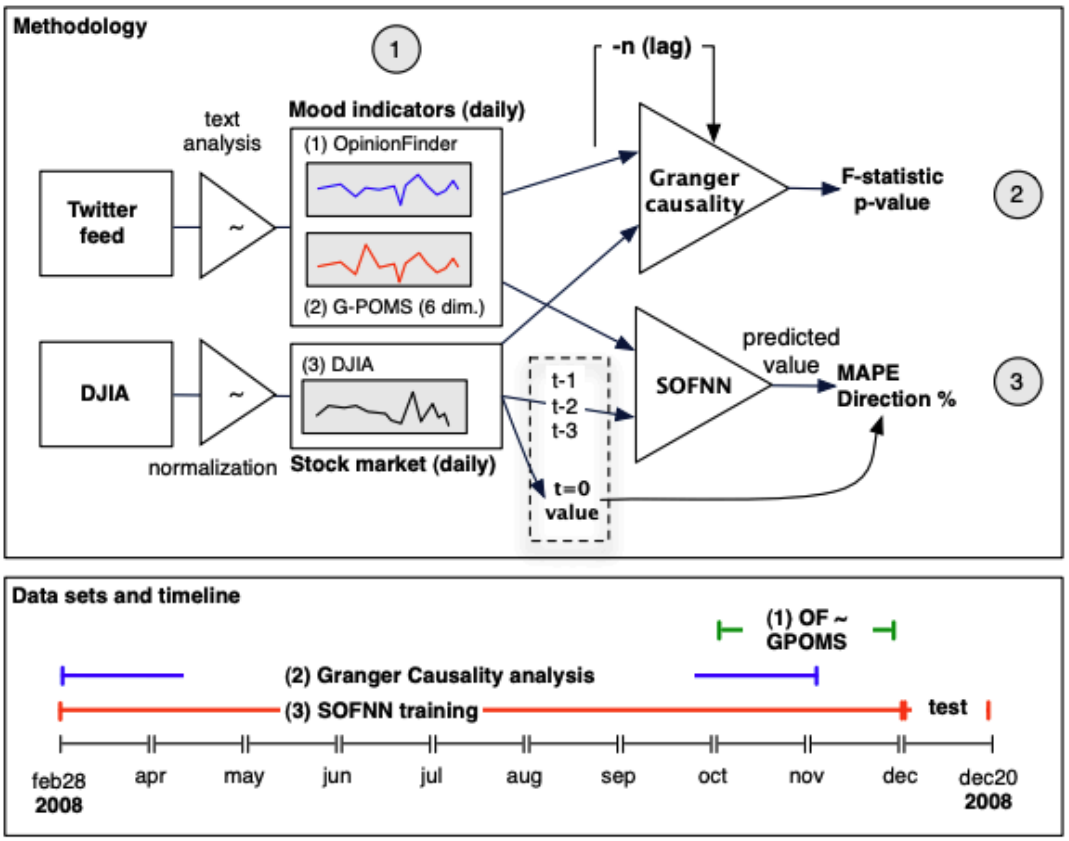
humor público pode oferecer alguns *insights* preditivos, notícias inesperadas continuam sendo um fator significativo na dinâmica do mercado de ações.

Os autores utilizaram a análise de causalidade de Granger para investigar a relação entre certas dimensões do humor e o DJIA. Embora a análise de Granger seja linear, a relação entre o humor público e os valores do mercado de ações provavelmente é não-linear. Assim, foi utilizado um modelo de Rede Neural Fuzzy Auto-Organizável (SOFNN) para fazer previsões do DJIA baseadas em (1) os últimos 3 dias de valores do DJIA e (2) a combinação desses valores com várias permutações da série temporal do humor.

Por fim, os resultados revelaram que, de todas as dimensões de humor, apenas "Calmo" e em certa medida "Feliz", têm uma relação significativa de causalidade Granger com o DJIA. A adição de "Calmo" ao modelo melhorou significativamente a precisão das previsões, enquanto algumas outras dimensões do humor, como "Certeza" e "Vital", não contribuíram positivamente para a previsão. Ao testar a linearidade da relação entre "Calmo" e "Feliz" com o DJIA, constatou-se que uma combinação linear dos dois produz resultados piores do que usar apenas "Calmo". Portanto, a relação entre essas dimensões do humor e o DJIA é não-linear.

A precisão da previsão foi avaliada pelos autores através do Erro Percentual Médio Absoluto (MAPE) e da precisão da direção durante um período de teste específico. A pesquisa concluiu que a análise de humor pode oferecer insights valiosos para prever movimentos do mercado de ações, mas os resultados são influenciados pela natureza não-linear das relações e pelas dimensões específicas do humor analisadas. A Figura 5, apresentada abaixo, contém uma representação visual do fluxo das etapas da metodologia de Bollen, Mao e Zeng.

Figura 5 - Etapas da metodologia



Fonte: Johan Bollen, Huina Mao, Xiao-Jun Zeng (2010).

A partir disso, os autores concluem que o estado geral de humor do público pode ser, sim, monitorado através do X e que a precisão de previsões do fechamento do DJIA pode ser significativamente melhorada através da inclusão das dimensões específicas de humor do GPMS. Também foi descoberta uma precisão de 87,6% na previsão de mudanças diárias para cima e para baixo dos valores de fechamento do DJIA, além de uma redução do erro percentual médio em mais de 6%.

3.2 Metodologias para Análise de Sentimentos de Tweets sobre o Mercado Financeiro

Neste artigo, Medeiros (2019) descreve e analisa duas metodologias diferentes para analisar sentimentos de publicações relacionadas a ativos listados em bolsas de valores. Ambas as metodologias compartilham algumas etapas, como a transformação dos *posts* em um espaço vetorial, a redução de dimensionalidade e a extração de tópicos. Também é realizado o agrupamento dos dados para fins de comparação entre os grupos

e os tópicos extraídos. Após a comparação, é realizada uma análise exploratória visual e, por fim, são utilizados algoritmos de *machine learning* para classificar os dados.

A primeira metodologia é focada na classificação de sentimentos de postagens em Língua Portuguesa sobre o mercado brasileiro e utiliza para isso, em grande parte, aprendizagem supervisionada. A partir do conjunto de dados, os *posts* são rotulados de acordo com os sentimentos principais da teoria psico evolucionária de Plutchik: Confiança, Desgosto, Alegria, Tristeza, Antecipação, Surpresa, Raiva e Medo.

Com essa metodologia, os autores encontraram uma forte relação entre os grupos das publicações e os tópicos. Os classificadores identificaram melhor a ausência do que a presença de um sentimento em um *post*, já que de acordo com o autor, *"existem mais posts não rotulados com um determinado sentimento do que rotulados com um determinado sentimento".*

Para chegar ao resultado da primeira metodologia, o autor utilizou as técnicas *Principal Component Analysis* e *t-Stochastic Neighbor Embedding* para reduzir a dimensionalidade e viabilizar a análise visual. Após a redução, foi realizado o agrupamento de publicações através do algoritmo *KMeans* e os tópicos foram extraídos com os métodos *Latent Dirichlet Allocation* e *Non-Negative Matrix Factorization*. Por fim, são utilizados os classificadores *NaiveBayes*, *Support Vector Machine* e *Random Forest* para analisar e comparar o desempenho entre eles. Com a primeira metodologia, foram atingidos os seguintes resultados: Precisão de 0,72, *Recall* de 0,52 e F1-score de 0,6.

Já a segunda metodologia trata-se de uma expansão da primeira e possui um foco maior na previsão de variação das ações. O autor utilizou um algoritmo de aprendizagem de máquina não supervisionada para analisar os sentimentos de *posts* escritos em Língua Inglesa sobre a empresa americana *Apple, Inc.* para prever, através de aprendizagem supervisionada, a variação diária de valor das suas ações na Bolsa de Valores Nasdaq. De acordo com o autor, foi possível notar uma relação entre os grupos dos *posts* e os tópicos extraídos, além de demonstrar um *F1-Score* acima de 0,5 (satisfatório) na classificação de variação das ações e uma taxa de acertos de 0,895. Ao contrário da primeira metodologia, nesta os classificadores possuem desempenho similar entre todos os sentimentos.

Para reduzir a dimensionalidade, nesta etapa Medeiros utilizou o algoritmo *Principal Component Analysis* e, para extrair os tópicos, o autor utilizou as técnicas *Latent Dirichlet Allocation* e *Non-Negative Matrix Factorization*. Os *posts* foram agrupados por meio dos algoritmos *K-Means* e *Expectation Maximization* e, por fim, para prever a

variação das ações da Apple, Inc, o autor utilizou Regressão Logística, *Naive-Bayes*, *Support Vector Machine* e *Random Forest*.

3.3 Sentimento de notícias e investimento estrangeiro em carteira no Brasil

Através de sua tese de dissertação, Cambará (2019) tem como objetivo analisar como o sentimento influencia no fluxo de investimento estrangeiro em carteira para o Brasil. A autora desenvolveu um índice a partir da análise de notícias do site do *Wall Street Journal*, o jornal diário de negócios mais lido nos EUA. Foram selecionadas 26.406 notícias econômicas, financeiras e políticas que mencionavam a palavra-chave "*Brazil"*, publicadas de janeiro de 1999 a maio de 2018.

A análise de sentimentos foi feita por Cambará usando o pacote *SentimentAnalysis*, na linguagem R, que avalia se um texto tem conotação positiva ou negativa com base em um dicionário. O método "*bag of words*" foi usado, tratando cada palavra individualmente e desconsiderando a ordem delas. Para classificar os sentimentos de cada palavra, foi utilizado o dicionário de uso geral *Harvard IV-4*, que classifica milhares de palavras em diversas categorias, incluindo emoções e sentimentos. A autora optou por empregar as categorias *Positiv* e *Negativ*, abrangendo mais de três mil palavras categorizadas como positivas ou negativas.

O sentimento de cada notícia foi determinado pela proporção de palavras positivas e negativas, resultando em um valor entre -1 e 1, tal qual pode ser visto na fórmula abaixo. O índice mensal representa a média de sentimento das notícias do mês, convertida em porcentagem.

$$Sentimento = \frac{número\ de\ palavras\ positivas - número\ de\ palavras\ negativas}{número\ de\ palavras\ positivas + número\ de\ palavras\ negativas}$$

O principal objeto de estudo na análise da autora foram os fluxos de investimento estrangeiro em carteira, que são transações com títulos de dívida e de participação no capital que não implicam em controle significativo sobre uma empresa. Estes fluxos foram analisados em relação ao Balanço de Pagamentos, especificamente a subconta de investimento em carteira da conta financeira. O estudo focou na participação de investidores estrangeiros no país, analisando passivos de ações locais e títulos de renda fixa. Os valores líquidos, que representam as alterações nas posições dos investidores, foram obtidos do site do BCB e ajustados pelo Índice de Preços ao Consumidor dos EUA.

Fundos de investimento foram excluídos da análise, pois seus dados começaram apenas em janeiro de 2010.

Diversas variáveis foram consideradas pelo estudo para entender os determinantes dos fluxos de investimento, incluindo indicadores econômicos e financeiros, como taxas de juros, crescimento do PIB, desempenho da bolsa de valores e risco-país. Além disso, foram considerados índices de incerteza, como o *Economic Policy Uncertainty Index for Brazil* (EPUBR), o *Global Economic Policy Uncertainty* (GEPU) e o Índice de Incerteza da Economia (IIE).

Para estudar o fluxo de capitais, Cambará utilizou a análise "*push-pull*", que categoriza os determinantes que influenciam decisões de investidores internacionais em fatores "*push*" (externos, que impulsionam investimentos em outros países) e "*pull*" (características internas de um país que influenciam o risco e retorno dos investimentos). Koepke (2015) identifica três determinantes principais para cada tipo de fator. Nos fatores "push", comuns são a aversão global ao risco (como o índice VIX, que mede a volatilidade dos preços das opções dos ativos do índice S&P500), taxa de juros (geralmente baseada na economia dos EUA) e crescimento do produto. Quanto aos fatores *pull*, normalmente são considerados o crescimento do produto, índice de retorno de ações e índice de risco-país

Cambará analisou o comportamento das variáveis internas (fatores *pull,* endógenas) e externas (fatores *push,* exógenas) do Brasil usando modelos VARX (vetores auto-regressivos com variável exógena), que permitem avaliar o comportamento das variáveis internas sem a necessidade de compreender completamente suas inter-relações.

Para isso, a autora descreve um modelo matemático específico, como visto na Figura 6 apresentada abaixo, no qual Yt representa variáveis domésticas e x representa variáveis externas. Cambará ressalta que, considerando o Brasil como uma economia menor e aberta, as variáveis internas não impactam as externas. Ao estimar o modelo, a autora utilizou o critério de Schwarz (SC) para determinar a defasagem ideal das variáveis, optando pelo modelo mais parcimonioso, descrito abaixo, sendo L o operador de defasagem.

Figura 6 - Modelo VARX

$$y_t = \Phi(L)B(L)x_t + \sum_{i=0}^{\infty} \Phi_i u_{t-i}$$

Fonte: Cambará (2019)

Além disso, Cambará empregou um modelo DCC-GARCH(1,1) para investigar a dinâmica conjunta da variância condicional. O DCC é um modelo que analisa correlações que variam com o tempo e, com isso, foi possível avaliar como essas correlações mudaram durante o período analisado e identificar características da dinâmica entre as variáveis.

Cambará também abordou um possível problema com o modelo VARX: ele pode não considerar efeitos simultâneos entre as variáveis. Assim, a autora optou por usar o MGM (método generalizado dos momentos) em duas etapas de Hansen como uma alternativa. Esta técnica depende da escolha correta de instrumentos, que devem estar correlacionados com as variáveis endógenas e ortogonais aos erros. Para determinar quais variáveis eram endógenas, a autora realizou um teste baseado em Hausman. Com um nível de significância de 10%, DCÂMBIO (diferença da série não estacionária CÂMBIO) foi identificado como endógeno para AÇÕES, SELIC para FIXA, e DCÂMBIO e DIBOV (diferença da série não estacionária IBOV) para CARTEIRA, e a autora usou suas duas primeiras defasagens como instrumentos.

Além disso, Cambará investigou o impacto do sentimento nas expectativas da taxa de câmbio e examinou diferentes variáveis de atividade econômica, como IBC-Br e PIB. A autora reestimou seus modelos considerando índices de incerteza, seja substituindo ou complementando o índice de sentimento.

Ao analisar a composição do dicionário usado para avaliar o sentimento das notícias, Cambará confirma o resultado encontrado por Baumeister et al. (2001): emoções negativas são mais representadas linguísticamente do que as positivas. No dicionário Harvard IV-4 utilizado, há 22,5% mais palavras negativas do que positivas. No entanto, essa predominância negativa não é refletida nos textos. Das 26.406 notícias analisadas, apenas 11,8% tinham índices negativos.

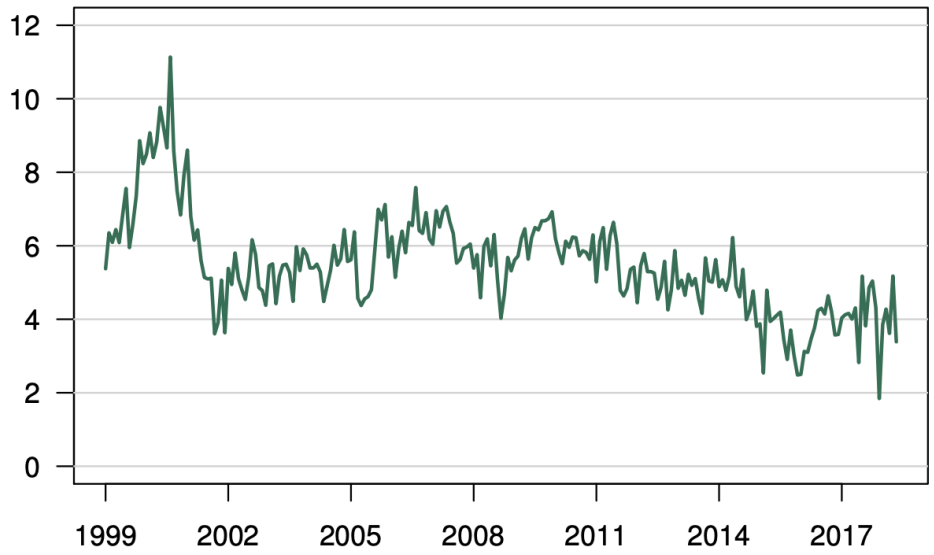
O índice de sentimento, que expressa a média mensal das notícias, teve uma média de 5,4847, variando entre 11,1353 e 1,8430. Ao longo do período estudado, que

começou no início do segundo governo de Fernando Henrique Cardoso e acompanhou eventos importantes no cenário brasileiro, o índice mostrou uma evolução.

Nos primeiros dois anos, houve um crescimento acentuado, atingindo um pico em 2000. No entanto, 2001 viu uma queda brusca devido a eventos como a crise energética, os ataques de 11 de setembro e a crise argentina. O índice aumentou novamente em 2002, apesar das preocupações sobre a eleição de Lula.

O sentimento melhorou até 2007, mas teve uma queda com a crise financeira de 2008. A eleição de Dilma Rousseff em 2010 viu uma queda significativa no índice, que continuou com sua reeleição em 2014. Durante esse período, eventos como a Operação Lava Jato, o impeachment de Dilma, e outros eventos políticos impactaram o sentimento. A evolução do sentimento ao longo do tempo pode ser observada na Figura 7 abaixo.

Figura 7 - Índice de sentimento



Fonte: Cambará (2019).

As análises estatísticas realizadas pela autora indicaram que o índice de sentimento é bastante previsível e meses com notícias similares tendem a seguir uns aos outros. Por exemplo, após um evento impactante, a volatilidade do sentimento leva, em média, cerca de um ano para retornar ao normal.

Quando o sentimento é positivo, há um aumento nos fluxos, com seu impacto diminuindo em aproximadamente dois anos. Algumas variáveis, como taxas de juros, tiveram impactos inesperados nos fluxos. Há uma volatilidade observada nos fluxos de capital, sendo os fluxos de ações mais instáveis do que os de renda fixa. Estes são

afetados tanto por eventos locais quanto globais, com fluxos de ações mais afetados por cenários internacionais e renda fixa por situações nacionais. Além do sentimento, a autora argumenta que indicadores como o IBOVESPA também são cruciais para o fluxo de ações.

Quanto à influência da incerteza e do sentimento sobre o investimento estrangeiro no Brasil, enquanto o índice IIE, que considera a volatilidade do mercado de ações, mostrou impacto significativo em certos fluxos, os índices EPU-BR e GEPU não apresentaram relevância. Ao incorporar o sentimento na análise, a importância da incerteza no investimento diminuiu. A autora conclui que, no contexto brasileiro estudado, sentimento e incerteza têm impactos diferentes no investimento estrangeiro e podem atuar como complementares ou substitutos, dependendo de sua definição.

Segundo Cambará, em relação às taxas de câmbio no Brasil, constatou-se que o sentimento não tem impacto significativo, assim como as transações correntes. Já o EMBI (risco-país) e o erro de previsão foram relevantes em todos os cenários. A influência da incerteza variou conforme o índice: a incerteza interna (EPUBR) não mostrou relevância, enquanto a incerteza global foi importante em horizontes curtos e a incerteza interna (IIE) em horizontes intermediários. O estudo realizado pela autora sugere que a incerteza pode afetar as expectativas cambiais, ao contrário do sentimento. Cambará conclui que, embora o sentimento impacte o investimento estrangeiro, não afeta as expectativas cambiais, mas a incerteza pode, dependendo de sua definição.

Por fim, a autora reafirma a importância do uso de variáveis macroeconômicas em adição à análise de sentimento, concluindo que a política econômica é de extrema importância para o fluxo de capital.

3.4 Outros trabalhos

Pagolu et al (2016), utilizando um conjunto de dados de 250.000 publicações do X relacionadas à Microsoft de agosto de 2015 a agosto de 2016, juntamente com os preços das ações da empresa, exploram como a opinião pública nas redes sociais impacta as tendências do mercado de ações. Os *posts* foram filtrados usando palavras-chave específicas e processados através de várias etapas, incluindo tokenização, remoção de palavras irrelevantes e eliminação de caracteres especiais. O estudo utilizou dois métodos de análise textual: N-gram e Word2vec. Esses métodos foram usados para analisar os sentimentos expressos nos *posts* e determinar sua correlação com os preços das ações da Microsoft.

A análise de sentimentos das publicações foi realizada usando um algoritmo Random Forest, e os resultados mostraram precisão semelhante para as representações N-gram e Word2vec, com preferência pela última devido à sua eficácia em conjuntos de dados maiores e significado consistente de palavras. Os preços das ações foram classificados com base em sua alta ou queda, e uma análise de correlação foi conduzida usando esses dados junto com os resultados da análise de sentimentos. O estudo alcançou uma precisão de cerca de 70% na previsão dos movimentos dos preços das ações com base nos sentimentos do X, indicando que esta correlação destaca o potencial do uso da análise de sentimentos nas redes sociais como uma ferramenta para analisar tendências do mercado de ações.

Mittal e Goel (2012) utilizam análise de sentimento e aprendizado de máquina, em particular Redes Neurais Fuzzy Auto-Organizáveis (SOFNN), para prever os movimentos do mercado de ações com base no humor público coletado através de publicações do X e nos valores históricos do Dow Jones Industrial Average. Eles alcançaram uma precisão de 75,56% para o período entre junho e dezembro de 2009. O trabalho estende a pesquisa de Bollen et al., que alcançou 87% de precisão com uma abordagem semelhante. A metodologia inclui o pré-processamento de dados do DJIA e do X, análise de sentimentos para determinação de humor (calmo, feliz, alerta, gentil) e aprendizado de modelo usando diferentes algoritmos, sendo o SOFNN o mais eficaz. Eles também propõem um novo método de validação cruzada para dados financeiros e implementam uma estratégia básica de gestão de carteira com base em suas previsões. O estudo conclui que o sentimento público, especificamente a calma e a felicidade, pode prever os valores do DJIA e que o uso dessas previsões para a gestão de carteiras pode ser interessante. No entanto, os autores reconhecem limitações, como o conjunto de dados não representar totalmente o sentimento público e a correlação indireta entre os usuários do X e os investidores do mercado de ações, sugerindo essas áreas para pesquisa futura.

3.5 Considerações

A tabela 1 contém uma comparação entre os cinco trabalhos mais relevantes relacionados discutidos anteriormente, contendo informações da análise, como a fonte dos dados, o mercado estudado, a amostra e os indicadores econômicos utilizados.

Tabela 1 - Comparação de informações utilizadas para a análise

Trabalho	Fonte dos dados	Mercado analisado	Amostra	Indicadores econômicos
Bollen, Mao e Zeng (2011)	publicações no X	Americano	> 9 milhões	DJIA
Medeiros (2019)	publicações no X	Americano e Brasileiro	84.369	Variação diária de ações da empresa Apple, Inc
Cambará (2019)	*Wall Street Journal*	Brasileiro	26.406	Indicadores apresentados na tabela 3 (p. 36)
Pagolu et al (2016)	publicações no X	Americano	250.000	DJIA
Mittal e Goel (2012)	publicações no X	Americano	> 476 milhões	DJIA

Fonte: Elaborado pelo autor (2024).

