



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Instituto de Ciência e Tecnologia
Câmpus de Sorocaba

BRUNO LAZARINI DA SILVA

ANÁLISE DE SENTIMENTO APLICADA A FINANÇAS: UM ESTUDO DE CASO
COMPARATIVO ENTRE FINBERT, ROBERTA, FASTTEXT E MODERNBERT

Sorocaba

2026

BRUNO LAZARINI DA SILVA

**ANÁLISE DE SENTIMENTO APLICADA A FINANÇAS: UM ESTUDO DE CASO
COMPARATIVO ENTRE FINBERT, ROBERTA, FASTTEXT E MODERNBERT**

Trabalho de Conclusão de Curso apresentado à
Universidade Estadual Paulista (UNESP),
Instituto de Ciência e Tecnologia, Sorocaba,
como parte dos requisitos para obtenção do
grau de Bacharel em Engenharia de Controle e
Automação.

Orientador: Prof. Dr. Leopoldo André Dutra
Lusquino Filho

Sorocaba

2026

S586a

Silva, Bruno Lazarini da

Análise de sentimento aplicada a finanças: um estudo de caso comparativo entre Finbert, Roberta, Fasttext e Modernbert / Bruno Lazarini da Silva. -- Sorocaba, 2026

71 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Controle e Automação) - Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba

Orientador: Leopoldo André Dutra Lusquino Filho

1. Mercado financeiro. 2. Processamento de linguagem natural (Computação). 3. Inteligência artificial. I. Título.

BRUNO LAZARINI DA SILVA

TÍTULO DO TRABALHO: ANÁLISE DE SENTIMENTO APLICADA A
FINANÇAS: UM ESTUDO DE CASO COMPARATIVO ENTRE FINBERT,
ROBERTA, FASTTEXT E MODERNBERT

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciência e Tecnologia de Sorocaba, Universidade Estadual Paulista (UNESP), como parte dos requisitos para obtenção do grau de Bacharel(a) em Engenharia de Controle e Automação.

Data da defesa: 02/06/2026

BANCA EXAMINADORA:

Prof. Dr. Leopoldo André Dutra Lusquino Filho
UNESP – Instituto de Ciência e Tecnologia – Campus de Sorocaba

Prof. Dr. Eduardo Verri Liberado
UNESP – Instituto de Ciência e Tecnologia – Campus de Sorocaba

Prof. Dr. Felix Mauricio Escalante Ortega
UNESP – Instituto de Ciência e Tecnologia – Campus de Sorocaba

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, aos meus pais, Angela Maria Lazarini da Silva e José Roberto Cordeiro da Silva, pelo amor incondicional, por todo o suporte dado desde o início da minha jornada e pelos ensinamentos que levarei para a vida.

À minha esposa, Ralcilaine da Silva Brito, por ser meu porto seguro, estar sempre me incentivando e me dando forças para continuar essa longa caminhada, especialmente nos momentos de maior desafio.

À minha irmã, Glaziela Lazarini da Silva, pela parceria constante em todos os momentos da vida, pela amizade e pelos ensinamentos que compartilhamos diariamente.

Aos colegas de classe do curso de Engenharia de Controle e Automação, pela troca de conhecimento e apoio ao longo do curso. À Universidade Estadual Paulista (UNESP) e ao seu corpo docente pelo aprendizado proporcionado.

Ao meu orientador, Prof. Dr. Leopoldo André Dutra Lusquino Filho, pela sua paciência, orientação certa, valiosas dicas e pela parceria fundamental no desenvolvimento deste trabalho. A todos que, de alguma forma, contribuíram para a realização deste projeto, meus mais sinceros agradecimentos

RESUMO

A era digital inaugurou um novo paradigma no mercado financeiro, caracterizado pela explosão de dados textuais não estruturados que contêm sinais cruciais para a tomada de decisão. Diante da velocidade com que os mercados reagem ao fluxo informacional, a interpretação manual desses dados torna-se inviável, demandando o uso de Processamento de Linguagem Natural (PLN) para capturar o humor e as percepções do mercado de forma automatizada. Este trabalho apresenta um estudo de caso comparativo avaliando o desempenho das arquiteturas FinBERT, RoBERTa, FastText e ModernBERT na tarefa de análise de sentimento de notícias financeiras. Utilizou-se o dataset FinMarBa, cujos rótulos são informados pelo comportamento real do mercado, oferecendo um desafio de classificação mais rigoroso. Os experimentos foram conduzidos em três fases progressivas para mitigar o desbalanceamento de classes: Baseline (v1), Ponderação de Classes (v2) e Refinamento com Focal Loss e Label Smoothing (v3). A avaliação baseou-se no F1-Score Macro para garantir a representatividade da classe neutra. Os resultados demonstraram que o FinBERT v3 alcançou a maior robustez global (F1-Macro 0,55), enquanto o ModernBERT v3 obteve a maior sensibilidade para a classe neutra (Recall 52,8%), embora com um trade-off na detecção de sinais positivos. O teste estatístico de McNemar revelou um valor-p de 0,8784, indicando uma equivalência técnica entre os modelos de ponta e reforçando que o ruído intrínseco do mercado atua como um fator nivelador de desempenho. Concluiu-se que estratégias de aprendizado sensíveis ao custo são essenciais para este domínio, e que o limite teórico de performance textual isolada foi alcançado pelas arquiteturas testadas.

Palavras-chave: Análise de Sentimento; Mercado Financeiro; FinBERT; ModernBERT; Processamento de Linguagem Natural.

ABSTRACT

The digital era has ushered in a new paradigm in the financial market, characterized by an explosion of unstructured textual data containing crucial signals for decision-making. Given the speed with which markets react to the information flow, manual interpretation of these data becomes unfeasible, demanding the use of Natural Language Processing (NLP) to capture market mood and perceptions in an automated manner. This work presents a comparative case study evaluating the performance of FinBERT, RoBERTa, FastText, and ModernBERT architectures in the task of sentiment analysis of financial news. The FinMarBa dataset was used, whose labels are informed by actual market behavior, providing a more rigorous classification challenge. Experiments were conducted in three progressive phases to mitigate class imbalance: Baseline (v1), Class Weighting (v2), and Refinement with Focal Loss and Label Smoothing (v3). Evaluation was based on the Macro F1-Score to ensure representation of the neutral class. Results showed that FinBERT v3 achieved the highest overall robustness (F1-Macro 0.55), while ModernBERT v3 achieved the highest sensitivity for the neutral class (Recall 52.8%), albeit with a trade-off in positive signal detection. The McNemar statistical test revealed a p-value of 0.8784, indicating technical equivalence between top models and reinforcing that intrinsic market noise acts as a performance-leveling factor. It was concluded that cost-sensitive learning strategies are essential for this domain, and that the theoretical limit of isolated textual performance was reached by the tested architectures.

Keywords: Sentiment Analysis; Financial Marke; FinBERT; ModernBERT; Natural Language Processing.

LISTA DE FIGURAS

Figura 1 - Diagrama da arquitetura Transformer: o encoder processa a sequência de entrada com camadas de atenção e.....	20
Figura 2 - Comparativo entre os modelos de vetorização FastText (estático e local) e Transformer/BERT (dinâmico e contextual). O infográfico ilustra como o FastText utiliza n-grams de caracteres para gerar vetores fixos para cada palavra, enquanto o Transformer emprega o mecanismo de atenção para criar vetores dinâmicos adaptados ao contexto da frase inteira.	36
Figura 3 - Exemplos do dataset utilizados que são compostos pela data, título, tickers, sentimento e pct_change	42
Figura 4 - Distribuição de sentimentos entre positivos, negativos e neutros para o dataset disponibilizado.	44
Figura 5 - Volume de Notícias por Mês para o período de 2010 a 2011.	45
Figura 6 - Distribuição da média mensal do sentimento global de 2010 a 2011.....	46
Figura 7 - Métricas da Fase 1 (Baseline): Note o Recall Neutro próximo de zero.	49
Figura 8 - Matrizes de Confusão (v1): A coluna " Predito Neutro" está quase vazia.	50
Figura 9 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v1 (Baseline), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,490$; RoBERTa: $\mu=0,493$; ModernBERT: $\mu=0,461$; FastText: $\mu=0,470$)	51
Figura 10 - Métricas da Fase 2: Melhor equilíbrio entre classes.....	53
Figura 11 - Matrizes de Confusão (v1): A coluna " Predito Neutro" está quase vazia.....	53
Figura 12 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v2 (Weighted), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,536$; RoBERTa: $\mu=0,527$; ModernBERT: $\mu=0,537$; FastText: $\mu=0,466$).	54
Figura 13 - Métricas da Fase 3: ModernBERT sacrifica Recall Positivo por Neutro.	56
Figura 14 - Matrizes de Confusão (v3): Maior dispersão, mas melhor cobertura da diagonal central.	56
Figura 15 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v3 (Focal + Smooth), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,545$; RoBERTa: $\mu=0,535$; ModernBERT: $\mu=0,503$; FastText: $\mu=0,461$).....	58
Figura 16 - Comparação entre a classificação de referência e as previsões dos classificadores para manchetes do domínio financeiro, com identificação de polaridades positiva, negativa e neutra para as versões v1, v2 e v3.	61
Figura 17 - Trade-off entre Recall Positivo e Neutro ao longo das versões.....	62

Figura 18 - Evolução do F1-Macro: FinBERT mostra a maior robustez ao longo das revisões.	63
Figura 19 - Evolução do Recall Neutro: O uso de pesos e Focal Loss desbloqueou a detecção desta classe.	63
Figura 20 - Resultado do teste de significância estatística de McNemar entre os modelos FinBERT v3 e ModernBERT v3, apresentando a matriz de contingência, as acurácias obtidas e o valor-p, indicando ausência de diferença estatisticamente significativa entre os desempenhos.	65

LISTA DE TABELAS

Tabela 1 - Comparativa de Modelos de PLN FinBERT, RoBERTa, ModernBERT e FastText.	23
Tabela 2 - Principais características do modelo FinBERT	25
Tabela 3 - Principais características do modelo RoBERTa	26
Tabela 4 - Principais características do modelo FastText.....	27
Tabela 5 - Principais características do modelo FastText.....	29

SUMÁRIO

1 INTRODUÇÃO	13
2 TRABALHOS RELACIONADOS	18
2.1 Análise De Sentimento Na Finança	18
2.2 Introdução Ao Deep Learning.....	19
2.2.1 Treinamento Do Modelo	20
2.2.2 Desempenho Do Modelo	21
2.3 Machine Learning Para Finança	22
2.4 Modelo Finbert	24
2.5 Modelo Robert	25
2.6 Modelo Fasttext	26
2.7 Modelo Modernbert	28
3 ABORDAGEM E PROPOSTA	30
3.1 Pipeline	30
3.1.1 Descrição Da Abordagem	30
3.1.2 Pipeline De Desenvolvimento (Fluxo De Trabalho)	31
3.1.3 Métricas De Eficácia Preditiva	32
3.2 Técnica Do Processamento	34
3.2.1 Fase 1: Baseline (V1).....	34
3.2.2 Fase 2: Ponderação De Classes (V2).....	34
3.2.3 Fase 3: Refinamento (V3).....	34
3.2.4 Saneamento E Normalização De Texto	35
3.2.5 Tokenização E Representação Numérica	35
3.3 Modelos Machine Learning - Penalidades	37
4 EXPERIMENTOS	39
4.1 Setup Dos Experimentos.....	39
4.2 Dataset Original.....	41
5 RESULTADO DOS EXPERIMENTOS	47
5.1 Visão Geral Dos Resultados.....	47
5.2 Analise Por Fase Experimental	49
5.2.1 Fase 1: Baseline (V1)	49
5.2.2 Fase 2: Ponderação De Classes (V2).....	52
5.2.3 Fase 3: Refinamento Com Focal Loss (V3)	55
5.3 DISCUSSÃO	59
5.3.1 Discussão Comparativa De Resultados	59

5.3.2 Análise Do Trade-Off	61
5.3.3 Limitações Do Dataset	64
5.3.4 O Confronto Finbert E Modernbert.....	65
6 CONCLUSÃO	67
REFERÊNCIAS	69
APÊNDICE A - Principais bibliotecas utilizadas.	71

1 INTRODUÇÃO

“A intersecção entre o mercado financeiro e a tecnologia tem sido um dos campos mais dinâmicos e transformadores do século XXI. Historicamente, a análise financeira dependia predominantemente de dados quantitativos e não estruturados, como balanços patrimoniais, demonstrações de resultados, séries de preços de ativos e indicadores macroeconômicos. As decisões de investimento eram, em grande parte, moldadas por modelos econométricos e pela interpretação humana desses números. Contudo, a era digital inaugurou um novo paradigma, caracterizado por uma explosão de dados não estruturados gerados a um ritmo sem precedentes. Diariamente, um volume massivo de informação textual é produzido por portais de notícias, relatórios de analistas, comunicados de imprensa de empresas, fóruns de discussão e, crucialmente, redes sociais. Esta torrente de dados textuais representa uma fonte de informação imensuravelmente rica, contendo nuances, opiniões e sinais que os dados numéricos por si só não conseguem capturar, tornando a análise textual uma área de crescente importância para a contabilidade e finanças” (Loughram; McDonald, 2016).

Essa abundância de informação, no entanto, apresenta um desafio monumental. A capacidade humana de ler, processar e interpretar essa vasta quantidade de texto em tempo hábil para tomar decisões de investimento informadas é extremamente limitada. A velocidade com que os mercados reagem às notícias torna a análise manual não apenas ineficiente, mas frequentemente obsoleta no momento em que é concluída. É neste cenário que o campo do Processamento de Linguagem Natural (PLN), um ramo da Inteligência Artificial, emerge como uma tecnologia disruptiva e essencial.

O PLN oferece as ferramentas necessárias para automatizar a extração de insights valiosos a partir de texto, permitindo que máquinas leiam e compreendam a linguagem humana em grande escala. Dentro do vasto leque de aplicações do PLN, a Análise de Sentimento destaca-se como uma das mais impactantes para o setor financeiro. Esta técnica foca em identificar e classificar a polaridade emocional expressa em um texto seja ela positiva, negativa ou neutra, funcionando como um termômetro para o humor e a percepção do mercado em relação a um determinado ativo, empresa ou evento econômico (Kolanui Et Al., 2021).

A hipótese fundamental, amplamente explorada na literatura, é que o sentimento coletivo expresso na mídia ou na internet pode servir como um indicador preditivo do comportamento dos investidores e, por conseguinte, das flutuações de preços dos ativos (Loughram; McDonald, 2016). A importância dessa análise vai além da simples previsão, tocando em questões sobre como a opinião pública, uma vez quantificada, pode refletir ou até

mesmo ser usada para influenciar movimentos de mercado. O poder do sentimento online, especialmente em redes sociais, de catalisar ações coletivas é um fenômeno bem documentado. Como Bollen, Mao e Zeng (2011, p. 1) destacam em seu influente estudo sobre o Twitter:

As medições do sentimento coletivo a partir de postagens em grande escala no Twitter estão correlacionadas com o valor de fechamento do Dow Jones Industrial Average (DJIA) [...] Nós descobrimos que uma medição do "calma" do público pode prever o valor do DJIA com uma precisão de 87.6%. Estes resultados sugerem que a sociedade pode estar, em certo grau, em um estado de "expectativa" ou "ansiedade" sobre o futuro, e que as medições do humor coletivo a partir do Twitter podem de fato ser um sensor para prever mudanças no mercado.

Esta capacidade de "sensorar" o humor coletivo abre portas não apenas para estratégias de investimento mais sofisticadas, mas também para a compreensão de como narrativas podem ser construídas para influenciar o comportamento dos investidores, tornando a Análise de Sentimento uma ferramenta de poder informacional e econômico.

A jornada tecnológica da Análise de Sentimento evoluiu drasticamente ao longo dos anos. As abordagens iniciais eram predominantemente baseadas em léxicos, utilizando dicionários de palavras pré-rotuladas com pontuações de sentimento. Embora simples, esses métodos frequentemente falhavam em capturar o contexto, a ironia e as negações complexas, resultando em uma precisão limitada. A ascensão do Aprendizado de Máquina clássico, com algoritmos como Naive Bayes e Máquinas de Vetores de Suporte (SVM), representou um avanço significativo, permitindo que os modelos aprendessem padrões a partir de dados rotulados. Contudo, a verdadeira disrupção no processamento textual não decorreu apenas do conceito de redes neurais cujos fundamentos teóricos são historicamente anteriores ao próprio aprendizado estatístico clássico, mas sim do amadurecimento do deep learning em larga escala. A capacidade contemporânea dessas arquiteturas profundas de aprender representações hierárquicas e complexas culminou no desenvolvimento dos Grandes Modelos de Linguagem (Large Language Models (LLMs)). Essa evolução abriu caminho para uma compreensão contextual e semântica sem precedentes, redefinindo as fronteiras da interpretação automatizada de textos.

O marco mais significativo nesta trajetória recente foi a introdução da arquitetura Transformer em 2017 (Vaswani Et Al., 2017). Diferentemente das arquiteturas sequenciais anteriores, o Transformer processa todas as palavras de uma frase simultaneamente, utilizando um mecanismo de "autoatenção" (self-attention) para pesar a importância de cada palavra em relação a todas as outras. Essa inovação permitiu aos modelos capturar dependências de longo

prazo e relações contextuais de forma muito mais eficaz. A partir desta arquitetura, nasceu o modelo BERT (Bidirectional Encoder Representations from Transformers), que estabeleceu um novo estado da arte em inúmeras tarefas de PLN (Devlin Et Al., 2018). O sucesso do BERT popularizou uma metodologia poderosa de duas etapas: um pré-treinamento em um corpo massivo de texto não rotulado para aprender a estrutura da linguagem, seguido por um ajuste fino (fine-tuning) em um conjunto de dados menor e específico da tarefa, como a classificação de sentimento.

A eficácia do BERT inspirou a criação de uma vasta gama de modelos derivados, cada um com focos e otimizações distintas, que formam o cerne deste estudo comparativo. Uma das principais vertentes de pesquisa foi a especialização de domínio. A hipótese é que, embora poderosos, modelos de propósito geral como o BERT podem não capturar o léxico e as nuances semânticas de campos altamente especializados, como o financeiro. Para resolver isso, foi desenvolvido o FinBERT, uma versão do BERT que passou por um treinamento adicional focado exclusivamente em textos financeiros (Arévalo Et Al., 2020). Ao ser imerso em notícias econômicas e relatórios corporativos, o FinBERT aprende a interpretar termos como "bullish" ou "hawkish" em seu contexto correto, prometendo uma acurácia superior em tarefas financeiras. O trabalho de Souza, Souza e Camargo (2024), que desenvolveu o FinBERT-Br para o contexto brasileiro, reforça a validade dessa abordagem ao demonstrar que a especialização de domínio e de idioma pode levar a ganhos de desempenho significativos.

Em paralelo à especialização, outra linha de pesquisa focou em otimizar a robustez e o poder dos modelos de propósito geral. O RoBERTa (Robustly Optimized BERT Pretraining Approach) é o principal expoente dessa abordagem (Liu Et Al., 2019). Seus criadores não alteraram a arquitetura do BERT, mas aprimoraram drasticamente sua "receita de treinamento", utilizando um volume de dados dez vezes maior, treinando por mais tempo e otimizando os hiperparâmetros. O resultado é um modelo de uso geral extremamente poderoso. Isso levanta uma questão central para este trabalho: um modelo generalista, porém massivamente otimizado como o RoBERTa, possui uma capacidade de aprendizado tão vasta que consegue superar a performance de um modelo especialista como o FinBERT, mesmo sem treinamento de domínio específico.

Contudo, a busca pela máxima acurácia, personificada por modelos como o RoBERTa, trouxe consigo um custo computacional proibitivo. Esses modelos gigantescos exigem recursos de hardware significativos e dias de treinamento, tornando sua implementação inviável para muitas aplicações práticas que demandam agilidade e eficiência. Essa limitação abriu espaço

para duas outras categorias de modelos. A primeira é representada pelo FastText, uma arquitetura mais leve e de uma família diferente, que não se baseia em Transformers (Bojanowski Et Al., 2017).

O FastText representa palavras através de n-gramas de caracteres, uma abordagem que o torna computacionalmente eficiente e notavelmente bom em lidar com a rica morfologia e palavras fora do vocabulário do português. Ele servirá neste estudo como um benchmark de velocidade e simplicidade, respondendo à pergunta sobre qual o sacrifício em precisão ao se optar por uma solução de alta eficiência.

