LUCAS LEME SANTOS

FINBERTPTBR: ANÁLISE DE SENTIMENTOS DE TEXTOS EM PORTUGUÊS REFERENTES AO MERCADO FINANCEIRO

São Paulo
2022

LUCAS LEME SANTOS

FINBERTPTBR: ANÁLISE DE SENTIMENTOS DE TEXTOS EM PORTUGUÊS REFERENTES AO MERCADO FINANCEIRO

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro da Computação.

São Paulo
2022

LUCAS LEME SANTOS

FINBERTPTBR: ANÁLISE DE SENTIMENTOS DE TEXTOS EM PORTUGUÊS REFERENTES AO MERCADO FINANCEIRO

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro da Computação.

Orientadora:
Anna Helena Reali Costa

São Paulo
2022

AGRADECIMENTOS

Agradeço aos meus pais, por todo o apoio e incentivo durante a minha vida.

Agradeço a minha orientadora, Professora Doutora Anna Helena Reali Costa, por todo o ensino e apoio durante o meu projeto. E também agradeço aos demais colaboradores pelos conselhos e auxílios ao longo do projeto: Julia Pocciotti, Kevin Ujiie, Thomas Ferraz, Vinícius Carmo, Professor Doutor Fábio Levy.

RESUMO

Esse trabalho se propõe a aprimorar o processo de tomada de decisões no mercado financeiro por meio do uso de inteligência artificial. Para isso, foi desenvolvido um modelo de análise de sentimentos de textos em português, utilizando a arquitetura de rede neurais BERT. Com a utilização desse modelo usuários podem processar grandes quantidades de dados de forma rápida e eficiente, obtendo informações relevantes para a tomada de decisões. O modelo foi treinado em duas principais etapas: modelagem de linguagem e modelagem de sentimentos. Na primeira etapa, foi treinado um modelo de linguagem com mais de 1,4 milhões de textos de notícias financeiras em português. A partir desse primeiro treinamento, foi possível construir um classificador de sentimentos com poucos textos rotulados (500) com uma convergência satisfatória. Ao final do trabalho, apresenta-se uma análise comparativa com outros modelos e também as possíveis aplicações do modelo desenvolvido. Na análise comparativa, foi possível observar que o modelo desenvolvido apresentou resultados superiores aos atuais modelos no estado da arte. Dentre as aplicações, foi demonstrado que o modelo pode ser utilizado para a construção de índices de sentimento, estratégias de investimento e análise de dados macroeconômicos, como a inflação.

Palavras-Chave – Processamento de Linguagem Natural, *Transformers*, Finanças Quantitativas.

ABSTRACT

This work aims to improve the decision-making process in the financial market through the use of artificial intelligence. To do this, a model of sentiment analysis of texts in Portuguese was developed, using the BERT neural network architecture. With the use of this model, users can process large amounts of data quickly and efficiently, obtaining relevant information for decision-making. The model was trained in two main stages: language modeling and sentiment modeling. In the first stage, a language model was trained with more than 1.4 million texts of financial news in Portuguese. From this first training, it was possible to build a sentiment classifier with few labeled texts (500) that presented a satisfactory convergence. At the end of the work, a comparative analysis with other models and the possible applications of the developed model are presented. In the comparative analysis, it was possible to observe that the developed model presented better results than the current models in the state of the art. Among the applications, it was demonstrated that the model can be used to build sentiment indices, investment strategies and macroeconomic data analysis, such as inflation.

Keywords – Natural Language Processing, *Transformers*, Quantitative Finance.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de capitais é algo vivo e muito dinâmico; ele reflete todas as informações e emoções da população, assim como dita a hipótese de mercados adaptativos [1]. Que diz que os preços dos ativos refletem todas as informações e emoções da população, e que os preços são adaptativos, ou seja, mudam de acordo com as novas informações e emoções. Sendo assim, os agentes que atuam nessa área devem ter uma excelente capacidade de processamento de informações, a fim de tomar as melhores decisões, evitando vieses comportamentais.

Com o aumento da quantidade de informações, uma ferramenta muito importante para auxílio na tomada de decisões é a inteligência artificial (IA). Essa tecnologia visa implementar sistemas que imitam a inteligência humana. Existem diversas aplicações de IA: visão computacional, processamento de linguagem natural (PLN), sistemas de recomendação, entre outros. O interesse desse projeto reside em aplicar PLN no contexto do mercado financeiro, facilitando assim a tomada de decisões em vista da abundante quantidade de informações textuais existentes.

Atualmente, os algoritmos no estado da arte em PLN são baseados em redes neurais profundas com mecanismos de atenção [2] e a maioria desses modelos são treinados na língua inglesa. Tais algoritmos se mostram bastante úteis para a realização de *fine-tuning* e atingem desempenho no estado da arte em diversas tarefas [3]. Porém, no contexto de finanças, esses modelos de propósito geral não convergem tão bem por conta do vocabulário específico do mercado financeiro. Portanto, para atingir resultados no estado da arte em finanças, é necessário o treinamento de modelos em tal domínio [4].

1.1 Objetivo

Este trabalho visa realizar o treinamento e análise de um modelo de linguagem do estado da arte no contexto do mercado financeiro em português. Para realizar esse objetivo, será necessária a aquisição e tratamento de dados textuais no domínio financeiro.

Por fim, para a avaliação do modelo será realizado um *fine-tuning* para a tarefa de análise de sentimentos, que posteriormente poderá ser utilizado para a construção de sinais para análise e construção de estratégias de investimentos.

1.2 Justificativa

Esse projeto tem como o seu principal foco o processamento de informações textuais para a análise e tomada de decisões no contexto de mercado de capitais. Para isso serão construídas bases de dados e algoritmos de PLN em linguagens com poucos recursos como português. A solução proposta por esse trabalho visa facilitar a obtenção de retornos acima da média por agentes do mercado financeiro e aumentar a disponibilização de modelos e bases de dados para PLN em português brasileiro. Portando, as principais justificativas são baseadas em finanças comportamentais e a escassez de algoritmos e dados em linguagens com recursos reduzidos, conforme explicado a seguir.

1.2.1 Finanças comportamentais

No mercado de capitais, define-se a busca de alfa como os retornos gerados acima de um *benchmark*. Assim, os agentes financeiros buscam constantemente por formas para "bater o mercado", e com isso gerar alfa para os seus investidores. Durante o processo decisório, com objetivo de obter lucro, os investidores realizam operações com base em suas expectativas futuras acerca dos ativos. Desse modo, para obter alfa, os agentes devem possuir expectativas melhores que o mercado, ou seja, adiantar a precificação dos ativos a partir de suas hipóteses de investimentos. A Figura 1 apresenta a formas de obtenção de alfa de cada gestor de investimentos.

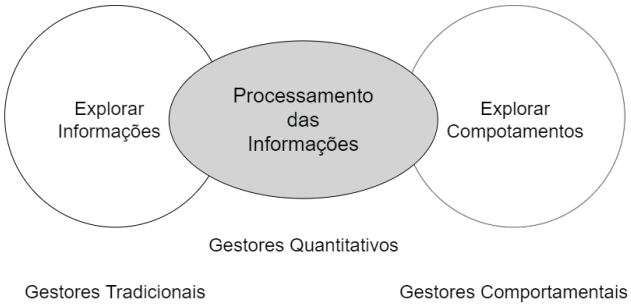

Figura 1: As 3 fontes de alfa no mercado de capitais.

A literatura de finanças comportamentais aponta que existem 3 principais fontes de

alfa [5]: informação superior, vieses comportamentais e melhor processamento de informação, conforme descrito a seguir.

Informação superior – Utilizada por gestores tradicionais que por possuírem um profundo conhecimento acerca do mercado podem realizar estimativas acerca do futuro dos ativos, gerando assim uma informação superior. Um exemplo da aplicação dessa metologia é a realização de projeção de lucratividade de um setor da economia.

Vieses comportamentais – Baseado na teoria dos mercados adaptativos, gestores comportamentais assumem que os agentes que operam no mercado nem sempre são racionais, e por conta disso cometem erros comportamentais que podem ser explorados financeiramente. Tais erros geralmente ocorrem em momentos de grande euforia, quando os investidores agem pela emoção e realizam uma precificação errada a acerca dos ativos.

Melhor processamento de informações – Alguns investidores assumem que as informações públicas disponíveis são suficiente para o processo decisório de investimento, e focam em melhorar a utilização desses dados. Gestores quantitativos seguem esse processo utilizando técnicas de modelagem matemática e estatística, assim podem identificar anomalias comportamentais ou realizar uma larga escala de estimativas para todas as empresas da bolsa.

1.2.2 Modelos de processamento de linguagem natural em português

Um problema atual em PLN é a construção de bases de dados e algoritmos em linguagens com poucos recursos. Por outro lado, com o avanço das pesquisas de PLN, muitas técnicas que eram baseadas em regras estáticas passaram a ser baseadas em modelos estatísticos. Entretanto, a maioria das pesquisas de modelagem focam na construção de algoritmos de apenas 20 línguas das mais de 7000 línguas do mundo, [6], e essas linguagens carentes de estudos são denominadas linguagens de poucos recursos (ou, em inglês, *low-resource languages*).

O português, apesar de possuir diversos avanços em PLN, ainda é considerada uma linguagem de poucos recursos, pois existe uma deficiência de bases de dados para modelos supervisionados e uma carência de modelos treinados unicamente na língua portuguesa. Pelo *Hugging Face* [7], uma plataforma de disponibilização de modelos de PLN, pode-se estimar a quantidade de modelos no estado da arte publicados na língua inglesa e

portuguesa: por exemplo, existem cerca de 6000 algoritmos disponíveis, já em português há apenas 200, sendo a maioria deles treinados em múltiplas linguagens. Portando, para aumentar o número de recursos disponíveis em português, esse trabalho visa construir bases de dados e algoritmos no contexto do português brasileiro.

1.3 Organização do Trabalho

A seguir, no capítulo 2 estão expostos os conceitos fundamentais para o desenvolvimento do trabalho. São abordados os conceitos técnicos dos algoritmos, os processos aplicados nos dados utilizados no projeto e por fim a descrição de tópicos relacionados a finanças quantitativas.

Na sequência, no capítulo 3 é descriminada a metodologia empregada no desenvolvimento do projeto, sendo descritos os principais requisitos do projeto, divididos em requisitos funcionais e não funcionais.

No capítulo 4, são apresentados os principais passos adotados durante o decorrer do desenvolvimento do projeto, desde as tecnologias utilizadas até os resultados obtidos a partir da modelagem de dados.

Por fim, no capítulo 5 são discutidas as considerações finais do projeto e trabalhos futuros.

2 ASPECTOS CONCEITUAIS

Nesta seção serão apresentados os conceitos fundamentais para o desenvolvimento do trabalho. Serão expostos tantos os conceitos técnicos dos algoritmos abordados, quanto os processos aplicados nos dados utilizados no projeto.

Na primeira parte estão cobertos os principais conceitos sobre redes neurais, na segunda serão discutidos os principais avanços em processamento de linguagem natural, na seguinte estão cobertos conceitos referentes a anotação de bases de dados. E por fim, na última parte, são abordados tópicos relevantes sobre o mercado de capitais.

2.1 Redes Neurais Artificiais

Redes neurais artificiais são poderosos modelos de aprendizado inspirados em estruturas de organismos inteligentes, em que o seu objetivo é aproximar uma função $y = f(x; \theta)$, que mapeia dados independentes (x) em variáveis dependentes (y), onde θ representa os parâmetros do modelo [8]. A estrutura desse algoritmo é constituída de diversas camadas de neurônios, as unidades básicas do modelo.

Um neurônio é composto por uma ponderação das entradas e uma função de ativação. Os sinais de entrada da rede são ponderados por pesos, que indicam a contribuição de cada sinal de entrada para o neurônio. Assim, com objetivo de capturar relações não lineares entre os dados, é aplicada uma função de ativação não linear no sinal ponderado, resultando assim no valor de saída (Figura 2), que pode ser utilizado para tarefas de regressão ou classificação.

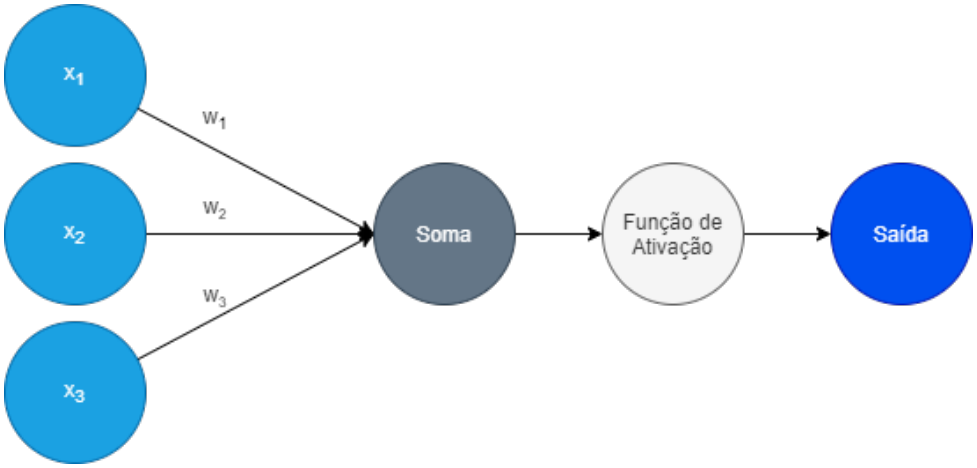

Figura 2: Neurônio de uma rede neural artificial.

A arquitetura básica de uma rede neural possui três categorias de camadas: *Input Layer*, *Hidden Layer* e *Output Layer*. Os neurônios presentes na *Input Layer* cumprem um papel passivo de repassar os sinais iniciais para as demais camadas da *Hidden Layer*. Por fim, a última camada da rede neural é chamada de *Output Layer*, interface onde que os sinais da *Hidden Layer* são transmitidos para a camada de saída.