A tabela 2 compara os métodos de análise de sentimento, demais técnicas estatísticas utilizadas e os resultados significativos dos trabalhos

Tabela 2 - Comparação dos resultados e métodos e técnicas utilizadas para a análise

Trabalho	Método de Análise de Sent.	Algoritmo	Métricas utilizadas	Resultados Significativos
Bollen, Mao e Zeng (2011)	OpinionFinder GPOMS	SOFNN Granger causality	Acurácia	86,7%
			MAPE	1,79%
			P-value	< 0,05 (calmo)
Medeiros (2019)	Aprendizagem supervisionada NaiveBayes SVM Random Forest PCA Atribuição de	Naive-Bayes SVM Random Forest Regressão Logística	Precisão	<= 0,72 (I);
			Recall	<= 0,52 (I);
			F1-score	<= 0,6 (I); > 0,5 (II)
			Taxa de acertos	<= 0,895 (II)

	sentimentos a rótulos usando nuvens de palavras e análise semântica.			
Cambará (2019)	Pacote SentimentAnalysis para R dicionário Harvard IV-4	Bag of Words	VARX (sentimento)	Ações: 0,1543 Renda fixa: 0,2645 Carteira: 0,6030
Pagolu et al (2016)	N-gram Word2vec	Random Forest	Acurácia	70,18%
			Precisão	0,711
			Recall	0,702
			F1-score	0,690
Mittal e Goel (2012)	Profile of Mood States expandido, filtragem, contagem de palavras para mapear os sent. dos *posts* em estados de ânimo	SOFNN Granger causality	MAPE	11.03%
			Acurácia	75.56%
			P-Values	0.0069 (calmo); 0.0658 (feliz)

Fonte: Elaborado pelo autor (2024).

A Tabela 3, apresentada a seguir, contém a lista de indicadores utilizados na análise de Cambará, bem como suas respectivas descrições.

Tabela 3 - Indicadores econômicos utilizados na dissertação de Cambará (2018)

Indicador	**Descrição**
CARTEIRA	Passivos líquidos de Investimento em Carteira como porcentagem do PIB
AÇÕES	Passivos líquidos de ações negociadas no país
FIXA	Passivos líquidos de títulos de renda fixa negociados no país
EMBI	Índice de risco-país calculado pelo J. P. Morgan Chase
CÂMBIO	Índice da taxa de câmbio real
SELIC	Taxa de juros interna
PIB	Índice do produto interno bruto

IBCBR	Índice de Atividade Econômica do Banco Central do Brasil
NFSP	Necessidade de Financiamento do Setor Público
IBOV	IBOVESPA
TB3M	Taxa de juros do título de três meses do Tesouro dos Estados Unidos no mercado secundário
GDP	Índice do produto interno bruto dos Estados Unidos
VIX	Índice de volatilidade utilizado para calcular a aversão global ao risco
EPUBR	Índice de Incerteza da Política Econômica para o Brasil
GEPU	Incerteza na Política Econômica Global
IIE	Índice de Incerteza da Economia
$EC\hat{A}MBIO_{t+h}$	Série das expectativas da taxa de câmbio sem transformação, em que h é o número de períodos à frente e t, o instante do tempo em que foi observada
TC	Série do saldo de Transações Correntes
$ERRO_h$	Erro de previsão dado pela diferença entre a expectativa h meses à frente para o mês t e o valor observado da taxa de câmbio nominal no mesmo mês

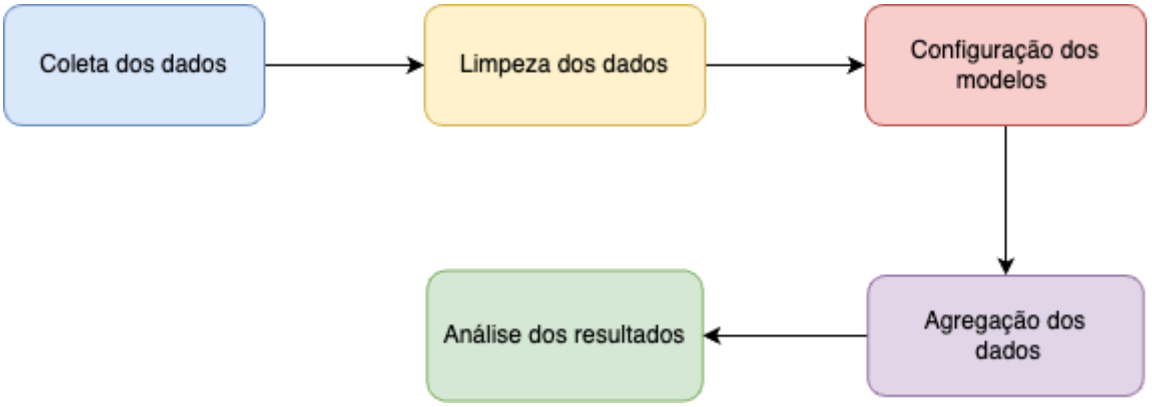
Fonte: Cambará (2019).

A maior parte dos trabalhos está relacionado ao mercado de ações americano e análise de sentimentos na língua inglesa, e nenhum deles utilizou algum *Large Language Model*, como o BERT ou GPT, que hoje em dia são os tipos de modelos mais avançados para análise de sentimento disponíveis. Dessa forma, o presente trabalho tem como objetivo coletar publicações do X em língua portuguesa sobre o IBOVESPA e analisar o desempenho dos modelos GPT e FinBERT-PT-BR na tarefa de classificação de sentimentos.

4. DESENVOLVIMENTO

Conforme descrito no Capítulo 2, o objetivo deste trabalho é comparar o desempenho de diferentes *LLMs* (Grandes Modelos de Linguagem) para classificar sentimentos oriundos de publicações do X sobre a Bolsa de Valores de São Paulo. Para cumprir esse objetivo, o trabalho foi dividido em 5 etapas principais: Coleta dos dados, limpeza dos dados, configuração dos modelos, agregação dos dados e análise dos resultados. O fluxo dessas etapas pode ser observado na Figura 8, apresentada abaixo.

Figura 8 - Etapas de desenvolvimento do trabalho



Fonte: Elaborado pelo autor (2024).

Na primeira etapa, as publicações foram extraídas através da API do X em um período de 30 dias, entre 30 de julho a 30 de agosto de 2024. Na etapa de limpeza dos dados, foram removidas postagens irrelevantes, links, menções, emojis, hashtags redundantes, e houve normalização de texto e datas, além de exclusão de duplicatas, resultando em 7511 publicações preparadas para análise. Na etapa de configuração dos modelos, foi realizada a análise de sentimento de cada publicação com os modelos FinBERT-PT-BR e GPT-3.5 Turbo. Já na etapa de agregação dos dados, o conjunto de dados com as publicações e sentimentos foi organizado em uma nova tabela contendo o sentimento predominante de cada dia e os respectivos valores de fechamento da BOVESPA. Por fim, foi criado um capítulo específico contendo a análise dos resultados, no qual foi analisada uma amostra de exemplos representativos das publicações e foi classificado 20% dos textos manualmente para calcular e avaliar as métricas de desempenho de cada modelo. Além disso, foram gerados gráficos para analisar a distribuição dos sentimentos ao longo do tempo e o nível de concordância entre os modelos.

A organização deste capítulo segue a sequência lógica das etapas desenvolvidas para alcançar o objetivo da pesquisa, apresentando cada fase de forma detalhada e em ordem cronológica.

4.1 Coleta dos dados

Nesta fase inicial do trabalho, os dados foram coletados diretamente do X por meio da API oficial da plataforma, com o objetivo de obter publicações relacionadas ao índice IBOVESPA. Para isso, utilizou-se a versão básica da API X v2, cujo acesso tem um custo mensal de 100 dólares americanos e permite a extração de até 10.000 *posts* publicados nos últimos sete dias a partir do momento da requisição.

Após o processo de registro e obtenção do token de acesso, foi desenvolvido um *script* em Python para coletar as postagens em português que continham as palavras-chave "ibov" ou "ibovespa". As publicações analisadas foram restritas a usuários com contas públicas. O período de coleta abrangeu do dia 30 de agosto ao dia 30 de setembro de 2024, totalizando 9.013 postagens.

Cada *post* coletado foi processado de modo a extrair as seguintes informações: o identificador único da postagem (ID), o texto da mensagem, a data e o horário da publicação e o identificador e nome de usuário do autor.

Os *posts* coletados foram organizados em uma tabela, facilitando o processamento e a análise dos dados. A Tabela 4, apresentada a seguir, oferece uma amostra dos dados extraídos:

Tabela 4 – Amostra da tabela de *posts* coletados

post_id	text	created_at	author_username
1818874183096373646	Pra toda ação que eu olho é só surpresa positiva, quase todas com operação azeitada com a margem lá em cima, IBOV deve disparar é só uma questão de tempo	2024-08-01T04:59:32.000Z	Backpack29
1822031447743750345	O Ibovespa terminou o dia com alta de 1,52%. O #MinutoB3 traz os principais destaques que tiveram impacto nas bolsas do Brasil, Estados Unidos e China. #pregão #EducaçãoFinanceira #SiteB3 #BoraInvestir #Ibovespa https://t.co/zuoBasPvgx	2024-08-09T22:05:23.000Z	B3_Oficial
1822033099972022294	Ibovespa volta aos 130 mil pontos, após diminuir a chance de recessão dos EUA, dólar cai a R$ 5,51 https://t.co/hFlJlXzNBv	2024-08-09T22:11:57.000Z	BandHoje
1822033558954762747	@MisesVsCerize alguem sabe me informar quando é o crash do Ibovespa	2024-08-09T22:13:46.000Z	moralmentecris

1822036406426071056	Com falas de Galípolo e IPCA, dólar recua a R$ 5,51; Ibovespa fecha em alta https://t.co/73pFjwAezb	2024-08-09T22:25:05.000Z	OGlobo_Economia
1821907313109696675	De volta ao mercado após um período de férias e surpreso pela entrada de gringo no IBOV nos últimos dias. Mas no mercado americano não acho que veremos novas máximas históricas tão cedo.	2024-08-09T13:52:07.000Z	Mestre_Don
1820425596452684029	Por enquanto o crash tá só nas bolsas asiáticas. Europa caindo menos de 3% em média ainda... Acho que o ibov também não vai muito além disso (hoje) https://t.co/o7IeTUXesW	2024-08-05T11:44:18.000Z	LucasHolifer

Fonte: Elaborado pelo autor (2024)

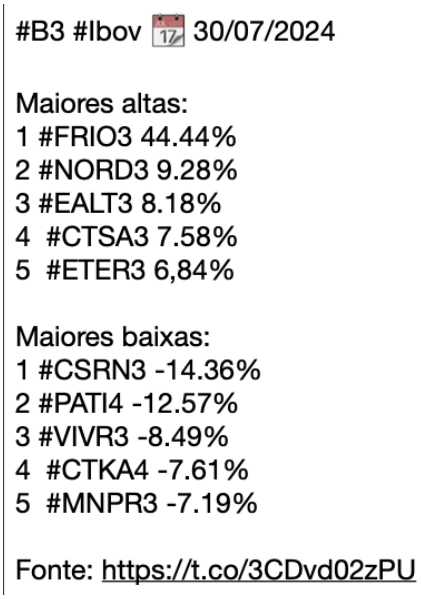
4.2 Limpeza dos dados

Após a coleta das publicações na fase inicial, é necessário realizar um processo de preparação e limpeza dos dados para garantir que os textos estejam adequados para a análise de sentimentos. Essa fase de preparação tem como objetivo remover ruídos e inconsistências nos dados.

A primeira ação realizada foi a remoção de postagens irrelevantes ou de baixa qualidade. Para isso, foram excluídos da base postagens publicadas por autores previamente identificados como indesejáveis, como bots ou perfis que constantemente publicam conteúdo automatizado ou *spam*. Esses perfis foram listados em uma tabela de exclusão e identificados através de uma análise exploratória inicial.

Além disso, as publicações que contém mais de 3 símbolos de porcentagem (%) também foram removidas. O motivo dessa remoção é que muitas postagens são realizadas automaticamente por *bots* e contém um panorama geral do mercado financeiro no dia, mas são apenas dados que não contém nenhuma opinião. Exemplo na Figura 9 abaixo:

Figura 9 - Texto de publicação automatizada com dados gerais sobre o mercado em determinado dia



Fonte: Elaborado pelo autor (2024).

Os *posts* passaram por um processo de remoção de links, menções a outros usuários e emojis, que poderiam distorcer os resultados da análise de sentimentos. O uso de emojis, por exemplo, embora possa ter significado em contextos informais, não contribui de maneira confiável para análises automatizadas neste contexto, uma vez que essas ferramentas priorizam a interpretação do texto escrito.

Outro aspecto importante foi o tratamento das hashtags. Embora as hashtags possam fornecer contexto relevante em algumas análises, neste caso, foram removidas aquelas que não agregam valor, como as relacionadas ao próprio tema de pesquisa (ex: #ibov ou #ibovespa). Isso foi feito para evitar que elas gerassem redundância ou distorcessem a compreensão do conteúdo real do *post*. Em contrapartida, hashtags que faziam parte de uma frase foram mantidas, sem o símbolo “#”, preservando o contexto sem comprometer a análise semântica do texto. Para isso, foram utilizadas expressões regulares para identificar hashtags em sequência ou em uma linha separada.

Foi também realizada a normalização dos dados temporais, convertendo a data e hora de publicação dos *posts* do padrão UTC para o fuso horário local de Brasília para garantir a correta interpretação da cronologia dos *posts* em relação aos eventos do mercado financeiro no Brasil.

Além disso, o texto das postagens passou por um processo de normalização, onde todo o conteúdo foi convertido para letras minúsculas e também foram removidos espaços em branco no início e no final de cada postagem, garantindo a consistência textual após as etapas de limpeza anteriores.

Em seguida, aplicou-se um filtro para garantir que as publicações incluídas na análise fossem diretamente relevantes ao IBOVESPA e à análise. Para isso, foram consideradas apenas as publicações que continham pelo menos uma das palavras-chave específicas: "ibov", "ibovespa", "bolsa" e "bovespa".

Por fim, foi feita a remoção de postagens duplicadas, tanto de *posts* com conteúdo idêntico quanto aqueles com o mesmo identificador único para garantir que cada *post* seja contabilizado apenas uma vez.

Ao fim da etapa de limpeza, obteve-se um total de 6471 publicações. A Tabela 4, apresentada a seguir, contém alguns exemplos de publicações antes e depois da limpeza.

Tabela 4 - Amostra para comparação das publicações antes e depois da limpeza dos dados

Pré-limpeza	**Pós-limpeza**
PETRÓLEO E MINÉRIO ARRASTAM O IBOV PARA O RISCO DA PERDA DOS 126MIL https://t.co/AZvglKqn7H	petróleo e minério arrastam o ibov para o risco da perda dos 126mil
@MisesVsCerize Sem cisne negro ou explosão do IBOV, essa aposta já está decidia	sem cisne negro ou explosão do ibov, essa aposta já está decidia
Suzano (SUZB3) conclui compra de ativos no Mato Grosso do Sul #economia #investimentos #negócios #investidor #ibovespa #ações #mercadofinaceiro https://t.co/UVDytc5YjO	suzano (suzb3) conclui compra de ativos no mato grosso do sul
Se o #LVOL11 estivesse disponível desde 2003, teria reduzido, em média, 20% da volatilidade quando comparado ao Ibovespa. Com a proposta de ser mais resiliente em momentos de adversidade da Bolsa, o #LVOL11 teria apresentado uma alta de +1.792%, contra +934% do Ibovespa.	se o lvol11 estivesse disponível desde 2003, teria reduzido, em média, 20% da volatilidade quando comparado ao ibovespa. com a proposta de ser mais resiliente em momentos de adversidade da bolsa, o lvol11 teria apresentado uma alta de +1.792%, contra +934% do ibovespa.

Quando o IBOV sobe muito até o santo desconfia 🤔	quando o ibov sobe muito até o santo desconfia
#IBOV segue fraco no curto prazo. Mercado sem comprador, gringo que vinha puxando o rally saiu fora. Pressão vendedora tem sido principalmente nas empresas ligadas a commodities. #VALE3 fazendo graça nos R$ 60. #EMBR3 #JBSS3: os papéis mais fortes de bolsa. https://t.co/Lb1EFKSeeg	ibov segue fraco no curto prazo. mercado sem comprador, gringo que vinha puxando o rally saiu fora. pressão vendedora tem sido principalmente nas empresas ligadas a commodities. vale3 fazendo graça nos r$ 60. embr3 jbss3: os papéis mais fortes de bolsa.

Fonte: Elaborado pelo autor (2024).

4.3 Configuração dos modelos

4.3.1 FinBERT-PT-BR

Nesta etapa do trabalho, utilizou-se o modelo pré-treinado FinBERT-PT-BR, uma versão pré-treinada do modelo BERT de língua portuguesa BERTimbau, especificamente treinado para realizar análise de sentimentos no contexto financeiro. O objetivo principal desta fase é aplicar o modelo FinBERT-PT-BR para classificar os *posts* coletados e filtrá-los em três categorias de sentimento: positivo, negativo e neutro.

Para a execução do modelo de análise de sentimentos, foi necessário carregar dois componentes principais:

1. Tokenizador: O tokenizador converte o texto bruto em uma sequência de tokens, que é a forma de representação do texto compreensível pelo modelo. No caso, foi utilizado o *BertTokenizer* associado ao modelo "FinBERT-PT-BR".
2. Modelo FinBERT-PT-BR: O modelo foi implementado utilizando a arquitetura *BertForSequenceClassification*, que é uma versão do BERT ajustada para tarefas de classificação de texto. O FinBERT-PT-BR foi pré-treinado com dados financeiros em português, tornando-o adequado para analisar textos financeiros em português.

Devido à limitação de recursos computacionais e para otimizar o tempo de processamento, os textos foram classificados em lotes de 16 publicações por vez, permitindo que o processamento ocorra de forma mais eficiente através da *GPU*.

Os textos são convertidos em tokens e, caso excedam o limite de 512 tokens — limite máximo permitido pelo modelo BERT —, são truncados para conter apenas 511 tokens, uma vez que o token especial (CLS), que indica o início da sequência, ocupa uma das posições disponíveis. Após a conversão e truncamento, os textos tokenizados são transformados em tensores e passados como entrada para o modelo FinBERT-PT-BR. Como saída, o modelo gera uma probabilidade para cada uma das três classes (positivo, negativo e neutro), com base nas saídas do modelo (*logits*). A probabilidade mais alta define a classe de sentimento atribuída a cada *post*.

Os textos que excedem o limite de 512 tokens são marcados e processados com truncamento. Essa informação é registrada em uma coluna adicional para que possa ser considerada em análises posteriores, evitando que resultados incorretos decorrentes da perda de parte do conteúdo não sejam ignorados. Ao fim da análise, apenas duas publicações foram truncadas por exceder o limite de 512 tokens, e portanto removidas do conjunto de dados..

Após a classificação, os resultados são adicionados à tabela com os seguintes campos:

1. predicted_label: O rótulo de sentimento previsto para cada *post* (positivo, negativo ou neutro).
2. positive, negative, neutral: As probabilidades associadas a cada classe de sentimento.

A tabela atualizada é então salva em um novo arquivo CSV, que contém tanto o conteúdo original dos textos quanto as predições geradas pelo modelo. A Tabela 5, apresentada abaixo, contém uma amostra dos resultados após a análise.

Tabela 5 - Amostra dos resultados da classificação com o FinBERT-PT-BR

post_id	text	created_at	author_username	predicted_label	positive	negative	neutral
1819159261001535538	apesar das dificuldades, o ibovespa continuou melhorando as condições de suas tendências nos prazos mais curtos, particularmente a dos volumes, que retomaram o viés de alta no curto prazo	2024-08-01T20:52:20.000000Z	alphafintec	POSITIVE	0.93	0.033	0.035
1829215851209470236	ibovespa ao vivo: bolsa cai e tenta manter os 136 mil pontos; azul4 despenca via	2024-08-29T14:53:38.000000Z	luishipolito	NEGATIVE	0.035	0.92	0.03
1824400153173807309	então quer dizer que as smalls só não acompanharam o ibovespa por causa da expectativa do aumento da selic ? e se não aumentar, o índice tende a se valorizar ?	2024-08-16T07:57:46.000000Z	Igoor_Serrano	NEUTRAL	0.2	0.38	0.40

1824408809495748930	novo dia e nova chance para recorde histórico do ibovespa	2024-08-16T08:32:10.000000Z	o_antagonista	POSITIVE	0.74	0.1	0.15
1819107153208398120	dólar sobe a r$ 5,73, maior patamar desde 2021, com cautela global; ibovespa recua	2024-08-01T17:25:17.000000Z	CNNEconomia	NEGATIVE	0.08887442	0.82745165	0.08367394

Fonte: Elaborado pelo autor (2024)

4.3.2 GPT-3.5 Turbo

Nesta etapa, foi utilizado o modelo GPT (Generative Pretrained Transformer) da OpenAI para realizar a análise de sentimentos das publicações coletadas e previamente limpas.

A análise de sentimentos foi realizada por meio da API da OpenAI, utilizando o modelo gpt-3.5-turbo. O fluxo básico do processo envolve enviar cada texto para a API, solicitando que o modelo classifique o sentimento contido na mensagem como "POSITIVE", "NEUTRAL" ou "NEGATIVE". Essa solicitação é realizada em forma de diálogo, onde o modelo recebe instruções claras sobre o que precisa fazer.

O modelo é instruído com um prompt inicial, que define o papel do sistema: "Você é um assistente que analisa sentimentos sobre postagens relacionadas ao IBOVESPA e a economia brasileira." Cada texto é passado como parte de uma mensagem, onde o modelo é solicitado a determinar o sentimento da frase enviada com o seguinte *prompt:* "Determine o sentimento desta frase: '{texto}'". O GPT responde com uma das três

classificações possíveis: POSITIVE, NEUTRAL ou NEGATIVE, que são extraídas da resposta e armazenadas no dataset.

Após a classificação dos sentimentos, os resultados são armazenados em uma nova tabela, contendo uma coluna adicional chamada *predicted_label*, que registra a classificação de sentimento para cada publicação. A Tabela 6, apresentada abaixo, contém uma amostra dos resultados após a classificação de sentimento utilizando o GPT com as mesmas publicações da Tabela 5.

Tabela 6 - Amostra dos resultados da classificação com o GPT-3.5 Turbo

post_id	text	created_at	author_username	predicted_label
1819159261001535538	apesar das dificuldades, o ibovespa continuou melhorando as condições de suas tendências nos prazos mais curtos, particularmente a dos volumes, que retomaram o viés de alta no curto prazo	2024-08-01T20:52:20.000000Z	alphafintec	POSITIVE
1829215851209470236	ibovespa ao vivo: bolsa cai e tenta manter os 136 mil pontos; azul4 despenca via	2024-08-29T14:53:38.000000Z	luishipolito	NEGATIVE
1824400153173807309	então quer dizer que as smalls só não acompanharam o ibovespa por causa da expectativa do aumento da selic ? e se não aumentar, o índice tende a se valorizar ?	2024-08-16T07:57:46.000000Z	Igoor_Serrano	NEUTRAL
1824408809495748930	novo dia e nova chance para recorde histórico do ibovespa	2024-08-16T08:32:10.000000Z	o_antagonista	POSITIVE
1819107153208398120	dólar sobe a r$ 5,73, maior patamar desde 2021, com cautela global; ibovespa recua	2024-08-01T17:25:17.000000Z	CNNEconomia	NEGATIVE

Fonte: Elaborado pelo autor (2024).