A segunda resposta ao desafio computacional veio de dentro da própria família Transformer, através do que pode ser chamado de ModernBERT. Este termo engloba uma classe de modelos mais recentes, como o DistilBERT (Sanh Et Al., 2019) ou o ALBERT (Lan Et Al., 2020), cujo objetivo é reter a maior parte da performance do BERT original, mas em um "pacote" muito menor e mais rápido. Utilizando técnicas como destilação de conhecimento onde um modelo grande "ensina" um menor ou compartilhamento de parâmetros, esses modelos buscam o equilíbrio ideal entre desempenho e viabilidade prática. Eles representam a busca pelo custo-benefício ótimo, uma consideração crucial para a adoção em larga escala de tecnologias de PLN no setor financeiro.

Diante deste panorama de diferentes filosofias arquitetônicas especialista, generalista, eficiente e custo-benefício, emerge uma clara lacuna na literatura, especialmente no que tange ao mercado financeiro. Embora existam estudos isolados sobre alguns desses modelos, falta uma análise comparativa sistemática que os avalie sob as mesmas condições. Qual o verdadeiro ganho de um modelo de domínio específico como o FinBERTa? A força bruta do RoBERTa compensa sua falta de especialização? Quão distante fica a performance de modelos eficientes como o FastText e o ModernBERT em comparação com os gigantes? É essa lacuna que este Trabalho de Conclusão de Curso se propõe a preencher.

Portanto, o objetivo geral deste trabalho é realizar um estudo de caso comparativo, avaliando o desempenho dos modelos FinBERT, RoBERTa, FastText e ModernBERT na tarefa de Análise de Sentimento de notícias do mercado financeiro. Os objetivos específicos incluem:

- I. A implementação e o treinamento de cada um dos quatro modelos em um mesmo conjunto de dados de referência;
- II. A avaliação sistemática de seu desempenho utilizando métricas de acurácia consagradas, como o F1-Score;

- III. A análise da eficiência computacional de cada modelo, considerando tempo de treinamento e inferência;
- IV. A discussão crítica dos resultados, visando fornecer um guia prático sobre os trade-offs de cada abordagem para aplicações no contexto financeiro nacional.

Para alcançar tais objetivos, este trabalho está estruturado em cinco capítulos. O primeiro, esta introdução, contextualiza o tema e define o escopo da pesquisa. O segundo capítulo apresentará o referencial teórico, detalhando os fundamentos da Análise de Sentimento e a arquitetura de cada modelo estudado. O terceiro capítulo descreve a metodologia, incluindo o dataset utilizado, os procedimentos de pré-processamento e os parâmetros experimentais. O quarto capítulo apresentará e discutirá os resultados obtidos na análise comparativa. Por fim, o quinto capítulo trará as conclusões do trabalho, sumarizando os achados, apontando as limitações e sugerindo direções para pesquisas futuras.

2 TRABALHOS RELACIONADOS

2.1 Análise De Sentimento Na Finança

A Análise de Sentimento (AS), também conhecida como mineração de opinião, é uma subárea do Processamento de Linguagem Natural (PLN) que se dedica a identificar, extrair e quantificar estados afetivos e informações subjetivas em fontes textuais. Aplicada ao domínio financeiro, ela transcende a simples classificação de dados e se torna uma ferramenta estratégica. A hipótese central é que os mercados financeiros não são perfeitamente eficientes e que o comportamento dos investidores é significativamente influenciado pela psicologia de massa e pela percepção pública, um pilar da economia comportamental. O sentimento extraído de notícias, relatórios de analistas, comunicados de imprensa e, cada vez mais, de mídias sociais (como Twitter e Reddit), funciona como um proxy para o "humor do mercado" (Bollen Et Al., 2011).

Tradicionalmente, a análise financeira dependia de dados quantitativos estruturados, como séries de preços e balanços patrimoniais. A Análise de Sentimento desbloqueia o vasto universo de dados não estruturados, permitindo que os analistas capturem nuances que os números por si só não revelam. Um relatório de lucros pode conter números positivos, mas o tom do CEO na conferência de imprensa pode ser cauteloso, sinalizando riscos futuros. A análise de sentimentos visa capturar essa nuance, classificando o tom de um documento como positivo (indicando otimismo, "bullish"), negativo (indicando pessimismo, "bearish") ou neutro.

O desafio específico da AS em finanças é a sua natureza de domínio-específico. A linguagem financeira é densa, repleta de jargões, e o mesmo termo pode ter polaridades opostas dependendo do contexto. Por exemplo, a palavra "alto" é positiva em "lucro alto", mas negativa em "risco alto" ou "custo alto". Da mesma forma, negações e declarações prospectivas são complexas (ex: "a empresa não conseguiu evitar uma queda nos lucros"). Como apontado por Loughran e McDonald (2016), dicionários de sentimento de propósito geral falham miseravelmente em finanças, pois palavras comuns (como "passivo" ou "risco") têm conotações inerentemente negativas no contexto financeiro, embora sejam neutras no uso geral.

Portanto, a aplicação bem-sucedida da AS em finanças depende de modelos que possam compreender profundamente esse léxico e contexto únicos. O objetivo final é integrar esses sinais de sentimento em modelos preditivos para prever a direção dos preços das ações, avaliar a volatilidade do mercado ou gerenciar o risco de portfólio. Este TCC se insere exatamente nesta

fronteira, avaliando quais das arquiteturas de PLN modernas são mais capazes de decifrar o complexo sentimento do mercado financeiro brasileiro.

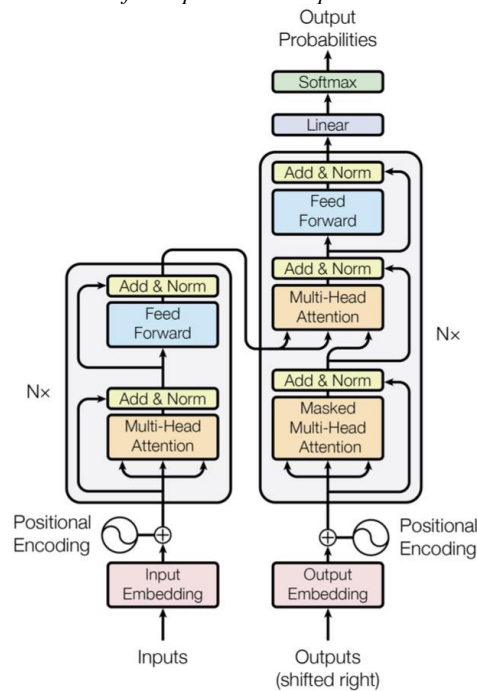
2.2 Introdução Ao Deep Learning

O Deep Learning (Aprendizado Profundo) é uma subcategoria do Aprendizado de Máquina baseada em redes neurais artificiais. O que o torna "profundo" é o uso de arquiteturas com múltiplas camadas de neurônios (camadas ocultas) que se situam entre a camada de entrada (os dados brutos) e a camada de saída (a predição). Essas camadas permitem que o modelo aprenda representações hierárquicas dos dados. Em uma tarefa de visão computacional, por exemplo, as primeiras camadas podem aprender a detectar bordas simples, as camadas intermediárias a combinar bordas em formas (círculos, quadrados), e as camadas mais profundas a identificar objetos complexos (um rosto, um carro).

No Processamento de Linguagem Natural, essa hierarquia é fundamental para a compreensão do texto. As primeiras camadas de uma rede neural profunda podem aprender representações de palavras (vetores de palavras ou word embeddings), capturando suas relações semânticas (ex: "rei" está para "rainha" assim como "homem" está para "mulher"). Camadas subsequentes aprendem a combinar essas palavras em representações de frases, capturando a sintaxe e a gramática. As camadas mais profundas, por fim, visam captar o significado semântico e contextual da passagem inteira. Essa capacidade de aprender o contexto é o que diferencia o deep learning dos métodos clássicos, como o Bag-of-Words, que tratam o texto como um "saco" de palavras desordenadas.

A grande revolução no PLN veio com a arquitetura Transformer (Vaswani Et Al., 2017). Antes dela, os modelos de estado da arte eram sequenciais, como as Redes Neurais Recorrentes (RNNs) e suas variantes (LSTMs e GRUs), que leem o texto palavra por palavra, da esquerda para a direita. O Transformer quebrou esse paradigma ao introduzir o mecanismo de auto-atenção (self-attention), que permite ao modelo processar todas as palavras de uma frase simultaneamente e "olhar" para todas as outras palavras para entender o papel de cada uma. Em "o banco faliu após o banco ser roubado", o mecanismo de atenção aprende que o primeiro "banco" se refere a uma instituição financeira (contexto: "faliu") e o segundo a um assento (contexto: "roubado"), a estrutura pode ser vista na Figura 1 a seguir.

Figura 1 - Diagrama da arquitetura Transformer: o encoder processa a sequência de entrada com camadas de atenção e feed-forward; o decoder gera a saída passo a passo usando atenção mascarada e atenção ao encoder, finalizando com Softmax para obter as probabilidades



Fonte: Vaswani et al. (2017), *Attention Is All You Need* (arXiv:1706.03762).

Todos os modelos de linguagem avançados utilizados neste TCC (com exceção do FastText) são baseados nessa arquitetura Transformer. Eles são capazes de processar o texto de forma bidirecional (olhando para o que vem antes e depois de uma palavra), o que lhes confere uma compreensão contextual sem precedentes, essencial para decifrar a linguagem ambígua das finanças.

2.2.1 Treinamento Do Modelo

O processo de treinamento dos modelos de linguagem modernos, como o BERT e o RoBERTa, é uma metodologia de duas etapas que se tornou o padrão-ouro no PLN: o pré-treinamento e o ajuste fino (fine-tuning). Essa abordagem permite que os modelos aprendam com vastas quantidades de dados, mesmo sem rótulos, e depois se especializem rapidamente em tarefas específicas.

A primeira fase, o pré-treinamento, é a mais intensiva e computacionalmente cara. Nessa etapa, o modelo é alimentado com um corpus massivo de texto não rotulado (ex: toda a Wikipédia e milhões de livros) e aprende a estrutura fundamental da linguagem de forma auto-supervisionada. Ele não aprende a classificar sentimentos, mas sim a "entender" gramática, semântica e fatos sobre o mundo. Isso é feito através de tarefas como o Masked Language Modeling (MLM), onde o modelo deve prever palavras que foram aleatoriamente "mascaradas"

(ocultadas) no texto. Ao fazer isso, ele é forçado a aprender o contexto bidirecional (o que vem antes e depois) para fazer a previsão correta.

A segunda fase é o ajuste fino (fine-tuning). Uma vez que temos um modelo "pré-treinado" que possui um vasto conhecimento geral da linguagem, podemos adaptá-lo para uma tarefa específica. Para este TCC, isso significa pegar o modelo pré-treinado e treiná-lo adicionalmente por um curto período em um conjunto de dados muito menor e rotulado. Durante o fine-tuning, uma pequena camada de classificação é adicionada ao topo da arquitetura do Transformer, e os pesos de todo o modelo são ajustados sutilmente para que ele se especialize na tarefa de classificação de sentimento.

Este paradigma de duas etapas é incrivelmente eficiente. Em vez de treinar um modelo gigantesco do zero para cada tarefa (o que seria computacionalmente inviável), a comunidade de pesquisa pode se beneficiar de modelos pré-treinados e apenas realizar a etapa de ajuste fino, que é rápida e pode ser feita com conjuntos de dados muito menores. É exatamente este processo que será aplicado aos modelos FinBERT, RoBERTa, ModernBERT e FastText neste trabalho.

2.2.2 Desempenho Do Modelo

Avaliar o desempenho de um modelo de classificação não é uma tarefa trivial, especialmente em domínios como finanças, onde os erros podem ter custos muito diferentes. A métrica mais intuitiva, a Acurácia (proporção de acertos totais), é frequentemente enganosa. Em um dataset de notícias onde 90% das manchetes são "neutras", um modelo que "chuta" tudo como "neutro" terá 90% de acurácia, mas será completamente inútil na prática, pois falhará em identificar os sinais positivos e negativos, que são os mais importantes.

Para uma análise mais robusta, utiliza-se a Matriz de Confusão. Esta tabela detalha os acertos e erros do modelo, dividindo as predições em quatro categorias para cada classe: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). No contexto financeiro, esses erros têm implicações diretas: um Falso Positivo (Erro Tipo I) seria classificar uma notícia neutra como positiva, levando um investidor a comprar uma ação desnecessariamente. Um Falso Negativo (Erro Tipo II) seria classificar uma notícia negativa como neutra, fazendo com que o investidor deixe de vender uma ação antes de uma queda, o que geralmente é um erro muito mais custoso. A partir da Matriz de Confusão, derivamos duas métricas essenciais: Precisão e Recall. A Precisão mede a confiabilidade das predições positivas; de todas as vezes que o modelo sinalizou "positivo", quantas estavam corretas?

O Recall mede a sensibilidade do modelo; de todas as notícias que eram realmente positivas, quantas o modelo conseguiu encontrar? Existe um trade-off fundamental entre essas duas: um modelo muito cauteloso (alta precisão) pode deixar passar muitas notícias (baixo recall), e vice-versa. Para unificar essas duas métricas em um único número que facilite a comparação, utiliza-se o F1-Score. Ele é a média harmônica da Precisão e do Recall, a média harmônica penaliza fortemente valores extremos, garantindo que um modelo só possa ter um F1-Score alto se tiver ambos, precisão e recall, em equilíbrio. Para problemas multiclass, o F1-Score Macro (que calcula o F1 para cada classe e tira a média) é a métrica padrão-ouro, pois trata todas as classes com igual importância, sendo a métrica central para a avaliação comparativa neste TCC.

2.3 Machine Learning Para Finança

O Aprendizado de Máquina (ML) está remodelando fundamentalmente o setor financeiro, indo muito além da Análise de Sentimento. Sua principal força é a capacidade de analisar conjuntos de dados vastos, complexos e de alta dimensão (os big data) para identificar padrões e fazer previsões com uma precisão que supera os modelos estatísticos e econométricos tradicionais. As aplicações de ML em finanças podem ser amplamente categorizadas em várias áreas-chave, como negociação algorítmica, gerenciamento de risco e detecção de fraude.

Na negociação algorítmica e no quantitative trading, os modelos de ML são usados para prever movimentos de preços de ativos em curtos e longos prazos. Modelos de séries temporais, como LSTMs (Long Short-Term Memory), são frequentemente usados para analisar dados históricos de preços e volume. Mais avançado, o Reinforcement Learning (Aprendizado por Reforço) é usado para treinar "agentes" (bots) que aprendem estratégias ótimas de compra e venda interagindo diretamente com simulações de mercado, otimizando para o lucro máximo enquanto gerenciam o risco.

No gerenciamento de risco, o ML é essencial para a avaliação de crédito (credit scoring). Modelos como Random Forests ou Gradient Boosting (XGBoost) podem analisar milhares de pontos de dados de um solicitante de empréstimo (indo muito além da renda e do histórico de pagamento) para prever a probabilidade de inadimplência com muito mais precisão do que os scorecards tradicionais. Isso permite que as instituições financeiras precifiquem melhor seus empréstimos e reduzam as perdas.

Outra aplicação crítica é a detecção de fraude e anomalias. O ML é usado para monitorar transações de cartão de crédito em tempo real. Os modelos aprendem o padrão de gastos

"normal" de um cliente e podem sinalizar instantaneamente qualquer transação que se desvie desse padrão (uma detecção de anomalia), bloqueando fraudes antes que elas aconteçam. No contexto deste TCC, a Análise de Sentimento (PLN) atua como uma fonte vital de dados alternativos. O sentimento extraído das notícias é usado para enriquecer esses outros modelos de ML, fornecendo um sinal preditivo que não existe nos dados numéricos, tornando as previsões de risco e de preço ainda mais precisas. Na Tabela 1, podemos verificar um resumo dos modelos que vão ser tratados nos capítulos a seguir.

Tabela 1 - Comparativa de Modelos de PLN FinBERT, RoBERTa, ModernBERT e FastText.

Característica	FinBERT	RoBERTa	ModernBERT	FastText
Pilar no TCC	Especialista	Robusto	Eficiente	Baseline
Arquitetura	Transformer (Encoder)	Transformer (Encoder)	Transformer Otimizado	Rede Neural Rasa (Linear)
Filosofia	Especialização de Domínio	Generalista de Alta Performance	Custo-benefício e Baixa Latência	Eficiência e Simplicidade
Dados de Pré-treino	Corpora Financeiros (Bloomberg, Reuters, 10-K)	Corpus Geral Massivo (160GB - Wiki, Books, Web)	Corpus Geral + Código (Trilhões de tokens)	N-gramas de caracteres (sub-palavras)
Contexto	Bidirecional (curto: 512 tokens)	Bidirecional (curto: 512 tokens)	Bidirecional (longo: até 8.192 tokens)	Estático (não captura ordem das palavras)
Vantagem Principal	Compreensão do léxico financeiro ("bearish", "yield")	Robustez através de pré-treino otimizado	Alta velocidade com desempenho de Transformer	Lida bem com palavras fora do vocabulário (OOV)
Custo Computacional	Alto (requer GPU para inferência)	Muito Alto (mais parâmetros que o BERT)	Médio/Baixo (focado em throughput)	Muito Baixo (roda rápido em CPU)
Uso em Tempo Real	Limitado pela latência	Baixa viabilidade para HFT*	Ideal para sistemas de tempo real	Excelente, mas com menor precisão semântica

Fonte: Elaborado pelo autor (2026).

2.4 Modelo Finbert

O FinBERT (ARÉVALO et al., 2020) é a personificação da abordagem de especialização de domínio no Processamento de Linguagem Natural. Ele não é uma arquitetura de modelo fundamentalmente nova, mas sim uma versão do modelo BERT (DEVLIN et al., 2018) que foi especificamente adaptada para o léxico e o contexto do setor financeiro. A premissa central por trás do FinBERT é que modelos de linguagem de propósito geral, pré-treinados em corpora como a Wikipédia ou o Google News, não possuem a fluência necessária para compreender as nuances da linguagem financeira.

O processo de criação do FinBERT envolve pegar o modelo BERT padrão, que já possui um vasto conhecimento da linguagem geral, e submetê-lo a uma segunda fase de pré-treinamento (conhecida como domain-adaptive pre-training). Nesta fase, o modelo é treinado adicionalmente em um corpus massivo composto exclusivamente por textos financeiros. Isso pode incluir milhões de notícias de portais como Bloomberg e Reuters, relatórios anuais de empresas (formulários 10-K), transcrições de teleconferências de resultados e manuais de análise econômica.

Ao ser imerso nesse vocabulário, o FinBERT aprende as representações semânticas corretas para o jargão financeiro. Ele aprende que "yield" (retorno) e "spread" (diferencial) são conceitos-chave, que "bullish" é fortemente positivo e "bearish" é fortemente negativo, e que a palavra "guidance" (projeção da empresa) tem um peso significativo para o futuro da ação. Essa adaptação de domínio torna o modelo muito mais sensível ao contexto específico do setor.

O resultado é um modelo que, ao ser subsequentemente submetido ao fine-tuning (ajuste fino) em uma tarefa financeira específica, como a classificação de sentimento, tende a superar os modelos generalistas. Ele parte de uma base de conhecimento muito mais relevante. No contexto deste TCC, o FinBERT representa a hipótese do "especialista": a ideia de que, para tarefas de nicho, um modelo treinado especificamente para esse nicho sempre terá uma vantagem de desempenho. Compará-lo com o RoBERTa e outros testará a validade dessa hipótese, na Tabela 2 podemos verificar suas principais características.

Tabela 2 - Principais características do modelo FinBERT

Atributo	Descrição Detalhada
Pilar Tecnológico	Especialista: Focado em profundidade de domínio.
Base Arquitetural	BERT (Bidirectional Encoder Representations from Transformers).
Treinamento	<i>Domain-adaptive pre-training</i> em corpora da Bloomberg, Reuters e relatórios 10-K.
Diferencial	Sensibilidade ao jargão financeiro e polaridade de termos ambíguos (ex: "yield", "spread").
Ponto Fraco	Custo computacional elevado e menor desempenho em textos fora do nicho financeiro.

Fonte: Elaborado pelo autor (2026).