A composição das funções da rede neural podem ser descritas como um grafo acíclico de funções em cadeia [8]. Por exemplo, uma rede com 3 camadas pode ser representada pelas funções $f^{(1)}$, $f^{(2)}$ e $f^{(3)}$, que em forma de cadeia são representadas da seguinte forma $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. Em geral, quanto maior o tamanho da cadeia, maior é a profundidade da rede e maior é complexidade que o algoritmo consegue capturar. O termo *Deep Learning* surgiu dessa terminologia de profundidade.

O treinamento de uma rede neural é um processo iterativo de tentativa e erro composto por duas principais etapas: *feedfowarding* e *backpropagation*. Na etapa de *feedfowarding* são realizadas as predições, com os parâmetros de peso θ, e a partir de uma função de custo $J(\theta)$ são avaliadas as predições. E logo após, durante o *backpropagation* é executada a otimização dos pesos θ visando minimizar a função de custo $J(\theta)$. Assim é desempenhado o processo de alternância entre *feedfowarding* e *backpropagation* com o objetivo da convergência da rede neural.

2.2 Processamento de Linguagem Natural

Processamento de linguagem natural (PLN) é a área da ciência de dados focada na geração e compreensão das línguas humanas. Essa área de pesquisa teve seu início em

1950, como uma intersecção entre a inteligência artificial e a linguística [9]. Atualmente, o campo de estudo de PLN é vasto e complexo, e é pautando principal pela aplicação de técnicas de machine learning.

2.2.1 Modelos de Linguagem

As técnicas iniciais envolvendo modelagem de dados em PLN tratavam elementos textuais como unidades isoladas, isto é, não havia noção de similaridade ou relação semântica. Com o progresso do poder computacional tornou-se possível treinar modelos cada vez mais complexos, sendo um dos principais avanços em PLN o emprego de redes neurais para modelagem de textos [10], de forma que se tornou possível treinar modelos que aprendem informações de similaridade e semântica dos dados.

Modelo de linguagem é aquele modelo que dado um texto de contexto estima a probabilidade da próxima palavra ou frase. Existem diversas técnicas para o treinamento dessa categoria de modelo, dentre as principais são as probabilísticas e as baseadas em redes neurais, conforme explicado a seguir:

Modelos Probabilísticos – As técnicas probabilísticas visam estimar as probabilidades condicionais dos textos presentes em um corpus, sendo que, uma forma de estimar essas probabilidades é por meio do número de ocorrências dos textos diante dos contextos. Essa estimativa de probabilidade é gerada a partir de todos os pares (contexto, texto) presentes em um corpus. Um exemplo resultante dessa estimativa: dado o contexto "A maçã caiu da" a probabilidade da palavra árvore é dada por: $P(\text{Árvore}|\text{A maçã caiu da})$.

Uma técnica probabilística para estimar a probabilidade de um texto é o TFIDF (*Term Frequency - Inverse Document Frequency*). O TFIDF é uma técnica de normalização de frequência de palavras, cujo objetivo é diminuir a influência de palavras comuns em um corpus. A técnica é calculada a partir da multiplicação da frequência de uma palavra no contexto pelo inverso da frequência da palavra no corpus. A equação do TFIDF é:

$$\text{TFIDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t), \tag{2.1}$$

onde t é a palavra e d é o documento.

A frequência de uma palavra no contexto é dada por:

$$\mathrm{TF}(t,d) = \frac{\text{Número de ocorrências de } t \text{ no contexto } d}{\text{Número total de palavras no contexto } d} \tag{2.2}$$

e a frequência de uma palavra no *corpus* é dada por:

$$\mathrm{IDF}(t) = \log \frac{\text{Número total de documentos no corpus}}{\text{Número de documentos que contém } t}. \tag{2.3}$$

Modelos Baseados em Redes Neurais – Modelos de linguagem baseados em redes neurais também visam estimar as probabilidades condicionais, porém, realizam isso a partir da seguinte aproximação da rede neural: $texto = f(contexto; \theta)$. Assim, através dos valores resultantes da *Hidden layer*, é possível extrair a representação vetorial dos textos, o denominado *word embedding*.

Uma característica muito importante dos *word embeddings* é a possibilidade de operações vetoriais com os textos. No artigo do algoritmo Word2Vec [10], o autor demonstra o aprendizado das relações semânticas por meio das operações vetoriais como: $\vec{\text{Rei}} - \vec{\text{Homem}} + \vec{\text{Mulher}} \approx \vec{\text{Rainha}}$.

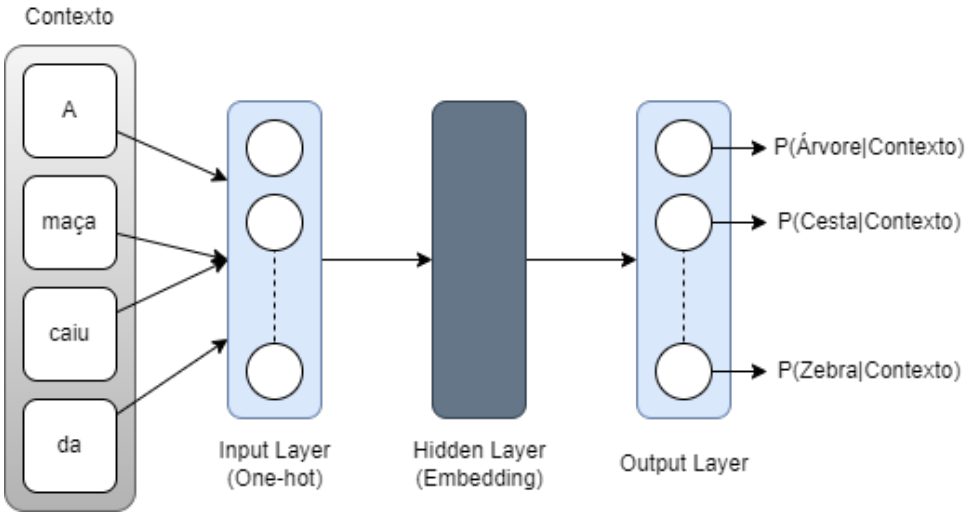

Figura 3: Modelo de linguagem baseado numa arquitetura de rede neural.

Uma das principais vantagens dos modelos de linguagem é o seu processo de treinamento, que é semi-supervisionado. Tal metodologia de treinamento possui uma variável alvo que pode ser construída a partir dos dados de entrada, que no caso de modelos de linguagem, os dados de treinamento são contexto e texto, e ambos são construídos a partir do corpus. Portando, esse método torna-se bastante útil, pois possibilita a utilização de grandes bases de dados, visto que não é necessária a anotação manual dos dados.

Outra característica muito importante dos modelos de linguagem é a sua grande utilidade para *transfer learning*. Durante o pré-treinamento a rede neural aprende diversas relações semânticas e sintáticas, consolidando assim um vasto conhecimento acerca da linguagem. Esse conhecimento facilita a convergência durante o treinamento para tarefas com diferentes escopos, o que é ideal para *transfer learning*.

2.2.2 Modelos de Atenção

Outra técnica muito importante que as redes neurais em estado da arte utilizam são os mecanismos de atenção, introduzidos no artigo *Attention is all you need* [2]. Essa técnica visa identificar partes mais importantes de um determinado texto para que a rede neural foque nas partes mais relevantes.

A técnica é fundamentada na operação de *Self Attention*, ilustrada na Figura 4, que recebe n vetores de *embedding* e retorna n vetores contextualizados com a importância de cada palavra para sentença. A fim de obter essa ponderação são utilizados três mecanismos de atenção baseados nos vetores de entrada: *Query*, *Key* e *Value*. A ideia central desses mecanismos é baseada em sistemas de busca, em que um usuário pesquisa um termo (*Query*) dentre diversas possibilidades (*Key*) sendo retornado o valor (*Value*) mais próximo ao procurado. Por analogia, no contexto de PLN o algoritmo visa aprender quais palavras do texto são mais relevantes para retornar o valor procurado. Assim, para que esse processo seja aprendido pela rede neural, cada mecanismo de atenção (*Query*, *Key* e *Value*) possui uma matriz de pesos associada, assim é possível realizar o processo de convergência das partes mais importante dos *embeddings* de uma sentença.

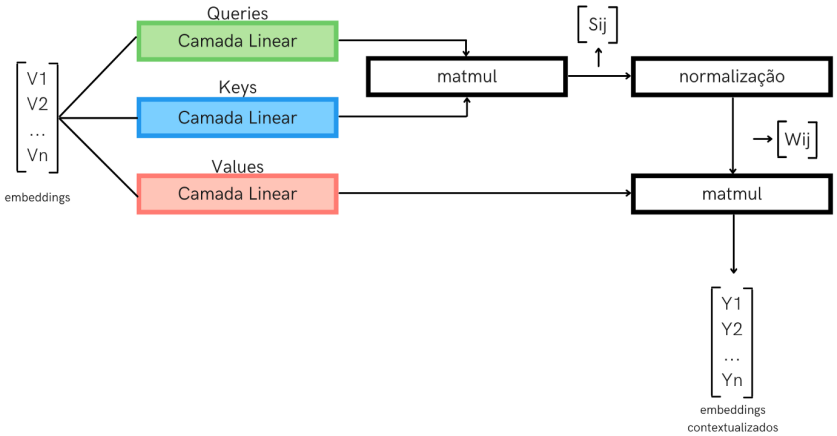

Figura 4: Arquitetura da camada de *Self Attention*.

Outro conceito introduzido pelo artigo *Attention is all you need* [2] são as camadas de

Multi-head Attention, que servem como uma expansão da contextualização do mecanismo de *Self Attention*. Ao realizar essa expansão, o modelo pode aplicar os mecanismos de atenção para diversas partes do texto paralelamente. Para isso, basta adicionar h camadas de *Self Attention*, o que resultará em h vetores com diferentes contextualizações. Com isso, ao realizar diferentes ponderações de importância das palavras, o modelo pode identificar múltiplas relações semânticas dentro de um texto.

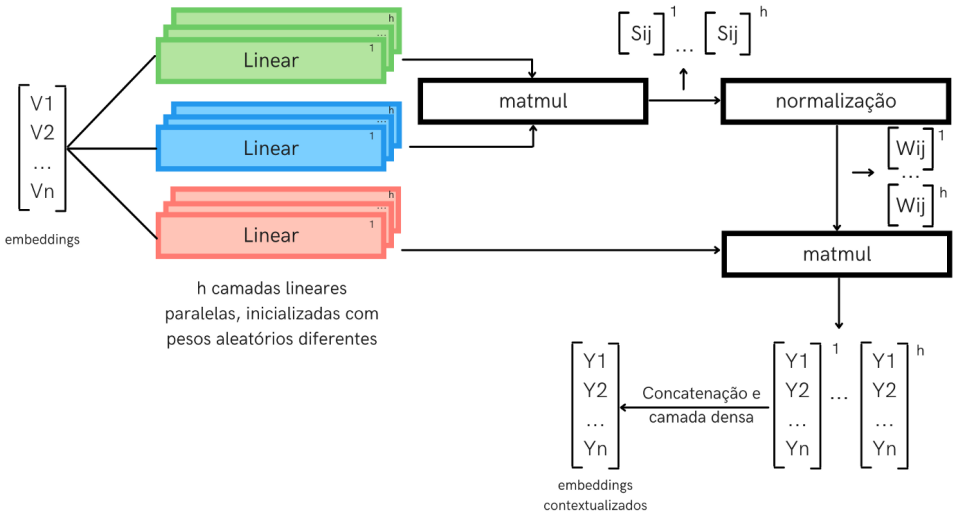

Figura 5: Arquitetura da camada de *Multi-head Attention*.

2.2.3 Análise de Sentimentos

Análise de sentimentos é um campo de PLN que implementa algoritmos para determinar sentimentos a partir de textos [11]. Geralmente essa técnica é bastante empregada para análises de *reviews* de produto do varejo, mas também pode ser empregada para análises de notícias e discursos políticos.

Existem 3 principais segmentações acerca da tarefa de classificação de sentimentos: nível de documento, nível de frase e ao nível de aspectos. A classificação ao nível de documento é responsável por determinar o sentimento de um conjunto de textos, já ao nível de frase a determinação de sentimento é para um único texto. Apesar da diferença dos textos alvos, não há nenhuma distinção fundamental para a construção de algoritmos de análise de sentimentos para nível de documento e sentença. A tarefa de análise de sentimento ao nível de aspecto visa definir o sentimento de uma frase em seus diversos pontos de vista, sendo um possível exemplo a frase "A qualidade do celular é boa, porém a bateria é ruim", essa frase possui dois sentimentos, sendo o positivo em relação à qualidade do aparelho e o negativo, em relação à bateria. Portanto, para construir algoritmos para

a tarefa ao nível de aspecto é necessário a identificação das principais entidades da frase.

Dentre as principais tarefas empregadas de análise de sentimentos estão [12]: composição de sentimentos, análise temporal e de emoções, conforme descrito a seguir.

Composição de sentimentos – Tarefas de composição de sentimentos consideram que cada constituinte de um texto pode representar uma orientação de sentimento, e portanto, o sentimento de uma sentença depende de todos os seus elementos. Além disso, um dos principais desafios na construção de algoritmos para essa categoria de tarefa são as polaridades presentes nos textos, como em frases "Não feliz" e "Um feliz acidente" temos sentimentos divergentes, dificultando a classificação do sentimento da frase.

Análise temporal – Outra dimensão de análise empregada é o sentimento ao longo do tempo. Com o decorrer do tempo, indivíduos têm seus pontos de vista alterados, e com o uso de algoritmos de análise de sentimento é possível analisar o passado ou até mesmo prever tendência de sentimentos.