4.4 Agregação dos dados

Nesta etapa do trabalho, os dados de sentimentos extraídos dos *posts* foram agregados por dia, e correlacionados com os preços de fechamento do índice IBOVESPA (IBOV). O objetivo foi organizar os dados de forma estruturada para permitir uma análise inicial da relação entre os sentimentos identificados pelos modelos e o desempenho diário do IBOVESPA. A seguir, são descritas as principais etapas do processo de agregação e organização dos dados.

Inicialmente, para os resultados de cada modelo, foi realizada a leitura dos arquivos CSV contendo as publicações e os sentimentos identificados. Em seguida, os sentimentos e as publicações foram agregados por dia. Primeiramente, foi identificada a categoria de sentimento dominante em cada dia. O sentimento dominante é determinado com base na frequência das categorias (positivo, negativo ou neutro), ou seja, o sentimento mais prevalente entre os *posts* publicados em um determinado dia. Além da identificação do sentimento dominante, foram calculadas as contagens diárias de publicações em cada categoria de sentimento e o número total de *posts* publicados em cada dia.

Em paralelo à agregação dos sentimentos, foi realizada a coleta dos preços de fechamento do índice IBOVESPA para o mesmo período em que os *posts* foram publicados. Os dados de preços foram obtidos por meio da API do *Yahoo! Finance* (yfinance), que fornece informações históricas sobre o desempenho do índice. Para cada dia de análise, o preço de fechamento do IBOVESPA foi adicionado à tabela de resultados.

Nos casos em que não havia dados de fechamento disponível (como em fins de semana ou feriados, quando a bolsa de valores está fechada), os dias foram marcados com uma *flag* indicativa (bolsa_fechada). Nestes casos, para garantir a continuidade dos dados, os valores de fechamento ausentes foram preenchidos utilizando o último valor de fechamento disponível, ou seja, o último dia em que a Bolsa estava aberta.

Após a agregação dos dados de sentimentos e a coleta dos preços de fechamento do IBOVESPA, as informações foram organizadas em um único conjunto de dados, contendo o sentimento predominante por dia, as contagens diárias de sentimentos (positivo, negativo e neutro), o número total de *posts* por dia, O preço de fechamento do IBOVESPA em cada dia e a indicação de dias em que a bolsa de valores esteve fechada.

Esse conjunto de dados foi salvo em um novo arquivo CSV, que contém todas as informações necessárias para as análises. A Tabela 8, apresentada abaixo, contém uma amostra dos dados agregados após análise com o FinBERT-PT-BR. A tabelas do modelos GPT.

Tabela 8 - Amostra da tabela com os dados agregados

data	sentimento	qtd_negativo	qtd_neutro	qtd_positivo	qtd_total	fechamento	bolsa_fechada
2024-07-30	NEGATIVE	83	38	15	136	126139.0	False
2024-07-31	POSITIVE	49	47	60	156	127652.0	False
2024-08-01	NEGATIVE	64	41	36	141	127395.0	False
2024-08-02	NEGATIVE	86	36	20	142	125854.0	False
2024-08-03	NEUTRAL	9	16	9	34	125854.0	True

Fonte: Elaborado pelo autor (2024).

5. RESULTADOS

Os resultados foram analisados por meio de uma análise exploratória dos dados e através da aplicação de métricas quantitativas de avaliação de desempenho, como precisão, *recall* e *F1-score* em uma amostra de publicações rotulada manualmente, com o objetivo de comparar a performance dos modelos GPT e FinBERT-PT-BR na tarefa de classificação de sentimentos em publicações sobre o IBOVESPA. Adicionalmente, foi realizada uma análise comparativa dos padrões de classificação e divergência entre os modelos, buscando identificar as particularidades de cada um na interpretação de textos. A análise também inclui um breve estudo entre entre o sentimento predominante nas publicações e o comportamento do índice IBOVESPA.

5.1 Análise Exploratória

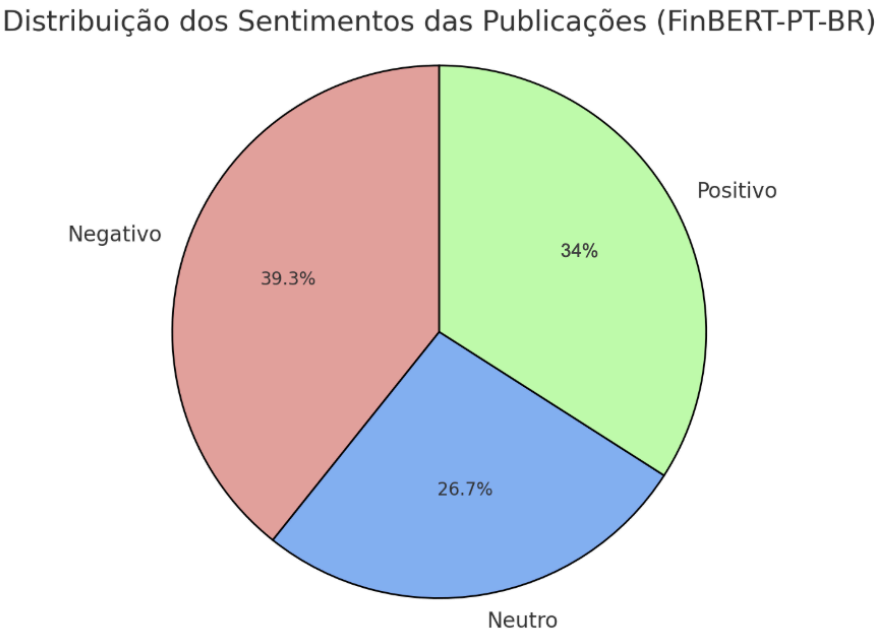
A seção de análise exploratória focou na distribuição dos sentimentos classificados pelos modelos, permitindo identificar sentimentos predominantes, como a prevalência de sentimentos positivos e neutros nas publicações. As distribuições foram visualizadas em gráficos, possibilitando observar as diferenças entre os modelos e entender como cada um lida com o contexto das publicações sobre o IBOVESPA. Também foram analisados exemplos representativos e matrizes de confusão para avaliar a concordância e divergência entre as previsões dos modelos, além de explorar possíveis motivos para diferenças nas classificações, especialmente em relação ao tom neutro e ao uso de ironia ou jargões financeiros.

5.1.1 Distribuição de sentimentos

Das 6471 publicações classificadas com o FinBERT-PT-BR, 2542 (39,28%) expressam um sentimento negativo, enquanto 2204 (34,06%) foram classificadas como

positivas. Além disso, 26,66% dos textos apresentaram um sentimento neutro, como pode ser visto na Figura 10, apresentada abaixo.

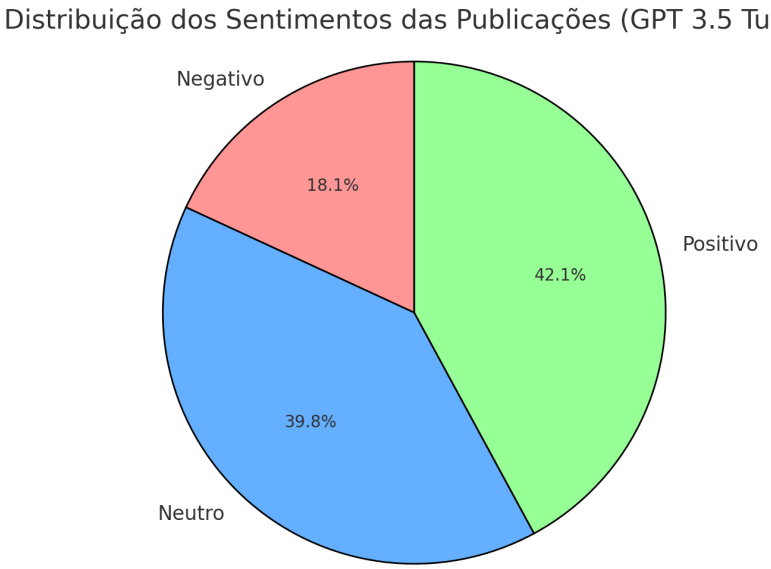
Figura 10 - Distribuição do sentimento das publicações classificadas com o FinBERT-PT-BR



Fonte: Elaborado pelo autor (2024).

A análise das publicações utilizando o modelo GPT 3.5 Turbo revelou uma predominância de sentimentos positivos, correspondendo a 42,10% (2724) do total. As publicações classificadas como neutras representaram 39,79% (2575), enquanto 18,11% (1172) foram categorizadas como negativas. A figura 11 abaixo apresenta a distribuição dos sentimentos classificados pelo GPT.

Figura 11 - Distribuição do sentimento das publicações classificadas com o GPT.



Fonte: Elaborado pelo autor (2024).

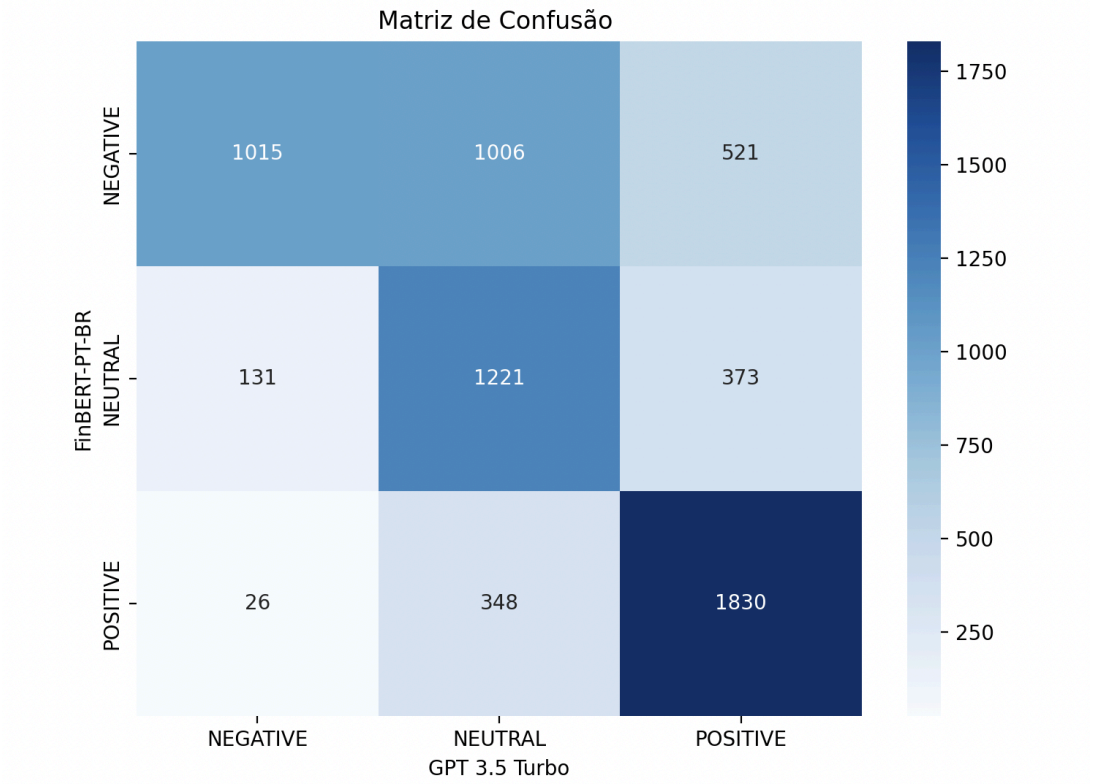
Quando comparado ao modelo FinBERT-PT-BR, o GPT apresentou uma classificação mais positiva, sugerindo que o modelo pode ter uma sensibilidade diferente em relação ao tom dos textos analisados (como identificação de sarcasmo, gírias e outros) e compreender melhor textos informais no formato de redes sociais como o X.

5.1.2 Matriz de confusão

A matriz de confusão foi utilizada para avaliar o nível de concordância e divergência entre os modelos FinBERT-PT-BR e GPT 3.5 Turbo na classificação de sentimentos das publicações. Essa matriz apresenta as previsões feitas por ambos os modelos para as categorias de sentimento, indicando onde os modelos concordam e discordam.

Para gerar a matriz de confusão, foi utilizada a função *confusion_matrix* da biblioteca *sklearn* para Python. A figura abaixo apresenta a matriz de confusão gerada entre os modelos.

Figura 12 - Matriz de confusão entre as classificações dos modelos



Fonte: Elaborado pelo autor (2024).

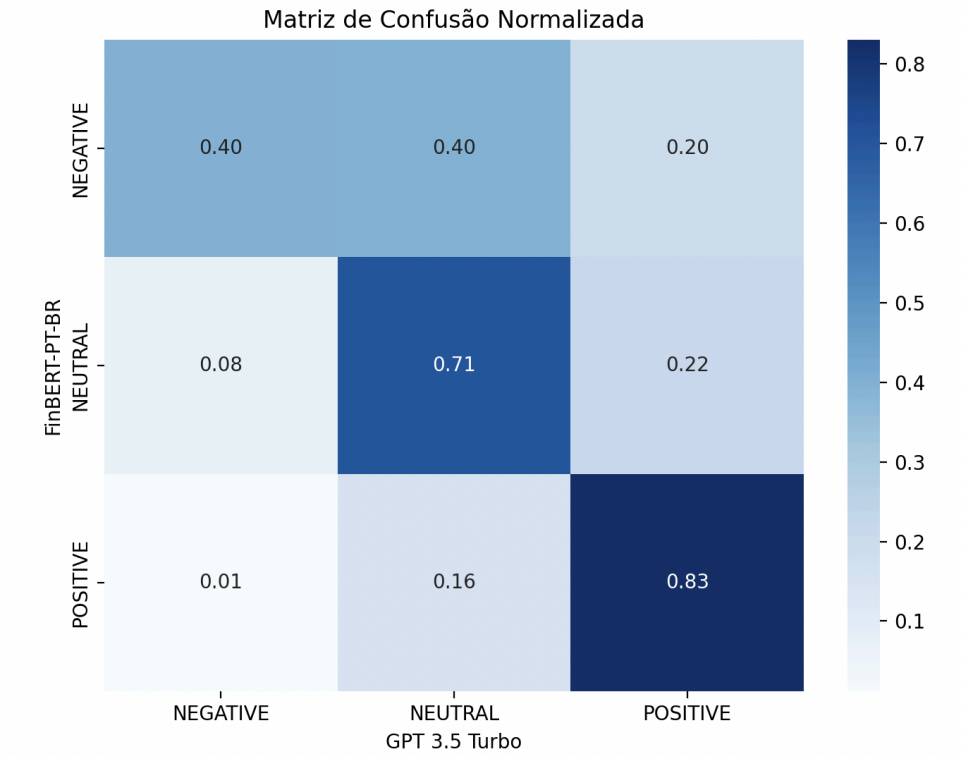
Os resultados da matriz de confusão demonstram que, das 6471 publicações, os modelos concordam em 1015 textos classificados como negativos, 1221 como neutros e 1830 como positivos, com um total de 4066 classificações em que ambos os modelos concordam. Esses valores indicam que, embora ambos os modelos apresentem divergências, há uma significativa concordância nas classificações, especialmente na categoria de publicações positivas.

A análise da matriz indica diferenças entre as classificações dos dois modelos, sobretudo nas publicações que o FinBERT-PT-BR classificou como negativas, mas que o GPT 3.5 Turbo rotulou como neutras, o que ocorreu em 1006 casos. Além disso, 521 publicações foram classificadas como positivas pelo GPT, enquanto o FinBERT-PT-BR as categorizou como negativas. Por fim, 373 publicações classificadas como neutras pelo FinBERT-PT-BR foram consideradas positivas pelo GPT.

Como o volume de publicações em cada categoria pode variar, foi realizada a normalização dos dados para obter uma comparação proporcional mais equilibrada entre as categorias de sentimento.

O processo de normalização consiste em dividir cada valor da matriz pela soma dos valores da sua respectiva linha, transformando os valores absolutos em proporções. Dessa forma, a soma de cada linha da matriz é igual a 1 (ou 100%), permitindo uma comparação entre as classes de sentimento independentemente da quantidade de publicações em cada categoria. Para normalizar os dados, foi adicionado o parâmetro *"normalize"* com valor igual a *"true"* ao gerar a matriz com o *sklearn.* A Figura 13 apresentada abaixo contém a matriz de confusão normalizada.

Figura 13 - Matriz de confusão normalizada



Fonte: Elaborado pelo autor (2024).

Os resultados da matriz de confusão com os dados normalizados indicam que, das publicações classificadas como negativas pelo FinBERT-PT-BR, cerca de 40% também foram classificadas como negativas pelo GPT 3.5 Turbo, mostrando uma concordância moderada entre os modelos nesta categoria. No entanto, uma divergência é observada nas classificações de publicações negativas: 40% dessas publicações foram consideradas neutras pelo GPT e 20% foram classificadas como positivas.

Nas publicações classificadas como neutras pelo FinBERT-PT-BR, houve uma concordância significativa, com aproximadamente 71% das publicações sendo igualmente classificadas como neutras pelo GPT 3.5 Turbo. No entanto, 22% das publicações neutras

foram classificadas como positivas pelo GPT, o que aponta uma diferença na maneira como os modelos interpretam nuances de textos moderados.

Por fim, nas publicações classificadas como positivas pelo FinBERT-PT-BR, observou-se a maior concordância entre os modelos, com 83% das publicações sendo também classificadas como positivas pelo GPT. Esse resultado sugere que para sentimentos claramente positivos, ambos os modelos conseguem identificar o sentimento de maneira similar, independentemente de suas diferenças de treinamento.

Esses resultados mostram que, embora exista uma alta concordância para as publicações positivas, há divergências na forma como os modelos classificam publicações como negativas e neutras, com o GPT apresentando uma tendência a classificar publicações como neutras ou positivas em situações onde o FinBERT-PT-BR identificou um sentimento negativo. Parte dessas diferenças podem ser explicadas pela diferença nos conjuntos de dados utilizados para treinar cada modelo. O FinBERT-PT-BR, apesar de ser especializado no domínio financeiro, foi treinado com dados de portais de notícias, enquanto o GPT possui uma abordagem mais generalista. Isso pode causar diferenças na forma em que os modelos identificam e lidam com sarcasmo, contexto, erros de escrita e gírias, além de diferenças entre o tipo de linguagem utilizado em redes sociais e portais de notícias especializados.

A subseção "5.1.3 Exemplos representativos" contém uma amostra das publicações onde os modelos concordaram e discordaram, além de mais comentários sobre as diferenças observadas nas classificações.

5.1.3 Exemplos representativos

Para entender melhor as diferenças nas classificações dos modelos, foram selecionadas aleatoriamente algumas publicações para análise individual, discutidas a seguir.

A Tabela 9, apresentada abaixo, contém uma amostra de publicações classificadas como negativas por ambos os modelos. Nos exemplos analisados, os modelos concordam na identificação de publicações com conteúdo objetivamente negativo, especialmente aquelas que descrevem quedas do IBOVESPA, desvalorização do real em relação ao dólar ou sinais e preocupações com crises econômicas.

Tabela 9 - Publicações classificadas como negativas por ambos os modelos

text	bert_prediction	gpt_prediction
a única coisa que caiu de ontem para hj foi o ibovespa	NEGATIVE	NEGATIVE
elon musk, o homem mais rico do mundo, está divulgando que o brasil não é seguro para estrangeiro investir. a notícia já está tendo repercussão nos mercados. o dólar disparou e o real caiu, ibovespa caiu muito! o mundo em uma direção e o brasil na contramão.	NEGATIVE	NEGATIVE
se o ibovespa não está no fundo do poço, está bem perto, diz o jp morgan	NEGATIVE	NEGATIVE

Fonte: Elaborado pelo autor (2024).

Um exemplo dessa concordância pode ser observado na publicação: "*a única coisa que caiu de ontem para hj foi o ibovespa*". Ao mencionar explicitamente a queda do índice, a publicação expressa sentimento claramente negativo, levando ambos os modelos a classificá-la como negativa.

De maneira semelhante, o texto "*elon musk [...] está divulgando que o brasil não é seguro para estrangeiro investir. [...] o dólar disparou e o real caiu, ibovespa caiu muito!*" reflete múltiplos sinais negativos para a economia – insegurança para investidores estrangeiros, alta do dólar e queda da bolsa –-, resultando em classificações negativas.

Outro exemplo relevante é a publicação *"se o ibovespa não está no fundo do poço, está bem perto, diz o jp morgan*", em que a expressão "fundo do poço" reforça um sentimento pessimista sobre a situação do mercado, justificando a atribuição de uma classificação negativa por ambos os modelos.

Esse padrão demonstra que, em publicações que expressam claramente e objetivamente sentimentos negativos e pessimistas em relação ao mercado, os modelos conseguem identificar de maneira eficiente o tom negativo das publicações.

Nas publicações analisadas, os modelos também concordaram em textos que expressam um tom neutro, com conteúdo informativo ou especulativo e que não expressam uma opinião ou sentimento claro. Nestes exemplos, apresentados na tabela abaixo, as publicações não possuem uma opinião ou juízo de valor explícito e algumas delas expressam dúvidas e tom de incerteza.

Tabela 10 - Publicações classificadas como neutras por ambos os modelos

text	bert_prediction	gpt_prediction
dólar e ibovespa hoje: o que esperar dos negócios na b3 nesta sexta	NEUTRAL	NEUTRAL
ibov a alta do índice vai depender...	NEUTRAL	NEUTRAL
renato, depois desses balanços pode se desfazer sem medo de ibovespa? ou mantém algo?	NEUTRAL	NEUTRAL

Fonte: Elaborado pelo autor (2024).

Um exemplo dessa situação ocorre na publicação *"ibov a alta do índice vai depender..."*. Tanto o FinBERT-PT-BR quanto o GPT classificaram esse texto como neutro, uma vez que ele apenas apresenta uma consideração sobre a evolução do índice, sem conter um tom ou sentimento específico. Outro exemplo é *"dólar e ibovespa hoje: o que esperar dos negócios na b3 nesta sexta"*, em que a publicação não contém nenhum julgamento e possui caráter informativo (a publicação original contém um link para um portal de notícias). A publicação "*renato, depois desses balanços pode se desfazer sem medo de ibovespa? ou mantém algo?*" possui uma expressão clara de incerteza e indecisão, e foi rotulada como neutra por ambos os modelos. Esse padrão encontrado nas amostras indica que em textos descritivos, sem emoção evidente e com demonstrações claras de incerteza os modelos são capazes de identificar o sentimento neutro.

Além disso, nas publicações selecionadas, os modelos concordaram em classificar como positivos os textos que expressam otimismo ou expectativas sobre o mercado, principalmente em situações que destacam recordes e altas no mercado financeiro. A tabela abaixo contém uma amostra das publicações classificadas como POSITIVE por ambos os modelos.