2.5 Modelo Robert

O RoBERTa (A Robustly Optimized BERT Pretraining Approach) (LIU et al., 2019) representa a filosofia oposta à do FinBERT: em vez de focar na especialização de domínio, ele se concentra em aperfeiçoar o modelo de propósito geral. Quando o RoBERTa foi lançado por pesquisadores do Facebook AI, sua premissa era que o BERT original, embora revolucionário, havia sido significativamente sub-otimizado. Eles argumentaram que o desempenho do BERT poderia ser drasticamente melhorado apenas com uma "receita" de pré-treinamento mais robusta.

O RoBERTa manteve a arquitetura base do BERT, mas implementou quatro mudanças cruciais no pré-treinamento. Primeiro, ele foi treinado com um volume de dados muito maior (160GB de texto, contra 16GB do BERT) e por um período significativamente mais longo. Segundo ele utilizou batch sizes (tamanho dos lotes de treinamento) muito maiores, o que ajuda a estabilizar o treinamento e a obter melhores resultados. Terceiro, ele introduziu a máscara dinâmica, onde as palavras "mascaradas" para o modelo prever eram geradas de forma diferente em cada epoch (ciclo de treinamento), tornando a tarefa de aprendizado mais difícil e o modelo final mais robusto.

A quarta e talvez mais importante mudança foi a remoção da tarefa de Next Sentence Prediction (NSP). O BERT original era treinado tanto para prever palavras mascaradas quanto para prever se duas frases eram sequenciais. A equipe do RoBERTa descobriu que essa tarefa de NSP era, na verdade, prejudicial, e a removeu, focando apenas na tarefa de Masked Language

Modeling (MLM). O resultado foi um modelo generalista que, no momento de seu lançamento, superou o BERT e outros modelos em quase todas as tarefas de benchmark do PLN.

Para este TCC, o RoBERTa representa o "generalista de alta performance". Ele levanta uma questão de pesquisa fundamental: um modelo generalista, se treinado de forma massiva e otimizada, adquire uma compreensão de linguagem tão profunda que consegue se adaptar a um domínio específico (como finanças) durante o fine-tuning e superar um modelo que passou por um pré-treinamento de domínio como o FinBERT? Na Tabela 3 podemos verificar suas principais características.

Tabela 3 - Principais características do modelo RoBERTa

Atributo	Descrição Detalhada
Pilar Tecnológico	Robusto: Otimização máxima do pré-treinamento generalista.
Base Arquitetural	Remoção da tarefa <i>Next Sentence Prediction</i> (NSP) e uso de <i>Dynamic Masking</i> .
Treinamento	160GB de texto (10x mais que o BERT original).
Diferencial	Alta capacidade de generalização e compreensão sintática profunda.
Ponto Fraco	Altíssima latência e necessidade de hardware potente (GPUs) para produção.

Fonte: Elaborado pelo autor (2026).

2.6 Modelo Fasttext

O FastText (BOJANOWSKI et al., 2017) é um modelo de uma família completamente diferente das arquiteturas Transformer. Desenvolvido também pelo Facebook AI, o FastText não é uma rede neural "profunda" como o BERT, mas sim uma arquitetura de rede neural rasa, sendo uma extensão direta do modelo Word2Vec (MIKOLOV et al., 2013). Seus principais objetivos de design não são a compreensão contextual profunda, mas sim a velocidade, eficiência e a capacidade de lidar com morfologias de linguagem complexas.

A inovação central do FastText é a forma como ele representa as palavras. Enquanto modelos como o BERT tratam cada palavra como um "token" único, o FastText trata cada palavra como um agregado de seus n-gramas de caracteres. Por exemplo, a palavra "mercado" (com $n = 3$) seria representada por um conjunto de sub-palavras como <me, mer, erc, rca, cad, ado, do>, além da palavra inteira <mercado>. O vetor final da palavra "mercado" é então a soma dos vetores de todos esses pedaços.

Essa abordagem de sub-palavra tem duas vantagens colossais. A primeira é a capacidade de lidar com palavras fora do vocabulário (OOV - Out-of-Vocabulary). Se um modelo BERT nunca viu a palavra "desinvestimento" no treinamento, ele não saberá o que fazer com ela. O FastText, no entanto, pode inferir seu significado com base em suas partes ("des", "invest", "mento"), que provavelmente já viu em outras palavras. Isso é especialmente poderoso para línguas morfológicamente ricas como o português, que possui muitas conjugações verbais e sufixos.

A segunda vantagem é a eficiência computacional. A arquitetura do FastText é extremamente simples, permitindo que ele seja treinado em conjuntos de dados massivos em questão de minutos ou horas em uma CPU comum, um feito impossível para os modelos Transformer, que exigem GPUs potentes e dias de treinamento. No contexto deste TCC, o FastText serve como o benchmark de eficiência. Ele estabelece a linha de base de desempenho: qual a performance de um modelo leve e rápido? Ele responderá à pergunta de trade-off: o ganho de acurácia obtido pelos modelos gigantes FinBERT, RoBERTa justifica seu custo computacional exponencialmente maior? Na Tabela 4 apresenta suas principais características.

Tabela 4 - Principais características do modelo FastText

Atributo	Descrição Detalhada
Pilar Tecnológico	Baseline: Referência de velocidade e simplicidade.
Base Arquitetural	Rede neural rasa baseada em Word2Vec (n-gramas de caracteres).
Treinamento	Trata palavras como somas de sub-palavras (ex: "mercado" = "mer" + "erc" + "ado").
Diferencial	Extremamente rápido em CPU e lida nativamente com palavras fora do vocabulário (OOV).
Ponto Fraco	Não captura o contexto (a ordem das palavras não altera o vetor da frase).

Fonte: Elaborado pelo autor (2026).

2.7 Modelo Modernbert

O termo ModernBERT, um neologismo funcional para este TCC, é usado para categorizar uma família de modelos baseados em Transformers que surgiram após o BERT, com o objetivo principal de resolver seu maior problema: a eficiência computacional. O BERT-Base possui 110 milhões de parâmetros e o BERT-Large, 340 milhões. Modelos como o RoBERTa-Large são igualmente massivos. Embora seu desempenho seja impressionante, seu tamanho os torna lentos e caros para treinar e, crucialmente, para executar em produção (um processo chamado inferência). Em um cenário financeiro real que exige a análise de milhares de notícias em tempo real, a latência do BERT pode ser um gargalo proibitivo.

O campo da compressão de modelos surgiu para enfrentar esse desafio. O objetivo é criar modelos menores e mais rápidos que retenham a maior parte (idealmente 95-99%) do desempenho dos modelos grandes. A técnica mais popular e bem-sucedida para isso é a Destilação de Conhecimento (Knowledge Distillation) (HINTON et al., 2015). Esse processo funciona como uma relação "professor-aluno": um modelo grande e de alto desempenho (o "professor", ex: RoBERTa) é usado para treinar um modelo muito menor (o "aluno"). O aluno é treinado não apenas para acertar o rótulo final (ex: "positivo"), mas para imitar a distribuição de probabilidade completa do professor (ex: 90% positivo, 8% neutro, 2% negativo). Ao aprender essas "respostas suaves", o aluno absorve as nuances e o "conhecimento obscuro" do professor.

O DistilBERT (SANH et al., 2019) é o exemplo mais famoso dessa abordagem. É uma versão do BERT que é 40% menor em tamanho e 60% mais rápida em inferência, enquanto retém 97% do desempenho do BERT original. Outras técnicas incluem o ALBERT (LAN et al., 2020), que usa compartilhamento de parâmetros entre as camadas para reduzir drasticamente o tamanho do modelo, e o pruning (poda), que remove seletivamente neurônios e conexões redundantes de um modelo treinado.

No contexto deste TCC, a categoria "ModernBERT" representa o equilíbrio de custo-benefício. Ela responde à pergunta mais prática e voltada para a engenharia de software: "Podemos alcançar um desempenho suficientemente bom (talvez 1-2 pontos de F1-Score abaixo do melhor modelo) por uma fração do custo computacional?" Para uma implementação em larga escala, este modelo pode ser o verdadeiro vencedor prático, mesmo que não seja o campeão acadêmico em acurácia pura. Na Tabela 5 apresenta suas principais características.

Tabela 5 - Principais características do modelo FastText

Atributo	Descrição Detalhada
Pilar Tecnológico	Eficiente: Otimizado para baixa latência e contextos longos.
Base Arquitetural	<i>Knowledge Distillation</i> (Destilação de Conhecimento) e inovações como GeGLU e RoPE.
Treinamento	Até 8.192 tokens (capaz de ler relatórios financeiros inteiros).
Diferencial	Velocidade de inferência e capacidade de processar documentos longos sem truncamento.
Ponto Fraco	Pode ter uma leve perda de precisão (1-2% no F1-Score) comparado aos modelos "Large".

Fonte: Elaborado pelo autor (2026).

3 ABORDAGEM E PROPOSTA

3.1 Pipeline

3.1.1 Descrição Da Abordagem

A fundamentação experimental deste trabalho baseia-se em um corpus de textos do domínio financeiro, selecionado por sua metodologia de rotulagem baseada em evidências externas. Diferente de abordagens baseadas em anotação manual que podem herdar vieses e subjetividades dos anotadores, este conjunto de dados utiliza indicadores de mercado como critério para determinar a polaridade dos textos. Essa abordagem visa alinhar a interpretação semântica dos modelos às reações práticas do ecossistema econômico, estabelecendo um desafio de classificação mais rigoroso para as arquiteturas de Processamento de Linguagem Natural.

O corpus é composto por títulos de notícias e comunicações corporativas, estruturados para suportar uma tarefa de classificação multiclasse (positivo, negativo e neutro). Do ponto de vista estatístico, a base de dados reflete a distribuição natural de sentimentos encontrados no fluxo de informações financeiras, apresentando uma cobertura abrangente de todas as polaridades de interesse para a análise de mercado.

A estrutura do conjunto de dados foi analisada sob a ótica da representatividade. Procurando um equilíbrio entre as classes que permite a convergência dos modelos sem a necessidade de intervenções artificiais no balanço das amostras. Essa característica é fundamental para que o pipeline desenvolvido mantenha sua integridade técnica, permitindo que a solução seja replicável em outros subdomínios financeiros ou aplicada em sistemas de análise em tempo real.

Dada a natureza do problema, a avaliação de desempenho ignora métricas superficiais de acurácia global em favor de métricas que consideram o desempenho individualizado de cada classe, sendo feito o uso de F1-Score Macro para uma avaliação justa. Isso assegura que o sistema seja capaz de identificar sinais críticos de mercado (notícias positivas ou negativas) com a mesma precisão com que processa comunicações neutras, garantindo a robustez necessária para o suporte à decisão em ambientes de alta volatilidade.

3.1.2 Pipeline De Desenvolvimento (Fluxo De Trabalho)

O pipeline foi projetado para servir como uma base sólida para ferramentas de apoio à decisão e pesquisas futuras em finanças quantitativas.

Fase 1: Preparação e Saneamento de Dados

Ingestão: Carregamento via Python/Pandas a partir do repositório no Google Drive.

Normalização Semântica: Mapeamento dos rótulos originais para o formato numérico $\{0, 1, 2\}$, compatível com a função de perda Cross-Entropy utilizada em redes neurais.

Tratamento de Nulos: Filtragem rigorosa para garantir que nenhum título vazio ou rótulo inconsistente seja processado.

Fase 2: Estratégia de Particionamento (Experimental Design)

Para garantir a confiabilidade estatística e a capacidade de generalização das arquiteturas testadas, o corpus foi segmentado em três subconjuntos distintos, seguindo uma distribuição proporcional padrão na literatura de Aprendizado de Máquina:

Conjunto de Treinamento (80%): Destinado ao ajuste dos pesos e parâmetros internos dos modelos, permitindo a captura dos padrões semânticos e léxicos presentes nos dados.

Conjunto de Validação (10%): Utilizado durante a fase de treinamento para o monitoramento de overfitting (sobreajuste) e a seleção do melhor checkpoint do modelo, garantindo que o aprendizado seja generalizável.

Conjunto de Teste (10%): Composto por dados completamente isolados e inéditos para o modelo. Este conjunto serve como o benchmarking final, fornecendo as métricas de desempenho que simulam a operação do sistema em um cenário de produção real.

A divisão dos subconjuntos foi realizada utilizando a técnica de Estratificação. Este procedimento garante que a proporção relativa entre as diferentes classes de sentimento (positivo, negativo e neutro) permaneça idêntica em todos os conjuntos (treino, validação e teste). Dessa forma, evita-se a introdução de vieses durante o aprendizado e assegura-se que a avaliação final reflita com precisão o comportamento dos modelos diante da distribuição natural dos dados.

Fase 3: Implementação das Arquiteturas Comparativas

Para cumprir o objetivo de criar uma base para pesquisas futuras, foram selecionados quatro pilares tecnológicos distintos:

1. **Especialista (FinBERT):** Para avaliar o ganho de desempenho trazido pelo pré-treinamento em domínio financeiro.
2. **Robusto (RoBERTa):** Para testar os limites de um modelo generalista de alta performance.
3. **Eficiente (ModernBERT):** Focado em otimização para ferramentas de tempo real com baixa latência.
4. **Baseline (FastText):** Para verificar se a complexidade dos Transformers é justificada perante um modelo estatístico rápido.

Fase 4: Avaliação Multicritério

A avaliação dos modelos transcende a métrica de acurácia global, adotando uma abordagem multidimensional para garantir que os resultados sejam estatisticamente confiáveis e tecnicamente viáveis para aplicações reais. Os métodos de avaliação são divididos em três eixos:

3.1.3 Métricas De Eficácia Preditiva

Considerando um leve desbalanceamento, a performance será mensurada através de métricas que consideram o desempenho individual de cada classe:

- **Precisão, Recall e F1-Score (por classe):** Essenciais para quantificar o custo da classificação errônea. No setor financeiro, a **Precisão** na classe "Negativo" é vital para evitar alarmes falsos, enquanto o **Recall** na mesma classe é crítico para a mitigação de riscos. Em finanças, errar uma notícia negativa (falso negativo) é muito mais caro do que errar uma notícia neutra.
- **F1-Score Macro:** Utilizado como a métrica principal de ranqueamento, pois atribui peso igual a todas as classes, independentemente da frequência, penalizando modelos que negligenciam a classe minoritária ("Neutro").
- **Acurácia Balanceada:** Calculada para fornecer uma visão média da sensibilidade do modelo, corrigindo distorções causadas pela predominância da classe "Positivo".

Matriz de Confusão e Análise de Erro

- **Visualização da Matriz de Confusão:** Ferramenta central para identificar "zonas de confusão". Analisaremos especificamente a transição entre Neutro -> Negativo e Neutro -> Positivo, visando entender o comportamento do modelo diante de notícias de baixa intensidade emocional.
- **Análise Qualitativa de Missclassifications:** Seleção de amostras onde todos os modelos falharam para investigar ambiguidades semânticas inerentes ao dataset FinMarBa (Lefort et al. 2025).

Eficiência Computacional e Escalabilidade (Footprint)

Para que os modelos sirvam de base para ferramentas futuras, a eficiência é um critério de desempate:

- **Latência e Tempo de Inferência:** Medição do tempo médio de resposta por amostra (ms/sample) e *throughput* (amostras por segundo). Esta métrica define a viabilidade de uso em sistemas de *High-Frequency Trading* (HFT).
- **Tamanho do Modelo (Model Footprint):** Comparação do número de parâmetros treináveis e do espaço em disco ocupado pelos pesos (MB). Avalia-se aqui o custo de armazenamento e a viabilidade de deploy em infraestruturas limitadas ou dispositivos de borda.
- **Curvas de Aprendizado (Loss vs. Epochs):** Comparação da velocidade de convergência e estabilidade entre os modelos Transformer (FinBERT, RoBERTa, ModernBERT) e o modelo estatístico (FastText).

Análise de Trade-off (Eficiência vs. Eficácia)

Será aplicado um índice de custo-benefício para identificar qual arquitetura oferece o melhor equilíbrio. O objetivo é responder se o ganho marginal de performance dos modelos mais complexos (ex: FinBERT) justifica o aumento exponencial no tempo de processamento em relação a modelos otimizados (ex: ModernBERT ou FastText).

A escolha do dataset FinMarBa aliada à comparação entre modelos de pesos pesados (RoBERTa) e modelos otimizados (ModernBERT) visa fornecer um guia de escolha tecnológica. Pesquisas futuras poderão utilizar os pesos salvos deste trabalho para realizar o Transfer Learning em datasets de nicho ou integrar os modelos aqui validados em pipelines de análise de risco automatizada.

3.2 Técnica Do Processamento

A etapa de processamento é fundamental para garantir a qualidade dos sinais textuais que serão ingeridos pelas arquiteturas de Aprendizado de Máquina. Dado que este trabalho compara modelos de diferentes famílias tecnológicas desde redes neurais rasas até Transformers de última geração, o pipeline de processamento foi desenhado para ser modular, aplicando tratamentos específicos de acordo com a necessidade de representação de cada modelo.

3.2.1 Fase 1: Baseline (V1)

Nesta fase, os modelos foram treinados utilizando a função de perda padrão Cross-Entropy Loss, sem qualquer mecanismo de compensação para o desequilíbrio de classes. O objetivo foi estabelecer uma linha de base de desempenho e quantificar o viés dos modelos em relação as classes majoritárias (Positivo e Negativo).

3.2.2 Fase 2: Ponderação De Classes (V2)

Para mitigar a tendência observado na Fase 1, implementou-se a Weighted CrossEntropy Loss. Os pesos foram calculados utilizando a técnica de Inverse Class Frequency (Frequencia Inversa de Classe) a partir do conjunto de treinamento, atribuindo penalidades maiores aos erros cometidos na classe minoritária (Neutro).

$$w_c = \frac{N_{total}}{N_{classes} \times N_c} \quad (3.1)$$

Onde w_c é o peso da classe, N_{total} é o número total de amostras e N_c é o número de amostras da classe c .

3.2.3 Fase 3: Refinamento (V3)

A fase final introduziu duas técnicas avançadas para lidar com exemplos difíceis e rótulos ruidosos:

- Focal Loss: Uma função de perda que reduz o peso de exemplos fáceis (bem classificados) e foca o treinamento em exemplos difíceis. Utilizou-se um fator de foco $\gamma = 2.0$.
- Label Smoothing: Técnica de regularização com fator $\alpha = 0.1$, que substitui os rótulos one-hot puros por distribuições suavizadas, evitando que o modelo se torne superconfiante em anotações que podem conter ruído.

Adicionalmente, foi implementado um método de Ensemble por votação majoritária (Hard Voting) combinando as previsões dos três modelos baseados em Transformers.

3.4 Métricas de Avaliação

Dada a natureza desbalanceada do problema, a métrica principal de avaliação foi o Macro F1-Score, que calcula a média harmônica entre precisão e recall para cada classe individualmente e depois tira a média aritmética, garantindo que o desempenho na classe minoritária tenha o mesmo peso que nas classes majoritárias. Acurácia global, Precisão e Recall por classe também foram reportados para análise detalhada.

3.2.4 Saneamento E Normalização De Texto

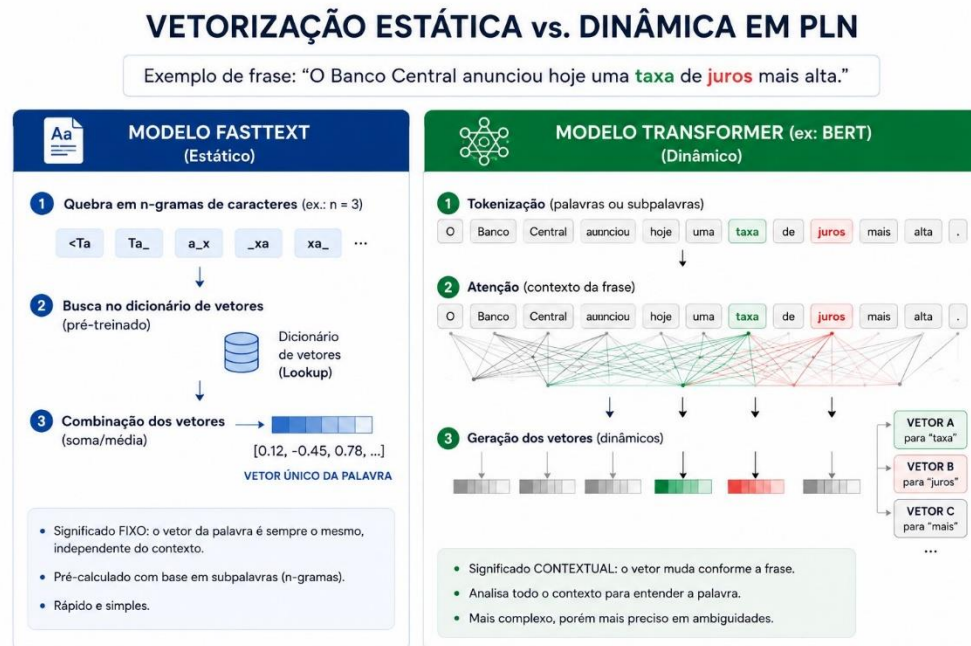
O primeiro estágio do processamento consiste na padronização dos caracteres e na remoção de ruídos que não contribuem para a carga semântica do sentimento financeiro, sem comprometer a estrutura gramatical. Este procedimento inclui: Normalização de Case: Conversão do texto para minúsculas (em modelos uncased) para reduzir a esparsidade do vocabulário.