Análise de emoções – Outro nível de análise que se aproxima de sentimentos e pensamentos humanos é a tarefa de análise de emoções. Algoritmos construídos para essa categoria de tarefa visam classificar as principais emoções humanas como são: amor, felicidade, alegria, surpresa, raiva, triste e medo.

2.3 Técnicas de validação de dados e modelos

Para realizar o treinamento de um modelo de aprendizado supervisionado é necessária uma boa definição da variável a ser predita. Para casos em que está variável não está classificada é preciso realizar o processo de anotação manual, em que avaliadores catalogam a base de dados. Esse processo deve seguir uma metodologia de anotação que será descrita nesta seção.

Outro fator relevante é a validação de modelos para garantir boa qualidade dos modelos de aprendizado de máquina. Estes aspectos também são considerados nesta seção.

2.3.1 Anotação de dados

Para a construção de uma base por **anotação manual** é ideal que o processo de classificação seja realizado por mais de uma pessoa para evitar vieses na base de dados.

Esse processo possui uma metodologia recomendada [13], que pode ser aplicada pelas seguintes etapas:

1. Seleção de textos representativos.
2. Definição clara sobre as classes a serem classificadas.
3. Anotação inicial coletiva de uma amostra da base de textos.
4. Avaliação do nível de concordância dos avaliadores.
   (a) Caso o nível de concordância não seja suficiente, retornar para a etapa 2;
   (b) Caso o nível de concordância seja suficiente, prosseguir para etapa 5.
5. Anotação do "corpus" inteiro conduzindo análises de concordância contínuas.
6. Treinamento de modelo realizando técnicas de validação cruzada para avaliação da base de textos classificados.

2.3.1.1 Seleção de Textos representativos

Um processo que pode auxiliar a tarefa de anotação manual é modelagem de tópicos. Com ela, é possível selecionar textos representativos antes da classificação pelos anotadores. Artigos apontam haver um aumento de eficiência quando anotadores utilizam informações de modelos de tópicos [14].

Técnicas tradicionais de classificação de textos requerem uma abundante quantidade de textos classificados, e em linguagens com pouco recursos o treinamento de modelos e a construção de bases de dados se torna mais difícil. Com isso, a modelagem de tópicos é uma técnica de aprendizado não supervisionado que visa identificar os principais tópicos presentes em um conjunto de documentos, ou seja, é um processo de classificação automática. O processo de classificação dessa técnica pode ser efetuado de duas maneiras: não supervisionada, onde o modelo determina os principais tópicos apenas a partir dos documentos, e *Zero-shot learning* em que são definidas as classes que o modelo deve agrupar os documentos.

A configuração de *Zero-shot learning* em PLN visa realizar a seguinte tarefa: dados n documentos $D = \{d_1, d_2, ...d_n\}$ e m classes a serem classificadas $L = \{l_1, l_2, ...l_m\}$ estima-se a probabilidade de pertencimento do documento d_i na classe l_i. Existem dois modelos que abordam essa categoria de problema: XLM-R Transformer [15] e ZeroBERTo

[16]. O primeiro é treinado em um escopo de linguagens com grandes recursos de dados, como o inglês, já o segundo tem uma arquitetura baseada em clusterização e modelos de linguagem, permitindo sua implementação em linguagem com poucos recursos.

Assim, ao utilizar tais modelos é possível definir a probabilidade de pertencimento para cada classe a ser avaliada pelos anotadores. Portando, com essas informações extras, fica facilitada a tomada de decisão para definir a classe de cada texto.

2.3.1.2 Métricas de concordância

Uma importante etapa no processo de anotação manual é a realização de análises do nível de concordância dos avaliadores. A análise do nível de concordância pode ser realizar por 3 métricas [17]: porcentagem de concordância, *Fleiss's Kappa* e *Krippendorff's alpha*, descritas a seguir.

Porcentagem de concordância: A porcentagem de concordância é a métrica mais simples, e visa calcular dentre todos os textos anotados qual foi a porcentagem de concordância entre os avaliadores. Por conta da sua simplicidade, a porcentagem de concordância possui algumas desvantagens como: viés para bases de dados desbalanceadas e baixa robustez para mais de 2 avaliadores. Essa métrica pode ser calculada da seguinte forma:

$$\%\text{Concordância} = \frac{1}{N}\sum_{i=1}^{N} agr_i, \tag{2.4}$$

com

$$agr_i = \begin{cases} 1, & \text{Se os avaliadores classificaram o texto na mesma categoria} \\ 0, & \text{Se os avaliadores não classificaram o texto na mesma categoria} \end{cases}. \tag{2.5}$$

Fleiss's Kappa A métrica *Fleiss's Kappa* utiliza lógica combinatória para medir a concordância em projetos que possuem mais de 2 anotadores. Uma das principais vantagens dessa métrica é a possibilidade de um número variável de anotadores para um determinado texto. A métrica de *Fleiss's Kappa* para um texto anotado é dada por:

$$agr_i = \frac{1}{c(c-1)}\sum_{k\in K} n_{ik}(n_{ik}-1), \tag{2.6}$$

onde c é o número de avaliadores, K o conjunto de classes de anotação e n_{ik} o número de classes anotadas como k no texto i. A concordância para todos os textos

classificados pode ser computada pela média do agr_i de todos os textos. Para o caso de um número de avaliadores constante para todos os textos, o nível de concordância é dado por:

$$\%\text{Concordância} = \frac{1}{N}\sum_{i \in N}\frac{1}{c(c-1)}\sum_{k \in K} n_{ik}(n_{ik}-1). \tag{2.7}$$

***Krippendorff's alpha*:** O *Krippendorff's alpha* visa estimar o nível de concordância para o caso de múltiplos avaliadores e sua metodologia é baseada apenas na distribuição das anotações, possibilitando o seu cálculo diversas categorias de dados além de categóricos. Sua metodologia é baseada em duas principais variáveis, D_o que é a variança amostral das respostas dos anotadores e D_e que é a variança esperada dos anotadores computada a partir das frequências das classes presentes na base dados:

$$\alpha = 1 - \frac{D_o}{D_e}. \tag{2.8}$$

Quando todos os avaliadores concordam, a variança amostral será $D_o = 0$, resultando assim num $\alpha = 1$, indicando uma alta confiabilidade na base de anotações. Um $\alpha = 0$ ocorre quando não possível distinguir as anotações de dados puramente aleatórios, pois a variança amostral das classificações será alta, indicando assim uma baixa confiabilidade.

2.3.2 Validação de modelos

A validação de modelos é um processo importante para garantir a qualidade dos modelos de aprendizado de máquina. Tal garantia de qualidade é importante para a verificação da capacidade de generalização do modelo, ou seja, a capacidade do modelo de generalizar para dados não vistos durante o treinamento. A validação de modelos pode ser realizada de diversas formas, sendo as mais comuns a validação cruzada e a validação por *holdout*. Nesta seção serão discutidos os conceitos de viés e variância e técnicas de validação de modelos.

2.3.2.1 Viés e variância

O viés e a variância são duas medidas de erro que podem ocorrer durante o treinamento de um modelo de aprendizado de máquina. O viés é a diferença entre o valor esperado do modelo e o seu valor real, enquanto a variância é a variação do modelo em relação ao valor esperado. Idealmente o modelo deve ter um viés baixo e uma variância baixa, pois isso indica que o modelo está generalizando bem para dados não vistos.

Outro conceito importante relacionado ao viés e a variância é o de *overfitting* e *underfitting*. O *overfitting* ocorre quando modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para dados não vistos, assim o modelo apresenta uma variância alta por conta do sobre-ajuste aos dados de treinamento. Já o *underfitting* ocorre quando o modelo não consegue se ajustar bem aos dados de treinamento, assim o modelo apresenta um viés alto por não conseguir generalizar bem para dados não vistos. O cenário ideal de uma modelagem é quando o modelo consegue se ajustar bem aos dados de treinamento e generalizar bem para dados não vistos, assim o modelo apresenta um viés baixo e uma variância baixa, evitando assim o *overfitting* e o *underfitting*.

2.3.2.2 Técnicas de validação de modelos

Visando evitar o *overfitting* e o *underfitting* é importante realizar a validação de modelos para verificar a capacidade de generalização do modelo. A validação de modelos pode ser realizada de diversas formas, sendo as mais comuns a validação cruzada e a validação por *holdout*, que serão discutidas a seguir e são apresentadas na figura 6.

Figura 6: Validação por *holdout* e validação cruzada.

Validação por *holdout*: A validação por *holdout* é uma técnica de validação de modelos que consiste em dividir a base de dados em duas partes, uma para treinamento, e outra para teste. A base de dados é dividida em duas partes de forma aleatória, sendo que a proporção entre elas é definida pelo usuário. A base de dados de treinamento é utilizada para treinar o modelo e a base de dados de teste é utilizada para avaliar a capacidade de generalização do modelo a partir de métricas de avaliação.

Validação cruzada: A validação cruzada é uma técnica que consiste em dividir a base de dados em k partes, sendo $k-1$ partes utilizadas para treinamento e a restante para teste. O processo de validação cruzada é repetido k vezes, sendo que cada

parte é utilizada uma vez para teste e $k-1$ vezes para treinamento. Geralmente a validação cruzada é preferida em relação à validação por *holdout*, pois a validação cruzada é mais robusta e possui uma estimativa mais precisa da capacidade de generalização do modelo, visto que utiliza todos os dados para treinamento e teste. Porém, a validação cruzada é mais custosa computacionalmente, visto que o modelo é treinado k vezes, sendo que k é o número de partições da base de dados.

2.3.2.3 Métricas de avaliação

Além da técnica de divisão dos dados, é importante escolher as métricas que serão utilizadas para teste do modelo. As métricas são utilizadas para avaliar a desempenho do modelo, sendo que as métricas mais comuns são a acurácia, a precisão, a revocação e F1-Score.

A acurácia é a proporção de acertos do modelo, ou seja, a proporção de dados classificados corretamente pelo modelo. No entanto, a principal desvantagem dessa métrica é a avaliação modelos de classificação desbalanceados, pois ela não considera a distribuição de classes.

Em alguns casos, além de classificar corretamente uma classe, também importante reduzir o número de falsos positivos e falsos negativos, e para isso é importante avaliar a precisão e a revocação. A precisão é a proporção de acertos positivos, ou seja, a proporção de dados classificados como positivos que realmente são positivos, enquanto a revocação é a proporção de acertos verdadeiros, ou seja, dentre as amostras positivas quantas foram classificadas corretamente. Portanto, a precisão é utilizada quando é importante reduzir o número de falsos positivos, já a revocação é usada quando é importante reduzir o número de falsos negativos.

Para computar tais métricas utiliza-se a matriz de confusão, sendo esta uma tabela que mostra a quantidade de acertos e erros do modelo, como apresentado na figura 7. A partir da matriz de confusão, a precisão pode ser calculada da seguinte forma:

$$\text{precisão} = \frac{VP}{VP + FP}, \tag{2.9}$$

onde VP é o número de verdadeiros positivos e FP é o número de falsos positivos. Já a revocação pode ser calculado da seguinte forma:

$$\text{revocação} = \frac{VP}{TP + FN}, \tag{2.10}$$

onde FN é o número de falsos negativos.

Figura 7: Matriz de confusão.

A partir da precisão e revocação pode-se avaliar um enfoque na redução de falsos positivos ou falsos negativos, mas e no caso que a redução de ambos é importante? Para esse caso é utilizado o F1-Score, o qual é a média harmônica entre a precisão e a revocação, calculado da seguinte forma:

$$\text{F1-Score} = \frac{2 \times \text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}}. \tag{2.11}$$

Portanto, o F1-Score é uma métrica que considera a precisão e a revocação, sendo que quanto maior o F1-Score melhor é o modelo.

2.4 Finanças Quantitativas

Uma das áreas do mercado de capitais que mais cresce é a de **finanças quantitativas**. Os profissionais desta área visam aplicar modelagem matemática e programação para a resolução de problemas financeiros. Por exemplo, a concepção de um portfólio de investimentos, a avaliação de risco de crédito, a precificação de derivativos, a avaliação de ativos, entre outros problemas, podem ser resolvidos por meio de técnicas de finanças quantitativas. Para isso, é necessário um profissional com uma base de formação voltada à pesquisa.

Um dos tópicos de finanças quantitativas com mais pesquisa é a descrição da origem dos retornos dos ativos financeiros, ou seja, a descrição dos fatores que explicam os retornos dos ativos. A área de concentração desse tipo de pesquisa é a literatura de investimento de fatores — fortemente baseada nos trabalhos de William F. Sharpe [18] — que visa representar de maneira quantitativa o retorno dos ativos. Artigos apontam que existem

6 principais fatores que explicam os retornos [19]: valor, qualidade, alto rendimento, tamanho, baixa volatilidade e momentum. A seguir serão apresentados um resumo de cada um desses fatores.

- Valor: consiste em comprar ativos com baixo preço relativo ao seu valor intrínseco
- Qualidade: consiste em comprar ativos com alta qualidade, independente do seu valor intrínseco, ou seja, ativos com finanças saudáveis.
- Alto rendimento: consiste em comprar ativos com alta lucratividade.
- Tamanho: consiste em comprar ativos com alto valor de mercado, então, em geral são ativos de grandes empresas com alta liquidez e geralmente líderes de mercado.
- Baixa volatilidade: consiste em comprar ativos com baixa volatilidade, portanto com baixa variabilidade de preços, ou seja, ativos com baixo risco.
- Momentum: consiste em comprar ativos que tiveram alta rentabilidade no passado, ou seja, com a premissa de que ativos que tiveram alta rentabilidade no passado tendem a ter alta rentabilidade no futuro.