Tabela 11 - Publicações classificadas como positivas por ambos os modelos

text	bert_prediction	gpt_prediction
grandes companhias vêm batendo recorde de valor de mercado e analistas já projetam ibovespa a 150 mil pontos	POSITIVE	POSITIVE
o ibovespa avança nesta terça-feira, em uma sessão de ganhos para as commodities e com investidores buscando pistas sobre o rumo dos juros no brasil e nos estados unidos. leia mais:	POSITIVE	POSITIVE

bolsa ultrapassa os 136 mil pontos pela primeira vez na história embalada pelo otimismo de investidores estrangeiros diante da expectativa de queda nos juros dos eua, ibovespa renova máxima histórica e fecha o dia com alta de 0,23%, a 136.087 pontos.	POSITIVE	POSITIVE
enquanto isso o ibovespa bate recordes!!	POSITIVE	POSITIVE

Fonte: Elaborado pelo autor (2024).

Por exemplo, na publicação *"grandes companhias vêm batendo recorde de valor de mercado e analistas já projetam ibovespa a 150 mil pontos"*, ambos os modelos atribuíram uma classificação positiva. Da mesma forma, a publicação "*enquanto isso o ibovespa bate recordes*!!" também foi classificada como positiva pelos modelos. Essas publicações destacam o desempenho favorável do mercado e expressam uma visão otimista, o que justifica a classificação positiva dos modelos. A concordância nesses casos indica que os modelos são capazes de reconhecer sinais de valorização e otimismo no mercado como indicadores de sentimento positivo.

No entanto, como a matriz de confusão aponta, os modelos discordaram em aproximadamente 37% das publicações, principalmente em relação a publicações classificadas como neutras ou negativas por algum dos modelos. A tabela abaixo contém uma amostra dos casos em que houve discordância entre os modelos:

Tabela 12 - Amostra de publicações em que os modelos discordaram

text	**bert_prediction**	**gpt_prediction**
ibovespa cai seguindo commodities, esperando decisões de juros	NEGATIVE	NEUTRAL
ontem o ibovespa caiu 0,95%, fechando aos 136.041 pontos, mas agosto ainda promete ser o melhor mês do ano, com avanço de 6,57%. a queda foi um ajuste após dias de alta.	NEGATIVE	POSITIVE
ibovespa com maior índice da história e agora dólar despencando. mas o brasil não estava falido?	NEGATIVE	NEUTRAL
ibovespa sobe 0,94% e atinge 135.208 pontos, impulsionado por balanços e expectativas do federal reserve, mas investidores mantêm cautela.	POSITIVE	NEUTRAL
ibov por volta de 70 no fim do ano. iceberg à vista…	NEGATIVE	NEUTRAL
ibovespa a 135 mil pontos, para a frustração dos pessimistas.	NEGATIVE	POSITIVE
essa última alta do ibovespa foi perfeita; é hora de zerar a posição compradora. é melhor zarpar, tirar o cavalinho da chuva, dar no pé e se mandar. saia de fininho, pule fora.	NEUTRAL	NEGATIVE
ibovespa em máxima histórica! ... em reais bolivarianos. agora em dólar só não estamos piores que o méxico.	POSITIVE	NEUTRAL
ibovespa em julho dá forte sinal de que investidor pode esperar virada de mesa em agosto. veja por quê	NEUTRAL	POSITIVE

wall street mergulhou e levou o ibovespa no arrasto!!	POSITIVE	NEGATIVE
ibovespa hoje pode recuperar o posto de pior bolsa do mundo, que foi o seu lugar em grande para do ano de 2024	POSITIVE	NEGATIVE
segunda feira vai ser feliz pra quem tá shortando ibov.	NEUTRAL	NEGATIVE
exato, por isso aposto no ibov	NEUTRAL	POSITIVE
dólar cai abaixo de r$ 5,50 à espera de dados de inflação nos eua; ibovespa avança cnnbrasil | notícias da	NEGATIVE	POSITIVE

Fonte: Elaborado pelo autor (2024).

Muitas das discordâncias ocorrem quando as publicações apresentam elementos tanto positivos quanto negativos, ou quando tratam de projeções futuras. O FinBERT-PT-BR tende a cometer mais erros ao classificar publicações que tratam de previsões, enquanto o GPT, em muitos casos, capta um otimismo implícito.

Por exemplo, na publicação "*ibovespa em julho dá forte sinal de que investidor pode esperar virada de mesa em agosto. veja por quê*", o FinBERT-PT-BR classificou o texto como neutro, possivelmente por não entender o significado de "virada de mesa" como uma expressão de otimismo, enquanto o GPT classificou como positivo influenciado pelo otimismo expresso no expectativa de recuperação do mercado.

Outro exemplo ocorre na frase "*ontem o ibovespa caiu 0,95%, fechando aos 136.041 pontos, mas agosto ainda promete ser o melhor mês do ano, com avanço de 6,57%. a queda foi um ajuste após dias de alta.*" O FinBERT-PT-BR focou na queda imediata do índice, resultando em uma classificação negativa, enquanto o GPT focou no desempenho positivo do IBOVESPA durante o mês de agosto e classificou o texto como positivo. Isso evidencia que o FinBERT-PT-BR pode priorizar eventos concretos e atuais, enquanto o GPT pode ser mais sensível às nuances textuais.

Publicações que tratam de cenários que envolvem riscos ou incertezas econômicas também revelam divergências entre os modelos. O FinBERT-PT-BR em várias ocasiões se mostrou mais sensível aos sinais de risco, mesmo que o texto também mencionasse resultados favoráveis.

Por exemplo, no texto "*Ibovespa sobe 0,94% e atinge 135.208 pontos, impulsionado por balanços e expectativas do Federal Reserve, mas investidores mantêm cautela*.", o FinBERT-PT-BR classificou a publicação como positiva, destacando a alta do índice. Entretanto, o GPT classificou o texto como neutro, provavelmente devido ao tom de cautela mencionado no final da frase, o que poderia indicar incerteza em relação ao mercado.

Da mesma forma, a publicação "*Ibovespa cai seguindo commodities, esperando decisões de juros*" foi classificada como negativa pelo FinBERT-PT-BR, que focou na queda do índice. Já o GPT classificou como neutro, possivelmente interpretando o texto como uma especulação ou expectativa futura sobre as decisões de juros, o que sugere uma abordagem mais equilibrada em relação ao impacto de eventos que ainda não ocorreram.

As publicações que utilizam ironia ou expressões figurativas também geram divergências entre os modelos. Um exemplo disso está na publicação "*ibovespa com maior índice da história e agora dólar despencando. mas o brasil não estava falido?*". O FinBERT-PT-BR classificou o texto como negativo, possivelmente focando no impacto negativo das palavras "despencando" e "falido", enquanto o GPT captou o tom irônico da publicação e classificou-a como neutra, ao entender que o sarcasmo da frase não indicava necessariamente um sentimento negativo. No entanto, a classificação real do texto pode ser considerada positiva, uma vez que o usuário expressa dois fatores positivos (recorde da bolsa e queda do dólar) e ainda ironiza a ideia de que o Brasil está falido. A classificação neutra do GPT pode ter sido causada pela dificuldade em captar o real sentimento por trás do sarcasmo da frase, ou até mesmo pelo contraste entre os resultados positivos do mercado e o questionamento sobre a situação econômica do Brasil. Como a linguagem irônica requer uma interpretação mais subjetiva e implícita, o modelo pode ter optado por uma classificação neutra.

Outro exemplo ocorre na publicação "*ibovespa em máxima histórica! ... em reais bolivarianos. agora em dólar só não estamos piores que o méxico.*" O FinBERT-PT-BR classificou o texto como positivo, provavelmente devido à menção da "máxima histórica". O GPT, no entanto, captou o tom irônico da segunda parte da frase e classificou-a como neutra. Assim como no caso anterior, o GPT possivelmente não captou o real sentimento negativo por trás do sarcasmo e optou por uma classificação neutra.

Outro ponto importante está em como os modelos lidam com cenários que envolvem informações mistas, que podem transmitir diferentes tipos de sentimentos dependendo da ênfase dada a certos elementos. Um exemplo disso é a publicação "*ibovespa a 135 mil pontos, para a frustração dos pessimistas.*" Nesse caso, o FinBERT-PT-BR classificou o texto como negativo, provavelmente atribuindo um sentimento negativo à menção da "frustração" e "pessimismo". Já o GPT interpretou o texto de maneira positiva, destacando o fato de que o índice atingiu um marco de 135 mil pontos, o que contradiz a expectativa dos pessimistas.

Uma situação similar ocorre na publicação "*essa última alta do ibovespa foi perfeita; é hora de zerar a posição compradora. é melhor zarpar, tirar o cavalinho da chuva, dar no pé e se mandar. saia de fininho, pule fora.*". O FinBERT-PT-BR classificou o texto como neutro, possivelmente devido à menção da "alta perfeita". O GPT, entretanto, capturou o tom de preocupação e alerta da recomendação de venda, classificando o texto como negativo.

A utilização de jargões do mercado financeiro e gírias também pode representar um desafio para o FinBERT-PT-BR, que foi treinado para interpretar textos formais e técnicos de portais de notícias. Na publicação *"segunda feira vai ser feliz pra quem tá shortando ibov."*, o GPT classificou corretamente como negativo, uma vez que "*shortar*" se refere a uma estratégia de venda a descoberto, associada a expectativas de queda. O FinBERT-PT-BR, no entanto, classificou o texto como neutro, possivelmente não captando o jargão "*shortar*" e seu significado no contexto de especulação de mercado.

Por outro lado, no texto *"exato, por isso aposto no ibov."*, o GPT classificou a frase como positiva, interpretando a palavra "apostar" como um indicativo de confiança no IBOVESPA. O FinBERT-PT-BR optou por uma classificação neutra ao não captar o otimismo implícito no verbo "apostar" que reflete um sentimento positivo em relação ao desempenho futuro do índice.

Os exemplos selecionados demonstram como os modelos FinBERT-PT-BR e GPT se comportam em diferentes contextos de análise de sentimento. Em casos onde o sentimento negativo é claro e objetivo, ambos os modelos tendem a concordar na classificação, identificando corretamente o tom pessimista de publicações que mencionam quedas do IBOVESPA, desvalorização do real e crises econômicas. Da mesma forma, em textos neutros, que se limitam a descrever eventos ou apresentam incertezas, os modelos também concordam, reconhecendo a ausência de sentimento explícito. No entanto, em casos que envolvem elementos subjetivos, como previsões, sentimentos mistos, ironia e linguagem informal, surgem diferenças importantes. O FinBERT-PT-BR, em geral, faz uma análise mais individualizada, sem levar em conta contextos semelhantes e sem processar ironia, apresentando uma tendência de realçar palavras soltas, sem interpretar o significado completo na frase. Por outro lado, o GPT consegue captar o lado subjetivo das frases com mais clareza ao lidar com nuances de sarcasmo levando em conta o contexto.

5.1.4 Comparação de Desempenho entre Modelos

Nesta seção, são apresentados e analisados os resultados de desempenho dos modelos de linguagem FinBERT-PT-BR e GPT na tarefa de classificação de sentimentos de publicações sobre o IBOVESPA. Para realizar a avaliação, foi selecionada uma amostra de 20% do conjunto de dados total (1290 das 6471 publicações), posteriormente rotulada manualmente pelo autor, que possibilitou a comparação entre as previsões dos modelos e os rótulos verdadeiros.

O conjunto de dados apresenta uma distribuição desbalanceada entre as classes de sentimentos, conforme observado na matriz de confusão da Figura 12, localizada na subseção “5.1.1 Distribuição de sentimentos”. Nas publicações em que ambos os modelos concordaram, observa-se uma predominância de textos com sentimento positivo, seguidos de textos neutros e, em menor quantidade, negativos. Os 1290 textos utilizados para a amostra foram selecionados de forma que essa distribuição original seja preservada para manter a característica real dos dados. A tabela abaixo contém uma amostra do conjunto de dados selecionados para classificação manual, em que a coluna *“real_sentiment”* foi preenchida pelo autor com o sentimento identificado na frase.

Tabela 13 - Amostra dos dados rotulados manualmente

post_id	text	created_at	finbertptbr_prediction	gpt_prediction	real_sentiment
1820607435553607963	erouuuuu. ibov em dólar está no low do ciclo. não aposte contra o brasil.	2024-08-05T20:46:52.000000Z	NEUTRAL	POSITIVE	POSITIVE
1824083597935935633	este ambiente parece corroborar a continuidade dos ajustes positivos dos ativos locais e pode levar ao oitavo pregão consecutivo de valorização e novas pontuações recordes para o ibovespa.	2024-08-15T10:59:54.000000Z	POSITIVE	POSITIVE	POSITIVE
1824113224981070257	ibovespa encosta em máxima histórica com foco em dados nos eua; dólar cai	2024-08-15T12:57:37.000000Z	POSITIVE	POSITIVE	POSITIVE

Fonte: Elaborado pelo autor (2024).

Para uma avaliação imparcial dos modelos, as métricas foram calculadas considerando alguns ajustes para lidar com o desbalanceamento do conjunto. Foram utilizadas duas abordagens:

- Médias Macro: Nesse cálculo, cada classe de sentimento (positiva, neutra e negativa) recebe o mesmo peso, independentemente do número de exemplos. Esse ajuste permite avaliar o desempenho dos modelos de forma igualitária entre as classes, sem que a predominância de uma delas influencie os resultados.
- Médias ponderadas: A média ponderada ajusta as métricas de acordo com a proporção de cada classe no conjunto de dados, refletindo a composição do conjunto original. Essa abordagem oferece uma visão precisa do impacto das previsões de cada modelo sobre a realidade do conjunto desbalanceado.

A Tabela 14 resume as métricas de precisão, *recall* e *F1-score* para cada modelo, calculadas com as médias macro e ponderadas. A acurácia não foi incluída como métrica de avaliação neste estudo devido ao desbalanceamento das classes presentes nos dados, o que poderia distorcer a interpretação dos resultados.

Verifica-se que o modelo GPT apresenta desempenho superior ao FinBERT-PT-BR em todas as métricas, corroborando o que foi analisado na seção *5.1.3 Exemplos representativos* e evidenciando uma capacidade aprimorada de identificar corretamente os sentimentos expressos nos textos.

Tabela 14 - Métricas de desempenho entre o GPT e o FinBERT-PT-BR

Modelo	**Precisão (Macro)**	***Recall* (Macro)**	***F1-Score* (Macro)**	**Precisão (Ponderada)**	***Recall* (Ponderada)**	***F1-Score* (Ponderada)**
FinBERT-PT-BR	0.642	0.620	0.629	0.702	0.644	0.652
GPT	0.861	0.841	0.850	0.867	0.864	0.863

Fonte: Elaborado pelo autor (2024).

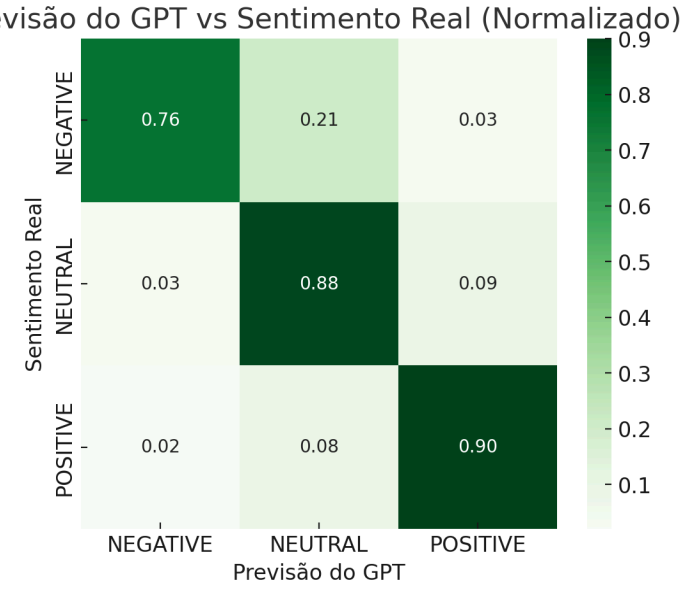
A precisão mede a exatidão das previsões de cada modelo ao classificar uma determinada classe de sentimento (HAN; KAMBER; PEI, 2011). O GPT atingiu uma precisão de 0,861 (macro) e 0,867 (ponderada), enquanto o FinBERT-PT-BR atingiu 0,642 (macro) e 0,702 (ponderada), indicando que o GPT é mais preciso na classificação correta de cada classe de sentimento, independentemente da quantidade de exemplos por classe. Além disso, a precisão mais elevada do GPT sugere uma menor propensão a falsos positivos em cada classe, o que o torna mais confiável para análise de sentimentos sobre o mercado financeiro nas redes sociais.

Segundo Han, Kamber e Pei (2011), o *recall* indica a capacidade do modelo em identificar todas as instâncias reais de cada classe de sentimento. O GPT obteve um *recall* de 0,841 (macro) e 0,864 (ponderada), enquanto o FinBERT-PT-BR apresentou 0,620 (macro) e 0,644 (ponderada). O maior *recall* do GPT indica que ele possui uma habilidade superior em capturar corretamente as publicações das três classes, reduzindo a quantidade de falsos negativos. Já o menor *recall* do FinBERT-PT-BR indica que o modelo não conseguiu identificar corretamente todas as instâncias de sentimentos negativos e neutros na amostra, indicando que o modelo apresenta dificuldades em capturar o tom exato desses sentimentos em algumas publicações.

O *F1-score* é a média harmônica entre precisão e recall e oferece uma visão equilibrada do desempenho dos modelos ao combinar a precisão e a abrangência das previsões (HAN; KAMBER; PEI, 2011). O GPT alcançou *F1-scores* de 0,850 (macro) e 0,863 (ponderada), indicando um desempenho balanceado entre precisão e recall. O FinBERT-PT-BR, por outro lado, obteve *F1-scores* menores, de 0,629 (macro) e 0,652 (ponderada), o que aponta para um desempenho menos robusto e equilibrado. Esses valores corroboram o que foi analisado anteriormente: enquanto o GPT possui um bom desempenho em classificar todas as classes, o FinBERT-PT-BR é menos eficaz em capturar sentimentos em contextos de linguagem informal, como nas redes sociais.

Para analisar o desempenho das classificações para cada classe, foi gerado o gráfico abaixo, que apresenta a matriz de confusão entre os sentimentos reais e os sentimentos identificados pelo GPT. Os dados foram normalizados para evitar que a predominância de uma classe afetasse classes com menor representatividade.

Figura 14 - Matriz de confusão entre as classificações do GPT e o sentimento real



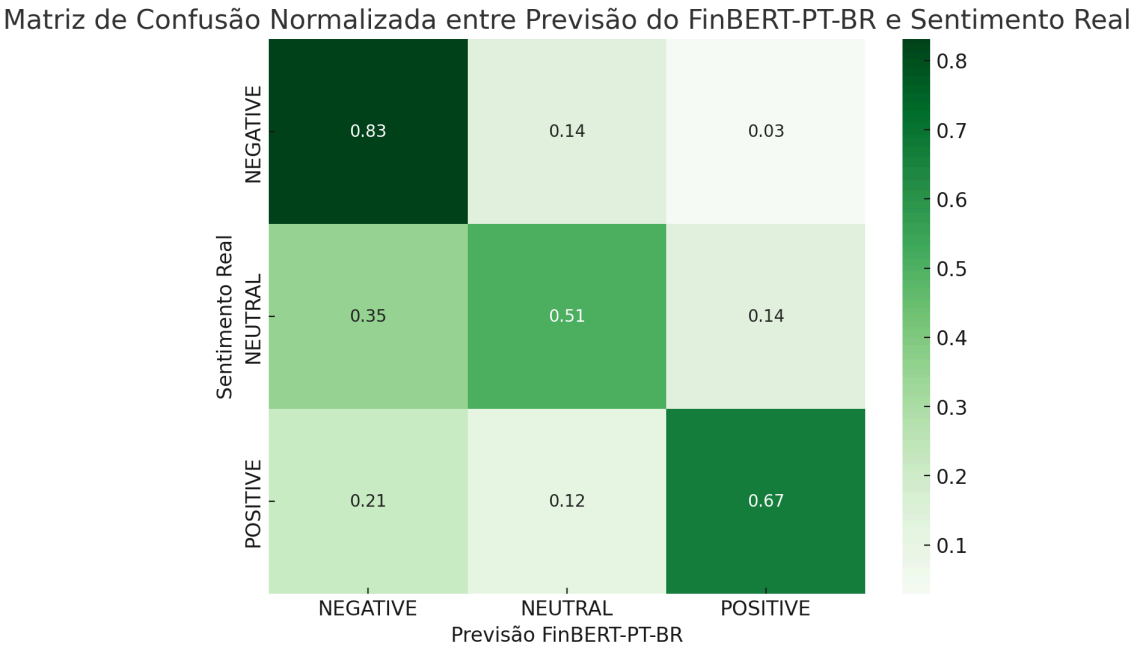
Fonte: Elaborado pelo autor (2024).

A matriz de confusão indica que o GPT possui um alto desempenho na classificação de sentimentos positivos e neutros, com acurácia de 90% e 88%, respectivamente. Para o sentimento negativo, a precisão, apesar de ainda ser considerável, é um pouco menor – 76%. Essa leve diferença na precisão ocorre pois algumas das publicações de tom negativo foram classificadas como neutras pelo GPT, o que ocorre em 21% dos casos. Essa diferença, conforme os exemplos representativos analisados, comprova que o modelo enfrenta alguns desafios na distinção entre sentimentos negativos e neutros, provavelmente devido à ambiguidade em certos textos. Isso ocorre na publicação "*ibovespa em máxima histórica! ... em reais bolivarianos. agora em dólar só não estamos piores que o méxico.*", classificada como neutra pelo GPT, mas que não captou o real sentimento negativo (desvalorização do IBOVESPA em dólar) por trás da ironia.

Dessa forma, embora o GPT mostre robustez em identificar sentimentos positivos e neutros, a performance para sentimentos negativos poderia ser aprimorada para reduzir a confusão entre sentimentos neutros e negativos e lidar melhor com o uso de ironia.

Da mesma forma, foi gerada uma matriz de confusão comparando as classificações do FinBERT-PT-BR e o sentimento real, como pode ser observado na Figura 15 abaixo.

Figura 15 - Matriz de confusão entre as classificações do FinBERT-PT-BR e o sentimento real



Fonte: Elaborado pelo autor (2024).

A matriz de confusão para o FinBERT-PT-BR revela um desempenho satisfatório na classificação de sentimentos negativos com uma precisão de 83%, ligeiramente maior que a precisão para os sentimentos classificados como negativos pelo GPT. No entanto, o modelo apresenta uma taxa de acerto mais baixa para sentimentos positivos (67%), com uma tendência a confundir esses sentimentos como negativos, o que ocorre em 21% dos casos. Para o sentimento neutro, a precisão é de apenas 51%, com uma confusão significativa entre os sentimentos negativos (35%) e positivos (14%). Esse resultado sugere que o modelo enfrenta dificuldade em identificar neutralidade, possivelmente por interpretar certos elementos neutros como negativos ou positivos dependendo do contexto.