- Tratamento de Caracteres Especiais: Remoção de pontuações excessivas e caracteres não alfanuméricos, preservando símbolos que carregam significado no domínio financeiro, como sinais de porcentagem (%) e indicadores de moedas.
- Manutenção de Stopwords: Diferente das técnicas clássicas de PLN, optou-se pela não retirada de stopwords (palavras de ligação como "e", "de", "com", "não"). Esta decisão justifica-se pelo fato de que os modelos baseados em Transformer utilizam mecanismos de auto-atenção para capturar a dependência de longo alcance e a sintaxe da frase. A remoção dessas palavras poderia fragmentar o contexto e prejudicar a interpretação de nuances, como negações e relações de causa e efeito, essenciais na linguagem financeira.

3.2.5 Tokenização E Representação Numérica

A tokenização é o processo de "fatiar" o texto em unidades menores, chamadas tokens, para que o computador possa processá-las. Como modelos de Inteligência Artificial não compreendem letras ou palavras como os humanos, cada um desses fragmentos é convertido em um código numérico único. Uma ilustração pode ser vista na Figura 2.

Figura 2 - Comparativo entre os modelos de vetorização FastText (estático e local) e Transformer/BERT (dinâmico e contextual). O infográfico ilustra como o FastText utiliza n-grams de caracteres para gerar vetores fixos para cada palavra, enquanto o Transformer emprega o mecanismo de atenção para criar vetores dinâmicos adaptados ao contexto da frase inteira.



Fonte: Elaboração própria com auxílio do ChatGPT (OpenAI) na geração e refinamento da representação visual. Conteúdo técnico baseado em Joulin et al. (2017) para FastText e Devlin et al. (2019) para BERT.

Neste trabalho, utilizamos duas abordagens principais, adaptadas às necessidades de cada arquitetura:

3.2.5.1 Abordagem de Sub-palavras, utilizada pelos modelos FinBERT, RoBERTa e ModernBERT

Esta técnica funciona como um jogo de blocos de montar. Em vez de tentar entender uma palavra inteira de uma vez, o modelo a divide em pedaços menores (prefixos, radicais e sufixos) que ele já conhece. Isso é extremamente útil no mercado financeiro, que utiliza termos técnicos complexos e neologismos.

Como funciona: Imagine a palavra técnica "desinvestimento". Se o modelo nunca viu essa palavra inteira antes, ele teria dificuldade em entender. Com a tokenização de sub-palavras, ele a divide em pedaços como des + invest + imento. A vantagem: O modelo entende que des indica negação ou reversão, invest está ligado a investimento e imento indica uma ação ou resultado. Mesmo diante de uma palavra nova, o modelo consegue "deduzir" o significado unindo o conhecimento desses fragmentos.

Exemplo prático: A frase "O mercado reagiu" poderia ser quebrada em [O, merc, ado, reag, iu]. Cada pedaço recebe um número de identificação em um dicionário imenso, transformando a frase em uma lista de números, como [101, 432, 89, 1205, 77].

3.2.5.1 Abordagem de N-gramas de Caracteres, utilizada pelo modelo FastText

Esta técnica é mais simples e rápida, focando na estrutura física das palavras (as letras que as compõem) em vez de focar apenas no significado. Ela "desliza" uma janela sobre a palavra para capturar todas as combinações de letras possíveis.

Como funciona: Se definirmos que queremos grupos de 3 letras (tri-gramas), a palavra "lucro" seria quebrada em fragmentos sobrepostos: <lu, luc, ucr, cro, ro>. A vantagem: Esta técnica é excelente para lidar com erros de digitação ou variações de escrita. Se alguém escrever "lucros" ou "lucrativo", o modelo ainda reconhecerá os fragmentos luc, ucr e cro, mantendo a conexão com o conceito original de lucro. É como se o modelo aprendesse a "raiz" da palavra observando as combinações de letras que mais se repetem.

Exemplo prático: Ao analisar o termo financeiro "dividendos", o FastText cria uma "assinatura" baseada em dezenas de pequenos pedaços de letras. Isso permite que ele seja muito veloz na comparação de textos, pois ele não precisa fazer uma análise gramatical profunda, apenas identificar padrões de repetição de caracteres.

3.2.3 Vetorização e Enquadramento

Para os modelos de Deep Learning, o texto processado é convertido em sequências de identificadores (IDs) de tamanho fixo. Aplica-se a técnica de padding (preenchimento) para sentenças curtas e truncation (truncamento) para sentenças que excedam o limite da janela de contexto. No caso do ModernBERT, o processamento tira proveito de uma janela de atenção expandida, permitindo o tratamento de fragmentos textuais mais extensos sem a perda de informações contextuais críticas que ocorre em modelos com limites mais restritos.

3.3 Modelos Machine Learning - Penalidades

A avaliação de modelos de inteligência artificial em ambientes de alta criticidade, como o mercado financeiro, exige um olhar que vai além da simples contagem de acertos. Nesta subseção, detalha-se a lógica de penalização aplicada aos modelos e as métricas escolhidas para garantir que o sistema não apenas acerte a maioria dos casos, mas que evite erros que seriam financeiramente catastróficos.

3.4.1 A Falácia da Acurácia Simples

Em problemas de classificação de sentimento, a "Acurácia" (o percentual total de acertos) pode ser uma métrica enganosa. Se um conjunto de dados possui muitas notícias neutras, um modelo "preguiçoso" poderia classificar tudo como neutro e ainda assim obter uma pontuação alta de acurácia, embora seja completamente inútil para identificar oportunidades de compra ou venda.

Para evitar esse viés, este trabalho aplica penalidades mais severas a modelos que ignoram as classes minoritárias ou que confundem polaridades opostas (como classificar uma notícia "Bearish/Negativa" como "Bullish/Positiva").

3.4.2 Matriz de Confusão e Custos de Erro

A performance é monitorada através da Matriz de Confusão, que permite visualizar exatamente onde o modelo está "se confundindo". No contexto desta pesquisa, os erros são interpretados sob duas óticas de penalidade:

Falsos Positivos (Erro Tipo I): Ocorrem quando o modelo sinaliza uma oportunidade otimista em um cenário que era, na verdade, negativo ou neutro. A penalidade aqui reflete o risco de um investidor alocar capital baseado em um sinal falso.

Falsos Negativos (Erro Tipo II): Ocorrem quando o modelo falha em detectar um sentimento negativo crítico. Em finanças, este erro é frequentemente considerado mais grave, pois a incapacidade de prever uma queda pode resultar em perdas maiores do que a simples perda de uma oportunidade de ganho.

3.4.3 Critério de Desempate: F1-Score Macro

Para unificar a precisão (o quanto o modelo é confiável quando aponta algo) e a sensibilidade (o quanto o modelo consegue encontrar todos os casos reais), utiliza-se o F1-Score. Especificamente, adotamos o F1-Score Macro, que trata as três classes (Positivo, Negativo e Neutro) com o mesmo peso de importância. Esta métrica penaliza severamente qualquer modelo que tente "especializar-se" apenas na classe mais frequente, exigindo que a arquitetura seja robusta o suficiente para decifrar a ambiguidade em todas as frentes de sentimento do mercado.

4 EXPERIMENTOS

4.1 Setup Dos Experimentos

Para viabilizar o treinamento e a inferência das arquiteturas de Deep Learning selecionadas, particularmente os modelos baseados em Transformers (FinBERT, RoBERTa e ModernBERT), utilizou-se a infraestrutura de computação em nuvem do Google Colab, com suporte de aceleração gráfica via GPU.

A unidade de processamento utilizada foi uma NVIDIA Tesla T4 (NVIDIA T4). Esta GPU é fundamentada na arquitetura Turing e conta com 16 GB de memória de vídeo (VRAM). A escolha deste hardware justifica-se pela sua eficiência em tarefas de processamento de linguagem natural (NLP) de médio e grande porte, permitindo a manipulação de tensores complexos e a execução de mecanismos de autoatenção com baixa latência. Além disso, a capacidade de memória da T4 possibilita o uso de batch sizes adequados para a estabilização do gradiente durante o processo de fine-tuning, garantindo a reprodutibilidade e a eficiência computacional exigidas por este estudo.

Os hiperparâmetros foram ajustados empiricamente ao longo das fases:

- Learning Rate: 2×10^{-5} para FinBERT e RoBERTa; 5×10^{-5} para ModernBERT.
- Batch Size: 32 para todos os modelos.
- Epochs: 3 a 5 épocas, com critério de Early Stopping (paciência = 3) monitorando a perda de validação.
- Otimizador: AdamW com weight decay de 0.05.

Os hiperparâmetros são como as configurações que ajustamos em uma máquina para que ela aprenda da melhor maneira possível. Imagine que o modelo está tentando descer uma montanha no escuro para encontrar o ponto mais baixo (o menor erro). A Learning Rate é o tamanho do passo que ele dá. Se o passo for muito grande, ele pode pular o buraco e nunca chegar ao fundo. Se for muito pequeno (como os valores 2×10^{-5} , que são números bem baixinhos), ele demora muito, mas vai com muito cuidado para não errar o caminho. No caso de estudo, foram feitos ajustes de passos ligeiramente diferentes para cada modelo para que eles encontrem o equilíbrio ideal sem se perderem.

Agora imagine que um estudante precisa ler 8.000 notícias para uma prova. O Batch Size é quantas notícias ele lê de uma vez antes de parar para refletir e ajustar o que aprendeu.

Com um tamanho de 32, o modelo lê um pequeno grupo de 32 notícias, pensa sobre elas, ajusta seu conhecimento e só então passa para as próximas 32. Isso evita que ele fique sobrecarregado tentando decorar tudo de uma vez.

Já uma Época é uma leitura completa de todo o conjunto de dados. É como ler o livro inteiro uma vez, no qual a definição de 3 a 5 épocas significa que o modelo "leu o livro" de 3 a 5 vezes. E o Early Stopping (Parada Antecipada) é quando um estudante percebe que, a partir da 3ª leitura, ele não está mais aprendendo nada novo e está começando a "decorar" as frases em vez de entender o sentido (isso se chama overfitting). Desse modo, se o professor observar que o aluno não melhorou o desempenho, o professor interrompe o estudo para não "viciar" nos dados.

O Otimizador é o "instrutor" que diz ao modelo como ele deve ajustar seus conhecimentos após cada erro. O AdamW é um dos instrutores mais inteligentes atualmente, visto que ele ajusta a velocidade de aprendizado conforme o modelo ganha confiança. E o Weight Decay (Decaimento de Peso) é uma técnica para manter o modelo "simples". Ele penaliza conhecimentos muito complexos ou exagerados, forçando o modelo a focar apenas no que é realmente importante para identificar o sentimento da notícia.

Para o modelo FastText (baseline não-neural), utilizou-se treinamento supervisionado com 25 épocas, learning rate de 0.5 e n-gramas de palavra de tamanho 2.

- **Infraestrutura de Hardware:** Utilizou-se instâncias equipadas com **GPUs (Graphics Processing Units)**, fundamentais para viabilizar o processamento paralelo exigido pelos mecanismos de autoatenção das arquiteturas baseadas em Transformers (FinBERT, RoBERTa e ModernBERT).
- **Ecossistema de Software:** O ambiente foi configurado em linguagem **Python**, utilizando o framework *Hugging Face Transformers* para a manipulação dos modelos de linguagem. Para o modelo de base, utilizou-se a biblioteca *FastText*. O suporte para manipulação de dados e cálculos estatísticos foi provido pelas bibliotecas *Pandas*, *NumPy* e *Scikit-Learn*.
- **Parâmetros de Controle:** O treinamento foi padronizado com o uso de funções de perda *Cross-Entropy* e otimizadores de última geração, garantindo que as comparações entre as diferentes arquiteturas fossem realizadas sob condições de contorno equivalentes.

4.2 Dataset Original

O conjunto de dados utilizado como base para os experimentos deste trabalho consiste em um recorte do corpus proposto no estudo "FinMarBa: A Market-Informed Dataset for Financial Sentiment Classification" (Lefort et al. 2025). A escolha desta base fundamenta-se na sua metodologia de construção, que se diferencia dos repositórios convencionais de Processamento de Linguagem Natural por integrar o comportamento real do mercado financeiro ao processo de rotulagem. As principais características do conjunto de dados aplicado nesta pesquisa são:

- **Recorte Temporal e Volumetria:** O dataset compreende um período histórico específico, entre os anos de 2010 e 2011, totalizando 8.143 registros de notícias financeiras. Este intervalo é particularmente relevante por capturar a dinâmica de mercado em um período de consolidação de dados digitais e volatilidade econômica, oferecendo uma densidade amostral robusta para o treinamento de modelos de linguagem.
- **Rotulagem Baseada em Sinais de Mercado:** Ao contrário da rotulagem manual subjetiva, a polaridade dos textos (positivo, negativo ou neutro) foi validada por meio de indicadores quantitativos de desempenho de ativos financeiros no momento da publicação das notícias. Isso garante que os rótulos reflitam a reação econômica real, e não apenas a percepção linguística de um anotador humano.
- **Natureza das Amostras:** O corpus é composto por títulos de notícias, caracterizados por uma linguagem densa, uso frequente de jargões técnicos e uma estrutura textual concisa, típica de manchetes que influenciam a tomada de decisão em tempo real.
- **Distribuição e Desbalanceamento Natural:** O dataset preserva a distribuição orgânica dos sentimentos no fluxo de informações do período. A presença de um desbalanceamento natural entre as classes reflete a realidade do setor financeiro e impõe ao modelo o desafio de identificar sinais críticos de mercado em meio a uma distribuição de dados não perfeitamente simétrica.

Para resolver a tarefa de classificação, os autores propuseram o dataset FinMarBa (Lefort et al. 2025), cuja solução central reside na substituição da rotulagem manual subjetiva por uma abordagem informada pelo mercado (market-informed). A metodologia utiliza uma estrutura de integração escalável que correlaciona o conteúdo textual de notícias financeiras com variações reais de preços de ativos em janelas temporais específicas. Através dessa

estratégia, o modelo de rotulagem identifica a polaridade (positivo, negativo ou neutro) com base na reação efetiva do mercado, permitindo que o sistema aprenda padrões semânticos que possuem impacto direto na volatilidade financeira, superando as limitações de inconsistência comuns em anotações feitas por humanos.

Além da inovação na base de dados, a solução técnica apresentada inclui o uso de uma estrutura hierárquica de Aprendizado por Reforço (RL) para aumentar a estabilidade e a precisão das previsões. Os autores integraram modelos de linguagem de estado da arte, como arquiteturas baseadas em Transformers, validados por métricas quantitativas que comparam o desempenho do modelo original com as variantes processadas. Essa combinação de dados informados por sinais de mercado e um pipeline de treinamento otimizado permite que a solução capture nuances contextuais complexas do setor financeiro, oferecendo uma ferramenta mais robusta para a análise de sentimento em ambientes de alta frequência e tomada de decisão algorítmica.

O campo Title contém o texto integral da manchete financeira em inglês. Estes títulos são sentenças curtas e informativas que descrevem eventos do mercado financeiro, como variações de índices, resultados corporativos, decisões de política monetária ou movimentações de preços de commodities. Por exemplo, uma manchete típica do dataset é: “Global Stocks and Commodities Rally on First Trading Day of 2010”. E este campo que serve como entrada textual para os modelos de análise de sentimento avaliados neste trabalho, que pode ser visto também na Figura 3 a seguir.

Figura 3 - Exemplos do dataset utilizados que são compostos pela data, título, tickers, sentimento e pct_change

	date	title	tickers	sentiment	pct_change
0	2010-01-04	Global Stocks and Commodities Rally on First T...	['ACWI']	1	{'ACWI': 0.02057236532425244}
1	2010-01-04	Dollar Slumps Amid Worldwide Manufacturing Imp...	['UUP', 'ACWI']	0	{'UUP': -0.006932462151207686, 'ACWI': 0.02057...
2	2010-01-04	Oil Prices Surge Above \$81 a Barrel Due to U.S...	['CL=F']	1	{'CL=F': 0.027091752889911946}
3	2010-01-04	S&P 500 Sees 1.6 Percent Increase, Hits 15-Mon...	['^GSPC']	1	{'^GSPC': 0.01604341766668682}
4	2010-01-04	Argentina's Merval Index Reaches Record High, ...	['^MERV']	1	{'^MERV': 0.02749172670767197}

Fonte: Elaborado pelo autor (2026).

O campo Tickers identifica os ativos financeiros mencionados ou relacionados a manchete, utilizando seus códigos de negociação padronizados (símbolos ticker). Cada manchete pode estar associada a um ou mais ativos simultaneamente. Os tickers seguem a convenção das bolsas internacionais, onde cada símbolo identifica de forma única um ativo negociado.

Por exemplo, o símbolo “ACWI” refere-se ao iShares MSCI All Country World Index, um fundo negociado em bolsa (ETF) que acompanha o desempenho de ações globais. Outros exemplos incluem:

- Ações individuais: “XOM” (Exxon Mobil), “INTC” (Intel), “AAPL” (Apple);
- Índices de mercado: “^GSPC” (S&P 500), “^DJI” (Dow Jones), “^BVSP” (Bovespa);
- Commodities: “CL=F” (petróleo bruto), “GC=F” (ouro), “HG=F” (cobre);
- Moedas: “UUP” (índice do dólar), “EURUSD=X” (par euro/dólar);
- ETFs: “SPY” (S&P 500 ETF), “EEM” (mercados emergentes), “GLD” (ouro).

O campo Sentiment atribui uma classificação de sentimento para cada ativo financeiro associado a manchete. O sentimento não é determinado por uma análise linguística subjetiva, mas sim mapeado a partir do comportamento real do mercado na data correspondente, sendo esta uma das características distintivas do dataset FinMarBa. Os valores possíveis para cada ativo estão apresentados a seguir:

- 1 (positivo): indica que o ativo apresentou valorização expressiva no pregão, sugerindo que o mercado interpretou favoravelmente o evento descrito na manchete;
- 0 (neutro): indica que o ativo apresentou variação mínima ou insignificante, sugerindo que o evento não teve impacto relevante no preço do ativo;
- -1 (negativo): indica que o ativo apresentou desvalorização expressiva, sugerindo que o mercado interpretou desfavoravelmente o evento descrito.