A maioria dos fatores aqui apresentados depende de informações externas do mercado, como, por exemplo, o valor intrínseco ou a lucratividade de um ativo, que são informações presentes no balanço patrimonial das empresas. Dada a dependência de informações externas, os fatores de investimento são considerados exógenos, ou seja, não podem ser explicados apenas pelas variáveis do sistema, que no caso são os preços dos ativos. Os ativos que apresentam uma exceção são *momentum* e baixa volatilidade, pois eles não dependem de informações externas, portanto são endógenos. Nosso interesse aqui consiste em analisar a relação do sentimento de notícias com os fatores de investimento endógenos, ou seja, *momentum* e baixa volatilidade. Esses fatores endógenos serão explicados mais a fundo a seguir.

2.4.1 Momentum

Momentum é uma palavra de origem latina, que na física clássica significa quantidade de movimento ou embalo, ele indica duas informações: a direção e a magnitude de um corpo em movimento. Além disso, outro elemento importante é a lei da inércia, a tendência natural de um objeto manter seu estado de repouso ou movimento. Com esses

dois conceitos, o *momentum* pode indicar para onde e o quão rápido um objeto está se movendo, e que se não houver interferência externa, será constante.

No contexto de finanças, o *momentum* é uma estratégia de investimento fundamentada no embalo dos preços, que considera a tendência dos preços para uma tomada de decisão de um investimento. Esse fator pode ser descrito por meio da equação do retorno de um ativo,

$$r_{t-k,t} = \frac{\text{Preço}_t - \text{Preço}_{t-k}}{\text{Preço}_{t-k}}, \tag{2.12}$$

onde k é o período do sinal.

A ideia central da estratégia de *momentum* [20] é que as tendências dos preços tendem a continuar para cima ou para baixo. O racional da estratégia é baseado em vieses comportamentais que apontam que investidores tendem a investir em ativos que estão se valorizando ou desinvestir naqueles que estão se desvalorizando. Artigos apontam que essa categoria de estratégia apresenta uma baixa correlação com o mercado e funciona há mais de 100 anos [21].

O retorno das estratégias de *momentum* pode ser representando pela equação do retorno de uma carteira de investimento:

$$r_{\text{carteira},t} = \sum_{s=1}^{S_t} w_t^s r_{t,t+1}^s, \tag{2.13}$$

onde S_t representa o número de ativos disponíveis até o tempo t; w_t^s o peso de cada ativo s na carteira. E $r_{t,t+1}^s$ é o retorno do ativo s entre os períodos t e $t+1$.

Existem duas principais categorias de estratégias de momentum: *Time Series Momentum* [22] e *Cross Sectional Momentum*. A principal diferença entre elas é a seleção dos ativos, sendo que na primeira todos os ativos são inseridos na carteira de investimento, e já na segunda apenas os ativos com os maiores momentums em absoluto são inseridos, conforme ilustra a Tabela 1, sendo σ_t^s a métrica de volatilidade do ativo (desvio padrão), e n_u e n_d, o número de ativos no quartil superior e inferior do sinal de momentum, respectivamente.

O desempenho histórico do *momentum* é bastante positivo, como pode ser visto na Figura 8, onde o gráfico mostra o retorno acumulado de uma carteira de investimento comprada na estratégia de momentum, isto é apenas o retorno de se investir em ativos com maior *momentum* positivo. O gráfico mostra que o retorno acumulado é bastante positivo, e que a estratégia de *momentum* é bastante robusta, pois apresenta um retorno acima do mercado. Um ponto negativo a ser destacado é que a estratégia de *momentum*

Tabela 1: Tipo de *momentum* e respectivo modelo de alocação

Tipo de Momentum	Modelo de alocação.
Time Series	$w_t^s = \frac{sign(r_{t-12,t}^s)}{S_t \sigma_t^s}$
Cross Sectional	$w_t^s = \begin{cases} \frac{1}{n_u}, \text{para ativos com } momentum \text{ no última quartil} \\ \frac{-1}{n_d}, \text{para ativos com } momentum \text{ no primeiro quartil} \\ 0, \text{para o restante dos ativos} \end{cases}$

comprada apresenta uma alta correlação com o mercado.

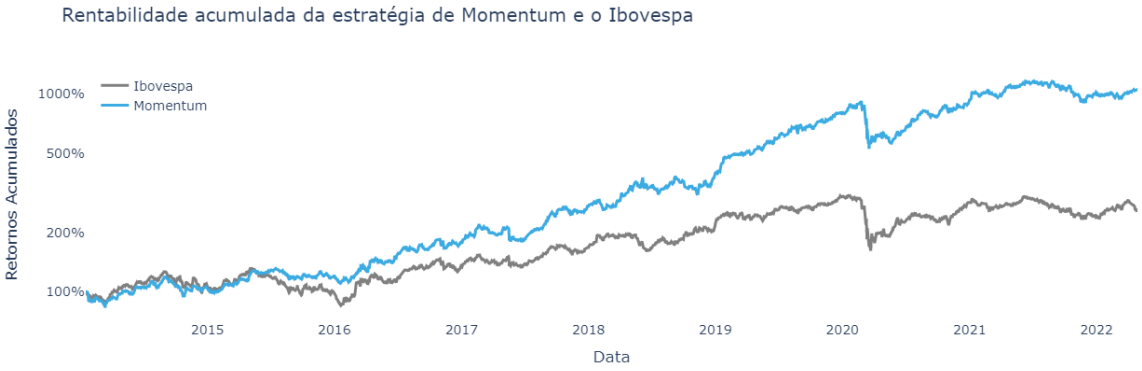

Figura 8: Retorno acumulado de uma carteira de investimento comprada na estratégia de momentum.

2.4.2 Baixa Volatilidade

Baixa Volatilidade é um fator de investimento que consiste em selecionar ativos com baixa volatilidade, ou seja, ativos que apresentam baixa variação nos preços. A ideia central é que ativos com baixa volatilidade tendem a apresentar retornos mais estáveis e menos risco para o investidor, visto que ao investir nesses ativos o investidor não arrisca perder muito dinheiro em um curto período.

A volatilidade é uma medida de risco, que indica a variação dos preços de um ativo e é geralmente medida pelo desvio padrão. Pois, o desvio padrão é uma medida de dispersão dos dados em relação a sua média, ou seja, quanto maior o desvio padrão, maior a variação dos preços. Portanto, ativos com baixa volatilidade são aqueles que apresentam menor variação nos preços.

A estratégia de baixa volatilidade também pode ser representada pela equação do retorno de uma carteira de investimento 2.13. A diferença é que, ao invés de selecionar

os ativos a partir dos retornos, selecionam-se os ativos a partir da volatilidade dos preços. Essa seleção é dada por meio dos quantis da volatilidade de todos os ativos,

$$w_t^s = \begin{cases} \dfrac{1}{n_u}, \text{para ativos com menor volatilidade no último quartil} \\ \dfrac{-1}{n_d}, \text{para ativos com menor volatilidade no primeiro quartil} \\ 0, \text{para o restante dos ativos} \end{cases} \tag{2.14}$$

sendo n_u e n_d, o número de ativos no quartil superior e inferior da volatilidade, respectivamente.

O desempenho histórico do fator de baixa volatilidade também é bastante positivo, como pode ser visto na Figura 9, onde o gráfico mostra o retorno acumulado de uma carteira de investimento comprada na estratégia de baixa volatilidade, isto é apenas o retorno de se investir em ativos com menor volatilidade. Diferente do momentum, o retorno dessa estratégia comprada apresenta uma menor correlação com o mercado.

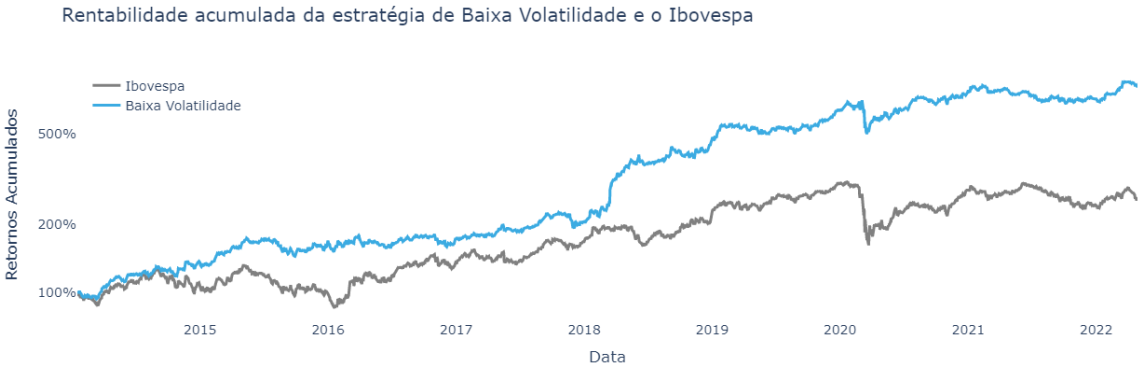

Figura 9: Retorno acumulado de uma carteira de investimento comprada na estratégia de baixa volatilidade

Outro tópico bastante pesquisado em finanças quantitativas é a avaliação de ativos, ou seja, como avaliar se um investimento é bom ou ruim. A avaliação de ativos é um tópico bastante importante, pois é a partir dela que se pode tomar decisões de investimento. Para isso, pode-se avaliar o investimento individualmente a partir de métricas ou pode-se avaliar o investimento relativo a outros ativos, ou seja, comparar o investimento com outros ativos.

Métricas de avaliação de ativos As métricas de avaliação de ativos avaliam o investimento individualmente, ou seja, não é necessário compará-lo com outros ativos. As métricas de avaliação de ativos mais comuns são o retorno, o risco e uma junção das duas

métricas, o *Sharpe Ratio*, o retorno ajustado ao risco. Tais métricas serão apresentadas a seguir.

- **Retorno:** Ganho ou perda de um investimento em um determinado tempo. Muito útil para determinar o ganho de capital, porém não avalia o risco incorrido ao longo do período.

- **Risco:** Medida de incerteza de um investimento, ou seja, quanto maior o risco, maior a incerteza de um investimento. No entanto, olhando apenas o risco não é possível avaliar se o investimento é bom ou ruim, pois o risco pode ser alto, mas ele não indica informações sobre o retorno. Para quantificar o risco, pode-se utilizar o desvio padrão, sendo uma medida de dispersão dos dados em relação a sua média, ou seja, quanto maior o desvio padrão, maior a variação dos preços.

- **Retorno Ajustado ao Risco:** Avalia o investimento a partir de unidades de retorno em relação ao risco, ou seja, quanto maior o retorno ajustado ao risco, melhor o investimento. Essa métrica é muito utilizada em finanças quantitativas, pois é uma métrica que avalia o investimento de forma mais completa. Uma forma de calcular tal métrica é a partir do *Sharpe Ratio*, que é determinado pela seguinte equação: $\frac{R_p - R_f}{\sigma_p}$, onde R_p é o retorno do investimento, R_f é o retorno livre de risco e σ_p é o desvio padrão do investimento. O retorno livre de risco é o retorno de um investimento sem risco, como, por exemplo, um investimento em títulos públicos. Esse retorno é utilizado para ajustar o retorno do investimento ao risco, pois se um ativo com alto risco apresentar um retorno menor que um ativo livre de risco, não faz sentido investir nesse ativo com alto risco, pois o retorno não compensa o risco.

Comparação de ativos A avaliação de ativos relativa a outros ativos é uma forma de avaliar o investimento comparando-o com outros ativos. Para essa análise pode-se avaliar os seguintes aspectos: rentabilidade relativa, correlação entre os ativos e decomposição da rentabilidade.

- **Rentabilidade relativa:** Avalia o investimento comparando-o com outros ativos. Portanto, se o investimento apresenta um retorno maior que o mercado, então o investimento é bom, caso contrário, o investimento é ruim.

- **Correlação entre os ativos:** Avalia a relação entre os ativos, ou seja, se eles apresentam uma correlação alta, então os ativos são similares, portanto se um apresenta um bom desempenho, o outro também apresentará um bom desempenho. Essa

análise é bastante útil para avaliar a diversificação de uma carteira de investimentos. Se um ativo é muito correlacionado com outros ativos investidos da carteira, não faz sentido investir nesse ativo, pois ele não diversifica a carteira, ou seja, quando os outros ativos apresentarem uma rentabilidade ruim, o ativo correlacionado também apresentará uma rentabilidade ruim.

- **Decomposição da rentabilidade:** A decomposição de retornos visa entender todas as fontes de rentabilidade de um determinado ativo, caso a fonte de rentabilidade de um ativo seja fortemente influenciada pelo mercado talvez faça mais sentindo investir no mercado. A técnica mais famosa para decompor a rentabilidade é a decomposição a partir do alfa de Jensen [23], determinado pela seguinte equação: $\alpha = R_p - R_f - \beta(R_m - R_f)$, onde R_p é o retorno do investimento, R_f é o retorno livre de risco, R_m é o retorno do mercado e β é o beta do investimento. O racional dessa equação é que se o alfa for positivo, então o investimento apresenta um retorno maior que o mercado, portanto o investimento é bom, caso contrário, o investimento é ruim. Investidores também podem adicionar os fatores de risco além do mercado na equação, dentre esses fatores estão os descritos no início dessa seção: momentum, baixa volatilidade, dentre outros.

3 METODOLOGIA DO TRABALHO

Neste capitulo será apresentada a metodologia do projeto, apontando os principais passos e requisitos para a realização.

3.1 Aquisição de Dados

No projeto serão utilizados dois tipos de dados: os textuais para o processamento de linguagem natural e os financeiros para a modelagem de estratégias quantitativas. Para a aquisição de dados textuais serão extraídas notícias de veículos de comunicação que possuem o foco em finanças. Já para os dados financeiros serão utilizadas bases de dados públicas de preços de ações.