Essa dificuldade também comprova o que foi discutido na seção de exemplos representativos, principalmente em publicações que contém ironia ou expressões negativas e positivas na mesma sentença, reforçando a dificuldade na interpretação de ironia e nuances.

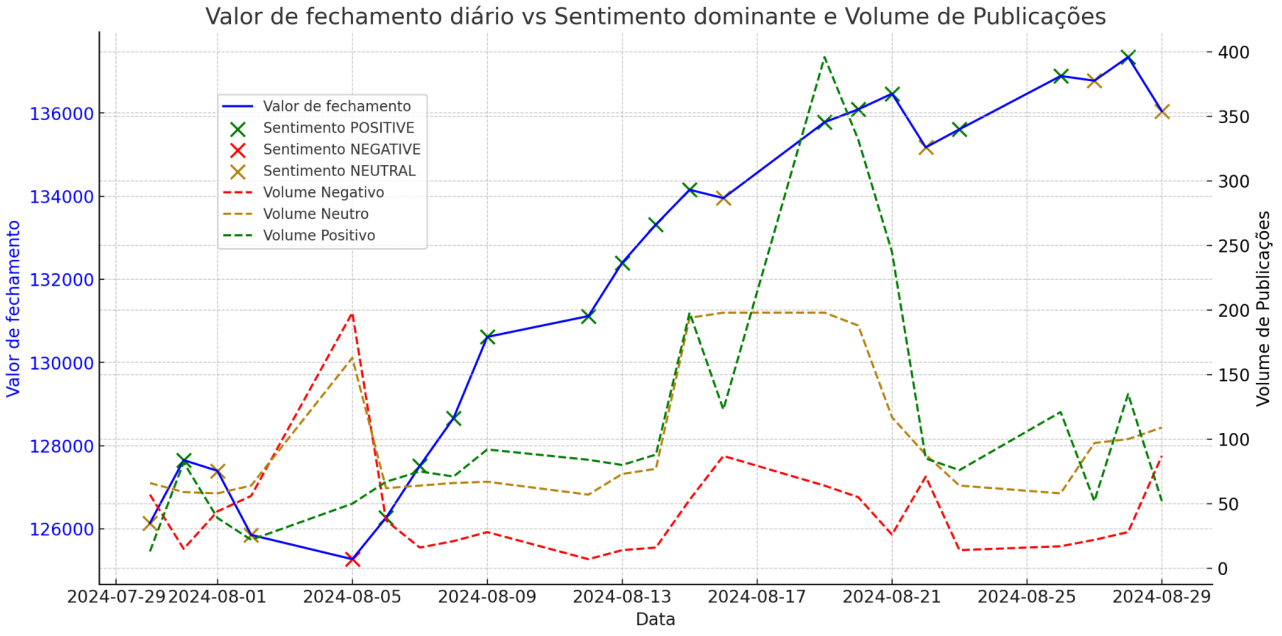
Em resumo, o FinBERT-PT-BR mostra um um desempenho superior ao GPT para publicações claramente negativas, mas poderiam ser realizadas melhorias para refinar a identificação de sentimentos neutros e positivos. A confusão entre publicações neutras e negativas abre também espaço para aprimoramento no ajuste fino do modelo para

distinguir melhor os tons sutis de neutralidade e negatividade. Como o FinBERT-PT-BR é treinado com textos de portais de notícias financeiras, uma alternativa viável seria realizar um ajuste fino e treinar o modelo com dados de publicações de redes sociais, criando um novo modelo pré-treinado com textos de redes sociais, o que facilitaria a classificação e interpretação de expressões informais e ironia.

5.1.5 Evolução Diária dos Sentimentos e Pontos do IBOVESPA

Para validar as classificações de ambos os modelos e comparar com o desempenho da bolsa de valores, foi gerado o gráfico da Figura 16 contendo o sentimento predominante de cada dia identificado pelo GPT, o volume de publicações para cada classe de sentimento e o valor de fechamento do índice da bolsa de valores. Para facilitar a análise da evolução do sentimento, os dias em que a bolsa estava fechada foram removidos do gráfico, já que o volume de publicações nestes dias é muito menor que os demais (abaixo de 50).

Figura 16 - Valor de fechamento diário vs Sentimento dominante e volume de publicações classificadas pelo GPT



Fonte: Elaborado pelo autor (2024).

Dos 6471 *posts* analisados e classificados pelo GPT, 1172 foram considerados negativos, 2575 neutros e 2724 positivos. Em diversos dias de amostra, o sentimento predominante positivo coincide com a valorização no fechamento da bolsa. Essa relação sugere que, em momentos de alta no mercado, o sentimento positivo tende a predominar.

No entanto, quando a bolsa apresenta queda, o sentimento predominante é o neutro na maioria dos casos, com exceção do dia 05/08. Isso pode ser explicado pela matriz de confusão da Figura 14, o que corrobora com o fato de que o GPT classifica erroneamente alguns textos negativos como neutros. Também é válido salientar que agosto de 2024 foi um mês de altas sucessivas na BOVESPA, o que pode contribuir com um maior otimismo expresso nas publicações apesar de algumas quedas.

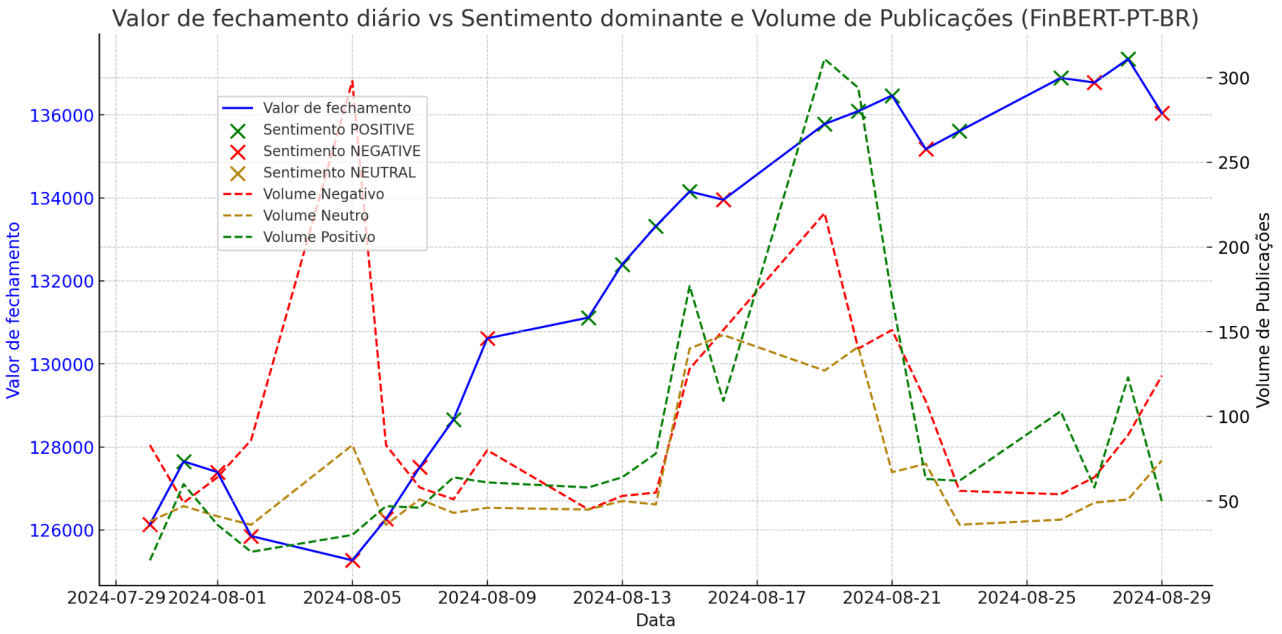
O pico de sentimentos positivos, como visto no gráfico, ocorreu no dia 18/08, que coincide com a data em que a Bolsa atingiu o topo histórico de seu valor nominal (Petry, 2024). Já o pico de sentimentos negativos ocorre no dia 05/08, que coincide com uma queda do índice e uma grande queda em várias bolsas ao redor do mundo (Crepaldi e Quesada, 2024).

Outro ponto importante a ser observado é o volume de publicações categorizadas em cada sentimento. Grande parte das publicações foram classificadas como neutras pelo GPT, indicando que boa parte das discussões sobre o IBOVESPA é informativa, sem expressar emoções, como seria o caso em respostas a eventos de maior impacto como recordes da bolsa. Entretanto, dias de volume elevado de sentimentos positivos ou negativos refletem oscilações mais acentuadas no fechamento da bolsa, sugerindo uma possível influência dos sentimentos mais carregados.

Em períodos onde predominam publicações neutras, o valor de fechamento da bolsa tende a se manter mais estável, com leves quedas ou ganhos. Esse comportamento pode indicar a ausência de eventos impactantes e, consequentemente, uma menor volatilidade no mercado. No período analisado, a predominância do sentimento neutro parece coincidir com um cenário de relativa tranquilidade, onde as publicações são mais descritivas, e menos emocionais.

A figura 17, apresentada abaixo, contém o mesmo gráfico gerado com os dados das publicações analisadas com o FinBERT-PT-BR.

Figura 17 - Valor de fechamento diário vs Sentimento dominante e volume de publicações classificadas pelo FinBERT-PT-BR



Fonte: Elaborado pelo autor (2024).

Dos 6471 posts analisados e classificados pelo FinBERT-PT-BR, 2542 foram considerados negativos, 1725 neutros e 2204 positivos. Para as publicações classificadas com o FinBERT-PT-BR, é possível observar que a maioria dos dias em que o sentimento positivo predomina coincide com uma valorização nos pontos da bolsa. Da mesma forma que as publicações classificadas pelo GPT, isso sugere que as publicações positivas estão associadas a um otimismo geral, o que possivelmente reflete boas expectativas em relação ao mercado. No entanto, em alguns dias em que a valorização da bolsa foi positiva, o sentimento predominante foi o negativo, como é possível observar nos dias 06, 07 e 09 de agosto.

Em contrapartida ao GPT, a bolsa apresenta uma tendência de queda nos preços nos dias em que o sentimento negativo identificado pelo FinBERT-PT-BR é predominante, o que pode sinalizar um receio generalizado em relação ao IBOVESPA, possivelmente motivado por fatores externos, como incertezas econômicas ou políticas. Esse fato é corroborado pela maior precisão do FinBERT-PT-BR na classificação de publicações com tom negativo em relação ao GPT.

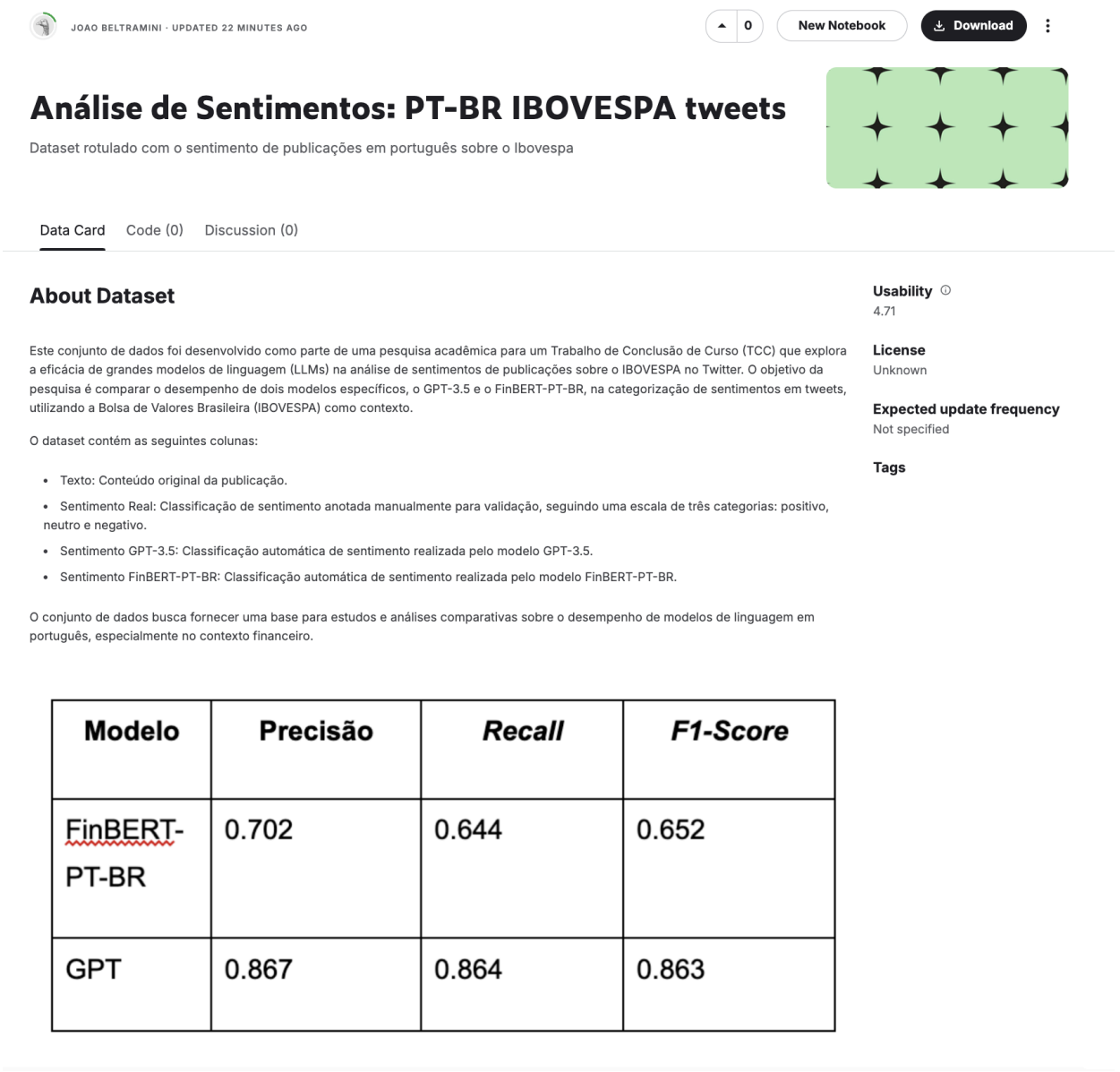
A comparação entre o GPT e o FinBERT-PT-BR na classificação de sentimentos revela nuances interessantes na relação entre o sentimento predominante nas publicações e o comportamento do IBOVESPA. Enquanto o GPT tende a identificar um sentimento neutro como predominante em dias de queda, o FinBERT-PT-BR capta um sentimento predominantemente negativo, indicando uma sensibilidade maior a publicações com tom pessimista. Em ambos os modelos, observa-se uma correlação entre o sentimento positivo e dias de alta na bolsa, sugerindo que o otimismo do mercado se reflete nas publicações. No entanto, o FinBERT-PT-BR apresentou uma relação mais direta entre sentimentos negativos e quedas no índice, o que demonstra a importância de considerar diferentes abordagens para capturar nuances de sentimentos associados às variações da bolsa.

5.2 Compartilhamento do *dataset* rotulado

Para contribuir com o avanço de pesquisas e o desenvolvimento de novos modelos em português, o conjunto de dados rotulados foi disponibilizado publicamente na plataforma Kaggle para consulta e análise (BELTRAMINI, 2024). O Kaggle é uma plataforma online de ciência de dados que oferece um ambiente colaborativo para compartilhamento de datasets e desenvolvimento de projetos de análise e aprendizado de máquina, sendo amplamente utilizado por pesquisadores e profissionais da área (KAGGLE, 2024).

As Figuras 18 e 19 abaixo apresentam uma captura de tela da página do Kaggle onde os dados foram publicados, permitindo uma visão geral das informações disponibilizadas para consulta pública.

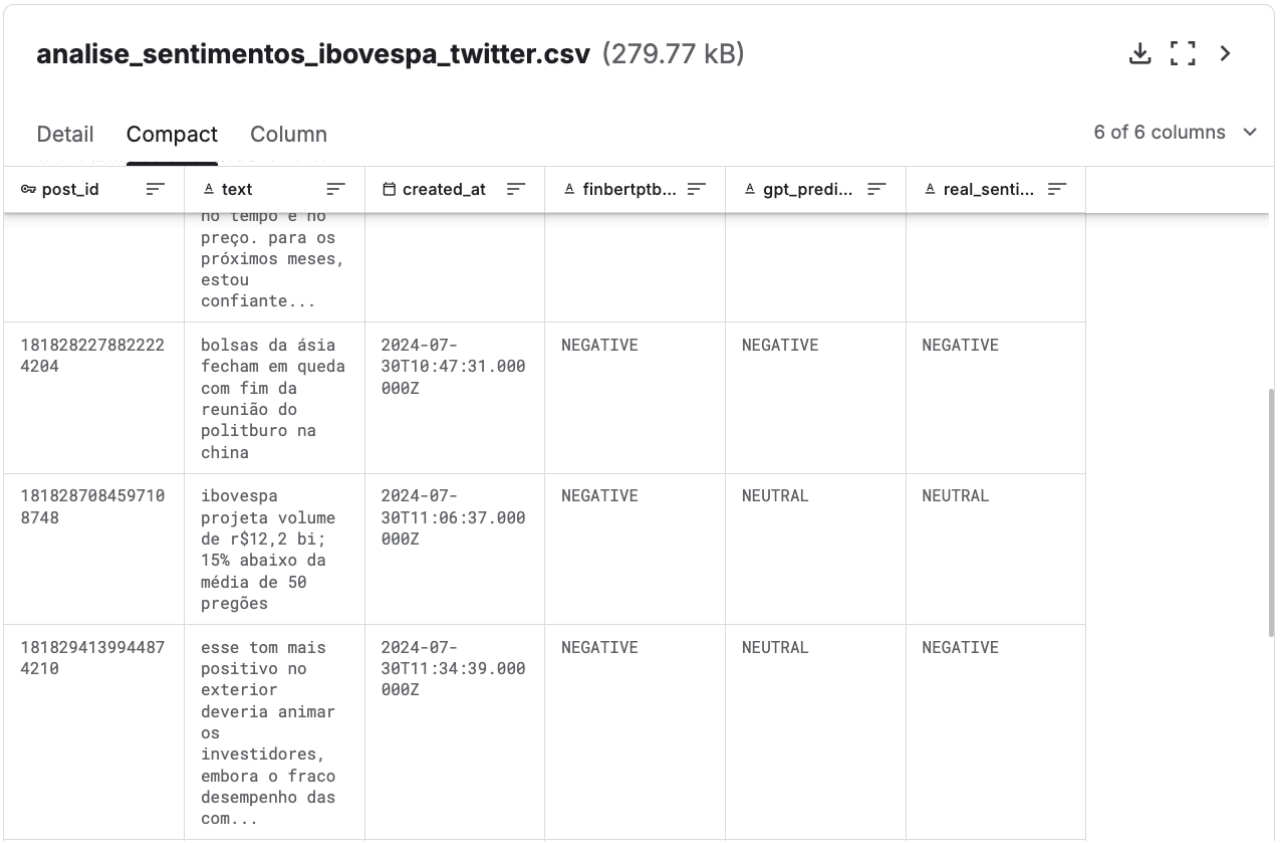
Figura 18 - Página do Kaggle com o conjunto de dados publicado



Modelo	Precisão	*Recall*	*F1-Score*
FinBERT-PT-BR	0.702	0.644	0.652
GPT	0.867	0.864	0.863

Fonte: Kaggle (2024).

Figura 19 - Continuação da página do Kaggle com o conjunto de dados publicado



analise_sentimentos_ibovespa_twitter.csv (279.77 kB)

Detail | Compact | Column | 6 of 6 columns

post_id	text	created_at	finbertptb...	gpt_predi...	real_senti...
	no tempo e no preço. para os próximos meses, estou confiante...				
1818282278822224204	bolsas da ásia fecham em queda com fim da reunião do politburo na china	2024-07-30T10:47:31.000000Z	NEGATIVE	NEGATIVE	NEGATIVE
1818287084597108748	ibovespa projeta volume de r$12,2 bi; 15% abaixo da média de 50 pregões	2024-07-30T11:06:37.000000Z	NEGATIVE	NEUTRAL	NEUTRAL
1818294139944874210	esse tom mais positivo no exterior deveria animar os investidores, embora o fraco desempenho das com...	2024-07-30T11:34:39.000000Z	NEGATIVE	NEUTRAL	NEGATIVE

Fonte: Kaggle (2024).

Todas as etapas do desenvolvimento podem ser encontradas no link https://github.com/jvbeltra/GPT-vs-FinBERT-PT-BR-Analysis, onde estão todos os scripts utilizados e os dados coletados.

6. CONSIDERAÇÕES FINAIS

Este trabalho teve o propósito de desenvolver uma análise comparativa entre modelos de linguagem natural aplicados à análise de sentimentos sobre o IBOVESPA, utilizando publicações em português coletadas da rede social X (anteriormente conhecida como Twitter). Diferenciando-se de estudos anteriores que focaram em análises de sentimento no mercado financeiro com dados predominantemente em inglês e sem o uso de LLM, este trabalho buscou aplicar e validar modelos para dados em português, aproximando a pesquisa do contexto de investidores brasileiros. Além disso, em comparação aos trabalhos relacionados, este trabalho teve como diferencial o uso de LLM para analisar os sentimentos, levando em conta a recente evolução dessas tecnologias e seu potencial para oferecer análises mais precisas e contextualmente informadas em

comparação a abordagens tradicionais Comparado ao principal trabalho relacionado, *Twitter mood predicts the stock market* de Bollen, Mao, e Zeng (2011), que obteve uma precisão de 87,6%, o GPT atingiu uma precisão de 86,7%, evidenciando um desempenho competitivo na classificação de sentimentos, mesmo considerando as diferenças contextuais e metodológicas entre os estudos. Essa proximidade nos resultados ressalta o potencial do GPT e dos *Large Language Models* para tarefas de análise de sentimentos relacionadas ao mercado financeiro, especialmente no contexto de redes sociais.

Assim, conclui-se que o objetivo geral deste trabalho foi atingido. Em relação aos objetivos específicos, todos foram cumpridos: a revisão do estado da arte em análise de sentimentos e processamento de linguagem natural foi realizada com base em trabalhos relacionados; os modelos FinBERT-PT-BR e GPT-3.5 Turbo foram configurados e testados em uma série de experimentos, possibilitando a comparação de desempenho e a análise de eficácia na detecção de sentimentos nas postagens coletadas; o conjunto de dados rotulados com o sentimento real foi publicado; e, por fim, os resultados foram analisados, possibilitando uma discussão sobre padrões e discrepâncias entre os modelos.

Durante o desenvolvimento, alguns desafios se destacaram. Primeiramente, não foi encontrado um *dataset* em português e adequado para o trabalho, o que demandou a implementação de um *script* para coletar as publicações por meio da API do X. O uso da API também causou outros desafios, como a quantidade de publicações coletadas e a abrangência do período de coleta, uma vez que o acesso mais básico da API custa 100 dólares americanos por mês.

Em relação aos principais resultados, o modelo FinBERT-PT-BR apresentou uma precisão de 70,2%, *recall* de 64,4% e F1-Score de 65,2%, enquanto o GPT-3.5 Turbo obteve uma precisão de 86,7%, *recall* de 86,4% e F1-Score de 86,3%, sendo superior em todas as métricas. O modelo GPT-3.5 Turbo, portanto, mostrou-se mais adequado à tarefa de análise de sentimentos para o mercado financeiro brasileiro, superando o FinBERT-PT-BR, o que sugere o potencial de sua aplicação prática, ainda que algumas limitações possam ser observadas na classificação de textos negativos com linguagem coloquial, ironia e gírias, as vezes sendo classificados como neutros.