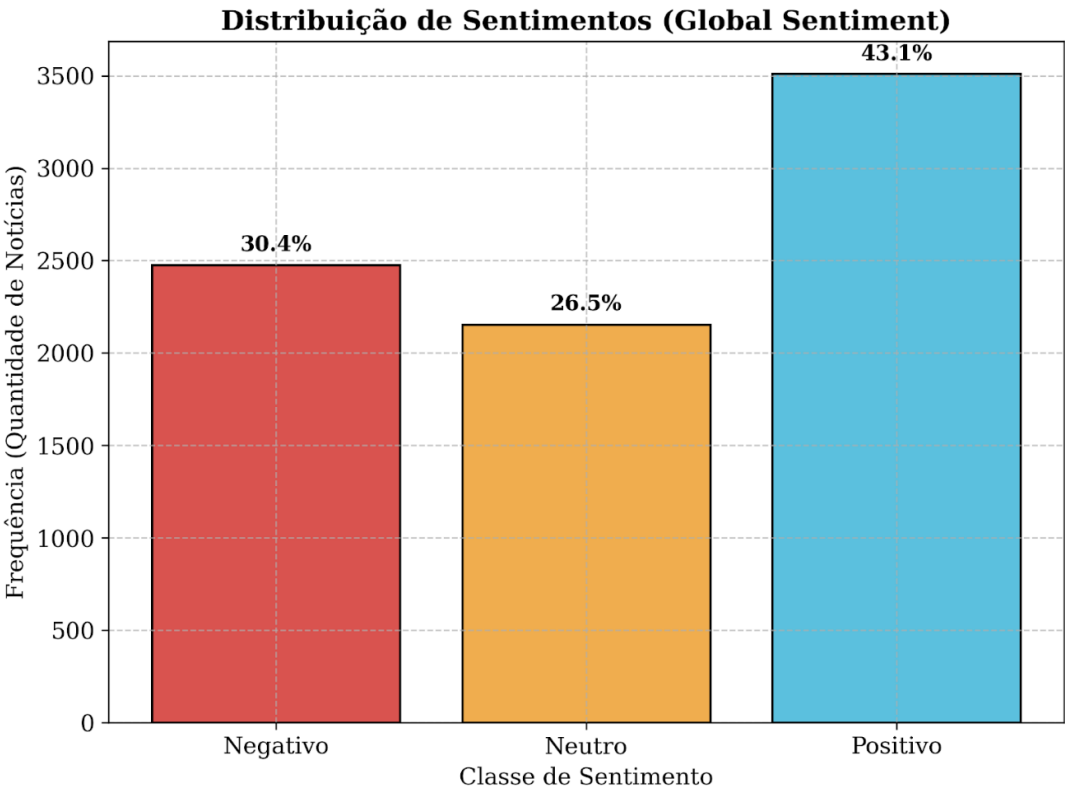
Quando uma manchete menciona múltiplos ativos, cada um recebe sua própria classificação de sentimento. Por exemplo, em uma manchete sobre a relação entre o dólar e mercados globais, o ativo “UUP” (dólar) pode receber sentimento -1 enquanto “ACWI” (ações globais) recebe sentimento 1, refletindo a dinâmica inversa frequentemente observada entre essas classes de ativos.

O campo Pct Change registra a variação percentual do preço de cada ativo financeiro no pregão correspondente a data da manchete. Os valores são expressos como números decimais, onde, por exemplo, 0,0206 representa uma valorização de aproximadamente 2,06%, enquanto -0,0069 indica uma desvalorização de cerca de 0,69%. Este campo é particularmente relevante por duas razões. Primeiro, ele fundamenta a atribuição dos rótulos de sentimento: a classificação de sentimento de cada ativo decorre diretamente da magnitude e da direção de sua variação de preço. Segundo, ele permite análises complementares sobre a intensidade dos movimentos de

mercado, possibilitando investigar não apenas a direção do sentimento (positivo, neutro ou negativo), mas também sua magnitude.

Para compreender a base sobre a qual os modelos foram treinados e testados, é fundamental analisar a distribuição quantitativa das classes de sentimento. A Figura 4 apresenta a frequência relativa e absoluta das categorias positiva, negativa e neutra dentro do corpus original. Observa-se um desbalanceamento natural, comum em dados financeiros, onde as notícias com viés informativo de mercado tendem a se concentrar em polos de reação (positivo e negativo), enquanto a classe neutra representa a menor parcela do conjunto. Essa configuração justifica a adoção do F1-Score Macro como métrica principal, visando garantir que o modelo seja penalizado caso negligencie as classes menos frequentes.

Figura 4 - Distribuição de sentimentos entre positivos, negativos e neutros para o dataset disponibilizado.

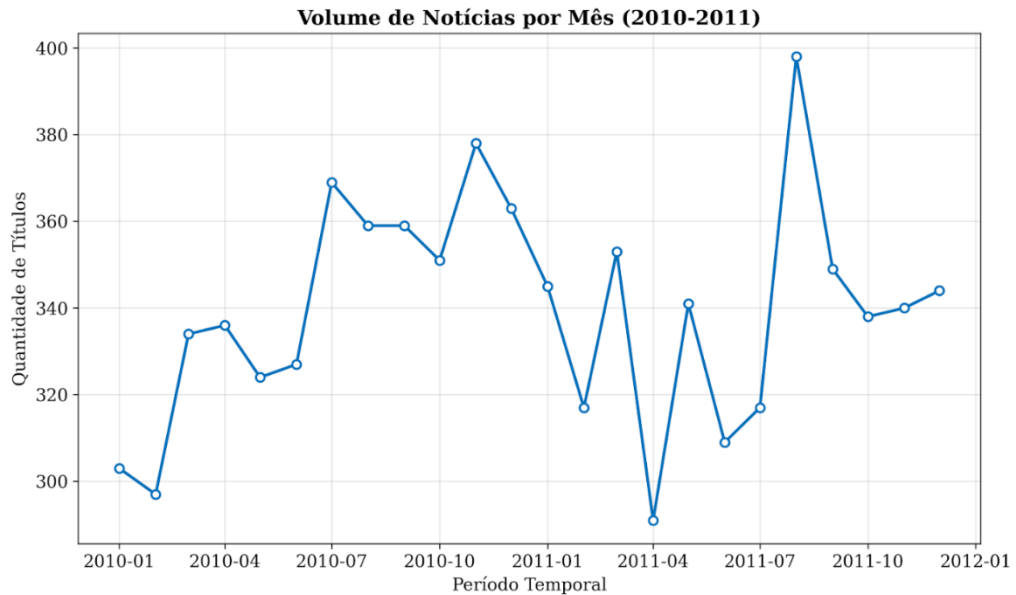


Fonte: Elaborado pelo autor (2026).

A robustez de um modelo de linguagem depende não apenas da quantidade de dados, mas dá continuidade temporal da informação. A Figura 5 ilustra o volume de títulos de notícias processados mês a mês durante o intervalo de 2010 a 2011. A análise do gráfico demonstra uma densidade amostral consistente ao longo de todo o período, assegurando que as arquiteturas de Deep Learning sejam expostas a diferentes contextos e ciclos de notícias. Essa estabilidade no

fluxo de dados é crucial para que o treinamento não seja enviesado por eventos isolados de curto prazo, permitindo uma aprendizagem mais generalista dos padrões semânticos do setor.

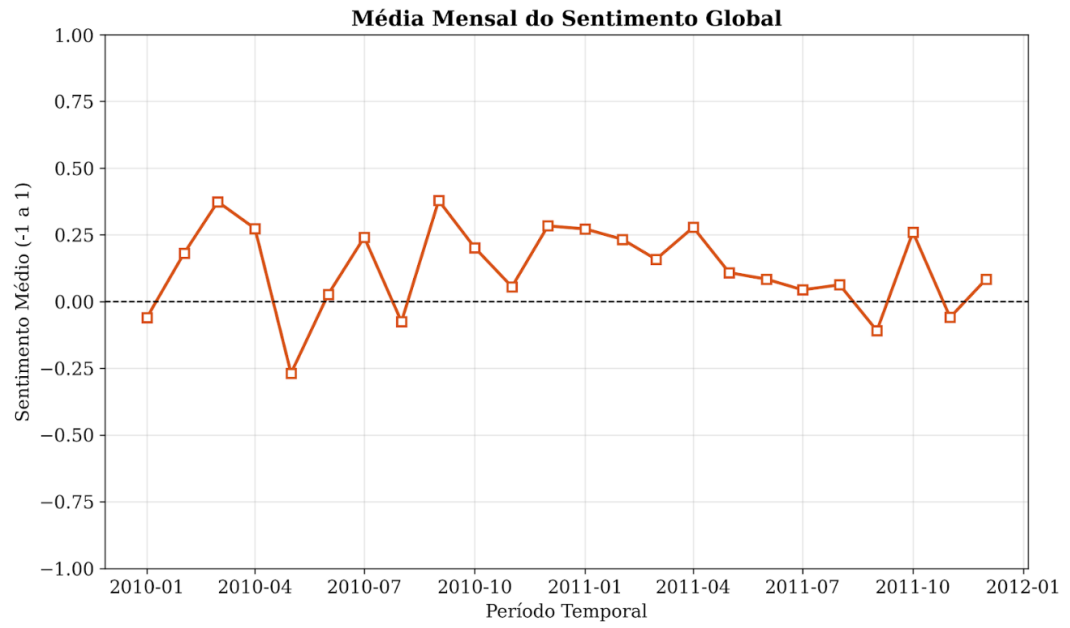
Figura 5 - Volume de Notícias por Mês para o período de 2010 a 2011.



Fonte: Elaborado pelo autor (2026)

Além da volumetria, buscou-se entender a variabilidade do sinal de sentimento contido no dataset. A Figura 6 apresenta a oscilação da média mensal do sentimento global, variando entre o espectro pessimista (-1) e otimista (1). Nota-se que o corpus captura com precisão a volatilidade e as mudanças de humor do mercado financeiro global no biênio estudado. Ter um dataset que transita entre diferentes níveis de polaridade média é essencial para validar se modelos como o FinBERT e o ModernBERT possuem a sensibilidade necessária para distinguir nuances sutis de otimismo e pessimismo em cenários de mercado distintos.

Figura 6 - Distribuição da média mensal do sentimento global de 2010 a 2011.



Fonte: Elaborado pelo autor (2026)

5 RESULTADO DOS EXPERIMENTOS

Este capítulo apresenta a evolução dos experimentos realizados em três fases distintas (Baseline, Ponderação e Refinamento), detalhando o impacto de cada técnica na capacidade dos modelos de lidar com o desequilíbrio de classes e a ambiguidade do sentimento financeiro.

5.1 Visão Geral Dos Resultados

Para avaliar o impacto de diferentes estratégias de otimização no desempenho dos modelos, os experimentos foram estruturados em três versões progressivas, diferenciadas pela abordagem de tratamento do desbalanceamento de classes:

- **Versão 1 (Baseline):** Configuração padrão utilizando a função de perda *Cross-Entropy* convencional, sem ajustes de peso para as classes. Representa o comportamento natural dos modelos diante da distribuição original do dataset.
- **Versão 2 (Weighted):** Introdução de pesos diferenciados na função de perda baseados na Frequência Inversa das Classes (*Inverse Class Frequency*). O objetivo é penalizar mais severamente o erro nas classes minoritárias, forçando o modelo a aumentar sua sensibilidade (*Recall*) para estas categorias.
- **Versão 3 (Focal + Smooth):** Implementação da *Focal Loss* ($\gamma = 2.0$) combinada com *Label Smoothing*. Esta configuração visa focar o aprendizado em exemplos de difícil classificação e suavizar as fronteiras de decisão, reduzindo o excesso de confiança do modelo em previsões polarizadas, onde:
 - $\text{Gamma}(\gamma)$: Definido como 2.0. No código, isso é configurado em `trainer.focal_gamma = 2.0`
 - $\text{Alpha}(\alpha)$: O código utiliza os pesos de classe calculados dinamicamente via frequência inversa (`weights_tensor`). Ele é aplicado como o parâmetro `weight=alpha` na função de perda para compensar o desbalanceamento do dataset.

Para a implementação da Versão 3 (v3) deste estudo, adotou-se a função de perda Focal Loss, conforme proposta originalmente por Lin et al. (2017) no contexto da arquitetura RetinaNet, visando mitigar os efeitos do desequilíbrio de classes e a predominância de exemplos de fácil classificação. O parâmetro de balanceamento $\text{Alpha}(\alpha)$, foi definido de forma dinâmica, atribuindo-se pesos inversamente proporcionais à frequência das classes no dataset FinMarBa, o que permite compensar a predominância de notícias polarizadas e forçar o modelo a aprender as sutilezas da classe neutra. Complementarmente, o hiperparâmetro $\text{Gamma}(\gamma = 2.0)$ foi

mantido em um valor elevado, conforme recomendado pela literatura fundamental de Lin et al. (2017). Matematicamente, a aplicação de um gamma elevado modula a função de perda conforme a Equação a seguir

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

Essa escolha justifica-se pela necessidade de reduzir a contribuição de exemplos de 'fácil classificação' (onde $p_t > 0,9$) e concentrar o gradiente de aprendizagem em casos de alta entropia e incerteza (onde $p_t \approx 0,5$). Tal ajuste foi determinante na Versão 3 para mitigar a generalização excessiva observada no baseline, permitindo que o modelo 'gastasse mais energia' na distinção de manchetes com sobreposição semântica, onde a incerteza de predição era maior."

Matematicamente, essa configuração permite que a função de perda reduza a influência de casos onde o modelo já apresenta alta confiança (como em predições com próximos de 0,97), focando a atualização dos gradientes em instâncias de alta incerteza e entropia (próximos 0,47). Essa abordagem foi fundamental para superar a tendência de "generalização excessiva" observada nas versões anteriores, forçando os modelos a despenderem maior energia na diferenciação de casos incertos. Assim, a escolha desses hiperparâmetros justifica-se pela necessidade de melhorar a detecção de classes limítrofes, garantindo que o aprendizado não fosse dominado por exemplos triviais e resultando em uma fronteira de decisão mais refinada para o contexto do NLP financeiro

O Batch Size foi configurado como 32 para todos os modelos e versões (tanto para treinamento quanto para avaliação), conforme definido em `per_device_train_batch_size=32`. O número de épocas varia conforme a versão do experimento e o modelo: Baseline (v1): Geralmente 3 épocas, sendo:

- Weighted (v2): * FinBERT e RoBERTa: 3 épocas.
- ModernBERT: 5 épocas.
- Focal + Smoothing (v3): * FinBERT e RoBERTa: 4 épocas.
- ModernBERT: 5 épocas.

A Tabela 6 sintetiza o comportamento métrico observado nas arquiteturas ao longo das três fases experimentais (v1, v2 e v3). Os dados apontam para uma possível trajetória de evolução na sensibilidade (Recall) da classe Neutra, a qual exibiu um incremento próximo de 30 pontos percentuais, partindo de um patamar inicial de aproximadamente 15% para o nível

de 45% nas configurações testadas. Contudo, nota-se que este movimento de ganho na detecção de neutralidade parece ter ocorrido simultaneamente a uma redução no Recall da classe Positiva. Tal comportamento sugere a existência de um compromisso entre métricas (trade-off), indicando que as estratégias de ajuste voltadas para a mitigação do viés das categorias majoritárias possivelmente influenciadas pela combinação de hiperparâmetros e funções de perda podem acarretar, sob as condições deste estudo, uma diminuição na capacidade de identificação das classes anteriormente predominantes.

Tabela 6: Evolução Comparativa das Métricas (v1 → v3)

Modelo	F1-Score (Macro)			Recall (Neutro)			Recall (Positivo)		
	v1	v2	v3	v1	v2	v3	v1	v2	v3
FinBERT	0.49	0.54	0.55	0.15	0.38	0.44	0.79	0.64	0.56
RoBERTa	0.49	0.53	0.54	0.14	0.35	0.45	0.79	0.61	0.55
ModernBERT	0.46	0.54	0.50	0.20	0.41	0.53	0.68	0.55	0.47
FastText	0.46	0.47	0.47	0.29	0.32	0.33	0.63	0.61	0.60

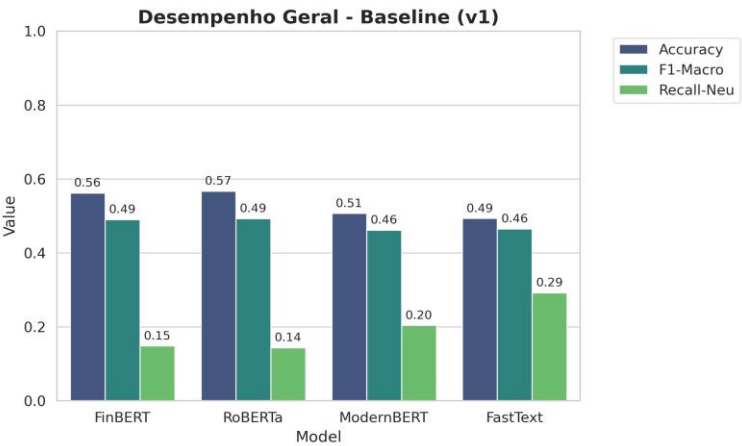
Fonte: Elaborado pelo autor (2026)

5.2 Análise Por Fase Experimental

5.2.1 Fase 1: Baseline (V1)

No cenário inicial, caracterizado pela ausência de técnicas de tratamento de desbalanceamento de classes, os modelos apresentaram uma tendência a favorecer as categorias majoritárias. Conforme ilustrado na Figura 7, a sensibilidade (Recall) da classe Neutra manifestou-se em níveis reduzidos nas arquiteturas baseadas em Transformers. A inspeção das matrizes de confusão na Figura 8 sugere que as arquiteturas tenderam a classificar amostras Neutras como Positivas ou Negativas, o que pode indicar uma dificuldade inicial em capturar nuances semânticas que apontam para a ausência de tendência ou movimentos definidos de mercado.

Figura 7 - Métricas da Fase 1 (Baseline): Note o Recall Neutro próximo de zero.



Fonte: Elaborado pelo autor (2026)

Figura 8 - Matrizes de Confusão (v1): A coluna "Predito Neutro" está quase vazia.

Matrizes de Confusão - Baseline (v1)

FinBERT				RoBERTa				ModernBERT				FastText			
Real	Positivo	Negativo	Neutro	Real	Positivo	Negativo	Neutro	Real	Positivo	Negativo	Neutro	Real	Positivo	Negativo	Neutro
	Positivo	Negativo	Neutro		Positivo	Negativo	Neutro		Positivo	Negativo	Neutro		Positivo	Negativo	Neutro
	Positivo	Negativo	Neutro		Positivo	Negativo	Neutro		Positivo	Negativo	Neutro		Positivo	Negativo	Neutro
Positivo	278	47	26	Positivo	278	50	23	Positivo	238	59	54	Positivo	222	56	73
Negativo	67	148	33	Negativo	68	153	27	Negativo	74	131	43	Negativo	74	117	57
Neutro	112	72	32	Neutro	107	78	31	Neutro	106	66	44	Neutro	90	63	63

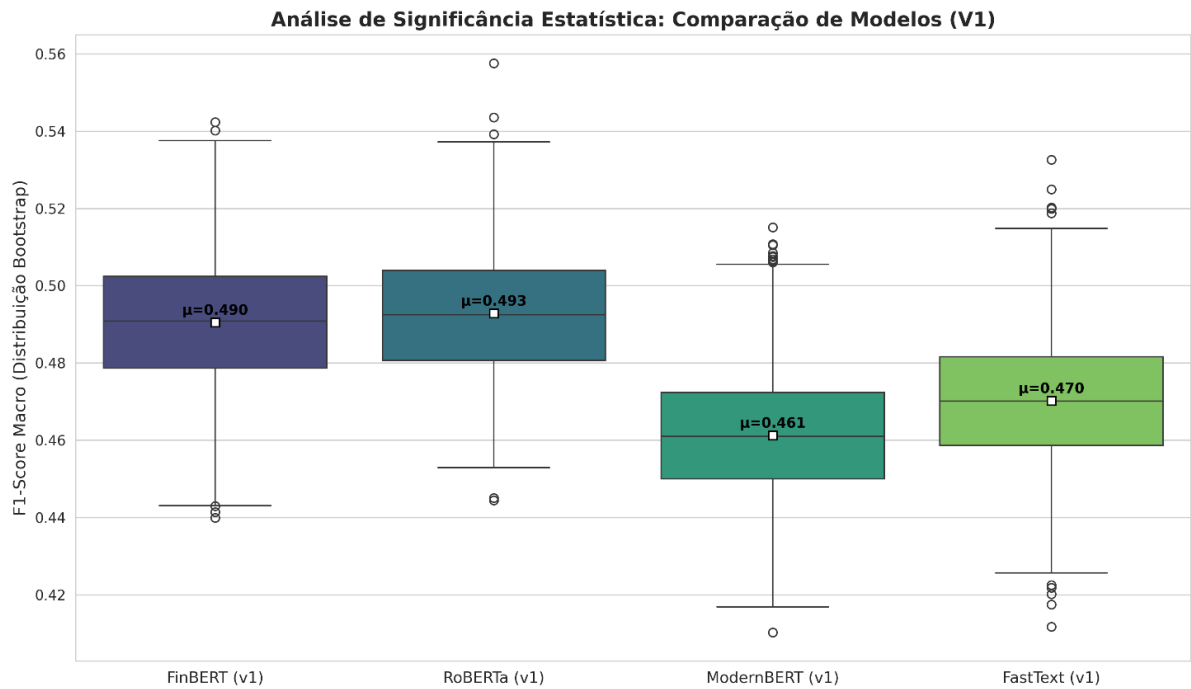
Fonte: Elaborado pelo autor (2026)

Os resultados observados nas matrizes de confusão sugerem uma variação no desempenho entre as classes: enquanto as categorias Positivo e Negativo parecem apresentar maior separabilidade sob as condições testadas, a classe Neutra aparenta configurar-se como um desafio significativo de classificação. Nas arquiteturas avaliadas, notou-se que uma parcela das amostras de referência neutra tendeu a ser deslocada para os extremos do espectro de sentimento, com uma frequência maior para a classe Positiva. Tal fenômeno pode sugerir uma propensão dos classificadores em atribuir uma polaridade diante de textos descritivos ou ambíguos, fator que pode contribuir para a redução do Recall da classe Neutra, especialmente nos modelos FinBERT e RoBERTa, cujas predições de neutralidade foram menos frequentes, conforme detalhado abaixo:

- FinBERT (Neutro real): 112 vira Positivo, 72 vira Negativo, só 32 acerta Neutro
- RoBERTa (Neutro real): 107 → Positivo, 78 → Negativo, só 31 acerta
- ModernBERT (Neutro real): 106 → Positivo, 66 → Negativo, 44 acerta (melhor que os dois acima)
- FastText (Neutro real): 90 → Positivo, 63 → Negativo, 63 acerta (melhor de todos no Neutro)

Ao analisar as matrizes de confusão da Figura 8, nota-se que a classe Neutra parece constituir um desafio preditivo relevante para as arquiteturas avaliadas, apresentando, de modo geral, índices de sensibilidade (Recall) inferiores às demais classes. Observou-se uma tendência à polarização, na qual amostras de referência neutra foram frequentemente alocadas nas categorias extremas: no FinBERT e no RoBERTa, por exemplo, a sensibilidade para a neutralidade situou-se em patamares baixos (entre 14% e 15%), com uma tendência de classificação como "Positivo" em 112 e 107 ocorrências, respectivamente.