3.2 Anotação de Dados

Para o treinamento do modelo de sentimento é preciso realizar a anotação manual dos sentimentos (positivo, neutro e negativo). Esse processo será efetuado segundo a metodologia recomendada [13], que visa a seleção de textos representativos e anotação dos sentimentos por mais de uma pessoa para evitar vieses.

3.3 Treinamento do Modelo

O modelo proposto por esse projeto é um modelo de linguagem baseado na arquitetura BERT [3]. Para realizar o treinamento do FinBERT PT BR há duas principais etapas (Figura 2):

1. Treinamento de modelo de linguagem semi-supervisionado: para treinar o modelo de linguagem no contexto de mercado financeiro será realizando *transfer learning* no BERTimbau [24], sendo uma rede neural treinada num escopo generalista. E visto

que um modelo de linguagem não precisa de classificação dos dados, o *fine-tuning* da rede neural será feito numa grande base de notícias no contexto do mercado financeiro.

2. Treinamento de modelo de sentimentos supervisionado: Tendo como base o modelo de linguagem no contexto do mercado financeiro será realizado o *transfer learning* para uma tarefa de classificação de sentimento. O segundo modelo será treinando numa base de notícias com os sentimentos classificados.

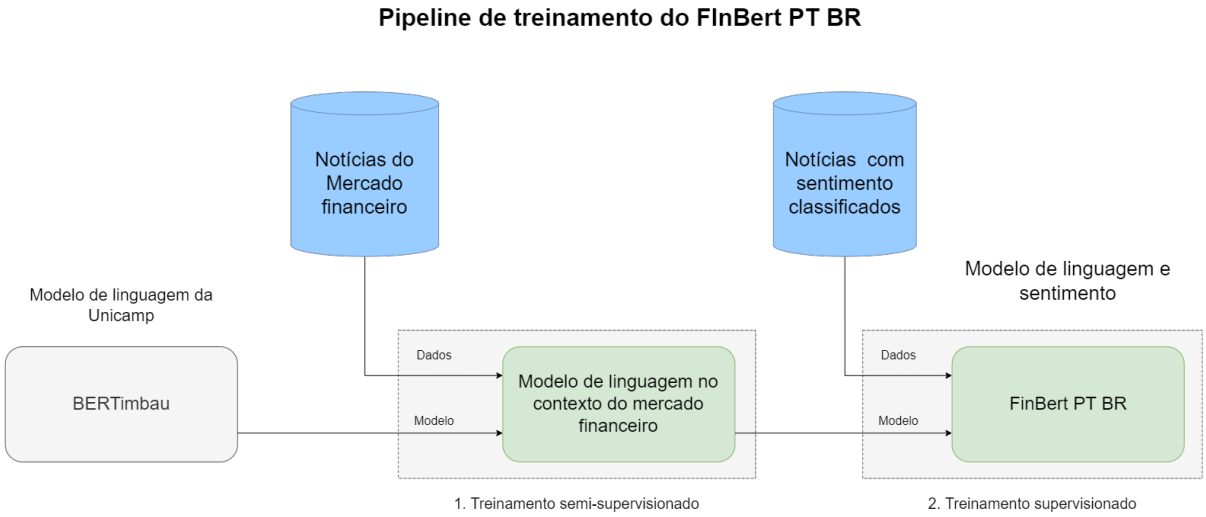

Figura 10: Arquitetura de treinamento do FinBERT PT BR.

3.4 Validação do Modelo

Os principais artigos de modelos de linguagem no estado da arte, além de apresentarem o desempenho do modelo, também apresentam a generalização para a aplicação de outras tarefas como: classificação, tradução, respostas a perguntas e entre outras. Visando avaliar o FinBERT PT BR em tarefas relacionadas a finanças, será construído um índice de sentimentos de mercado e avaliado sua relação com dados macroeconômicos e fatores de investimento.

3.4.1 Índice de Sentimentos

O índice de sentimento será uma série temporal construída a partir das classificações das notícias ao longo do tempo [25].

$$\text{Índice}_{t-k,t} = \frac{\text{Pos}_{t-k,t} - \text{Neg}_{t-k,t}}{\text{Pos}_{t-k,t} + \text{Neu}_{t-k,t} + \text{Neg}_{t-k,t}} \tag{3.1}$$

Onde $\text{Pos}_{t-k,t}, \text{Neg}_{t-k,t}, \text{Neu}_{t-k,t}$ representam o número de notícias positivas, negativas e neutras compreendido entre o intervalo de tempo $t-k$ e t respectivamente.

3.5 Requisitos do projeto

Nesta seção serão apresentados os requisitos funcionais e não funcionais do projeto.

3.5.1 Requisitos funcionais

O projeto possui diversos requisitos funcionais que devem ser cumpridos e serão descritos a seguir:

Aquisição de dados Em projetos de modelagem de dados com redes neurais é de suma importância a aquisição de uma grande base de dados para o treinamento do modelo. Neste projeto há dois principais categorias de dados: os textuais e financeiros (para validação do modelo).

Os dados textuais além de possuírem informações como título, corpo e autor também devem ter as informações sobre data da postagem. Já os dados financeiros devem possuir os preços diários históricos das ações da bolsa de valores brasileira.

Classificação de sentimentos O principal objetivo do modelo treinado será a predição de sentimentos a partir dos textos. Os textos poderão ser classificados em três categorias: positivo, neutro e negativo.

Construção de índice de sentimentos A partir dos textos classificados e suas respectivas datas de publicações, será construído um índice de sentimento do mercado, que será uma série temporal. Esse índice será calculado a partir do número de notícias positivas, negativas e neutras [25] ao longo do tempo.

3.5.2 Requisitos não funcionais

O projeto possui diversos requisitos não funcionais relacionados ao uso da aplicação que descritos a seguir:

Boas métricas de erro Ao longo do treinamento do modelo para a tarefa de classificação de sentimentos, é ideal que o classificador apresente boas métricas de classificação, para assim determinar os sentimentos das sentenças com uma maior assertividade.

As métricas utilizadas serão as métricas de algoritmos de classificação: Matriz de confusão, acurácia, revocação, precisão.

Alto nível de concordância na anotação de sentimentos A base de dados de sentimentos classificados é fundamental para o treinamento do modelo de sentimentos. E para uma modelagem sem vieses dos avaliadores é ideal que o nível de concordância dos avaliadores seja alto e sem vieses sistêmicos. Para garantir um alto nível de concordância dos sentimentos durante o processo de anotação é necessário o acompanhamento das métricas definidas na seção 2.3.1.2.

Base de dados balanceada Durante o aprendizado estatístico de modelos de classificação é ideal que as classes expostas durante do treinamento sejam balanceadas. Isto é o número de textos com sentimentos devem estar equilibrados em relação a todas as classes, assim a rede pode convergir bem para todas elas.

4 DESENVOLVIMENTO DO TRABALHO

Neste capítulo será apresentado o desenvolvimento do trabalho, desde a listagem de tecnologias utilizadas até a apresentação dos resultados e aplicações do projeto.

4.1 Tecnologias Utilizadas

Nesta seção serão descritas as principais tecnologias utilizadas durante o desenvolvimento do projeto.

4.1.1 Python

O Python é uma linguagem de programação de alto nível utilizada em diversos tipos de aplicações, como desenvolvimento web e aprendizado de máquina. No contexto de ciência de dados essa linguagem é extremamente popular e possui uma ampla comunidade. O que possibilita a construção de soluções de aprendizado de máquina por meio das bibliotecas e *frameworks* de código aberto.

4.1.2 Scrapy

Scrapy é um framework python completo para aquisição de dados de sites em larga escala. Com ele é possível percorrer todos os sites de uma plataforma de uma maneira eficiente. A ferramenta consegue realizar a coleta de forma distribuída, e respeita as limitações, como número de requisições simultâneas, da plataforma descrita pelo *robots.txt*. Para projetos de PLN que utilizam da metodologia de *web-as-corpus* [26] essa ferramenta se demonstra bastante útil para a aquisição textos.

4.1.3 *Hugging Face*

O *Hugging Face* é uma plataforma de *machine learning* que disponibiliza grandes bases de dados e modelos pré-treinados para a comunidade. Através do módulo python disponibilizado é possível utilizar, treinar e disponibilizar modelos no estado da arte.

4.1.4 PyTorch

O PyTorch é uma biblioteca de código aberto com um conjunto de ferramentas que auxiliam o treinamento de modelos de *machine learning*. Dentre as principais funcionalidade estão o cálculo de gradiente de forma automática e integração com GPU, o que acelera o processo de treinamento de modelos.

4.1.5 *Kedro*

O *Kedro* é um framework de código aberto que aplica conceitos de engenharia de software e engenharia de dados para o desenvolvimento de projetos de ciência de dados. Destaca-se a possibilidade de construção de fluxos de dados por meio de códigos modulares, o que facilita a manutenção e a reprodutividade dos experimentos.

4.1.6 *Kaggle*

O *Kaggle* é uma plataforma de ciência de dados que disponibiliza bases de dados e competições para a comunidade. Através da plataforma é possível participar de competições e compartilhar soluções para problemas de ciência de dados. A plataforma também disponibiliza notebooks para a execução de códigos com recursos computacionais como placas de vídeo.

4.1.7 *Wandb*

O *Wandb* é uma plataforma de código aberto que disponibiliza recursos para o monitoramento de experimentos de *machine learning*. Através da plataforma é possível acompanhar métricas de modelos, visualizar gráficos e compartilhar resultados.

4.2 Projeto e Implementação

4.2.1 Aquisição de dados

A aquisição de dados é um dos passos mais importantes de um projeto de ciência de dados. Visto que o projeto visa realizar a construção de modelos aplicados a finanças, foram coletadas notícias de canais de notícias financeiras. As plataformas escolhidas foram o *Valor Econômico*, *Exame* e *Infomoney*.

O processo de aquisição foi automatizado por meio da ferramenta Scrapy, utilizada para percorrer os sites de notícias financeiras e coletar os textos e metadados das notícias. O Scrapy consegue percorrer um site via um grafo de links entre os sites e também via *sitemap*, o qual é um arquivo que descreve todos os links possíveis para um site.

No total, foram coletados 2,51 milhões de sentenças de notícias do mercado financeiro entre 2006 e 2022. De forma mais específica, foi possível coletar um maior volume de notícias dos canais que disponibilizavam o *sitemap*, como o *Valor Econômico* e *Exame*, que retornaram mais de 1 milhão de textos cada. O *Infomoney* retornou um volume menor de textos, cerca de 460 mil, devido à falta de *sitemap*. A distribuição dos textos coletados por plataforma é apresentada na Figura 11.

Além dos textos das notícias, também foram coletados os metadados das notícias, como título, subtítulo, data de publicação, data de atualização da publicação, nome e página do autor e links.

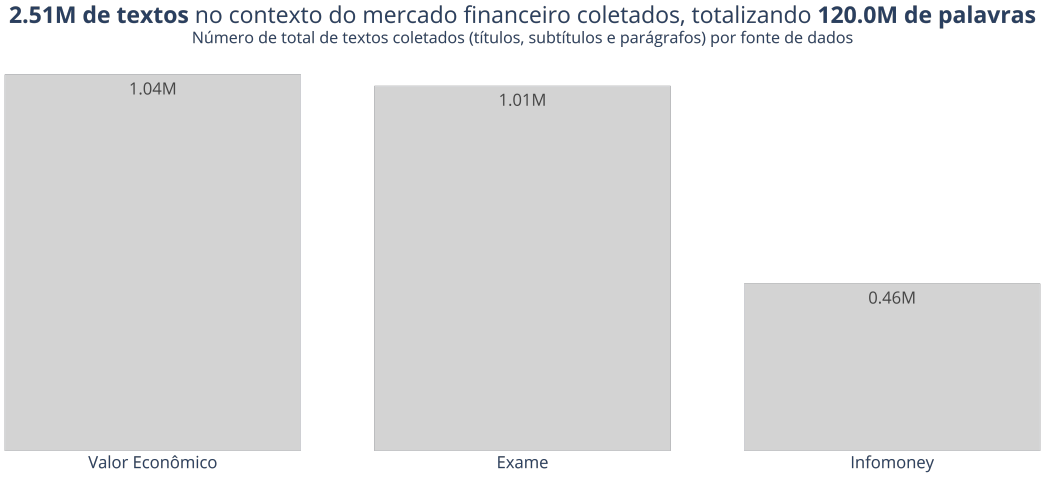

Figura 11: Distribuição de textos por fonte de dados

Após a coleta dos textos foi necessário realizar a **limpeza dos dados**. A limpeza dos

dados foi realizada por meio de expressões regulares que identificaram padrão de textos que não eram relevantes para o projeto. A maioria dos textos com má formação possuíam caracteres especiais e códigos de fonte. Após a limpeza dos dados, a base final ficou com 1,6 milhão de sentenças.

4.2.2 Treinamento de modelo de linguagem

O modelo de linguagem foi treinado utilizando o *PyTorch* e o *Hugging Face*, tendo como pesos iniciais os pesos do modelo BERTimbau [24]. Os principais aspectos acerca do treinamento estão descritos a seguir:

Dados de treinamento – Os dados utilizados foram textos das notícias coletadas, e no total de 1,6 milhão de sentenças de notícias, 1,4 milhão foram utilizadas. Visto que a arquitetura BERT suporta até 512 tokens na entrada do modelo, foi necessário realizar uma filtragem dos textos para que fossem utilizados apenas os textos com até 512 tokens, resultando em 1.428.867 milhões de sentenças de notícias.

Para fim de comparação, o modelo FinBERT (EN) [4] teve o seu modelo de linguagem treinado com 400 mil sentenças de notícias. A base de dados utilizada foi uma amostra da base TRC2 [27] da Reuters, que contém 1,8 milhões de notícias de 2008 a 2010. A base de dados utilizada no projeto é maior que a utilizada no modelo FinBERT (EN), e também possui um maior período, o que pode ser um fator que contribui para a melhoria do modelo.