O conjunto de dados criado para esta pesquisa inclui uma diversidade de textos coletados da plataforma X, com rotulagem manual e critérios de pré-processamento para redução de ruídos. Ele representa uma contribuição relevante para a área de PLN em

português, especialmente para estudos futuros sobre sentimentos no contexto financeiro e de redes sociais.

A partir deste estudo, identificaram-se diversas oportunidades para aprimoramento e desenvolvimento de futuras pesquisas. Entre elas, destacam-se a expansão do trabalho com uma base de dados mais ampla – mais publicações, maior período de coleta e múltiplos avaliadores para rotular o texto devido à natureza subjetiva e complexa da linguagem humana –, o desenvolvimento/ajuste-fino de um novo modelo focado em textos de redes sociais e investigações com outros modelos e classificadores.

Uma outra sugestão para trabalhos futuros seria aplicar os dois modelos e realizar uma ponderação das classificações para escolher o sentimento. Neste caso, seria possível atribuir um peso maior à classificação negativa do FinBERT-PT-BR, uma vez que o mesmo apresentou uma maior acurácia na classificação de textos negativos

Outra abordagem interessante para trabalhos futuros seria o uso das técnicas de *few-shot* e *zero-shot* para otimizar a análise de modelos como o GPT em contextos novos, com pouca ou nenhuma necessidade de rotulagem manual adicional. Essas abordagens permitiriam que o modelo fizesse inferências em tarefas sem exemplos específicos previamente fornecidos, viabilizando a classificação de sentimentos em português com base em seu treinamento geral e potencializando a aplicação prática de modelos de PLN em português.

Além disso, recomenda-se que, com um período mais extenso de coleta, a análise da relação entre os sentimentos e a evolução da bolsa, bem como outros indicadores econômicos, seja aprofundada, a fim de identificar possíveis oportunidades de investimento em ações específicas listadas no mercado.

Assim, esta pesquisa demonstra o potencial de modelos de linguagem natural aplicados à análise de sentimentos no mercado financeiro brasileiro e reafirma a importância da adaptação de técnicas de PLN ao contexto local, visando ampliar o monitoramento e a previsão de tendências econômicas a partir de interações nas redes sociais.

REFERÊNCIAS

AL QUDAH, Dana; AL-ZOUBI, Ala; CASTILLO, Pedro; FARIS, Hossam. **Sentiment analysis for e-payment service providers using evolutionary extreme gradient boosting.** IEEE Access. Disponível em: https://www.researchgate.net/publication/346318732_Sentiment_Analysis_for_e-Payment_Service_Providers_Using_Evolutionary_eXtreme_Gradient_Boosting. Acesso em dez. 2024.

ASSAF NETO, Alexandre. **Mercado financeiro**. 14. ed. São Paulo: Atlas, 2019.

BACCIANELLA, Stefano; ESULI, Andrea; SEBASTIANI, Fabrizio. **SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.** In: LREC, 2010.

BAFNA, P.; PRAMOD, D.; VAIDYA, A. **Document clustering: TF-IDF approach**. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, Chennai, India. Anais [...]. Chennai: IEEE, 2016. p. 61-66. doi: 10.1109/ICEEOT.2016.7754750.

BAHETI, Pragati. **The Essential Guide to Neural Network Architectures**. 2021. Disponível em: https://www.v7labs.com/blog/neural-network-architectures-guide. Acesso em: nov. 2023.

Bollen, Johan & Mao, Huina & Zeng, Xiao-Jun. **Twitter Mood Predicts the Stock Market**. Journal of Computational Science, 2011. Disponível em <https://www.sciencedirect.com/science/article/abs/pii/S187775031100007X>. Acesso em set. 2023.

BREIMAN, L.; et al. Classification and Regression Trees. Wadsworth, 1984.

Brownlee, J. (2019). **A Gentle Introduction to the Bag-of-Words Model**. Machine Learning Mastery. Disponível em: https://www.machinelearningmastery.com. Acesso em: nov. 2023

BROWN, T. B.; MANN, B.; RYDER, N.; et al. **Language Models are Few-Shot Learners**. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020. Disponível em: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. Acesso em: nov. 2023.

BOMMASANI, R.; HUDSON, D. A.; ADELI, E.; et al. **On the Opportunities and Risks of Foundation Models**. 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em: nov. 2023.

CAMBARA, Leilane de Freitas Rocha. **Sentimento de notícias e investimento estrangeiro em carteira no Brasil**. Dissertação (Mestrado em

Economia) - Universidade Federal de Santa Catarina, Florianópolis, 2019. Disponível em <https://repositorio.ufsc.br/handle/123456789/211499>. Acesso em: set. 2023.

CREPALDI, Rebecca; QUESADA, Beatriz. *Ibovespa hoje: 05/08/2024*. Exame, 5 ago. 2024. Disponível em: https://exame.com/invest/mercados/ibovespa-hoje-05-08-2024/. Acesso em: 06 nov. 2024.

Conheça a Fintwit, a comunidade do mercado financeiro no Twitter. Folha de São Paulo. Disponível em: <https://www1.folha.uol.com.br/mercado/2019/09/conheca-a-fintwit-a-comunidade-do-mercado-financeiro-no-twitter.shtml>. Acesso em jul. 2022

CORTES, C.; VAPNIK, V. Support-vector networks. Machine Learning, 1995. Disponível em https://link.springer.com/article/10.1007/BF00994018. Acesso em: nov. 2023

DAMODARAN, Aswath. **Avaliação de investimentos**. 3. ed. Porto Alegre: Bookman, 2017.

DANG, N. C.; MORENO-GARCÍA, M. N.; DE LA PRIETA, F. **Sentiment analysis based on deep learning: a comparative study**. *Electronics*, v. 9, n. 3, p. 483, 2020. Disponível em: https://doi.org/10.3390/electronics9030483. Acesso em: dez. 2024.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: **Pre-training of Deep Bidirectional Transformers for Language Understanding**. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2018. Disponível em: https://aclanthology.org/N19-1423/. Acesso em: nov. 2023

Foote, Keith D. **A Brief History of Natural Language Processing (NLP).** Dataversity, 2019. Disponível em <https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/#> Acesso em ago. 2022.

GRAHAM, Benjamin; DODD, David. **O investidor inteligente**. 1. ed. São Paulo: HarperCollins Brasil, 2014.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques.** 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann, 3rd edition, 2011. 744 p.

Hoang, Mickel, Oskar Alija Bihorac, and Jacobo Rouces. **Aspect-based sentiment analysis using bert**. Proceedings of the 22nd nordic conference

on computational linguistics. 2019.

Hochreiter, S., & Schmidhuber, J. (1997). **Long Short-Term Memory**. Neural Computation.

HUGGING FACE. **Perplexity**. 2024. Disponível em: https://huggingface.co/docs/transformers/perplexity. Acesso em: 7 abr. 2024.

HULL, John C. **Fundamentos dos mercados de futuros e de opções**. 9. ed. São Paulo: BM&FBOVESPA, 2018.

Influenciadores de investimentos 2. ANBIMA. Disponível em: <https://www.anbima.com.br/pt_br/especial/influenciadores-de-investimentos-2.htm>. Acesso em ago. 2022

Jabri, Siham, et al. **Ranking of text documents using TF-IDF weighting and association rules mining.** 2018 4th international conference on optimization and applications (ICOA). IEEE, 2018.

JOACHIMS, Thorsten. Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers, 2002. Disponível em https://link.springer.com/book/10.1007/978-1-4615-0907-3. Acesso em nov. 2023

BELTRAMINI, João Vitor. ***Análise de Sentimentos: PT-BR IBOVESPA tweets*.** Kaggle, 2024. Disponível em: <https://www.kaggle.com/datasets/jvbeltra/sentiment-analysis-pt-br-stock-market-tweets\>.

Jurafsky, Daniel; Martin, James H. **Speech and Language Processing**. Pearson, 2014.

Jwo, Dah-Jing & Wu, Jia-Chyi & Ho, Kuan-Lin. (2021). **Support Vector Machine Assisted GPS Navigation in Limited Satellite Visibility**. Computers, Materials & Continua. 69. 555-574. 10.32604/cmc.2021.018320. Disponível em https://www.researchgate.net/publication/352320015_Support_Vector_Machine_Assisted_GPS_Navigation_in_Limited_Satellite_Visibility. Acesso em: nov. 2023

Kaddour, Jean, et al. "Challenges and applications of large language models." arXiv preprint arXiv:2307.10169 (2023).

Kuncahyo Setyo Nugroho, Anantha Yullian Sukmadewa, and Novanto Yudistira. 2021. **Large-Scale News Classification using BERT Language Model: Spark NLP Approach**. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (SIET '21). Association for Computing Machinery, New York, NY, USA, 240–246. https://doi.org/10.1145/3479645.3479658

KWAK, H. et al. **What is twitter, a social network or a news media?** In: WWW'10. Proceedings of the 19th international conference on World wide web. [S.l.], 2010. p. 591–600.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. Nature, 2015. Disponível em https://www.nature.com/articles/nature14539. Acesso em: nov. 2023

Lee, Sung-Jick, and Han-Joon Kim. **Keyword extraction from news corpus using modified TF-IDF.** The Journal of Society for e-Business Studies 14.4 (2009): 59-73.

Liu, Bing. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, 2012.

LOUKIDES, M. **What is data science?**. O'Reilly. Disponível em <https://www.oreilly.com/radar/what-is-data-science/>. Acesso em jul. 2022

Manning, Christopher D.; Schütze, Hinrich. **Foundations of Statistical Natural Language Processing**. MIT Press, 1999.

MCCARTHY, John. **What is Artificial Intelligence**. Stanford: Stanford University, 2007. Disponível em <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf>. Acesso em jul. 2022.

Medallion, um dos fundos mais rentáveis da história. Forbes. Disponível em <https://forbes.com.br/forbes-money/2020/12/medallion-um-dos-fundos-mais-rentaveis-da-historia/>. Acesso em jul. 2022.

Medeiros, Murilo Cerqueira. **Metodologias para Análise de Sentimentos de Tweets**
sobre o Mercado Financeiro. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecânica) - Universidade de Brasília, Brasília, 2019. Disponível em <https://bdm.unb.br/bitstream/10483/29207/1/2019_MuriloCerqueiraMedeiros_tcc.pdf.> Acesso em: set. 2023.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). **Efficient Estimation of Word Representations in Vector Space.** Proceedings of the International Conference on Learning Representations (ICLR). Disponível em: https://arxiv.org/pdf/1301.3781.pdf. Acesso em: nov. 2023

Mittal, A.; Goel, A. **Stock prediction using twitter sentiment analysis**. Stanford University, CS229, 2012. Disponível em <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf> Acesso em: out. 2023

NEWBOLD, Paul; CARLSON, William L.; THORNE, Betty. **Statistics for Business and Economics**. 8. ed. Upper Saddle River: Pearson Education, 2013. Disponível em <http://ndl.ethernet.edu.et/bitstream/123456789/13768/1/1-Paul%20Newbold.pdf>.

NORVIG, Peter; Russel, Stuart J. **Artificial Intelligence: A Modern Approach**. Prentice Hall, 1995.

Pagolu, Sasank & Reddy, Kamal & Panda, Ganapati & Majhi, Babita. (2016). **Sentiment analysis of Twitter data for predicting stock market movements**. 1345-1350. 10.1109/SCOPES.2016.7955659. Disponível em <https://www.researchgate.net/publication/318327720_Sentiment_analysis_of_Twitter_data_for_predicting_stock_market_movements>. Acesso em: nov. 2023

Pak, Alexander; Paroubek, Patrick. **Twitter as a corpus for sentiment analysis and opinion mining**. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 2010, p. 1320-1326. Disponível em <https://aclanthology.org/L10-1263/>. Acesso em nov. 2023.

Pang, Bo; Lee, Lillian. **Opinion Mining and Sentiment Analysis**. Now Publishers, 2008. Disponível em <https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>. Acesso em nov. 2023

Pennington, J., Socher, R., & Manning, C. D. (2014). **GloVe: Global Vectors for Word Representation.** Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Disponível em: https://aclanthology.org/D14-1162/. Acesso em: nov. 2023

PETRY, Rodrigo. **Ibovespa rompe máxima, mira 138-140 mil pontos; correções à vista.** InfoMoney, 20 ago. 2024. Disponível em: https://www.infomoney.com.br/mercados/ibovespa-rompe-maxima-mira-138-140-mil-pontos-correcoes-a-vista/. Acesso em: 06 nov. 2024.

PIB. IBGE. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Acesso em jul. 2022.

QUINLAN, J. R. Induction of Decision Trees. Machine Learning, 1986. Disponível em: https://link.springer.com/article/10.1007/BF00116251. Acesso em nov. 2023

RADFORD, A.; et al. **Improving Language Understanding by Generative Pre-Training**. 2018. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: nov. 2023

RADFORD, A.; WU, J.; CHILD, R.; et al. Language Models are Unsupervised Multitask Learners. 2019. Disponível em:

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Acesso em: nov. 2023;

Resumo dos mercados. Brasil, Bolsa, Balcão. Disponível em <https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-de-derivativos/resumo-das-operacoes/resumo-dos-mercados/>. Acesso em jul 2022.

Salton, G., & McGill, M. J. (1983). **Introduction to Modern Information Retrieval**. McGraw-Hill.

SANTOS, Lucas L.; BIANCHI, Reinaldo A. C.; COSTA, Anna H. R.. **FinBERT-PT-BR: Análise de Sentimentos de Textos em Português do Mercado Financeiro.** In: BRAZILIAN WORKSHOP ON ARTIFICIAL INTELLIGENCE IN FINANCE (BWAIF), 2. , 2023, João Pessoa/PB. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 144-155. Disponível em: <https://doi.org/10.5753/bwaif.2023.231151>. Acesso em nov. 2023.

SCHÖLKOPF, B.; BURGES, C.; SMOLA, A. J. **Advances in Kernel Methods: Support Vector Learning**. MIT Press, 1999.

SCHÖLKOPF, B.; SMOLA, A. J. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. MIT Press, 2002.

Shannon, C. E. (1948). **A Mathematical Theory of Communication.** Bell System Technical Journal.

Schütz, Mina, et al. **Automatic fake news detection with pre-trained transformer models.** Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII. Springer International Publishing, 2021.

Sobkowicz, P; Kaschesky, M; & Bouchard, G. **Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web**. Government Information Quarterly, 29(4), 470-479, 2012. Disponível em <https://www.sciencedirect.com/science/article/abs/pii/S0740624X12000901>. Acesso em jul. 2022.

SOCHER, R.; et al. **Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank**. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.

TABOADA, Maite et al. **Lexicon-based methods for sentiment analysis**. Computational linguistics, 2011.

TURNEY, Peter D.; LITTMAN, Michael L. **Measuring praise and criticism: Inference of semantic orientation from association.** ACM Transactions on Information Systems (TOIS), 2003.

Vaswani, A., et al. (2017). **Attention Is All You Need.** Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS).

Análise Comparativa de Grandes Modelos de Linguagem na Avaliação de Sentimentos Sobre o IBOVESPA em Redes Sociais

João Vitor Beltramini

Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88010-970 – Florianópolis – SC – Brazil
jvbeltra2011@gmail.com

Abstract. *With the growing participation of Brazilian investors and financial discussions on platforms like X, understanding sentiment is crucial for decision-making. This study compares two large language models (LLMs) — FinBERT-PT-BR and GPT-3.5 Turbo — for sentiment analysis of IBOVESPA-related posts. Over 30 days, data was collected, pre-processed, and analyzed, with 20% of the dataset being manually labeled for validation. GPT-3.5 Turbo outperformed FinBERT-PT-BR, achieving an F1-score of 0.863 versus 0.652. GPT excelled in positive/neutral sentiment classification, while FinBERT struggled with subjective and colloquial texts, but outperformed GPT when it comes to negative texts. These findings contribute to the development of sentiment-based indices, offering valuable insights for assessing investor sentiment and potentially influencing financial decision-making in the Brazilian stock market.*

Resumo. Com a crescente participação de investidores brasileiros e o aumento das discussões financeiras em plataformas como o X, compreender o sentimento dos investidores torna-se crucial para a tomada de decisões. Este estudo compara dois grandes modelos de linguagem (LLMs) — FinBERT-PT-BR e GPT-3.5 Turbo — na análise de sentimentos em publicações relacionadas ao IBOVESPA. Durante 30 dias, os dados foram coletados, pré-processados e analisados, com 20% das publicações rotuladas manualmente para validação. O GPT-3.5 Turbo superou o FinBERT-PT-BR, alcançando um *F1-score* de 0,863 contra 0,652. O GPT se destacou na classificação de sentimentos positivos e neutros, enquanto o FinBERT apresentou dificuldades com textos subjetivos e coloquiais, mas superou o GPT em textos negativos. Esses resultados contribuem para o desenvolvimento de índices baseados em sentimentos, oferecendo *insights* valiosos sobre o sentimento dos investidores com potencial para influenciar decisões financeiras no mercado de ações brasileiro.

1. Introdução

O mercado financeiro brasileiro tem experimentado um crescimento expressivo, especialmente na Bolsa de Valores de São Paulo (BOVESPA), que movimentou mais de R$ 7 trilhões em 2021, representando 80% do PIB do país. Com o aumento do número de investidores individuais, que somaram 4,2 milhões em janeiro de 2022, as redes sociais, como o X (antigo Twitter), têm se tornado importantes plataformas para discussão e compartilhamento de informações sobre o mercado de ações.

A complexidade e a quantidade de dados disponíveis nas redes sociais demandam o uso de métodos quantitativos e qualitativos para auxiliar na tomada de decisões. Nesse contexto, a inteligência artificial (IA) e o Processamento de Linguagem

Natural (PLN) ganham destaque, permitindo que computadores entendam e analisem textos humanos. Uma aplicação relevante de PLN é a análise de sentimentos, que consiste em identificar e classificar emoções expressas em textos.

O presente trabalho propõe analisar publicações em língua portuguesa do X relacionadas ao IBOVESPA e comparar o desempenho de dois modelos de linguagem – FinBERT-PT-BR e GPT – na tarefa de análise de sentimentos. O estudo se diferencia de abordagens anteriores ao focar no mercado financeiro brasileiro e utilizar modelos avançados de PLN em um contexto específico em português brasileiro.

2. Trabalhos relacionados

Com o intuito de identificar trabalhos relacionados ao projeto, foram realizadas buscas por estudos que utilizassem métodos de processamento de linguagem natural e análise de sentimentos extraídos de redes sociais e sites de notícias relacionados ao mercado financeiro. Nesta seção, serão analisados, discutidos e comparados os três estudos mais relevantes encontrados em relação ao projeto atual. Foram selecionados trabalhos em três diferentes esferas: análise de sentimento de publicações do X em inglês; análise de sentimento de publicações do X em português; e uma dissertação sobre análise de sentimentos por meio de notícias.

2.1 Twitter mood predicts the stock market

O estudo de Bollen, Mao e Zeng (2011) investiga a relação entre o humor público, coletado através de publicações no X (antigo Twitter), e o desempenho do mercado de ações, representado pelo índice *Dow Jones Industrial Average* (DJIA).

Os autores coletaram 9.853.498 posts publicados entre fevereiro e dezembro de 2008 e aplicaram duas ferramentas de análise de sentimentos: OpinionFinder (OF), que classifica posts como positivos ou negativos, e GPOMS, que identifica o humor em seis dimensões (Calmo, Alerta, Certo, Vital, Gentil e Feliz). Os resultados mostraram que o GPOMS captura nuances emocionais mais detalhadas do que o OF.

Utilizando a análise de causalidade de Granger e um modelo de Rede Neural *Fuzzy* Auto-Organizável (SOFNN), os autores avaliaram se o humor público poderia prever o fechamento do DJIA. Constatou-se que apenas a dimensão "Calmo" apresentou uma relação preditiva significativa e não-linear com o DJIA, melhorando a precisão das previsões. O estudo alcançou 87,6% de acurácia na previsão da direção das mudanças diárias do DJIA e reduziu o erro percentual médio em 6%.

Os resultados indicam que o monitoramento do humor público no X pode fornecer insights valiosos para prever tendências do mercado de ações, embora eventos inesperados ainda representem desafios significativos.

2.2 Metodologias para Análise de Sentimentos de Tweets sobre o Mercado Financeiro

No estudo de Medeiros (2019), são apresentadas duas metodologias para análise de sentimentos em publicações do X relacionadas ao mercado financeiro. Ambas compartilham etapas comuns, como redução de dimensionalidade, extração de tópicos e classificação por machine learning.

Na primeira metodologia, o foco é a classificação de sentimentos de postagens em português sobre o mercado brasileiro, utilizando aprendizagem supervisionada e a teoria psicoevolucionária de Plutchik (Confiança, Alegria, Raiva, etc.). A redução de

dimensionalidade foi realizada com PCA e t-SNE, enquanto o agrupamento e a extração de tópicos utilizaram K-Means, LDA e NMF. Os classificadores Naive Bayes, SVM e *Random Forest* obtiveram desempenho de Precisão: 0,72, *Recall*: 0,52 e F1-score: 0,6.

A segunda metodologia expandiu a abordagem anterior, focando na previsão de variação de ações. Sentimentos de posts em inglês sobre a Apple, Inc. foram analisados com algoritmos não supervisionados e, posteriormente, usados para prever a variação diária das ações na Nasdaq. Técnicas como PCA, K-Means, EM e métodos de classificação (Regressão Logística, Naive-Bayes, SVM e *Random Forest*) foram aplicadas. O modelo apresentou um *F1-score* acima de 0,5 e uma taxa de acerto de 0,895 na previsão de variações.

Os resultados destacam a relação entre tópicos extraídos e variação de sentimentos, sugerindo que os métodos aplicados são eficazes tanto na classificação de sentimentos quanto na previsão de movimentos do mercado.

3.3 Sentimento de notícias e investimento estrangeiro em carteira no Brasil

Na dissertação de Cambará (2019), o objetivo é analisar como o sentimento de notícias impacta os fluxos de investimento estrangeiro em carteira no Brasil. Foram selecionadas 26.406 notícias econômicas, financeiras e políticas do *Wall Street Journal*, publicadas entre 1999 e 2018, com a palavra-chave "*Brazil*". A análise utilizou o pacote *SentimentAnalysis* (R) e o método *bag of words*, com base no dicionário Harvard IV-4, categorizando palavras em positivas ou negativas para criar um índice de sentimento mensal.

A autora explorou os fluxos de ações locais e títulos de renda fixa, considerando variáveis econômicas como taxas de juros, PIB, risco-país e índices de incerteza (EPU-BR, GEPU e IIE). A análise utilizou os modelos VARX (vetores auto-regressivos com variável exógena) e DCC-GARCH(1,1) para avaliar a relação entre variáveis internas (fatores pull) e externas (fatores push) que influenciam os investimentos.