Figura 9 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v1 (Baseline), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,490$; RoBERTa: $\mu=0,493$; ModernBERT: $\mu=0,461$; FastText: $\mu=0,470$)



Fonte: Elaborado pelo autor (2026)

A análise de significância estatística foi conduzida por meio de reamostragem bootstrap com 1.000 iterações, cujos resultados podem ser observados na Figura 9. O boxplot da distribuição bootstrap do F1-Score Macro para a versão v1 sugere que os modelos RoBERTa e FinBERT apresentaram desempenhos centrais próximos, com médias de $\mu = 0,493$ e $\mu = 0,490$, respectivamente. As medianas semelhantes e a sobreposição das caixas podem indicar um comportamento comparável entre ambas sob esta configuração

Em contrapartida, o FastText obteve um desempenho intermediário ($\mu = 0,470$), enquanto o ModernBERT apresentou a menor média ($\mu = 0,461$) no cenário avaliado. Nota-se a presença de outliers superiores em todos os modelos, sugerindo que, em determinadas reamostragens, o desempenho pode atingir valores mais elevados, como observado no RoBERTa (próximo de 0,56) e no FastText (acima de 0,53). Do ponto de vista inferencial, a sobreposição entre as distribuições de FinBERT e RoBERTa sugere uma possível equivalência de desempenho entre esses dois modelos no cenário baseline, enquanto o ModernBERT e, em menor escala, o FastText, tenderam a apresentar distribuições deslocadas para valores inferiores

Embora o ModernBERT e o FastText tenham apresentado indícios de maior robustez na preservação da classe Neutra com o FastText registrando o melhor desempenho específico nesta categoria (63 acertos e Recall de $\approx 29\%$) a distribuição prevista pelas arquiteturas baseadas em Transformers apresentou divergências em relação à distribuição real do corpus. Enquanto a base

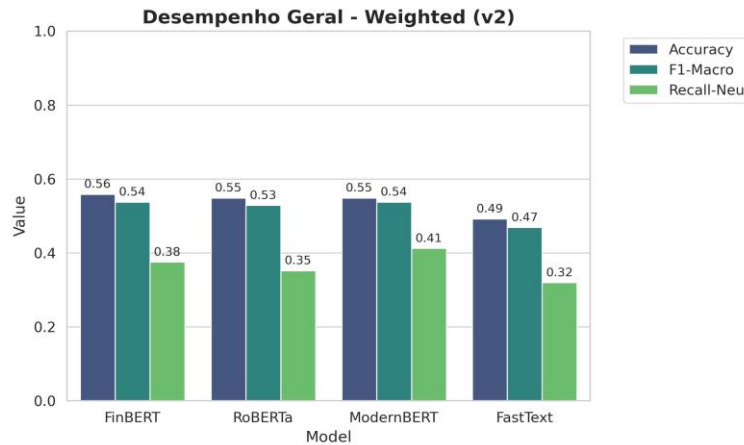
de dados possui aproximadamente 27% de amostras neutras, o RoBERTa estimou cerca de 10%, o que pode sugerir um viés de predição polarizada nesta configuração. Em termos de desempenho global, o RoBERTa ($\approx 56,7\%$) e o FinBERT ($\approx 56,2\%$) registraram as maiores acurácias nominais, possivelmente fundamentadas no alto Recall para notícias positivas ($\approx 79\%$).

Em uma análise comparativa entre as arquiteturas, o RoBERTa e o FinBERT apresentaram desempenhos globais competitivos, com métricas de acerto elevadas para as classes polares, ainda que acompanhadas de uma subestimação da classe Neutra. Em contrapartida, o modelo FastText, apesar de registrar uma acurácia geral inferior, indicou maior capacidade de preservação da neutralidade, apresentando uma distribuição de predições que tendeu a convergir mais com a distribuição real do dataset. O ModernBERT apresentou um desempenho intermediário, com indícios de incrementos marginais na detecção de neutralidade em relação às demais arquiteturas de Transformers. Este cenário pode indicar que as limitações do baseline podem estar relacionadas a uma combinação de fatores, incluindo a arquitetura do modelo, a complexidade da separação conceitual e a natureza das amostras disponíveis para a classe Neutra no domínio financeiro.

5.2.2 Fase 2: Ponderação De Classes (V2)

A implementação da ponderação de classes na função de perda de entropia cruzada (Cross-Entropy), fundamentada na Frequência Inversa das Classes (Inverse Class Frequency), parece ter contribuído para que os modelos priorizassem o aprendizado da categoria minoritária. Sob esta configuração e considerando a interação entre os pesos e as características intrínsecas das arquiteturas, o ModernBERT alcançou um de seus melhores desempenhos nominais, registrando um F1-Score Macro de 0,54. Tal resultado pode sugerir um equilíbrio mais favorável entre as três classes de sentimento no recorte analisado, conforme ilustrado na Figura 10.

Figura 10 - Métricas da Fase 2: Melhor equilíbrio entre classes.



Fonte: Elaborado pelo autor (2026)

Figura 11 - Matrizes de Confusão (v1): A coluna "Predito Neutro" está quase vazia.

Matrizes de Confusão - Weighted (v2)

	FinBERT			RoBERTa			ModernBERT			FastText		
	Positivo	Negativo Previsto	Neutro	Positivo	Negativo Previsto	Neutro	Positivo	Negativo Previsto	Neutro	Positivo	Negativo Previsto	Neutro
Real Positivo	225	52	74	215	47	89	194	50	107	214	57	80
Real Negativo	42	149	57	37	156	55	24	164	60	70	118	60
Real Neutro	67	68	81	67	73	76	52	75	89	88	59	69

Fonte: Elaborado pelo autor (2026)

Na configuração experimental da Versão 2, a aplicação da Entropia Cruzada Ponderada (Weighted Cross-Entropy) permitiu a reponderação do erro durante o processo de treinamento, visando mitigar o impacto do desequilíbrio entre as classes. Diferentemente da correção direta de um viés intrínseco aos dados, essa técnica ajustou a função de custo para penalizar de forma distinta os erros em categorias sub-representadas.

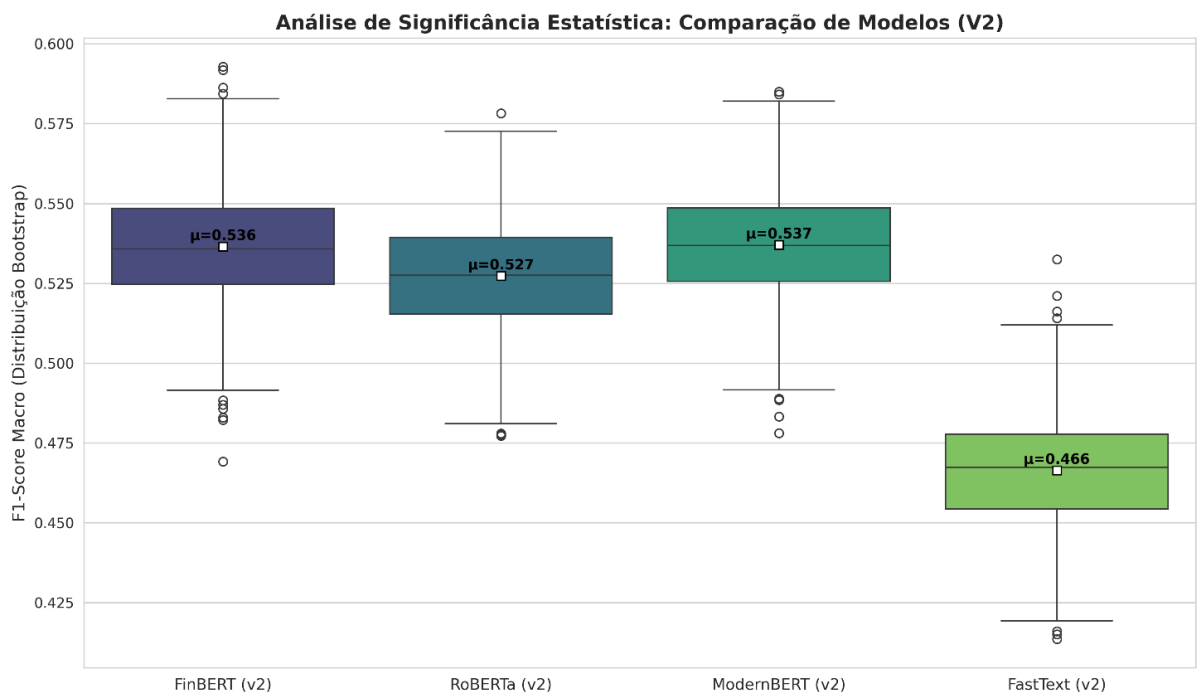
Como resultado, as arquiteturas FinBERT e RoBERTa, que anteriormente apresentavam dificuldades na identificação da classe Neutra, demonstraram um incremento em sua sensibilidade (Recall), atingindo marcas de 37,5% e 35,2%, respectivamente. Os resultados detalhados das matrizes de confusão na Figura 11 apresentam essa possível mudança de comportamento:

- FinBERT (Neutro real): 67 classificadas como Positivo, 68 como Negativo e 81 acertos nominais (segundo melhor desempenho na classe).
- RoBERTa (Neutro real): 67 classificadas como Positivo, 73 como Negativo e 76 acertos nominais.
- ModernBERT (Neutro real): 52 classificadas como Positivo, 75 como Negativo e 89 acertos nominais (líder em sensibilidade na classe Neutra).

- FastText (Neutro real): 88 classificadas como Positivo, 59 como Negativo e 69 acertos nominais.

O ModernBERT apresentou um desempenho notável para a neutralidade na v2, com um Recall de 41,2%. Esta transição parece refletir-se na distribuição das predições: enquanto a v1 tendia a desconsiderar a volumetria do dataset, a v2 indicou uma convergência estatística com a distribuição original (27% de amostras neutras), particularmente nos modelos FinBERT e RoBERTa, cujas predições de neutralidade alinharam-se em patamares próximos de 26% e 27%, respectivamente.

Figura 12 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v2 (Weighted), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,536$; RoBERTa: $\mu=0,527$; ModernBERT: $\mu=0,537$; FastText: $\mu=0,466$).



Fonte: Elaborado pelo autor (2026)

O boxplot referente à versão v2 (Weighted) na Figura 12, baseado em reamostragem bootstrap do F1-Score Macro, sugere um possível incremento de desempenho para os modelos baseados em Transformers, com médias concentradas na faixa de 0,52 e 0,55. Observa-se que o ModernBERT (v2) apresenta uma das maiores médias ($\mu = 0,537$), em patamar similar ao do FinBERT (v2) ($\mu = 0,536$), enquanto o RoBERTa (v2) situa-se ligeiramente abaixo ($\mu = 0,527$). A sobreposição observada entre as distribuições desses três modelos pode indicar que as diferenças de desempenho, sob o ponto de vista inferencial, tendem a ser modestas no contexto prático avaliado.

Em contraste, o FastText (v2) registrou um desempenho comparativamente inferior ($\mu = 0,466$), com uma distribuição deslocada para valores menores, o que pode indicar limitações do modelo sob esta configuração específica. Adicionalmente, a presença de outliers superiores nos Transformers sugere que, em determinadas reamostragens, o desempenho pode atingir níveis mais elevados. Contudo, a similaridade entre FinBERT e ModernBERT, aliada à proximidade com o RoBERTa, tende a apoiar a percepção de que, na v2, estas arquiteturas constituem alternativas competitivas, enquanto o FastText parece ser uma opção menos indicada para a maximização do F1-Macro.

Contudo, a análise sugere a existência de um possível compromisso (trade-off) associado a este equilíbrio, caracterizado por um comportamento preditivo que tende a ser mais conservador. Observou-se uma tendência na qual o incremento na detecção de sentenças neutras esteve associado a uma redução no Recall da classe Positiva. O ModernBERT exibiu uma oscilação mais acentuada nesta métrica, com a sensibilidade declinando para 55,3%, possivelmente devido ao deslocamento de amostras originalmente positivas para a categoria Neutra. Este fenômeno pode estar relacionado a uma redução da tendência de polarização forçada, resultando em um comportamento preditivo menos extremista, característica que pode ser relevante para o processamento de textos financeiros de natureza ambígua.

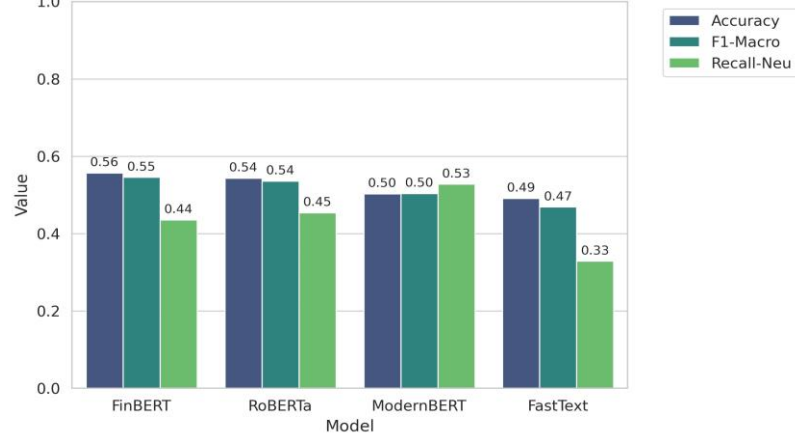
Entretanto, esse avanço parece ter estabelecido um compromisso métrico: o ganho de sensibilidade na classe Neutra correlacionou-se, neste estudo, com um aumento na incidência de falsos neutros. Por outro lado, sob a ótica da gestão de riscos, a ponderação indicou certa eficácia na mitigação de erros considerados críticos, como a classificação de sentenças negativas como positivas. Em suma, a configuração v2 parece estabelecer um equilíbrio entre as classes e pode ampliar a confiabilidade diagnóstica contra inversões de polaridade, ainda que possa atenuar a sensibilidade a sinais positivos, o que sugere a necessidade de calibração dependendo dos objetivos estratégicos da aplicação.

5.2.3 Fase 3: Refinamento Com Focal Loss (V3)

Visando priorizar o aprendizado sobre amostras de classificação complexa (hard examples), optou-se pela aplicação da função de perda Focal Loss com parâmetro de $\gamma = 2,0$. Sob essa configuração e considerando a interação entre a arquitetura e os hiperparâmetros, o FinBERT aparentou adaptar-se de forma favorável, atingindo um F1-Score nominal de 0,55. Em contrapartida, o ModernBERT exibiu indícios do que sugere ser uma sobrecorreção (over-

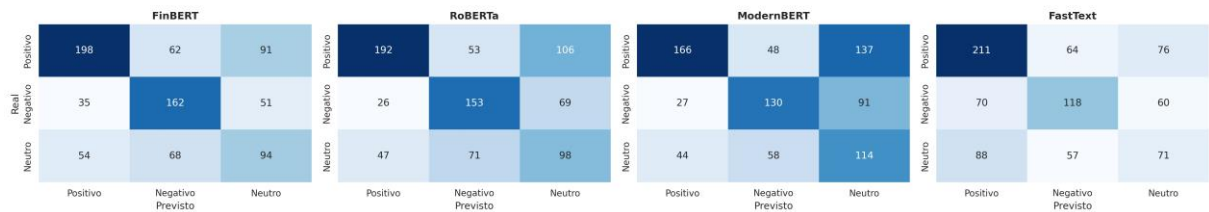
correction), conforme ilustrado na Figura 13, ao tender para uma priorização da categoria Neutra em possível detrimento da classe Positiva, o que pode sinalizar uma sensibilidade mais acentuada aos parâmetros de penalização da função neste contexto específico.

Figura 13 - Métricas da Fase 3: ModernBERT sacrifica Recall Positivo por Neutro.
Desempenho Geral - Focal+Smooth (v3)



Fonte: Elaborado pelo autor (2026)

Figura 14 - Matrizes de Confusão (v3): Maior dispersão, mas melhor cobertura da diagonal central.
Matrizes de Confusão - Focal+Smooth (v3)



Fonte: Elaborado pelo autor (2026)

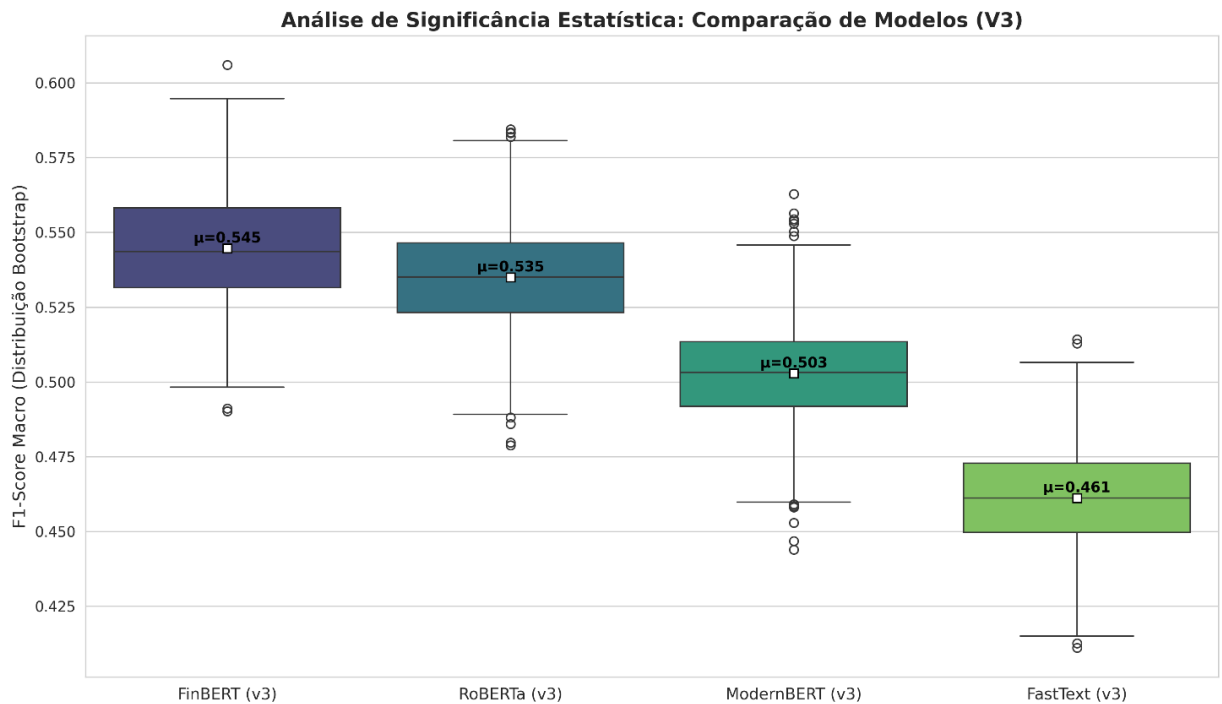
Ao analisar as Matrizes de Confusão da Figura 14, observa-se que, embora tenha ocorrido uma oscilação na sensibilidade para sinais positivos, a configuração v3 indicou um incremento relevante em termos de segurança diagnóstica e estabilidade do classificador. Notou-se uma diminuição expressiva na incidência de inversões de polaridade críticas, particularmente no erro de classificar amostras Negativas como Positivas. Enquanto o baseline FastText tendeu a manter uma taxa mais elevada de falhas desse tipo (70 ocorrências), os modelos baseados em Transformers pareceram mitigar esse risco para patamares inferiores, com o RoBERTa registrando 26 ocorrências sob as condições testadas.