Recursos Computacionais – O treinamento foi realizado em uma máquina do *Kaggle* que possuía 30 GB de RAM e 30 GB de GPU (2x Nvidia T4), o que limitou o tamanho do lote (*batch size*) e o carregamento de todos os textos em memória. O *batch size* foi definido como 16, por limitação de alocação do modelo e dos textos na GPU simultaneamente. Devida a abundância de textos foi necessário implementar o processo de treinamento com alocação dinâmica de memória, ou seja, ao longo do treinamento os textos são carregados do disco para a memória e depois para a GPU. Apesar das limitações, foi possível realizar o treinamento de duas épocas em 11 horas.

Parâmetros de treinamento – Para o treinamento do modelo de linguagem, os dois principais parâmetros que não dependem de recursos computacionais diretamente são: taxa de aprendizado (*learning rate*) e probabilidade de máscara. A taxa de aprendizado foi definida como 2e-5, o valor recomendado pela literatura [28] para

o *fine tunning* de modelos de linguagem que serão treinados para outras tarefas depois. A probabilidade de máscara, sendo a probabilidade da ocultação de tokens, foi definida como 15%, sendo a mesma utilizada no treinamento do modelo BERT original [3].

Resultados – Para avaliar o modelo de linguagem foi utilizada a métrica de perplexidade. A perplexidade é uma métrica que indica o quão bem o modelo consegue prever uma palavra dado o contexto. Essa métrica foi calculada numa amostra de 100 mil sentenças, que não foram utilizadas no treinamento do modelo. O modelo treinado apresentou uma perplexidade de 1,24, sendo uma métrica satisfatória, visto que o modelo BERTimbau original apresenta uma perplexidade de 1,51. A convergência do modelo pode ser visualizada na figura 12 e as métricas de perplexidade na tabela 2.

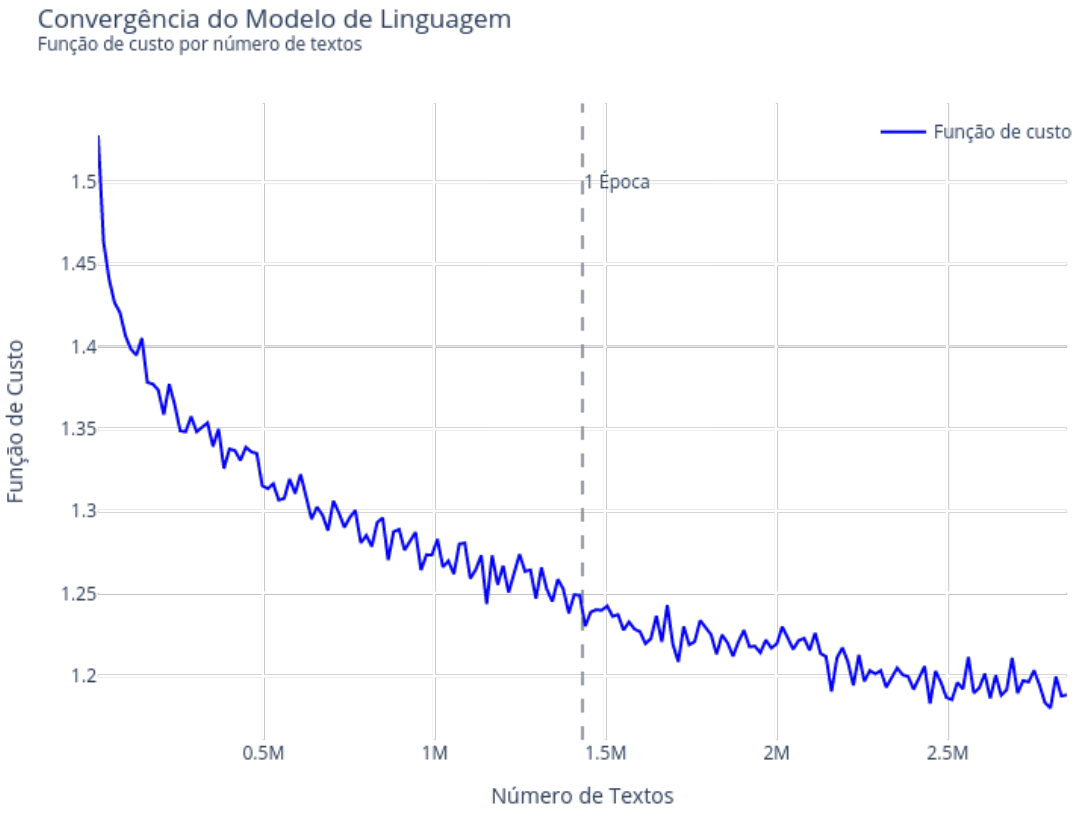

Figura 12: Treinamento do modelo de linguagem com textos no contexto do mercado financeiro.

Tabela 2: Perplexidade dos modelos de linguagem

Métrica	Perplexidade
BERTimbau	1.51
FinBERT PT BR	1.24

4.2.3 Anotação de dados

O processo de anotação foi realizado por 3 pessoas, para garantir a consistência das anotações, com todos os textos anotados por pelo menos 2 pessoas. A anotação teve como objetivo identificar o sentimento das sentenças de notícias do mercado financeiro. O processo de anotação e concordância da base final estão descritos a seguir:

Processo de anotação – As categorias de anotação definidas previamente foram: positivo, negativo, neutro e não se aplica. A definição de anotação foi: "Classifique a notícia considerando se o texto implicaria em uma rentabilidade Positiva, Negativa ou Neutra. "Não se aplica" para textos não relacionados a finanças, de políticos ou sem sentido." Após a definição da anotação foi realizada a anotação em uma amostra da base para verificar a concordância dos anotadores, assim que a concordância foi verificada, foi realizada a anotação completa da base.

No total foram anotados 1000 textos, sendo 497 textos descartados, pois foram classificados como "Não se aplica" ou não houve concordância entre os anotadores. Após a remoção dos textos descartados, a base de treinamento ficou com 503 textos, sendo 160 textos positivos, 203 textos negativos e 140 textos neutros. A Figura 13 mostra a distribuição dos textos por categoria.

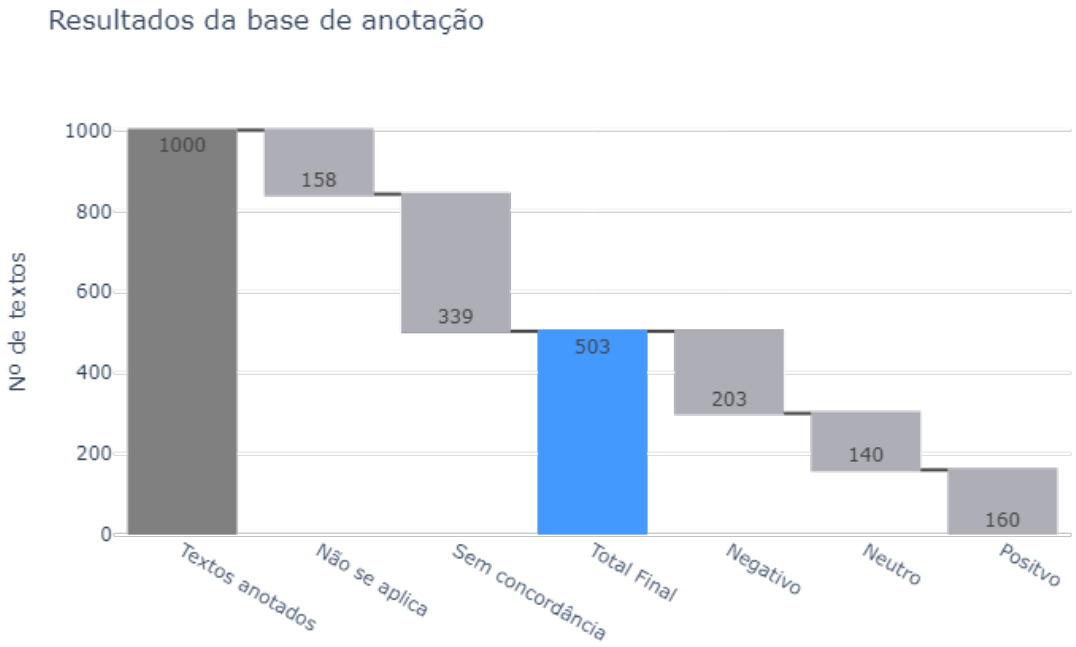

Figura 13: Resultado da anotação da base de sentimentos.

Análise de concordância – A concordância entre os anotadores foi verificada utilizando a porcentagem de concordância e o *Krippendorff's alpha* [17]. A porcentagem de concordância varia entre 0% e 100%, sendo 100% a concordância perfeita e a métrica obtida na base final de anotação foi de 90,4%. Já o *Krippendorff's alpha* varia entre -1 e 1, sendo 1 a concordância perfeita, -1 a discordância perfeita e 0 representa anotação aleatória, a métrica obtida na base final de anotação foi de 0,88. As métricas de concordância apontam que a anotação foi realizada consistentemente entre os anotadores.

Tabela 3: Métricas de concordância entre os anotadores.

Métrica	Valor
Porcentagem de concordância	90,4%
Krippendorff's alpha	0,88

4.2.4 Treinamento do modelo de sentimentos

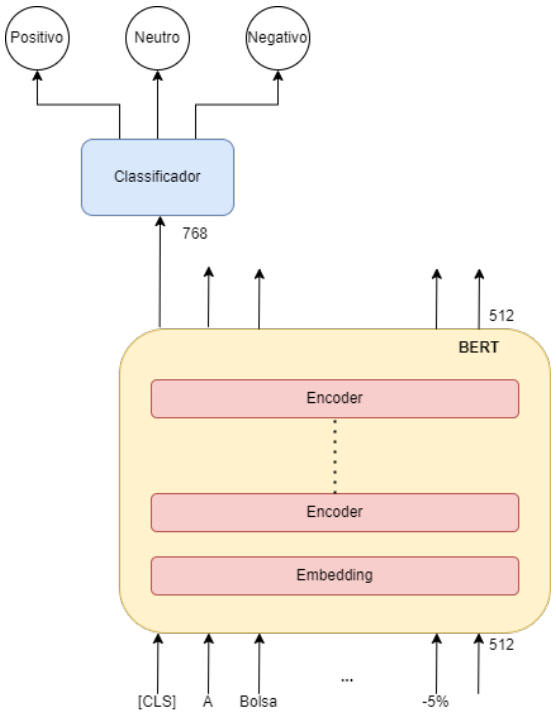

Figura 14: Arquitetura do modelo de classificação de sentimento

O modelo de sentimentos foi treinado utilizando o modelo de linguagem pré-treinado com os textos do mercado financeiro com a adição de uma camada de classificação de sentimentos. Essa camada foi adicionada na primeira dimensão de saída do BERT, seguindo a recomendação da literatura [3], conforme a figura 14. O processo de treinamento e validação dos resultados do modelo de sentimentos estão expostos a seguir:

Processo de treinamento – O processo de utilizar o modelo de linguagem para a tarefa de análise de sentimento é caracterizado como um *transfer learning*. Um problema comum durante a aplicação de *transfer learning* em modelos de linguagem é o esquecimento de informações durante o treinamento. Para evitar o esquecimento foi aplicada a técnica de *Gradual Unfreezing* [3] durante o treinamento do modelo de sentimentos. Essa técnica consiste em desbloquear a atualização dos pesos da rede neural gradativamente durante o treinamento, dessa forma a rede neural consegue aprender a tarefa de classificação de sentimento sem esquecer as informações aprendidas durante o treinamento do modelo de linguagem.

Visto que o BERT possui 11 camadas de *Encoder*, a técnica de *Gradual Unfreezing* foi aplicada para desbloquear as camadas de *Encoder* gradativamente. A taxa de aprendizado adotada foi de 5e-6 e o número de épocas foi 11.

Validação do modelo – Para realizar o treinamento do classificador de sentimentos foi utilizada a técnica de validação cruzada com 5 divisões. A base de treinamento foi dividida em 5 partes, sendo 4 partes para treinamento e 1 parte para validação. O processo de validação cruzada foi realizado 5 vezes, sendo que em cada iteração uma parte da base foi utilizada para validação e as outras 4 partes para treinamento. O modelo de sentimento com a melhor convergência (função de custo) da validação foi utilizado para realizar a predição dos textos da base de teste.

Modelos de compração – Para comparação serão utilizados os seguintes modelos de classificação de texto: Random Forest com TFIDF, FinBERT com textos traduzidos para o inglês como o modelo, BERT e BERT com modelagem de linguagem com textos no contexto do mercado financeiro (FinBERT PT BR).

- Random Forest com TFIDF: O modelo de Random Forest com TFIDF, técnica descrita na seção 2.2.1, foi treinado utilizando o algoritmo de classificação de texto do scikit-learn [29].
- FinBERT (EN): O modelo FinBERT [4] foi utilizando com textos traduzidos com o modelo multilingual do Facebook [30].
- BERTimbau: Transfer Learning do BERTimbau [24] para a tarefa de classificação de sentimentos.
- FinBERT PT BR: Utilizado o modelo de sentimento treinamento a partir do modelo de linguagem pré-treinado com textos do mercado financeiro.

Resultados – O FinBERT PT BR obteve o melhor resultado na base de teste, com uma acurácia de 0,76 para predição das 3 categorias de sentimento, conforme a tabela 4. O segundo modelo em desempenho foi o BERTimbau, com uma acurácia de 0,67, seguido pelo FinBERT (EN) com uma acurácia de 0,69 e o Random Forest com TFIDF com uma acurácia de 0,45. Em relação ao F1-Score, o modelo FinBERT PT BR obteve o melhor resultado com 0,73, seguido pelo FinBERT (EN) com 0,67, BERTimbau com 0,63 e Random Forest com TFIDF com 0,35.