Os resultados mostraram que o índice de sentimento é bastante previsível e que eventos importantes aumentam a volatilidade por cerca de um ano. Sentimentos positivos estão associados ao aumento nos fluxos de investimento, com impacto reduzido após dois anos. Fluxos de ações se mostraram mais voláteis e influenciados por fatores globais, enquanto títulos de renda fixa são mais sensíveis a eventos locais.

A incerteza também apresentou influência variável: enquanto a incerteza global teve impacto no curto prazo, a incerteza interna (IIE) afetou horizontes intermediários. Por outro lado, o sentimento não teve impacto significativo nas expectativas da taxa de câmbio, embora a incerteza tenha mostrado relevância.

Cambará conclui que o sentimento e a incerteza possuem impactos distintos e complementares no investimento estrangeiro, ressaltando a importância de incorporar variáveis macroeconômicas e políticas econômicas para compreender os fluxos de capital.

3.4 Considerações

A tabela 1 contém uma comparação entre os três trabalhos mais relevantes relacionados discutidos anteriormente, contendo informações da análise, como a fonte dos dados, o mercado estudado, a amostra e os indicadores econômicos utilizados.

Trabalho	Fonte dos dados	Mercado analisado	Amostra	Indicadores econômicos
Bollen, Mao e Zeng (2011)	publicações no X	Americano	> 9 milhões	DJIA
Medeiros (2019)	publicações no X	Americano e Brasileiro	84.369	Variação diária de ações da empresa Apple, Inc
Cambará (2019)	*Wall Street Journal*	Brasileiro	26.406	Câmbio, *GDP,* RF, SELIC, PIB, IBOV, etc..

Tabela 1. Comparação de informações utilizadas para a análise

A tabela 2 compara os métodos de análise de sentimento, demais técnicas estatísticas utilizadas e os resultados significativos dos trabalhos

Trabalho	Método de Análise de Sent.	Algoritmo	Métricas utilizadas	Resultados Significativos
Bollen, Mao e Zeng (2011)	OpinionFinder GPOMS	SOFNN Granger causality	Acurácia	86,7%
			MAPE	1,79%
			P-value	< 0,05 (calmo)
Medeiros (2019)	Aprendizagem supervisionada NaiveBayes SVM Random Forest PCA Classificação de sentimentos com nuvens de palavras e análise semântica	Naive-Bayes SVM Random Forest Regressão Logística	Precisão	<= 0,72 (I);
			Recall	<= 0,52 (I);
			F1-score	<= 0,6 (I); > 0,5 (II)
			Taxa de acertos	<= 0,895 (II)
Cambará (2019)	Pacote SentimentAnalysis para R dicionário Harvard IV-4	Bag of Words	VARX (sentimento)	Ações: 0,1543 Renda fixa: 0,2645 Carteira: 0,6030

Tabela 2. Comparação dos resultados e métodos e técnicas utilizadas para a análise

A maior parte dos trabalhos está relacionado ao mercado de ações americano e análise de sentimentos na língua inglesa, e nenhum deles utilizou algum *Large Language Model*, como o BERT ou GPT, que hoje em dia são os tipos de modelos mais avançados para análise de sentimento disponíveis. Dessa forma, o presente trabalho tem como objetivo coletar publicações do X em língua portuguesa sobre o IBOVESPA e analisar o desempenho dos modelos GPT e FinBERT-PT-BR na tarefa de classificação de sentimentos.

4. Desenvolvimento

Conforme descrito anteriormente, o objetivo deste trabalho é comparar o desempenho de diferentes *LLMs* (Grandes Modelos de Linguagem) para classificar sentimentos oriundos de publicações do X sobre a Bolsa de Valores de São Paulo. Para cumprir esse objetivo, o trabalho foi dividido em 5 etapas principais: coleta dos dados, limpeza dos dados, configuração dos modelos, agregação dos dados e análise dos resultados. O fluxo dessas etapas pode ser observado na Figura 1, apresentada abaixo.

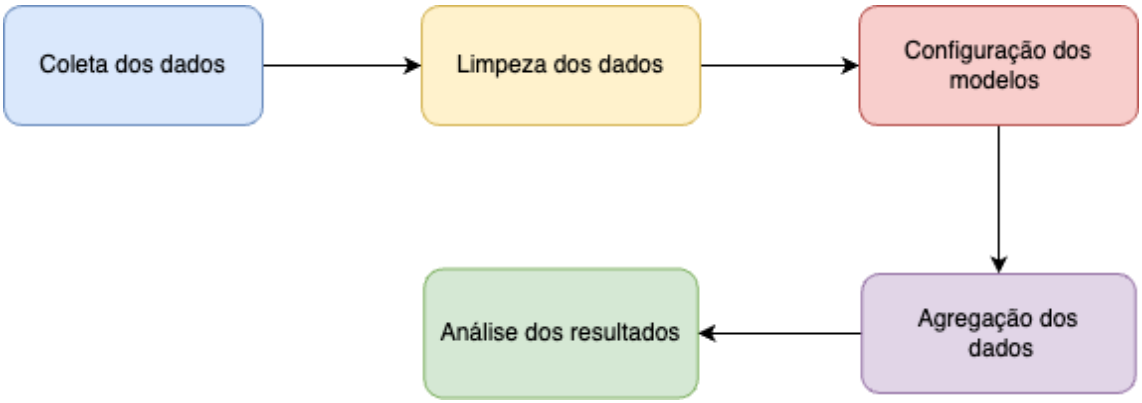


Figura 1. Etapas de desenvolvimento do trabalho.

Na primeira etapa, as publicações foram extraídas através da API do X em um período de 30 dias, entre 30 de julho a 30 de agosto de 2024. Na etapa de limpeza dos dados, foram removidas postagens irrelevantes, links, menções, emojis, hashtags redundantes, e houve normalização de texto e datas, além de exclusão de duplicatas, resultando em 7511 publicações preparadas para análise. Na etapa de configuração dos modelos, foi realizada a análise de sentimento de cada publicação com os modelos FinBERT-PT-BR e GPT-3.5 Turbo. Já na etapa de agregação dos dados, o conjunto de dados com as publicações e sentimentos foi organizado em uma nova tabela contendo o sentimento predominante de cada dia e os respectivos valores de fechamento da BOVESPA.

Por fim, foi realizada a análise dos resultados, na qual uma amostra de exemplos representativos das publicações foi examinada, e 20% dos textos foram rotulados manualmente para calcular e avaliar as métricas de desempenho de cada modelo. Além disso, foram gerados gráficos para analisar a distribuição dos sentimentos ao longo do tempo e o nível de concordância entre os modelos.

4.1 Coleta dos dados

Nesta etapa inicial, os dados foram coletados do X utilizando a API oficial v2, que permite extrair até 10.000 *posts* recentes por semana, ao custo de 100 dólares mensais. Um script em Python foi desenvolvido para coletar postagens públicas em português contendo as palavras-chave "ibov" ou "ibovespa", durante o período de 30 de agosto a 30 de setembro de 2024, totalizando 9.013 publicações. As informações extraídas incluíram ID do *post*, texto, data, horário, e dados do autor, sendo organizadas em uma tabela para facilitar a análise.

4.2 Limpeza dos dados

Após a coleta inicial, foi realizado um processo de preparação e limpeza dos dados para garantir a adequação dos textos à análise de sentimentos, eliminando ruídos e inconsistências. Postagens irrelevantes ou de baixa qualidade, como aquelas feitas por bots ou contendo mais de três símbolos de porcentagem (%), foram removidas, já que geralmente não apresentavam opiniões, apenas dados automatizados.

Links, menções a usuários e emojis foram excluídos para evitar distorções nos resultados. *Hashtags* genéricas, como "#ibov" ou "#ibovespa", foram removidas para evitar redundância, enquanto *hashtags* incorporadas a frases foram mantidas sem o símbolo "#", preservando o contexto. A normalização do texto incluiu a conversão para letras minúsculas e ajustes em espaços em branco.

As datas e horários foram convertidos do padrão UTC para o fuso horário de Brasília, assegurando a correta interpretação cronológica. Para garantir a relevância, foram consideradas somente publicações contendo palavras-chave como "ibov", "ibovespa", "bolsa" ou "bovespa". Por fim, postagens duplicadas, identificadas por conteúdo ou ID, foram removidas.

Ao término da etapa de limpeza, foram totalizadas 6.471 publicações. A Tabela 3 apresenta exemplos de publicações antes e depois da limpeza.

Pré-limpeza	Pós-limpeza
PETRÓLEO E MINÉRIO ARRASTAM O IBOV PARA O RISCO DA PERDA DOS 126MIL https://t.co/AZvglKqn7H	petróleo e minério arrastam o ibov para o risco da perda dos 126mil
Suzano (SUZB3) conclui compra de ativos no Mato Grosso do Sul #economia #investimentos #negócios #investidor #ibovespa #ações #mercadofinaceiro https://t.co/UVDytc5YjO	suzano (suzb3) conclui compra de ativos no mato grosso do sul
Se o #LVOL11 estivesse disponível desde 2003, teria reduzido, em média, 20% da volatilidade quando comparado ao Ibovespa. Com a proposta de ser mais resiliente em momentos de adversidade da Bolsa, o	se o lvol11 estivesse disponível desde 2003, teria reduzido, em média, 20% da volatilidade quando comparado ao ibovespa. com a proposta de ser mais resiliente em momentos de adversidade da bolsa, o

#LVOL11 teria apresentado uma alta de +1.792%, contra +934% do Ibovespa.	lvol11 teria apresentado uma alta de +1.792%, contra +934% do ibovespa.

Tabela 3. Amostra para comparação das publicações antes e depois da limpeza dos dados

4.3 Configuração dos modelos

4.3.1 FinBERT-PT-BR

Nesta etapa, foi utilizado o modelo pré-treinado FinBERT-PT-BR, uma adaptação do BERTimbau para análise de sentimentos no contexto financeiro em português. O modelo classificou os posts coletados em três categorias: positivo, negativo e neutro.

A análise foi realizada com dois componentes principais: o *tokenizer*, que converte os textos em tokens para interpretação pelo modelo, e o próprio FinBERT-PT-BR, implementado com a arquitetura *BertForSequenceClassification* para classificação de texto. Para otimizar o processamento, os textos foram analisados em lotes de 16 publicações utilizando GPU.

Textos com mais de 512 *tokens* foram truncados para 511, reservando um *token* especial (CLS) necessário para o modelo. Apenas dois posts ultrapassaram o limite e foram removidos. Os resultados da classificação incluem o rótulo de sentimento (*predicted_label*) e as probabilidades para cada classe (*positive, negative, neutral*). A tabela atualizada com as predições foi salva em um novo arquivo CSV, contendo o texto original e os resultados.

4.3.2 GPT-3.5 Turbo

Nesta etapa, foi utilizado o modelo GPT-3.5-Turbo, da OpenAI, para realizar a análise de sentimentos das publicações previamente limpas. A classificação foi realizada por meio da API da OpenAI, na qual cada texto foi enviado ao modelo com a instrução de classificar o sentimento como "*POSITIVE*", "*NEUTRAL*" ou "*NEGATIVE*".

O processo utilizou um prompt inicial que definia o papel do modelo: "Você é um assistente que analisa sentimentos sobre postagens relacionadas ao IBOVESPA e à economia brasileira." Em seguida, cada texto foi apresentado com a instrução: "Determine o sentimento desta frase: '{texto}'". A resposta gerada pelo modelo foi extraída e armazenada no conjunto de dados, com a classificação correspondente registrada na coluna *predicted_label*. Os resultados foram organizados em uma nova tabela, contendo as publicações e suas respectivas classificações de sentimento.

4.4 Agregação dos dados

Nesta etapa, os dados de sentimentos extraídos dos posts foram agregados por dia e correlacionados aos preços de fechamento do índice IBOVESPA (IBOV), com o objetivo de estruturar as informações para analisar a relação entre os sentimentos e o desempenho diário do índice.

Para cada modelo, os arquivos CSV contendo as publicações e os sentimentos foram lidos, e os dados foram agregados por dia. Identificou-se o sentimento predominante diário com base na frequência das categorias (positivo, negativo ou neutro), além de calcular as contagens diárias de sentimentos e o número total de posts publicados por dia.

Os preços de fechamento do IBOVESPA foram coletados utilizando a API Yahoo! Finance (yfinance) para o período analisado. Nos dias sem dados disponíveis (como fins de semana e feriados), os valores de fechamento ausentes foram preenchidos com o último preço disponível, e esses dias foram marcados com uma *flag* "bolsa_fechada".

As informações agregadas, incluindo sentimento predominante, contagens de sentimentos, total de posts, preços de fechamento e flag de dias sem operação da Bolsa, foram organizadas em um único conjunto de dados. O resultado foi salvo em um arquivo CSV. A Tabela 4 apresenta uma amostra dos dados agregados após análise com o FinBERT-PT-BR, sendo que as tabelas dos modelos GPT seguem o mesmo formato.

data	sentimento	qtd_ negativo	qtd_ neutro	qtd_ positivo	qtd_ total	fechamento	bolsa_ fechada
2024-07-30	NEGATIVE	83	38	15	136	126139.0	False
2024-07-31	POSITIVE	49	47	60	156	127652.0	False
2024-08-01	NEGATIVE	64	41	36	141	127395.0	False
2024-08-02	NEGATIVE	86	36	20	142	125854.0	False
2024-08-03	NEUTRAL	9	16	9	34	125854.0	True

Tabela 4. Amostra da tabela com os dados agregados

5. Análise dos resultados

Os resultados foram avaliados por meio de uma análise exploratória dos dados e da aplicação de métricas de desempenho, como precisão, *recall* e F1-score, utilizando uma amostra de 1290 publicações rotuladas manualmente pelo autor. O objetivo foi comparar a performance dos modelos GPT-3.5 Turbo e FinBERT-PT-BR na classificação de sentimentos em posts relacionados ao IBOVESPA.

Além disso, foi realizada uma análise comparativa dos padrões de classificação e das divergências entre os modelos, destacando as particularidades de cada um na interpretação de textos. Por fim, foi realizado um breve estudo sobre a relação entre o sentimento predominante das publicações e o comportamento do índice IBOVESPA.

5.1 Análise exploratória

Nesta etapa, foram analisadas as distribuições dos sentimentos classificados pelos modelos, destacando a prevalência de sentimentos positivos e neutros nas publicações. Gráficos foram utilizados para visualizar as diferenças entre os modelos e compreender como cada um interpreta o contexto das postagens sobre o IBOVESPA.

Além disso, exemplos representativos e matrizes de confusão foram analisados para avaliar a concordância e as divergências nas previsões dos modelos. A análise explorou possíveis razões para discrepâncias nas classificações, com foco em desafios como o tom neutro, ironia e o uso de jargões financeiros, que podem influenciar a interpretação dos textos pelos modelos.

5.1.1 Distribuição dos sentimentos

Das 6.471 publicações classificadas pelo FinBERT-PT-BR, 39,28% (2.542) apresentaram sentimento negativo, enquanto 34,06% (2.204) foram classificadas como

positivas. Sentimentos neutros representaram 26,66% dos textos. A Figura 2 abaixo ilustra a distribuição dos sentimentos identificados por esse modelo.

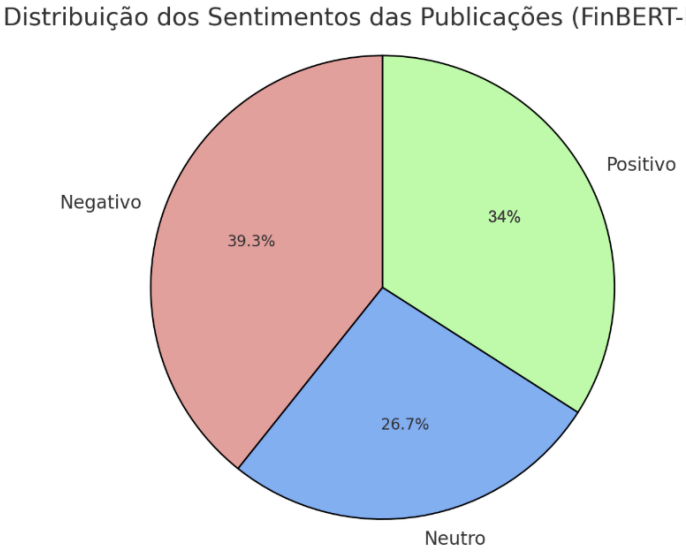


Figura 2. Distribuição do sentimento das publicações classificadas com o FinBERT-PT-BR

Por outro lado, a análise com o modelo GPT-3.5 Turbo revelou uma predominância de sentimentos positivos, representando 42,10% (2.724) do total, seguidos por 39,79% (2.575) de publicações classificadas como neutras e 18,11% (1.172) como negativas. A Figura 3 abaixo apresenta a distribuição dos sentimentos identificados pelo GPT.

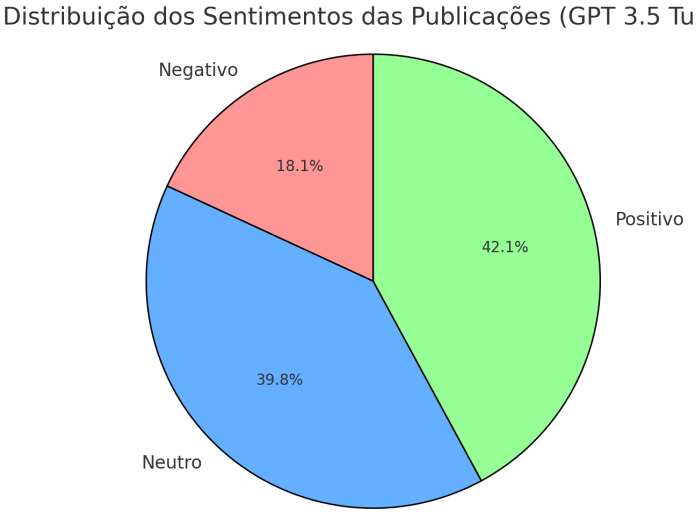


Figura 3. Distribuição do sentimento das publicações classificadas com o GPT

Ao comparar os dois modelos, observa-se que o GPT-3.5 Turbo apresentou uma tendência a classificar mais publicações como positivas, o que sugere diferenças na sensibilidade entre os modelos, especialmente em relação ao tom das publicações.

5.1.2 Matriz de Confusão

A matriz de confusão foi utilizada para avaliar o nível de concordância e divergência entre os modelos FinBERT-PT-BR e GPT-3.5 Turbo na classificação de sentimentos. Como o volume de publicações em cada categoria pode variar, foi realizada a normalização dos dados para obter uma comparação proporcional mais equilibrada entre as categorias de sentimento.

O processo de normalização consiste em dividir cada valor da matriz pela soma dos valores da sua respectiva linha, transformando os valores absolutos em proporções. Dessa forma, a soma de cada linha da matriz é igual a 1 (ou 100%), permitindo uma comparação entre as classes de sentimento independentemente da quantidade de publicações em cada categoria. A Figura 4, apresentada abaixo, contém a matriz de confusão gerada.

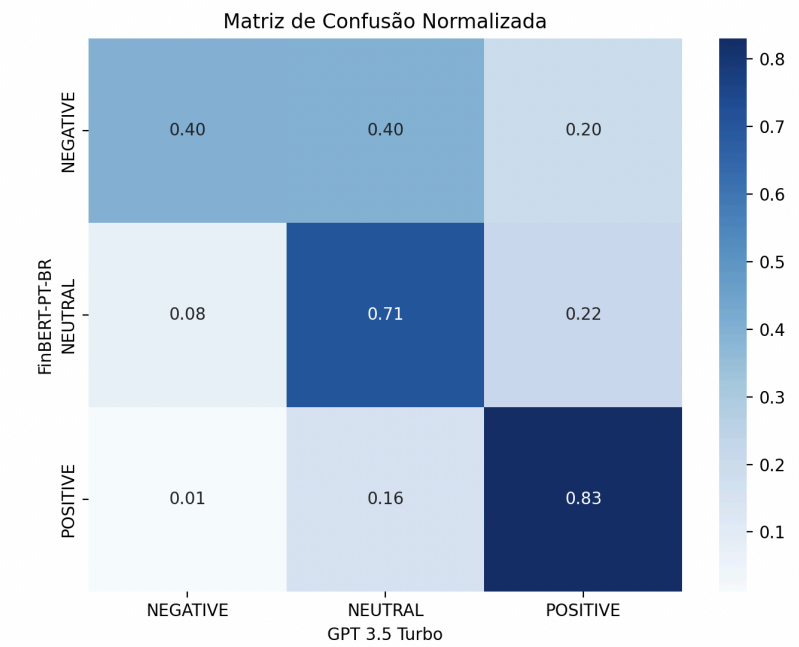


Figura 4. Matriz de confusão normalizada

A matriz normalizada revela que, das publicações negativas classificadas pelo FinBERT, 40% também foram negativas pelo GPT, enquanto 40% foram neutras e 20% positivas. Para publicações neutras, o GPT concordou em 71% dos casos, mas classificou 22% como positivas. A maior concordância foi para publicações positivas, com 83% sendo igualmente classificadas por ambos os modelos.

Esses resultados indicam que, embora haja alta concordância para sentimentos positivos, há divergências significativas em publicações negativas e neutras. O GPT-3.5 Turbo apresenta maior tendência a classificar textos como neutros ou positivos em casos onde o FinBERT-PT-BR identificou sentimentos negativos.

As diferenças podem ser atribuídas aos conjuntos de dados de treinamento dos modelos. O FinBERT-PT-BR, especializado no domínio financeiro, foi treinado com textos de portais de notícias, enquanto o GPT tem uma abordagem mais generalista, capaz de lidar melhor com textos informais, gírias, sarcasmo e erros de escrita, comuns em redes sociais como o X.

5.1.3 Exemplos Representativos

Para explorar as diferenças entre os modelos FinBERT-PT-BR e GPT-3.5 Turbo, foram analisadas publicações selecionadas aleatoriamente, destacando concordâncias e divergências em suas classificações nas categorias negativa, neutra e positiva.

Em publicações classificadas como negativas, ambos os modelos mostraram forte concordância na identificação de textos com conteúdos explicitamente pessimistas, como quedas do IBOVESPA, desvalorização do real ou sinais de crise econômica. Por exemplo, na publicação *"a única coisa que caiu de ontem para hj foi o ibovespa"*, a menção direta à queda do índice resultou em classificação negativa por ambos os modelos. Outro caso de concordância foi observado no texto *"se o ibovespa não está no fundo do poço, está bem perto, diz o jp morgan"*, em que a expressão "fundo do poço" reforça o sentimento negativo.

Para textos neutros, com caráter informativo ou especulativo e sem opiniões explícitas, os modelos também apresentaram alta concordância. Um exemplo disso é a publicação *"ibov a alta do índice vai depender..."*, que ambos classificaram como neutra devido à ausência de emoção ou julgamento no texto. Outra publicação neutra foi *"dólar e ibovespa hoje: o que esperar dos negócios na b3 nesta sexta"*, que tem um caráter meramente descritivo, sem expressar nenhum sentimento claro.

Textos positivos, por sua vez, também apresentaram alta concordância, especialmente quando expressavam otimismo ou destacavam recordes no mercado financeiro. Na publicação *"grandes companhias vêm batendo recorde de valor de mercado e analistas já projetam ibovespa a 150 mil pontos"*, ambos os modelos identificaram o tom otimista e classificaram a publicação como positiva. Da mesma forma, no texto *"enquanto isso o ibovespa bate recordes!!"*, os dois modelos reconheceram o otimismo e atribuíram a classificação positiva.