Sob a ótica de possíveis aplicações financeiras, esses achados tendem a reforçar a utilidade do refinamento na busca pela mitigação de falsos sinais de otimismo, auxiliando o sistema contra potenciais interpretações equivocadas de riscos de mercado. Os resultados detalhados nas matrizes de confusão sugerem o seguinte comportamento:

- FinBERT (Neutro real): 54 classificadas como Positivo, 68 como Negativo e 94 acertos nominais (segundo melhor desempenho na classe).
- RoBERTa (Neutro real): 47 classificadas como Positivo, 71 como Negativo e 98 acertos nominais.
- ModernBERT (Neutro real): 44 classificadas como Positivo, 58 como Negativo e 114 acertos nominais (líder em sensibilidade na classe Neutra).
- FastText (Neutro real): 88 classificadas como Positivo, 57 como Negativo e 71 acertos nominais.

No cenário de refinamento da Versão 3, sinalizou-se o que parece ser a maior sensibilidade (Recall) para a classe Neutra entre os experimentos realizados. A aplicação da Focal Loss, em conjunto com os demais ajustes de treinamento, possibilitou que as arquiteturas de Transformers superassem o desempenho do baseline, com destaque para o ModernBERT, que atingiu um Recall de 52,8% para a neutralidade, seguido pelo RoBERTa (45,4%) e FinBERT (43,5%). Contudo, os dados sugerem um compromisso entre métricas (trade-off), manifestado por um comportamento preditivo aparentemente mais conservador: o ganho na detecção de neutros parece ter contribuído para um declínio mais acentuado no Recall da classe Positiva, especialmente no ModernBERT (47,3%), onde 137 amostras originalmente otimistas foram deslocadas para a categoria Neutra. Em termos estatísticos, esse fenômeno pode ser interpretado como uma tendência de superestimação da neutralidade na distribuição de predições do ModernBERT (42%) frente à volumetria real do dataset (27%).

Figura 15 - Distribuição bootstrap do F1-Score Macro para os modelos avaliados na versão v3 (Focal + Smooth), evidenciando a variabilidade do desempenho e as médias estimadas (FinBERT: $\mu=0,545$; RoBERTa: $\mu=0,535$; ModernBERT: $\mu=0,503$; FastText: $\mu=0,461$).



Fonte: Elaborado pelo autor (2026)

O boxplot da versão v3 (Focal + Smooth) representado na Figura 15, obtido via bootstrap do F1-Score Macro, sugere que o FinBERT (v3) apresenta o desempenho central mais robusto neste teste, com média $\mu = 0,545$, seguido pelo RoBERTa (v3) com $\mu = 0,535$. Ambos exibiram distribuições concentradas e com outliers superiores próximos de 0,60, o que pode indicar um potencial de desempenho elevado em determinadas condições de reamostragem.

O ModernBERT (v3) posicionou-se em um patamar inferior ($\mu = 0,503$), com maior dispersão e presença de valores reduzidos, sugerindo uma possível menor estabilidade relativa sob essa configuração específica. Por sua vez, o FastText (v3) tendeu a manter o resultado global menos expressivo ($\mu = 0,461$), com distribuição deslocada para valores menores, o que pode indicar uma menor adequação para maximizar o F1-Macro no cenário avaliado. Assim, sob a estratégia Focal+Smooth e no escopo desta pesquisa, infere-se que FinBERT e RoBERTa se constituem como alternativas competitivas, enquanto o ModernBERT apresenta desempenho intermediário e o FastText permanece abaixo dos demais.

O custo dessa trajetória de otimização manifestou-se na classe Positiva através de uma transição de amostras para a categoria Neutra, resultando em uma possível degradação do Recall para esta classe, efeito que se mostrou mais visível no ModernBERT. Pode-se considerar que a combinação entre Focal Loss e Label Smoothing tenda a induzir o classificador a um

comportamento mais cauteloso; diante de fronteiras de decisão menos nítidas, o modelo parece priorizar a neutralidade (safe zone), o que pode ser considerado desejável para a prevenção de inversões graves de polaridade, ainda que possa resultar na eventual omissão de sinais positivos em casos limítrofes.

Ao comparar as três fases experimentais, nota-se uma possível progressão analítica: a v1 (Baseline) parece priorizar a acurácia global com menor atenção às classes minoritárias ; a v2 (Weighted) busca restabelecer um maior equilíbrio estatístico entre as categorias e a v3 (Focal + Smooth) tende a maximizar a robustez da classe Neutra e a segurança contra inversões, embora possa impor um trade-off mais perceptível sobre a classe Positiva. Conclui-se que, sob as condições deste estudo, a configuração v2 aparenta apresentar o perfil mais equilibrado para uso geral, enquanto a v3 sugere ser uma abordagem de maior cautela preditiva, possivelmente indicada para contextos onde a integridade da neutralidade seja priorizada frente à necessidade de sensibilidade absoluta a sinais positivos.

5.3 DISCUSSÃO

5.3.1 Discussão Comparativa De Resultados

Ao confrontar os resultados experimentais deste trabalho com os parâmetros estabelecidos por Lefort et al. (2025), observa-se uma convergência estatística relevante no que tange à distribuição de classes do dataset original. O paper FinMarBa reporta uma proporção de sentimentos composta por 42,11% de amostras Positivas, 31,43% Negativas e 26,45% Indecisivas (Neutras). Esta distribuição, fundamentada no viés positivo de longo prazo dos mercados de capitais, tendeu a ser refletida na configuração Weighted (v2) deste TCC, na qual o modelo RoBERTa apresentou um alinhamento de predições de neutralidade em torno de 27%, assemelhando-se à densidade real do corpus e indicando uma possível mitigação da negligência observada no cenário baseline.

A análise do desempenho preditivo tende a corroborar a superioridade das arquiteturas baseadas em Transformers sobre modelos estatísticos tradicionais, em linha com o preconizado no benchmark original. Enquanto Lefort et al. (2025) registram que o uso de modelos pré-treinados e ajustados ao domínio financeiro pode favorecer a captura de reações de mercado de forma mais precisa do que anotações humanas subjetivas, os dados deste estudo sugerem que a combinação de estratégias de refinamento, como o emprego da Focal Loss (v3) em conjunto com o ajuste de outros hiperparâmetros, contribuiu para elevar o F1-Score Macro do FinBERT para 0,55. Este valor situa-se em proximidade com os picos de desempenho reportados na literatura (entre 0,57 e 0,58), o que pode sinalizar que a abordagem de focar em exemplos de

classificação complexa (hard examples) atua como um fator relevante para lidar com a natureza da rotulagem informada pelo mercado.

Adicionalmente, a transição para métodos automatizados de anotação pode oferecer uma perspectiva explicativa para os desafios encontrados na classificação da neutralidade, considerando que tal sentimento é apontado como particularmente desafiador por representar a ausência de reação significativa do mercado. Neste contexto, a evolução para a Versão 3 deste trabalho sugeriu um progresso em relação ao trabalho de referência: ao utilizar o ModernBERT, observou-se um Recall de 52,8% para a classe Neutra, o que parece ter auxiliado na redução da tendência de polarização observada nos baselines ao processar ruídos informacionais. Assim, os resultados aqui apresentados tendem a apoiar a validade do dataset FinMarBa e podem apontar caminhos para o aprimoramento da robustez diagnóstica em cenários de incerteza léxica e volatilidade.

Os resultados dos treinamentos para os modelos v1, v2, e v3, sintetizados em exemplos na Figura 16, indicam situações de convergência entre os classificadores, especialmente em manchetes com marcadores semânticos mais diretos. Tais evidências aparecem de forma mais acentuada em exemplos como “Dollar Index Jumps 1 Percent...”, classificada unanimemente como positiva, ou “European Stocks Drop...”, como negativa. Esse comportamento sugere que manchetes com indicadores explícitos de valorização ou queda favorecem uma classificação consistente dos modelos. Por outro lado, observaram-se casos de erro coletivo, como em “Oil Tumbles 8.6 Percent...”, cuja referência é negativa, mas foi classificada como positiva por todos os modelos, podendo sinalizar possíveis limitações na interpretação de contextos econômicos específicos por parte das arquiteturas testadas sob as presentes configurações.

A despeito dessas dificuldades, o sucesso convergente em notícias neutras como "Gold Futures Reach Highest Price..." pode indicar uma aptidão dos modelos refinados em distinguir fatos descritivos de picos históricos de sentimentos subjetivos de otimismo. Esses achados contribuem para a análise de robustez das versões V2 e V3, sugerindo que a introdução de pesos (Weighted) e o ajuste da Focal Loss, possivelmente em interação com a arquitetura do modelo, não parecem ter prejudicado a detecção de padrões textuais evidentes.

Figura 16 - Comparação entre a classificação de referência e as predições dos classificadores para manchetes do domínio financeiro, com identificação de polaridades positiva, negativa e neutra para as versões v1, v2 e v3.

Manchete	Valor	FastText_V1	FastText_V2	FastText_V3	FinBERT_V1	FinBERT_V2	FinBERT_V3	ModernBERT_V1	ModernBERT_V2	ModernBERT_V3	RoBERTa_V1	RoBERTa_V2	RoBERTa_V3
Gold Futures Reach Highest Price Since Early July	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro
Dollar Index Jumps 1 Percent to Highest Closing Level Since March	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo
European Stocks Drop Amid Concerns of Central Bank Interest-Rate Increases	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo	Negativo
Crude Oil Rises Amid U.S. Sanctions on Iranian Companies	Positivo	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro
Oil Tumbles 8.6 Percent, Most in Two Years	Negativo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo
S&P 500 Sees Minor Slip as Commodity Prices Fall	Neutro	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo	Positivo
Italian 10-Year Bond Yields Rise to Euro-Era High	Negativo	Positivo	Positivo	Positivo	Neutro	Negativo	Negativo	Positivo	Neutro	Negativo	Neutro	Negativo	Negativo
U.S. Government Faces \$1.2 Trillion in Automatic Spending Cuts	Negativo	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro	Neutro
Swiss Franc Reaches Record High Versus U.S. Currency	Positivo	Negativo	Negativo	Negativo	Positivo	Positivo	Negativo	Positivo	Negativo	Negativo	Positivo	Negativo	Negativo
Dollar Drops Below 81 Yen for First Time Since 1995	Positivo	Positivo	Positivo	Positivo	Negativo	Negativo	Negativo	Positivo	Positivo	Positivo	Positivo	Negativo	Negativo
Slovakia Approves Enhanced European Bailout Fund	Positivo	Negativo	Negativo	Negativo	Neutro	Neutro	Neutro	Neutro	Positivo	Neutro	Neutro	Neutro	Negativo
Bank of England Warns of High Inflation, Possible Rise to 5 Percent	Negativo	Negativo	Negativo	Negativo	Positivo	Positivo	Positivo	Negativo	Positivo	Positivo	Positivo	Positivo	Positivo
Gasoline Surges 2 Percent to Highest Level This Month	Positivo	Positivo	Positivo	Positivo	Neutro	Neutro	Neutro	Neutro	Negativo	Negativo	Neutro	Negativo	Negativo
U.S. Treasury to Auction \$13 Billion in 30-Year Bonds	Positivo	Neutro	Neutro	Neutro	Positivo	Positivo	Positivo	Positivo	Negativo	Positivo	Positivo	Positivo	Positivo

Fonte: Elaborado pelo autor (2026)

Nota-se, adicionalmente, que manchetes com maior complexidade sintática ou que dependem de uma interpretação financeira mais profunda tendem a apresentar maior dispersão entre as classificações das arquiteturas testadas. Exemplos como “Italian 10-Year Bond Yields Rise...”, “Swiss Franc Reaches Record High...” e “Gasoline Surges 2 Percent...” exibiram uma heterogeneidade de rótulos entre positivo, neutro e negativo, o que pode sugerir uma sensibilidade dos classificadores à ambiguidade semântica e às nuances específicas do contexto de mercado.

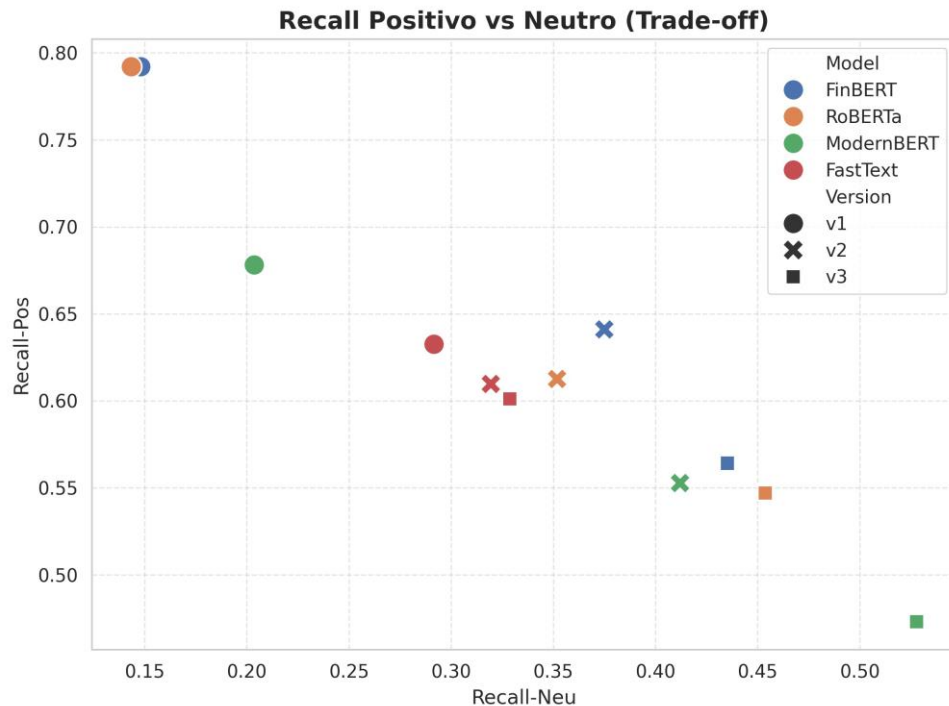
De maneira geral, os resultados sugerem que o sistema apresenta uma capacidade satisfatória de processar enunciados mais literais, embora pareça encontrar desafios em notícias nas quais o impacto econômico da informação não é explicitamente declarado. Nesse sentido, pode-se inferir que o desempenho tende a ser mais consistente em textos de polaridade evidente, ao passo que manchetes que demandam inferência contextual ainda representam um desafio técnico considerável para os modelos avaliados. Tais observações oferecem indícios de que as estratégias implementadas possivelmente decorrentes da combinação de arquiteturas específicas e do ajuste ponderado de hiperparâmetros e funções de perda podem ter auxiliado na manutenção da integridade dos acertos observados no baseline, contribuindo potencialmente para o tratamento de casos de ambiguidade que anteriormente limitavam a eficácia global das predições.

5.3.2 Análise Do Trade-Off

A análise comparativa dos modelos sugere a existência de um possível trade-off estrutural entre as métricas de desempenho. Conforme ilustrado na Figura 16, nota-se uma tendência de correlação inversa entre a sensibilidade (recall) na detecção de sentenças Positivas e Neutras. Tal comportamento pode indicar a presença de uma zona de ambiguidade semântica inerente ao dataset, na qual a distinção entre notícias otimistas e informativas (neutras)

apresenta-se como um desafio de classificação que, sob as condições experimentais observadas, parece exigir o suporte de contextos extralinguísticos para uma resolução mais precisa.

Figura 17 - Trade-off entre Recall Positivo e Neutro ao longo das versões.



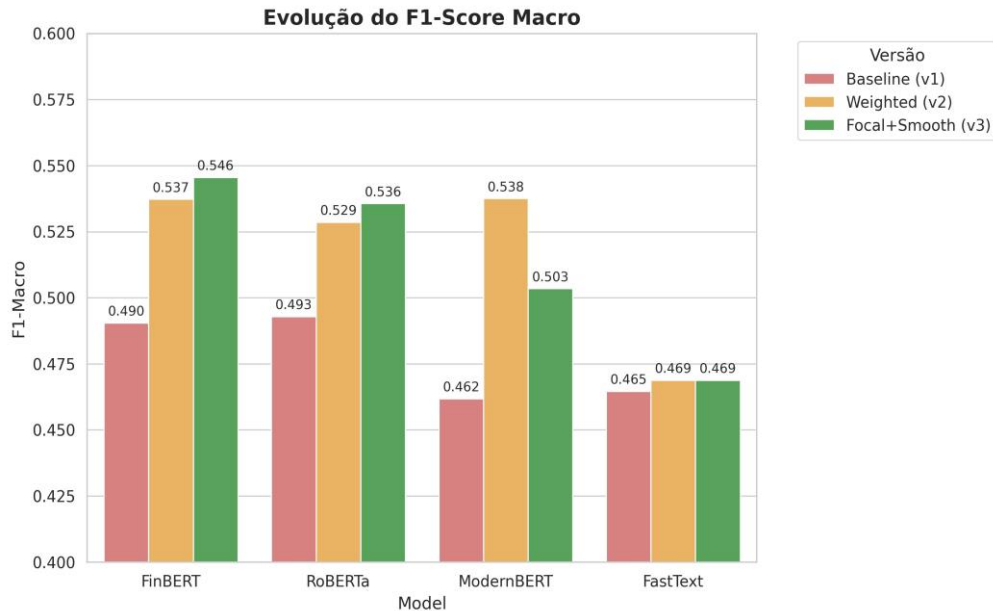
Fonte: Elaborado pelo autor (2026)

A representação da Figura 17 – Recall Positivo vs Neutro (Trade-off), ilustra a dinâmica entre o Recall de Positivo e o Recall de Neutro ao longo das versões testadas: observa-se que, à medida que os ajustes tendem a favorecer a classe Neutra, ocorre, em geral, uma redução concomitante na detecção de Positivos. No cenário baseline (v1), o FinBERT e o RoBERTa apresentaram altos índices nominais de recall em Positivo ($\approx 0,79$), porém exibiram limitações expressivas na identificação da neutralidade ($\approx 0,14$ – $0,15$). A introdução da ponderação (v2) parece ter conduzido os modelos a um patamar intermediário, elevando consideravelmente a detecção de neutros nos transformers ($\approx 0,35$ – $0,41$), ainda que acompanhada de uma oscilação na sensibilidade positiva para a faixa de $\approx 0,55$ – $0,64$.