Tabela 4: Resultados dos modelos de classificação de texto

Nome do Modelo	Acurácia	F1-Score
Random Forest com TFIDF	0.45	0,35
FinBERT (EN)	0.67	0.67
BERTimbau	0.69	0,63
FinBERT PT BR	**0.76**	**0,73**

Visando avaliar o intervalo de confiança das métricas dos modelos de classificação de texto, foi utilizado o método *bootstrapping* [31]. O método *bootstrapping* consiste em realizar a amostragem com reposição dos dados de teste, dessa forma é possível estimar o intervalo de confiança das métricas de avaliação dos modelos. Com um intervalo de confiança de 80% é possível afirmar que o modelo FinBERT PT BR é o melhor modelo de classificação de sentimento, visto que o intervalo de confiança da acurácia e do F1-Score do modelo FinBERT PT BR não se sobrepõe aos intervalos de confiança dos outros modelos, como pode ser visto pelas figuras 15 e 16.

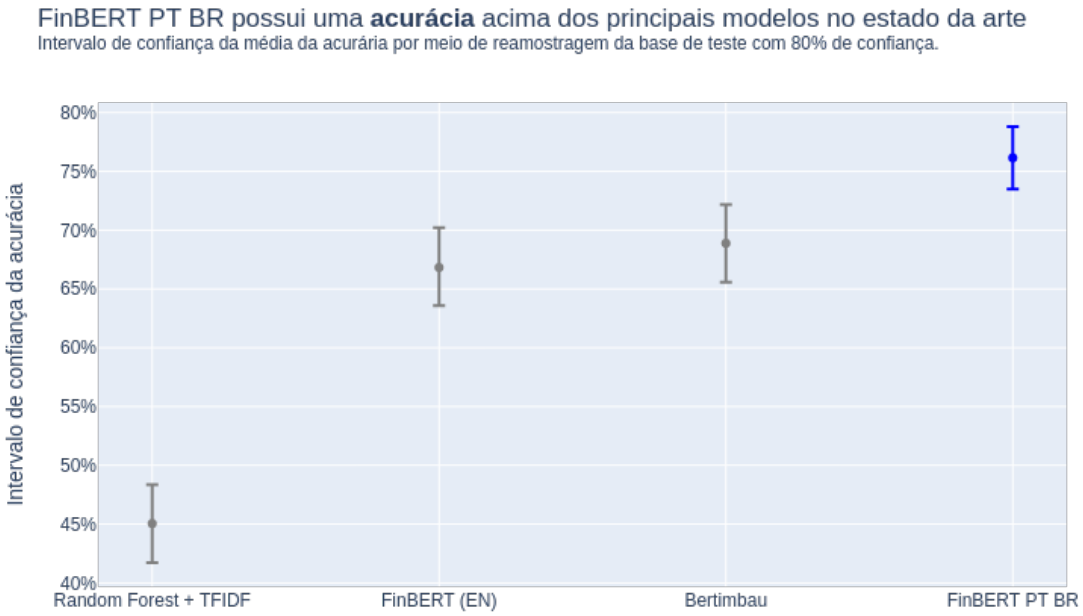

Figura 15: Intervalo de confiança da acurácia dos modelos de classificação de texto

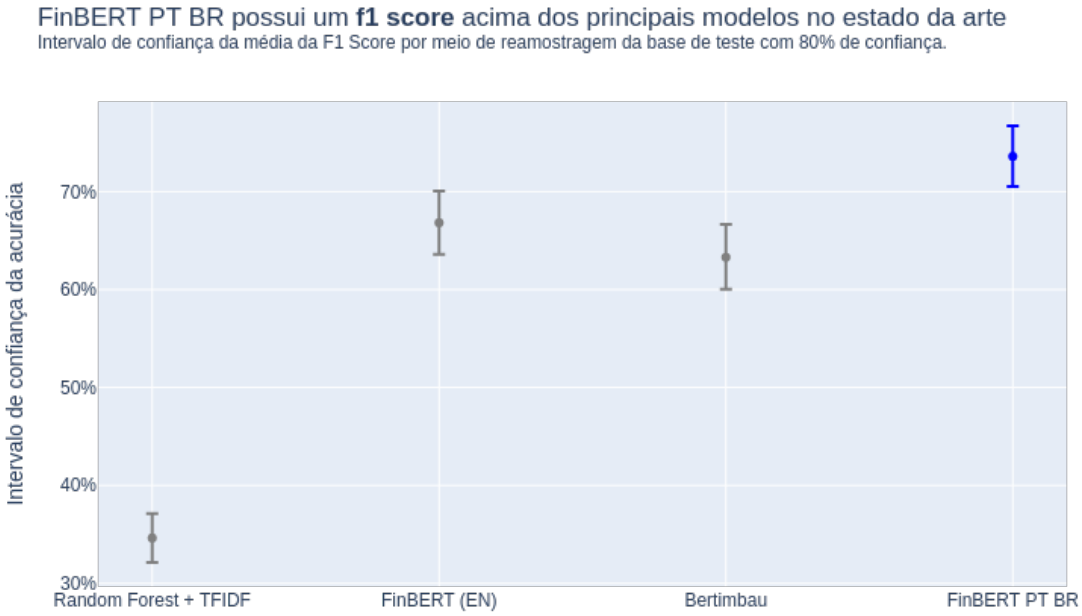

Figura 16: Intervalo de confiança do F1-Score dos modelos de classificação de texto

Por fim, foi realizado o teste de hipótese para verificar se existe diferença estatística entre os modelos de classificação de texto. O teste foi construído a partir da re-amostragem com reposição dos dados de validação do modelo. Dado que com a utilização de técnicas de re-amostragem as estatísticas tendem para uma gaussiana pelo teorema do limite central, é possível construir testes de hipótese na distribuição empírica da estatística. Então foram construídas distribuições empíricas da acurácia e *f1 score* de todos os modelos, e com um teste Z foi possui realizar o teste com as seguintes hipóteses:

A hipótese nula do teste de hipótese é que não existe diferença estatística entre as métricas dos modelos de classificação de texto. A hipótese alternativa do teste de hipótese é que existe diferença estatística entre as métricas dos modelos de classificação de texto. A conclusão do teste de hipótese é que existe diferença estatística entre as métricas dos modelos de classificação de texto, visto que o p-valor é numericamente igual a 0. Com isso, é possível afirmar que o modelo FinBERT PT BR é o melhor modelo de classificação de texto para a tarefa de classificação de sentimento de notícias do contexto do mercado financeiro.

4.3 Aplicações do modelo

4.3.1 Índice de sentimentos

O modelo de classificação de sentimento pode ser utilizado para gerar um índice de sentimentos, sendo um indicador que mede a percepção do mercado sobre o sentimento do mercado. O índice de sentimentos é calculado a partir da média dos sentimentos dos textos classificados como positivos, negativos e neutros ao longo do tempo. A fórmula do índice de sentimentos é definida pela equação 3.1.

Ao longo do tempo, o índice de sentimentos pode ser utilizado para identificar se o mercado ou um determinado setor está otimista, ou pessimista. No caso dessa aplicação, o índice é calculado a partir de uma amostra das notícias gerais coletadas e pode ser visto na figura 17 com alguns fatos relevantes da economia.

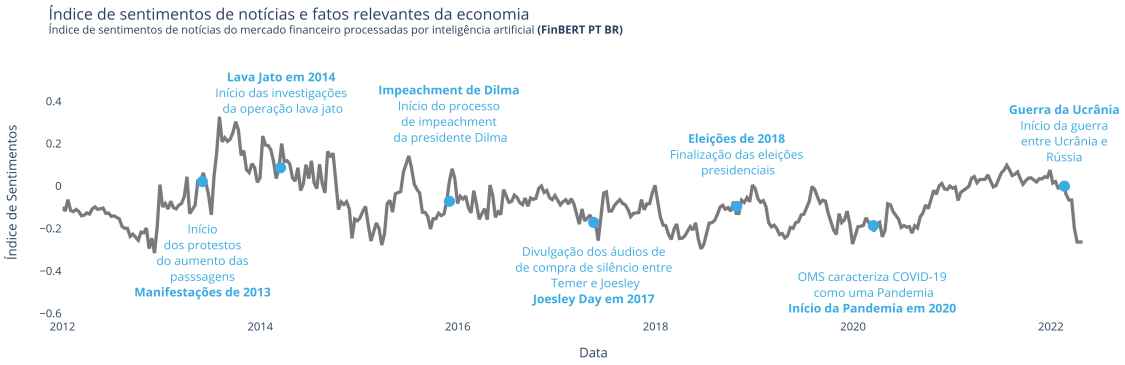

Figura 17: Índice de sentimentos e fatos relevantes da economia.

O primeiro fato relevante destacado no índice de sentimentos foram as manifestações de 2013, que ocorreram no Brasil por conta da alta do preço dos transportes públicos, marcadas por protestos e vandalismo. Avaliando o índice de sentimentos, é possível perceber que o sentimento do mercado teve uma grande variação negativa no mês de junho de 2013, quando ocorreu o primeiro protesto. Logo após a finalização das manifestações, o sentimento do mercado apresentou outra grande variação, dessa vez positiva, no mês de agosto.

O segundo fato relevante destacado no índice de sentimentos foi o início da operação Lava Jato, que ocorreu em março de 2014. Esse período foi mercado por uma sequência de meses com revelações de escândalos de corrupção envolvendo diversos políticos e empresas. Olhando para o índice de sentimentos, nota-se que no médio prazo houve uma tendência

de queda no sentimento do mercado, isso deve pelo acúmulo de notícias negativas sobre corrupção. O valor mínimo do índice de sentimentos foi atingido em março de 2015, quando grandes executivos da Petrobras foram presos.

O final de 2015 foi marcado pelo início do processo de impeachment da presidente Dilma Rousseff, finalizado em agosto de 2016. Logo após o início do processo de impeachment, o sentimento do mercado apresentou uma grande variação positiva seguida de uma queda no sentimento do mercado. Visto que o país estava em recessão, o sentimento do mercado tendeu a um valor positivo por uma expectativa de resolução da crise econômica, porém foi apenas uma pequena janela de euforia.

Após o impeachment, já no governo de Michel Temer, o sentimento do mercado apresentou uma grande variação negativa no denominado "Joesley Day". Esse dia, 17 de maio de 2017, foi marcado pela divulgação de uma gravação de uma conversa entre o empresário Joesley Batista e o presidente Michel Temer, em que o assunto da conversa era a compra do silêncio do ex-presidente da Câmara dos Deputados Eduardo Cunha, que estava preso por corrupção. Por envolver o presidente da República, houve muita incerteza sobre o futuro do governo, e o sentimento do mercado apresentou uma grande variação negativa. Visto que o presidente não foi destituído, o sentimento de mercado foi amenizado e voltou a um valor positivo.

Em 2018, o sentimento de mercado apresentou um leve aumento após a conclusão das eleições presidenciais. Tal aumento pode ser explicado pelo fato de que o mercado estava otimista com a vitória de Jair Bolsonaro, que era visto como um candidato favorável ao mercado. Porém, após a posse de Bolsonaro, o sentimento de mercado retornou a um valor negativo, que pode ser explicada pela incerteza sobre o futuro do governo. Depois, ao longo do primeiro ano de governo, o sentimento apresentou algumas sequências de alta que podem ser atribuídas a notícias positivas, como a aprovação da reforma da previdência.

O principal destaque de 2020 foi a pandemia da COVID-19, que começou em janeiro, classificada como pandemia pela OMS em março. A pandemia trouxe ao mundo muitas incertezas, e o sentimento de mercado não apresentou grandes variações justamente por conta das incertezas. Enquanto o vírus ainda não havia se espalhado pelo Brasil, alguns veículos notícias comunicavam que no Brasil a crise sanitária não seria tão grave por conta do clima tropical, já outros veículos noticiavam que o Brasil estava despreparado para lidar com a pandemia. Portanto, devido a essa incerteza, o sentimento de mercado não apresentou grandes variações.

Mais para o final de 2020, o sentimento de mercado apresentou uma grande variação

positiva, que pode ser explicada pelo desenvolvimento de vacinas e estímulos econômicos dos governos. O valor de pico do sentimento de mercado foi atingido no final do ano, quando alguns países começaram a vacinar a população. Outro fato, que pode ter contribuído para o aumento do sentimento de mercado, foram os estímulos econômicos dos governos, implementados para tentar minimizar os efeitos da pandemia e acarretaram grandes altas nos mercados.

Por fim, o último fato relevante destacado no índice de sentimentos foi início da guerra da Rússia e Ucrânia. De longe, esse foi o fato que mais impactou o sentimento de mercado, com o início da guerra vários países impuseram sanções econômicas à Rússia, e o sentimento de mercado apresentou uma grande variação negativa. Nesse período o mundo ainda estava se recuperando da pandemia, haviam sinais inflacionários por conta dos estímulos econômicos, e a guerra trouxe mais pressão inflacionária ao mundo. Por conta disso, o sentimento de mercado apresentou uma grande variação negativa.

4.3.2 Relação com dados macroeconômicos

O modelo de classificação de sentimento pode ser utilizado para identificar se existe correlação entre o sentimento do mercado e a inflação. Dado que a inflação é um indicador importante para a economia, e é fortemente influenciado pelas expectativas acerca do futuro da economia, é possível que o sentimento tenha alguma correlação com a inflação. Ou seja, se a população está pessimista em relação à inflação, ela tende a gastar menos, e consequentemente a inflação tende a ser menor. E visto que com o índice de sentimentos é possível captar sinais de otimismo ou pessimismo do mercado, é possível verificar se existe correlação entre o índice de sentimentos e a inflação.