Apesar das concordâncias, cerca de 37% das publicações analisadas apresentaram divergências nas classificações, especialmente em textos com sentimentos mistos, previsões futuras, ironias ou linguagem informal. Em textos que envolvem projeções, o GPT frequentemente capturou o otimismo implícito, enquanto o FinBERT-PT-BR apresentou interpretações mais literais. Por exemplo, na publicação *"ibovespa em julho dá forte sinal de que investidor pode esperar virada de mesa em agosto. veja por quê"*, o FinBERT-PT-BR classificou como neutro, possivelmente não interpretando "virada de mesa" como uma expressão de otimismo, enquanto o GPT considerou o contexto e classificou como positivo.

Casos de ironia também evidenciaram as diferenças. No texto *"ibovespa com maior índice da história e agora dólar despencando. mas o brasil não estava falido?"*, o FinBERT-PT-BR classificou como negativo, focando nas palavras "despencando" e "falido", enquanto o GPT identificou o tom irônico e classificou como neutro. No entanto, o sarcasmo do texto sugere um sentimento positivo, ao destacar fatores econômicos favoráveis, como recordes no IBOVESPA e queda do dólar, mas ambos os modelos mostraram dificuldade em captar completamente o tom implícito.

Publicações que utilizam jargões financeiros também causaram divergências. Na frase *"segunda feira vai ser feliz pra quem tá shortando ibov"*, o GPT classificou corretamente como negativo, já que "*shortar*" refere-se a uma estratégia de venda

associada a expectativas de queda. O FinBERT-PT-BR, no entanto, classificou como neutro, provavelmente devido à sua menor familiaridade com expressões informais do mercado financeiro.

Por outro lado, em textos que misturam sentimentos ou apresentam mensagens ambíguas, os modelos demonstraram abordagens diferentes. Na publicação *"ontem o ibovespa caiu 0,95%, fechando aos 136.041 pontos, mas agosto ainda promete ser o melhor mês do ano, com avanço de 6,57%. a queda foi um ajuste após dias de alta"*, o FinBERT-PT-BR focou na queda imediata do índice e classificou como negativo, enquanto o GPT considerou o desempenho positivo ao longo do mês e classificou como positivo.

Esses exemplos ilustram como os modelos diferem na interpretação de sentimentos. O FinBERT-PT-BR, treinado em textos formais de portais de notícias, apresenta análises mais literais e tende a priorizar palavras específicas, enquanto o GPT, com sua abordagem mais generalista, mostra maior capacidade de interpretar nuances, incluindo sarcasmo, jargões e o contexto geral das publicações. Essas características destacam as forças e limitações de cada modelo em diferentes cenários de análise de sentimentos.

5.1.4 Comparação de Desempenho entre Modelos

Nesta seção, são analisados os desempenhos dos modelos FinBERT-PT-BR e GPT-3.5 Turbo na classificação de sentimentos de publicações sobre o IBOVESPA. Para a avaliação, foi selecionada uma amostra de 20% do conjunto de dados total, correspondendo a 1.290 publicações, rotuladas manualmente pelo autor. Essa amostra preserva a distribuição desbalanceada dos sentimentos observada nos dados originais, predominando sentimentos positivos, seguidos por neutros e negativos.

A análise utilizou as métricas precisão, *recall* e *F1-score*, com duas abordagens principais:

A. Média macro: cada classe de sentimento tem o mesmo peso, independentemente de sua frequência.
B. Média ponderada: ajusta as métricas conforme a proporção de cada classe no conjunto, refletindo o impacto real das previsões sobre o conjunto desbalanceado.

Modelo	**Precisão (Macro)**	***Recall* (Macro)**	***F1-Score* (Macro)**	**Precisão (Pond.)**	***Recall* (Pond.)**	***F1-Score* (Pond.)**
FinBERT-PT-BR	0.642	0.620	0.629	0.702	0.644	0.652
GPT	0.861	0.841	0.850	0.867	0.864	0.863

Tabela 5. Métricas de desempenho entre o GPT e o FinBERT-PT-BR

Os resultados, apresentados na Tabela 5, mostram que o GPT supera o FinBERT-PT-BR em todas as métricas. A precisão do GPT foi de 0,861 (macro) e 0,867 (ponderada), indicando maior exatidão na classificação correta das classes, enquanto o

FinBERT-PT-BR atingiu 0,642 (macro) e 0,702 (ponderada). A maior precisão do GPT reflete uma menor tendência a falsos positivos, sendo mais confiável para análise de sentimentos no mercado financeiro.

No recall, o GPT também apresentou desempenho superior, com 0,841 (macro) e 0,864 (ponderada), frente a 0,620 (macro) e 0,644 (ponderada) do FinBERT-PT-BR. Isso demonstra que o GPT identificou melhor as instâncias reais de cada classe, com menos falsos negativos. Já o menor recall do FinBERT-PT-BR evidencia dificuldades em capturar sentimentos neutros e negativos, especialmente em contextos ambíguos.

O F1-score, métrica que combina precisão e recall, reforça o desempenho equilibrado do GPT: 0,850 (macro) e 0,863 (ponderada), enquanto o FinBERT-PT-BR obteve 0,629 (macro) e 0,652 (ponderada). Esses resultados destacam a robustez do GPT na classificação de todas as classes de sentimento, enquanto o FinBERT enfrenta limitações, especialmente em contextos informais ou mais subjetivos, como redes sociais.

A Figura 5 apresenta a matriz de confusão do GPT em relação aos rótulos reais. O modelo mostrou alta precisão para sentimentos positivos (90%) e neutros (88%), mas menor desempenho para sentimentos negativos (76%), com 21% das publicações negativas sendo classificadas como neutras. Um exemplo dessa dificuldade é o texto *"ibovespa em máxima histórica! ... em reais bolivarianos. agora em dólar só não estamos piores que o méxico.*", que o GPT classificou como neutro devido à dificuldade em captar a ironia do contexto negativo.

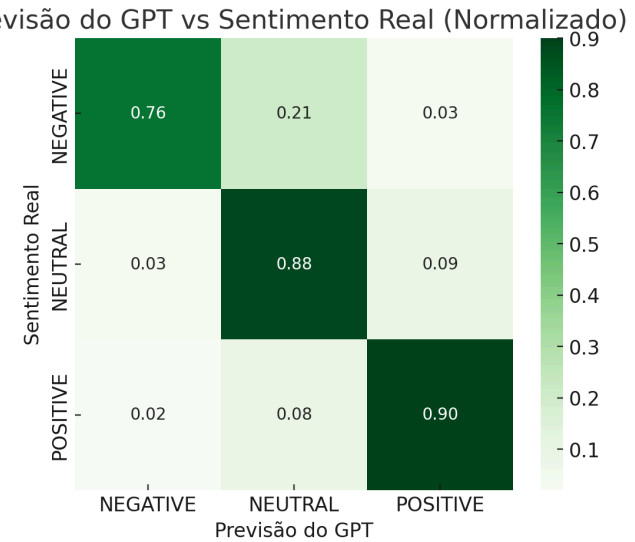


Figura 5. Matriz de confusão entre as classificações do GPT e o sentimento real

A Figura 6, com a matriz de confusão entre o FinBERT-PT-BR e o sentimento real, revela maior precisão em sentimentos negativos (83%), superando o GPT nesta classe. No entanto, o modelo apresenta menor precisão para sentimentos positivos (67%) e ainda mais baixa para neutros (51%), com confusões significativas entre sentimentos neutros e negativos (35%) e positivos (14%). Esses resultados reforçam as limitações do FinBERT-PT-BR na identificação de neutralidade, muitas vezes interpretando publicações neutras como negativas ou positivas, dependendo do contexto.

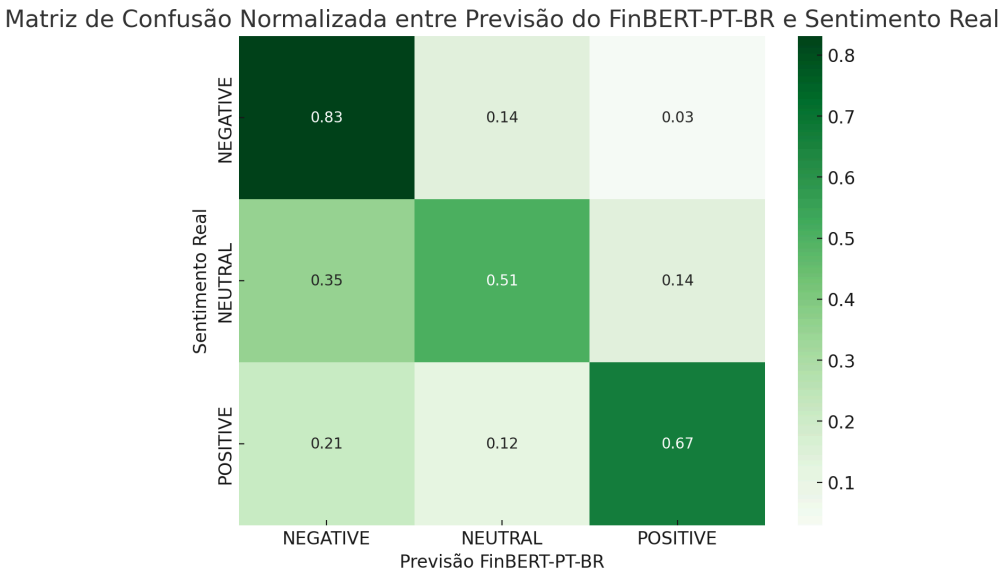


Figura 6. Matriz de confusão entre as classificações do FinBERT-PT-BR e o sentimento real

A análise confirma que o GPT é mais eficaz na classificação geral de sentimentos, especialmente em textos positivos e neutros, mas enfrenta desafios em alguns casos de ironia ou ambiguidades. Por outro lado, o FinBERT-PT-BR tem melhor desempenho em textos claramente negativos, mas tem dificuldade para lidar com nuances linguísticas comuns nas redes sociais.

Para melhorar o FinBERT-PT-BR, uma alternativa seria realizar um ajuste fino com dados provenientes de redes sociais, criando um modelo especializado em linguagem informal e identificar sentimentos implícitos, como sarcasmo e ironia. Isso poderia reduzir a confusão em textos ambíguos e ampliar sua eficácia em contextos mais variados.

5.1.5 Evolução Diária dos Sentimentos e Pontos do IBOVESPA

Para validar as classificações dos modelos e compará-las com o desempenho da bolsa de valores, foi gerado o gráfico da Figura 7, que relaciona o sentimento predominante identificado pelo GPT, o volume de publicações para cada classe de sentimento e o valor de fechamento do IBOVESPA. Dias em que a bolsa estava fechada foram excluídos do gráfico para facilitar a visualização, devido ao baixo volume de publicações nesses períodos e ao fato de que o preço da bolsa não foi alterado nestes dias.

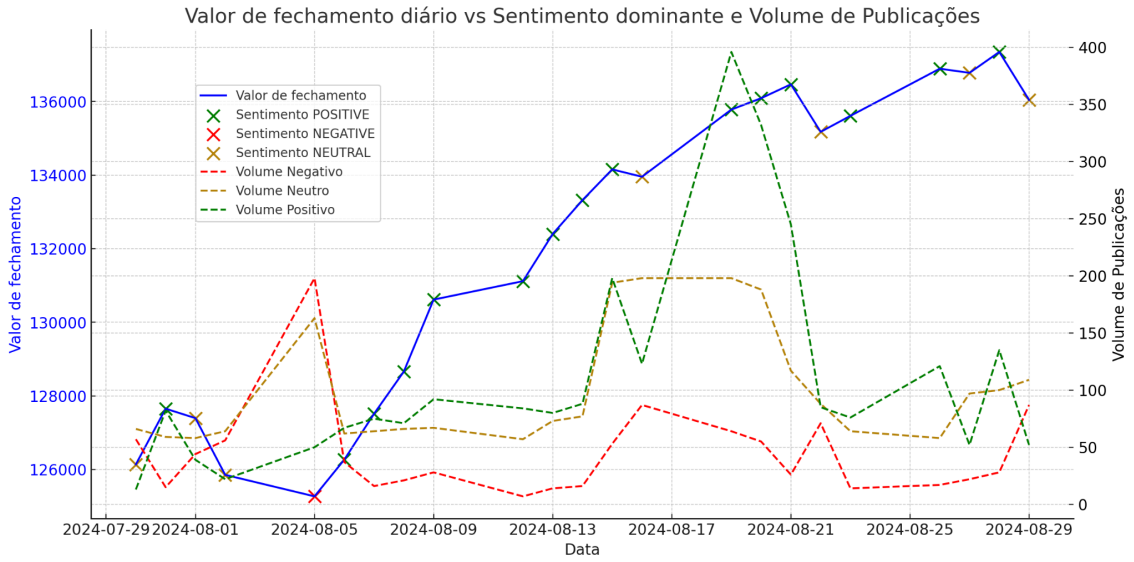


Figura 7. Valor de fechamento diário vs Sentimento dominante e volume de publicações classificadas pelo GPT

Dos 6.471 posts analisados pelo GPT, 1.172 foram classificados como negativos, 2.575 como neutros e 2.724 como positivos. A análise revelou que o sentimento predominante positivo frequentemente coincide com a valorização do índice IBOVESPA, sugerindo uma relação entre altas no mercado e otimismo nas publicações. Em contraste, em dias de queda no índice, o sentimento predominante foi, na maioria dos casos, neutro. Um exemplo é o dia 05/08, quando ocorreu um pico de publicações negativas, coincidindo com uma queda significativa no IBOVESPA e em outras bolsas globais.

O maior volume de sentimentos positivos ocorreu em 18/08, data do recorde histórico do IBOVESPA em valor nominal (Petry, 2024). Já o maior volume de sentimentos negativos foi observado em 05/08, durante uma queda global no mercado (Crepaldi e Quesada, 2024). Publicações neutras predominaram em dias de menor volatilidade no mercado, refletindo um tom mais descritivo e menos emocional.

O gráfico equivalente para o FinBERT-PT-BR, apresentado na Figura 17, fornece uma visão comparativa.

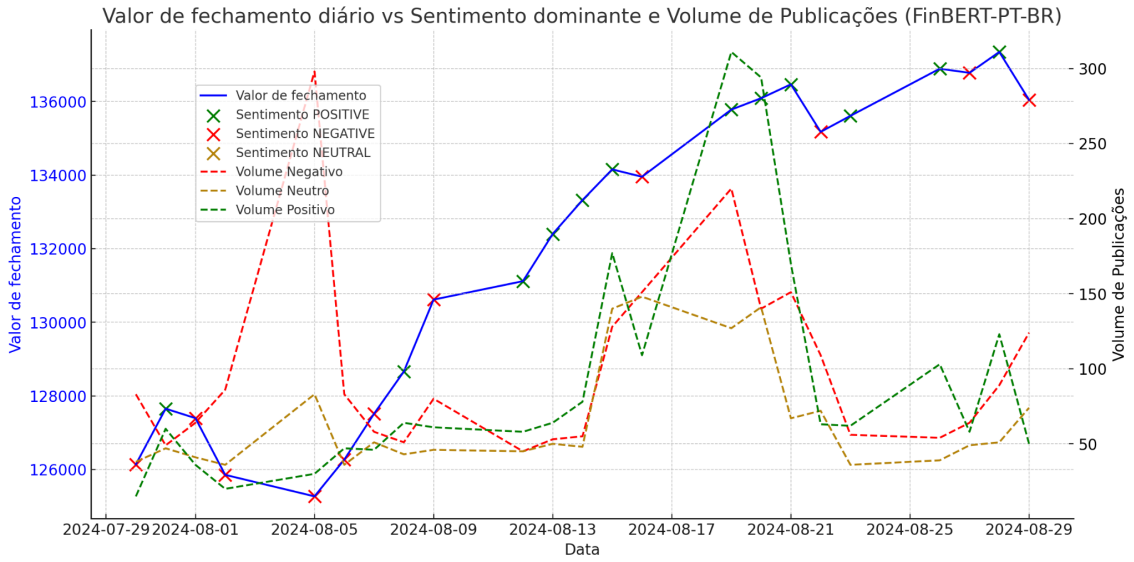


Figura 8. Valor de fechamento diário vs Sentimento dominante e volume de publicações classificadas pelo FinBERT-PT-BR

Dos 6.471 posts analisados pelo FinBERT-PT-BR, 2.542 foram classificados como negativos, 1.725 como neutros e 2.204 como positivos. A análise indica que a maioria dos dias com sentimento positivo predominante coincide com altas no IBOVESPA, similar ao observado com o GPT. No entanto, em alguns dias de valorização da bolsa, como 06, 07 e 09 de agosto, o sentimento predominante foi negativo, o que pode ser atribuído à maior precisão do FinBERT-PT-BR na identificação de tons pessimistas.

Uma diferença notável é que, enquanto o GPT frequentemente classifica o sentimento predominante como neutro em dias de queda, o FinBERT-PT-BR tende a identificar um sentimento predominantemente negativo. Essa maior sensibilidade às publicações pessimistas pode estar relacionada ao treinamento do FinBERT-PT-BR com textos financeiros formais, o que o torna mais eficaz na identificação de riscos e preocupações econômicas.

Ambos os modelos demonstraram uma correlação entre sentimentos positivos e altas no IBOVESPA, sugerindo que o otimismo no mercado é refletido nas publicações. No entanto, o FinBERT-PT-BR apresentou uma relação mais direta entre sentimentos negativos e quedas no índice, enquanto o GPT frequentemente classificou publicações com tons negativos como neutras, especialmente em casos de ambiguidades ou ironias.

Essas diferenças mostram a importância de considerar múltiplas abordagens para capturar nuances de sentimentos associados ao desempenho do mercado financeiro. Enquanto o GPT se destaca na análise geral de sentimentos positivos e neutros, o FinBERT-PT-BR apresenta maior precisão na identificação de sentimentos negativos, oferecendo *insights* complementares sobre o impacto do humor dos investidores no comportamento do IBOVESPA.

5.2 Compartilhamento do *dataset* rotulado

Para contribuir com o avanço de pesquisas e o desenvolvimento de novos modelos em português, o conjunto de dados rotulados foi disponibilizado publicamente na plataforma Kaggle para consulta e análise. O Kaggle é uma plataforma online de ciência de dados que oferece um ambiente colaborativo para compartilhamento de datasets e

desenvolvimento de projetos de análise e aprendizado de máquina, sendo amplamente utilizado por pesquisadores e profissionais da área (KAGGLE, 2024).

6. CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo comparar modelos de linguagem natural aplicados à análise de sentimentos sobre o IBOVESPA, utilizando publicações em português coletadas da rede social X (anteriormente Twitter). Diferenciando-se de estudos anteriores que utilizaram predominantemente dados em inglês e abordagens tradicionais, o presente estudo explorou o uso de LLMs (*Large Language Models*), como o GPT-3.5 Turbo, no contexto de investidores brasileiros.

O desempenho do GPT-3.5 Turbo foi comparável ao trabalho "*Twitter mood predicts the stock market*" de Bollen, Mao e Zeng (2011), que alcançou uma precisão de 87,6%, enquanto o GPT obteve 86,7%, mesmo considerando as diferenças metodológicas e contextuais entre os estudos. Esses resultados destacam o potencial do GPT e de outros *LLM* para análise de sentimentos no mercado financeiro, especialmente em redes sociais, onde o contexto e a linguagem informal desempenham papéis relevantes.

Os objetivos gerais e específicos foram cumpridos, incluindo a configuração e avaliação dos modelos FinBERT-PT-BR e GPT-3.5 Turbo e a construção de um conjunto de dados rotulados manualmente. Além disso, os resultados permitiram identificar padrões e discrepâncias entre os modelos, gerando uma discussão crítica sobre suas limitações e suas aplicações práticas.

Durante o desenvolvimento, alguns desafios foram enfrentados. A ausência de *datasets* adequados em português tornou necessária a implementação de uma coleta de publicações via API do X, o que, devido a limitações de custo e acesso, restringiu o volume de dados e o período analisado. Apesar disso, o conjunto de dados criado, com textos rotulados manualmente e pré-processados, representa uma contribuição para a área de PLN em português, especialmente no contexto de sentimentos relacionados ao mercado financeiro.

Os principais resultados demonstraram que o GPT-3.5 Turbo superou o FinBERT-PT-BR em todas as métricas avaliadas. O GPT alcançou precisão, *recall* e *F1-score* superiores, sendo mais eficiente na análise de textos subjetivos e informais característicos de redes sociais. Por outro lado, o FinBERT-PT-BR apresentou maior precisão na classificação de textos negativos, mas enfrentou dificuldades na identificação de sentimentos neutros e positivos. Essas limitações podem estar relacionadas ao treinamento do FinBERT-PT-BR com textos de portais de notícias financeiros, menos adequados para lidar com nuances linguísticas e gírias comuns nas redes sociais.

Este trabalho também identificou diversas oportunidades para trabalhos futuros. A ampliação da base de dados, incluindo mais publicações, períodos mais longos de coleta e múltiplos avaliadores para rotulagem, poderia melhorar a generalização dos modelos. O desenvolvimento de um modelo ajustado para textos de redes sociais, combinando o foco financeiro do FinBERT-PT-BR com a adaptabilidade do GPT para a linguagem informal, seria uma abordagem interessante. O uso de técnicas de *few-shot* e *zero-shot* learning poderia otimizar as análises, reduzindo a necessidade de rotulagem

manual e ampliando a aplicabilidade prática dos modelos. Além disso, a coleta de dados em períodos mais extensos permitiria aprofundar a análise da relação entre sentimentos expressos nas publicações e indicadores econômicos, como a evolução da bolsa e oportunidades de investimento.

Conclui-se que esta pesquisa reforça o potencial de modelos de linguagem natural aplicados à análise de sentimentos no mercado financeiro brasileiro e a importância de adaptar essas ferramentas ao contexto local. O uso de LLM mostrou-se promissor para monitorar sentimentos em relação à economia e explorar novas possibilidades no acompanhamento do mercado financeiro por meio das interações em redes sociais. O estudo destaca a relevância de aprimorar essas técnicas para otimizar previsões econômicas e melhorar a análise do comportamento do mercado em diferentes cenários.

References

Bollen, J., Mao, H., e Zeng, X. J. (2011) *Twitter Mood Predicts the Stock Market*. *Journal of Computational Science*.

Cambara, L. F. R. (2019) Sentimento de notícias e investimento estrangeiro em carteira no Brasil. Dissertação (Mestrado em Economia), Universidade Federal de Santa Catarina, Florianópolis.

Crepaldi, R., e Quesada, B. (2024) Ibovespa hoje: 05/08/2024. *Exame*, 5 ago.

Kaggle. (2024) Plataforma de datasets e competições de ciência de dados.

Medeiros, M. C. (2019) Metodologias para Análise de Sentimentos de Tweets sobre o Mercado Financeiro. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecânica), Universidade de Brasília, Brasília.

Petry, R. (2024) Ibovespa rompe máxima, mira 138-140 mil pontos; correções à vista. *InfoMoney*, 20 ago.