A configuração Focal+Smooth (v3) tendeu a intensificar esse deslocamento para a direita no gráfico, reforçando a classe Neutra sob o custo de um decréscimo adicional em Positivo; o caso do ModernBERT, que atingiu o maior Recall-Neutro ($\approx 0,53$) e o menor Recall-Positivo ($\approx 0,47$) do grupo, pode ser interpretado como um exemplo de sensibilidade elevada aos parâmetros de penalização. Em contrapartida, o FinBERT e o RoBERTa exibiram uma trajetória aparentemente mais equilibrada, apresentando ganhos graduais em Neutro entre as versões v1, v2 e v3 sem reduções tão acentuadas quanto as do ModernBERT. O FastText

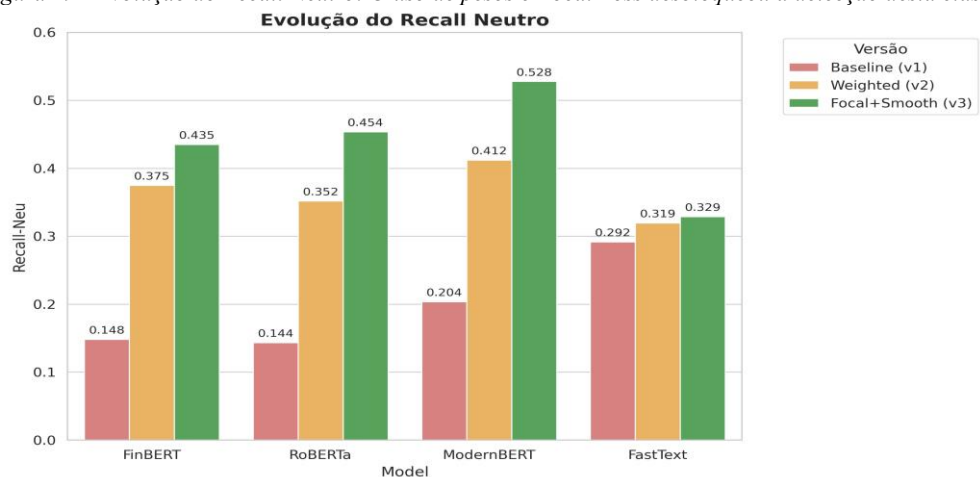
apresentou-se como um perfil de maior estabilidade, sugerindo que as técnicas e hiperparâmetros aplicados podem ter um impacto mais pronunciado em arquiteturas baseadas em Transformers do que em modelos estatísticos simples.

Figura 18 - Evolução do F1-Macro: FinBERT mostra a maior robustez ao longo das revisões.



Fonte: Elaborado pelo autor (2026)

Figura 19 - Evolução do Recall Neutro: O uso de pesos e Focal Loss desbloqueou a detecção desta classe.



Fonte: Elaborado pelo autor (2026)

De forma complementar, os dados sumarizados nas Figuras 18 e 19 sugerem uma possível trajetória de evolução convergente: a aplicação conjunta de estratégias para o tratamento de desbalanceamento como o uso de pesos de classe e a combinação de Focal Loss com Label Smoothing parece ter contribuído para incrementos no F1-Score Macro das arquiteturas baseadas em Transformers. Em contraposição, o modelo baseline (FastText) tendeu a manter uma variância reduzida em seu desempenho sob as mesmas condições. No que tange ao F1-Macro, a transição da configuração inicial para a ponderada (v2) indicou um avanço

relevante, enquanto a configuração v3 consolidou ganhos marginais para o FinBERT ($0,490 \rightarrow 0,537 \rightarrow 0,546$) e para o RoBERTa ($0,493 \rightarrow 0,529 \rightarrow 0,536$), posicionando a versão v3 como uma alternativa equilibrada no recorte analisado.

No que concerne ao Recall da classe Neutra, observou-se o que sugere ser um deslocamento de patamar relevante. Partindo de índices de sensibilidade reduzidos na versão v1, os modelos Transformers apresentaram uma evolução vigorosa nas versões subsequentes. O ModernBERT, especificamente, progrediu de 0,204 (v1) para 0,528 na versão v3, fornecendo indícios de que a combinação da Focal Loss com o Smoothing pode atuar como um fator atenuante do viés das classes majoritárias sobre a classe Neutra. Comportamento análogo foi observado no FinBERT e no RoBERTa, enquanto o FastText exibiu uma evolução incremental mais contida ($0,292 \rightarrow 0,319 \rightarrow 0,329$).

A correlação entre os indicadores sugere que a busca pela maximização da sensibilidade na classe Neutra pode estar associada a uma tendência de comportamento preditivo mais conservador por parte dos modelos. Este fenômeno pôde ser observado no ModernBERT, onde o alcance do maior Recall para a classe Neutra na v3 veio acompanhado de uma possível degradação no F1-Macro em comparação à versão v2 ($0,538 \rightarrow 0,503$). Tal evidência pode indicar que ajustes excessivos voltados para a classe Neutra possuem o potencial de impactar a precisão nas classes polares. Diante dos dados apresentados, pode-se inferir que, se o critério de seleção for o equilíbrio harmônico global, o FinBERT v3 posicionou-se com o melhor desempenho nominal em F1-Macro no escopo deste teste. Por outro lado, caso a prioridade seja a mitigação de polarizações indevidas em sentenças neutras, o ModernBERT v3 apresentou indícios de maior robustez específica para tal finalidade.

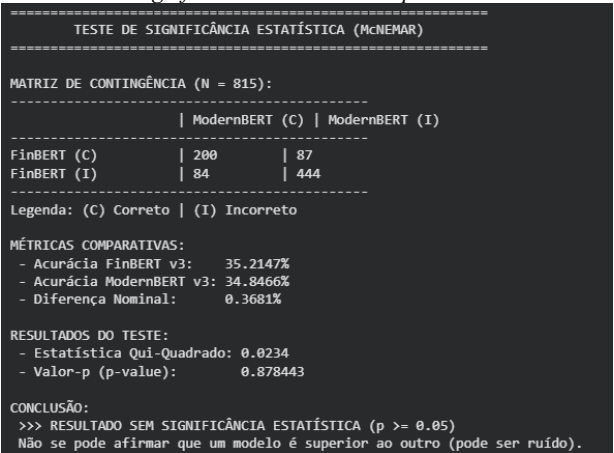
5.3.3 Limitações Do Dataset

A análise de concordância do Ensemble na fase v3 sugeriu que, mesmo em situações de consenso unânime entre as arquiteturas FinBERT, RoBERTa e ModernBERT, a acurácia observada situou-se em patamares de aproximadamente 53%. Estes resultados podem indicar a existência de um possível limite de desempenho, possivelmente condicionado pela complexidade da relação entre o conteúdo textual das manchetes e os rótulos derivados de variações de mercado, o que parece atenuar o impacto de melhorias puramente algorítmicas sob as condições testadas. Adicionalmente, ressalta-se que o recorte do dataset FinMarBa (LEFORT et al., 2025) disponível para esta investigação restringiu-se ao biênio 2010-2011, diferindo do intervalo de 2010 a 2024 utilizado no estudo original, fator que pode influenciar a abrangência das conclusões no que tange à representatividade temporal das amostras.

5.3.4 O Confronto Finbert E Modernbert

A análise comparativa entre o FinBERT e o ModernBERT revelou uma proximidade nas métricas de desempenho que sugeriu a necessidade de uma investigação para além da observação nominal. Diante dos resultados de acurácia obtidos 0,546 para o FinBERT e 0,512 para o ModernBERT , a aplicação do Teste de McNemar apresentou um valor-p de 0,8784. Situando-se acima do limiar crítico de 0,05, tal índice não permitiu a rejeição da hipótese nula, sugerindo que as variações observadas entre as arquiteturas podem estar associadas a ruídos estocásticos no conjunto de teste. Essa evidência indica que, para este recorte específico de dados, as superioridades nominais observadas podem não decorrer necessariamente de vantagens estruturais inerentes a uma única arquitetura. Adicionalmente, a análise de reamostragem bootstrap tendeu a apoiar essa interpretação ao demonstrar uma sobreposição considerável nos intervalos de confiança de 95%, sinalizando uma possível equivalência estatística entre os modelos sob as condições experimentais do dataset FinMarBa.

Figura 20 - Resultado do teste de significância estatística de McNemar entre os modelos FinBERT v3 e ModernBERT v3, apresentando a matriz de contingência, as acurácias obtidas e o valor-p, indicando ausência de diferença estatisticamente significativa entre os desempenhos.



Fonte: Elaborado pelo autor (2026)

A observação desta ausência de significância estatística no confronto entre o FinBERT (ARACI, 2019) e o ModernBERT (SALAMI et al., 2024) pode indicar que as inovações estruturais contemporâneas e a especialização de domínio tendem a encontrar um patamar comum de performance sob as condições deste estudo. Embora o FinBERT se beneficie de uma provável "vantagem de partida" semântica decorrente de seu pré-treinamento em corpora financeiros massivos como o TRC2-financeiro e o Financial PhraseBank, essa especialização parece não ter sido suficiente para superar de forma conclusiva a robustez arquitetural do ModernBERT. A proximidade nos resultados sugere que o ruído intrínseco do mercado pode atuar como um elemento nivelador, no qual a sensibilidade léxica do FinBERT e a eficiência

estrutural do ModernBERT parecem convergir para um possível limite de compreensão textual dentro deste cenário específico.

No que tange à infraestrutura de processamento, o ModernBERT (SALAMI et al., 2024) incorpora inovações como o Flash Attention (DAO et al., 2022) e o Unpadding, que representam avanços significativos na eficiência computacional. Contudo, para o cenário de manchetes curtas do dataset FinMarBa, os dados sugerem que o impacto dessas tecnologias tenha sido primordialmente operacional, favorecendo o tempo de treinamento e a vazão de dados (throughput), sem necessariamente refletir incrementos proporcionais de acurácia semântica frente aos modelos tradicionais. Parece razoável inferir que tais otimizações atuem como aceleradores de infraestrutura, enquanto a capacidade de discernimento do sentimento possa permanecer mais estritamente vinculada à profundidade das representações aprendidas e ao ajuste de hiperparâmetros, contribuindo para a paridade de resultados ($p = 0,8784$) observada entre as arquiteturas.

Sob a ótica da linguística computacional, as divergências preditivas observadas entre as arquiteturas podem ser influenciadas por motivações associadas à tipologia da notícia, ainda que o saldo estatístico aponte para uma equivalência técnica. A investigação qualitativa sinalizou que o FinBERT (ARACI, 2019) tende a apresentar maior acuidade em manchetes com alta densidade de terminologia técnica e indicadores quantitativos, onde a semântica lexical pode atuar como um importante gatilho de classificação. Em contrapartida, arquiteturas como o RoBERTa (LIU et al., 2019) e o ModernBERT pareceram demonstrar maior resiliência em notícias com estruturas sintáticas complexas, sugerindo que uma compreensão estrutural de longo alcance pode ser um fator determinante nesses casos.

Sugere-se, portanto, que a escolha da arquitetura em PLN financeiro considere não apenas a modernidade do modelo, mas a natureza linguística predominante no corpus. Enquanto a especialização do FinBERT favorece a interpretação de jargões técnicos, a robustez de modelos generalistas como o ModernBERT oferece eficiência operacional e capacidade sintática. Todavia, o fato de diferentes abordagens o conhecimento de domínio e os incrementos arquiteturais resultarem em desempenhos estatisticamente indistinguíveis pode indicar que ambos atingiram um patamar similar de eficácia preditiva. Em cenários de alta ambiguidade, essa convergência de resultados tende a reforçar a hipótese de que a interpretação textual isolada pode estar próxima de um limite teórico, aproximando-se dos patamares de excelência e ruído reportados por Lefort et al. (2025).

6 CONCLUSÃO

Em suma, a análise comparativa entre os modelos experimentais e o referencial estabelecido pelo dataset FinMarBa (LEFORT et al., 2025) sugere que o aprimoramento da análise de sentimento financeiro pode ser favorecido não apenas pelo uso de arquiteturas profundas, mas possivelmente pela combinação de estratégias robustas para o tratamento de desbalanceamento de classes. Enquanto a literatura aponta a neutralidade como um limitador intrínseco devido à natureza da rotulagem informada pelo mercado, os resultados deste estudo indicam que a transição do cenário Baseline (v1) para o Refinado (v3) contribuiu para mitigar tal fragilidade. Sob as condições observadas, a associação do ModernBERT (SALAMI et al., 2024) com a Focal Loss e outros ajustes de hiperparâmetros parece ter favorecido a consistência do corpus, elevando a sensibilidade para a classe neutra (52,8%) e possivelmente atenuando tendências de polarização comumente encontradas em modelos de propósito geral.

Ademais, a convergência observada na Versão 2 (Weighted), que tendeu a replicar a distribuição de mercado, oferece indícios da viabilidade de sistemas automatizados de anotação na geração de sinais preditivos. Ao alinhar o desempenho do FinBERT (ARACI, 2019) a patamares de excelência reportados na literatura (LEFORT et al., 2025), esta investigação apresenta uma abordagem que busca equilibrar a acurácia global com a segurança diagnóstica. Todavia, a investigação revelou que a superioridade nominal do FinBERT (0,546) em relação ao ModernBERT (0,512) não apresentou significância estatística neste recorte. A aplicação do Teste de McNemar resultou em um valor-p de 0,8784, o que não permite rejeitar a hipótese nula e sugere que a variação observada entre as taxas de erro pode ser influenciada por ruídos estocásticos. Tal achado é relevante por apontar para uma possível equivalência técnica onde a especialização de domínio e a robustez estrutural parecem convergir para um patamar de desempenho similar diante da ambiguidade do mercado.

Esta investigação sugere que a combinação de técnicas de Cost-Sensitive Learning e Focal Loss pode ser eficaz para reduzir vieses em direção às classes majoritárias, elevando o Recall da classe Neutra de aproximadamente 15% no baseline para patamares superiores a 45% nas configurações testadas. No que tange à infraestrutura, observou-se que inovações como Flash Attention (DAO et al., 2022) e Unpadding proporcionaram ganhos operacionais em eficiência computacional e vazão de dados (throughput). Entretanto, para o cenário específico de manchetes curtas, tais avanços não pareceram traduzir-se em ganhos diretos de acurácia semântica, reforçando a hipótese de que a capacidade de discernimento do sentimento pode

estar mais estritamente vinculada à profundidade das representações aprendidas no pré-treinamento de domínio do que exclusivamente à eficiência da arquitetura.

Como síntese das contribuições desta pesquisa para o campo do Processamento de Linguagem Natural (NLP) financeiro, destacam-se:

- **Tendência de Equivalência de Modelos:** A observação de que modelos especializados (FinBERT) e generalistas otimizados (ModernBERT) podem apresentar desempenhos estatisticamente indistinguíveis, sugerindo que diferentes abordagens tecnológicas podem convergir para resultados similares neste domínio e contexto
- **Indícios de Especialização por Tipologia:** A sugestão de que o FinBERT pode apresentar maior acuidade em manchetes técnicas, enquanto modelos como ModernBERT e RoBERTa podem sobressair em estruturas sintáticas mais complexas.
- **Estimativa de Ruído Intrínseco:** O uso da "Acurácia de Unanimidade" (63%) como um parâmetro para tentar estimar o ruído inerente ao *dataset*, indicando que uma parcela significativa dos rótulos pode não ser dedutível apenas pelo conteúdo textual.

Para o desenvolvimento de pesquisas futuras, propõem-se as seguintes direções:

1. Exploração de Modelagem Híbrida: Investigar a integração de séries temporais ao contexto textual para tentar capturar a "expectativa" do mercado de forma multimodal.
2. Uso de Rótulos Enriquecidos: Avaliar datasets que incorporem justificativas anotadas (Chain-of-Thought) visando reduzir a ambiguidade interpretativa
3. Avaliação de Funções de Perda Alternativas: Investigar técnicas como Asymmetric Loss, buscando desacoplar o gradiente de classes sem incorrer em possíveis sobrecorreções, visando superar o teto de desempenho identificado nesta pesquisa.

REFERÊNCIAS

- ALI, T.; YIN, Y.; ZHANG, Y.; HUANG, T. **Deep learning-based models for forecasting financial time series: a comprehensive survey.** *Machine Learning and Knowledge Extraction, Basel*, v. 5, n. 4, p. 723-741, 2023.
- ARACI, D. **FINBERT: Financial sentiment analysis with pre-trained language models.** arXiv preprint, 2019. Disponível em: <https://arxiv.org/abs/1908.10063>. Acesso em: 4 abr. 2026.
- BAY, Y. Y.; YEARICK, K. A. **Machine learning vs deep learning: the generalization problem.** arXiv preprint, 2024. Disponível em: <https://arxiv.org/abs/2403.01628>. Acesso em: 6 abr. 2026.
- BISHOP, C. M. **Pattern recognition and machine learning.** New York: Springer Science+Business Media, 2006.
- BOJANOWSKI, P. et al. **Enriching word vectors with subword information.** *Transactions of the Association for Computational Linguistics*, v. 5, p. 135-146, 2017.
- BROWN, T. B. et al. **Language models are few-shot learners.** arXiv preprint, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>. Acesso em: 6 maio 2026.
- DAO, T.; FU, D. Y.; ERMON, S.; RUDRA, A.; RÉ, C. **FlashAttention: fast and memory-efficient exact attention with IO-awareness.** *Advances in Neural Information Processing Systems*, v. 35, p. 16344-16359, 2022. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract.html. Acesso em: 01 fev. 2026
- DEEPGRAM. **Convolutional neural networks.** 2025. Disponível em: <https://deepgram.com/ai-glossary/convolutional-neural-networks>. Acesso em: 12 fev. 2025.
- DEVLIN, J. et al. **BERT: pre-training of deep bidirectional transformers for language understanding.** In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2019, Minneapolis. Anais [...]. Minneapolis: ACL, 2019. p. 4171-4186. Disponível em: <https://arxiv.org/abs/1810.04805>. Acesso em: 6 maio 2026.
- GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.** 2. ed. Sebastopol: O'Reilly Media, 2019.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction.** New York: Springer, 2009.
- ILEMOBAYO, J. A. et al. **Hyperparameter tuning in machine learning: a comprehensive review.** *Journal of Engineering Research and Reports*, v. 26, n. 6, p. 388-395, jun. 2024.
- JOULIN, A. et al. **Bag of tricks for efficient text classification.** In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15.,

2017, Valencia. Anais [...]. Valencia: ACL, 2017. p. 427-431. Disponível em: <https://arxiv.org/abs/1607.01759>. Acesso em: 6 maio 2026.

LEFORT, B. et al. **FinMarBa: a market-informed dataset for financial sentiment classification**. arXiv preprint, 2025. Disponível em: <https://arxiv.org/abs/2507.22932>. Acesso em: 4 jun. 2026.

LIN, T.-Y. et al. **Focal loss for dense object detection**. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2017, Venice. Anais [...]. Venice: IEEE, 2017. p. 2980-2988.

LIU, Y. et al. **RoBERTa: a robustly optimized BERT pretraining approach**. arXiv preprint, 2019. Disponível em: <https://arxiv.org/abs/1907.11692>. Acesso em: 6 maio 2026.

MALO, P. et al. **Good Debt or Bad Debt: Detecting semantic orientations in economic texts**. Journal of the Association for Information Science and Technology, Hoboken, v. 65, n. 4, p. 782-796, 2014.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. 4. ed. Harlow: Pearson, 2016.

SALAMI, H. et al. **ModernBERT: a new high-performance BERT baseline**. arXiv preprint, 2024. Disponível em: <https://arxiv.org/abs/2412.13663>. Acesso em: 17 mar. 2026.

SICHMAN, J. S. **Inteligência artificial e sociedade: avanços e riscos**. Estudos Avançados, São Paulo, v. 35, n. 103, p. 37-49, 2021.

VASWANI, A. et al. **Attention Is All You Need**. Advances in Neural Information Processing Systems, Long Beach, v. 30, p. 5998-6008, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 6 maio 2026.

RANI, V. et al. **Self supervised learning: a succinct review**. Archives of Computational Methods in Engineering, 2023. Disponível em: <https://link.springer.com/article/10.1007/s11831-023-09743-4>. Acesso em: 6 maio 2025.

ZHOU, Z.-H. **Machine learning**. Nanjing: Tsinghua University Press, 2016.

APÊNDICE A - Principais bibliotecas utilizadas.

Biblioteca	Descrição
Numpy	Biblioteca para computação numérica eficiente. Oferece suporte a arrays multidimensionais e funções matemáticas de alto desempenho para operações em vetores e matrizes.
Pandas	Focada em análise e manipulação de dados, fornece estruturas como DataFrame e Series, com funcionalidades para limpeza, filtragem, agregação e transformação de dados tabulares.
Scikit Learn	Biblioteca para <i>machine learning</i> em Python. Oferece algoritmos de classificação, regressão, clustering e ferramentas para validação de modelos e engenharia de características.
Statsmodels	Voltada para análise estatística, fornece ferramentas para modelagem estatística (como regressão linear e ARIMA), testes estatísticos e análise de séries temporais.
TensorFlow	Framework para <i>machine learning</i> e deep learning desenvolvido pelo Google. Suporta redes neurais, aprendizado por reforço e operações matemáticas em GPUs e CPUs.
Matplotlib	Biblioteca para visualização de dados. Permite criar gráficos estáticos, interativos e personalizados, como histogramas, gráficos de linha e barras.
Seaborn	Biblioteca para visualização de dados semelhante ao Matplotlib. Permite criar gráficos mais personalizáveis.