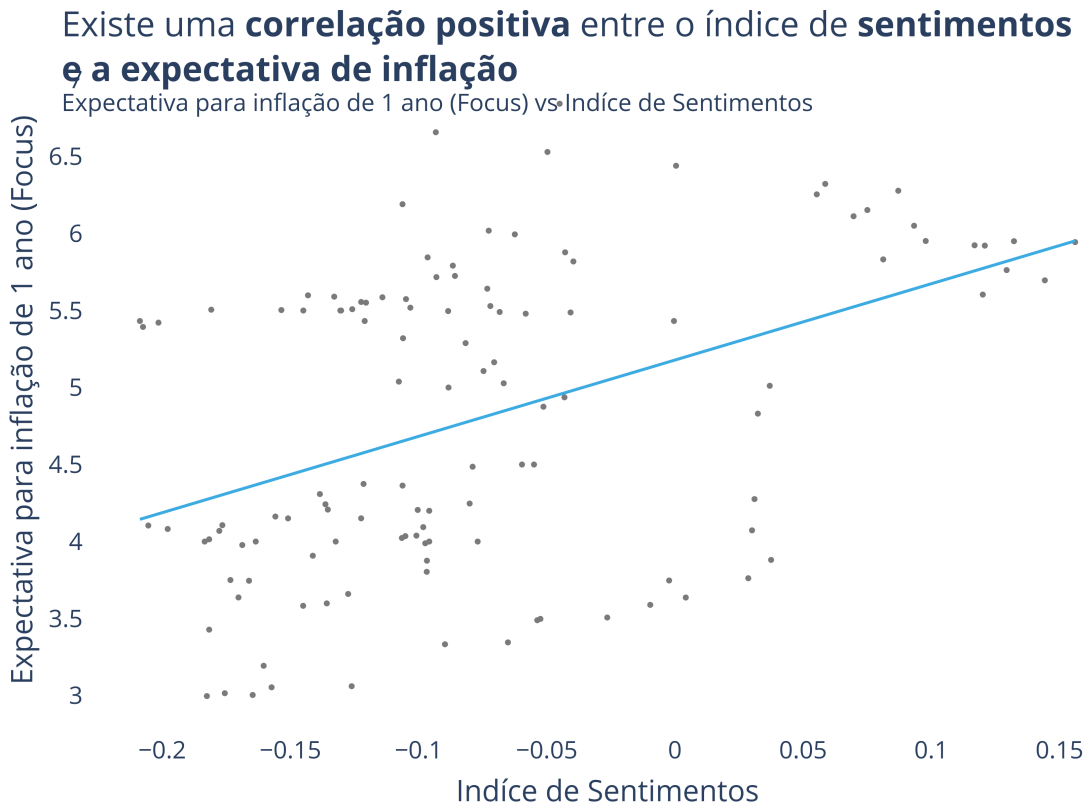

Figura 18: Correlação entre o índice de sentimentos e a inflação

A figura 18 mostra que existe uma relação entre a expectativa da inflação e o índice de sentimentos. A correlação linear entre o índice de sentimentos e a inflação é de 0,52, indicando uma correlação moderada. O fato da correlação linear ser positiva indica que quando o índice de sentimentos é maior, a expectativa da inflação é maior e vice-versa.

4.3.3 Apostando contra o sentimento

Nesta seção, introduzimos uma nova estratégia de investimento baseada no índice de sentimento construído, a estratégia é denominada como *apostando contra o sentimento*. Essa estratégia consiste em apostar contra o sentimento do mercado, ou seja, investir em ações que possuem alta correlação negativa com o índice de sentimentos. O racional dessa estratégia é que investidores pessimistas tendem a vender ações influenciados pelas notícias negativas. Porém, os fundamentos das ações com alta correlação negativa não são necessariamente afetados pelas notícias negativas do mercado na totalidade, mas são influenciados por fatores macroeconômicos e/ou intrínsecos da empresa. Então, quando investidores pessimistas vendem tais ações, há um aumento no prêmio de risco dessas ações. Portanto, investir em ações com alta correlação negativa com o índice de sentimentos é uma estratégia de investimento que pode gerar retornos positivos, visto que o

prêmio de risco dessas ações tende a ser maior que as demais ações.

Para testar essa estratégia, foi utilizado o índice de sentimentos construídos e uma base de preços de ativos da bolsa de valores brasileira. A base de preços de ativos foi obtida do site *Yahoo Finance* e contém os preços de fechamento das ações negociadas na bolsa de valores brasileira. A base de preços de ativos contém os preços de fechamento das ações de 2014 a 2022. Para cada ação, foi calculada a correlação entre o índice de sentimentos e os retornos históricos de cada ação, e as ações com alta correlação negativa com o índice de sentimentos foram selecionadas.

O resultado da simulação da estratégia de investimento *apostando contra o sentimento* é mostrado na figura 19. A estratégia proposta apresentou um retorno acumulado de 683% ao longo de 8 anos, enquanto o índice Bovespa apresentou um retorno acumulado de 254% no mesmo período. O retorno acumulado da estratégia *apostando contra o sentimento* foi 2,7 vezes maior que o retorno acumulado do índice Bovespa.

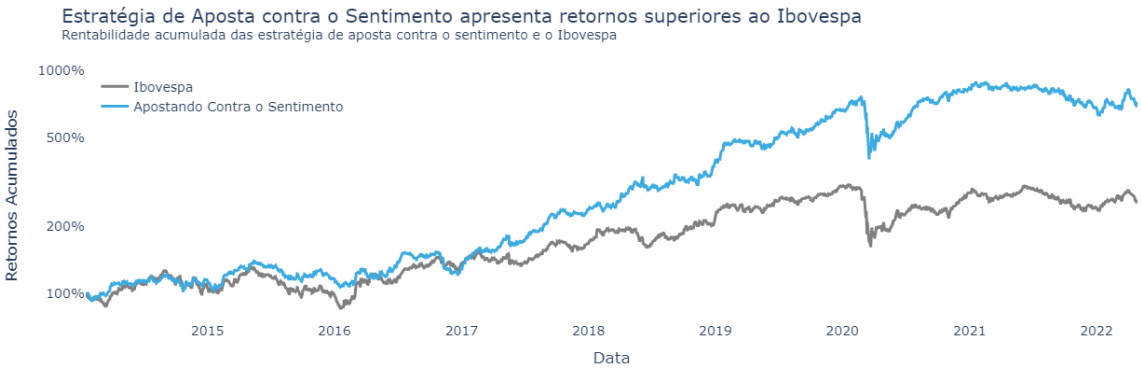

Figura 19: Simulação da estratégia de investimento *apostando contra o sentimento* e do índice Bovespa.

Outra análise realizada foi a comparação do retorno acumulado da estratégia *apostando contra o sentimento* com outros fatores de risco (Momentum e Baixa Volatilidade), e o resultado é mostrado na figura 20. E, para a comparação de desempenho das estratégias foi utilizado o índice de Sharpe. A estratégia *apostando contra o sentimento* apresentou um índice de Sharpe de 0,68, enquanto os demais fatores *momentum* e baixa volatilidade ambos apresentaram 0,86. Portanto, apesar da estratégia apresentar retornos acima da média de mercado, seu desempenho não é superior aos demais fatores de risco.

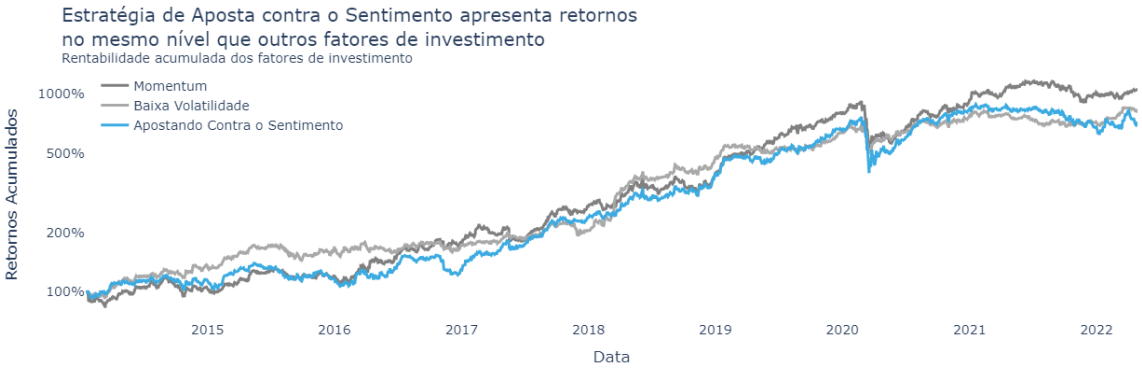

Figura 20: Simulação da estratégia de investimento *apostando contra o sentimento* e de outros fatores de risco.

Visto que os fatores de investimento visam explicar os retornos do mercado, também foi elaborada uma regressão linear para explicar os retornos do mercado com os fatores de investimento. A regressão linear foi realizada com os retornos do mercado, taxa livre de risco (SELIC) e os retornos das estratégias *apostando contra o sentimento*, *momentum* e *baixa volatilidade*. A regressão linear foi realizada com os retornos do mercado de 2014 a 2021, e os resultados são mostrados na tabela 5.

Tabela 5: Resultados da regressão linear com o índice de mercado como variável dependente e fatores de investimento como variáveis independentes

Fator de Risco	Coeficiente	P-Valor	T-Valor
Termo Constante	-0,0008	0,105	-1,621
Apostando Contra o Sentimento	0,3806	0,000	15,715
Momentum	0,2670	0,000	11,197
Baixa Volatilidade	0,2455	0,000	10,200
Taxa Livre de Risco Anualizada	0,0052	0,364	0,909

A regressão linear apresentou um R quadrado de 0,62, o que indica que os fatores de investimento explicam 66% dos retornos do mercado. Além disso, o termo constante e taxa livre de risco apresentaram P-valores acima de 5%, reforçando que os fatores de investimento explicam boa parte dos retornos do mercado. E, dentre os fatores de investimento, o *apostando contra o sentimento* apresentou o maior coeficiente e mais significativo, o que indica que a estratégia *apostando contra o sentimento* é o fator de investimento que mais explica os retornos do mercado. Portanto, a partir dos resultados da regressão linear, pode-se concluir que a estratégia *apostando contra o sentimento* é

um fator de investimento relevante e que pode ser utilizado para explicar os retornos do mercado.

5 CONSIDERAÇÕES FINAIS

Neste capítulo, serão apresentadas as considerações finais do trabalho, com a conclusão dos resultados e os trabalhos futuros.

5.1 Conclusão

O presente trabalho apresentou um modelo de análise de sentimentos de textos em português referentes ao mercado financeiro, utilizando a arquitetura de redes neurais BERT, denominado FinBERT PT BR. O modelo foi treinado em duas principais etapas: modelagem de linguagem com 1,4 milhão de textos no contexto do mercado financeiro e modelagem de sentimentos com 503 notícias com os sentimentos anotados.

A base de sentimentos anotados apresentou altas métricas de concordância, com porcentagem de concordância 90,4 e *Krippendorff's alpha* de 0,88. E os modelos treinados conseguiram superar as métricas dos principais modelos no estado da arte, com destaque para as métricas do modelo de análise de sentimento, com acurácia e F1-Score 0,76 e 0,73, respectivamente. Sendo que os modelos de referência apresentaram acurácia e F1-Score de 0,69 e 0,67, respectivamente como as maiores métricas Além do treinamento dos modelos, foram apresentadas aplicações relevantes para o mercado financeiro, como a análise de sentimento de notícias, construção de índices de sentimento e análise de dados macroeconômicos. Com o índice de sentimentos, foi possível verificar uma correlação positiva entre o sentimento de mercado e a expectativa de inflação, e também foi realizada uma análise qualitativa do índice com fatos relevantes da economia brasileira. Por fim, foi introduzido um fator de investimento denominado *apostando contra o sentimento*, que consiste em investir em ações apresentam alta correlação negativa com o índice de sentimento, além dessa estratégia apresentar um retorno ajustado ao risco satisfatório ela pode ser utilizada para explicabilidade do desempenho do índice Ibovespa.

5.2 Trabalhos futuros

Para trabalhos futuros, é possível aprimorar o modelo de análise de sentimentos, utilizando uma base maior e mais específica de textos financeiros para o treinamento do modelo de linguagem. Além disso, é possível aprimorar o modelo de análise de sentimentos, aumentando a quantidade de textos rotulados, e mantendo altas métricas de concordância. Em relação às aplicações, é possível aprimorar a fórmula de cálculo do índice de sentimento ou aplicar a metodologia para setores específicos da bolsa de valores. A análise de dados macroeconômicos pode ser expandida para outros indicadores além da inflação, como PIB e taxa de desemprego. Por fim, a explicabilidade dos retornos da bolsa de valores pode ser expandida para outros ativos financeiros, como ações e fundos de investimentos.

REFERÊNCIAS

[1] LO, A. W. The adaptive markets hypothesis. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 30, n. 5, p. 15–29, 2004.

[2] VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

[3] DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] ARACI, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[5] FULLER, R. J. Behavioral finance and the sources of alpha. *Journal of Pension Plan Investing*, v. 2, n. 3, p. 291–293, 1998.

[6] MAGUERESSE, A.; CARLES, V.; HEETDERKS, E. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[7] WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

[8] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016.

[9] NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, Oxford Academic, v. 18, n. 5, p. 544–551, 2011.

[10] MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[11] MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, Elsevier, v. 5, n. 4, p. 1093–1113, 2014.

[12] ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1253, 2018.

[13] HOVY, E.; LAVID, J. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, v. 22, n. 1, p. 13–36, 2010.

[14] POURSABZI-SANGDEH, F.; BOYD-GRABER, J. Speeding document annotation with topic models. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. [S.l.: s.n.], 2015. p. 126–132.

[15] GOYAL, N. et al. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.

[16] ALCOFORADO, A. et al. Zeroberto: Leveraging zero-shot text classification by topic modeling. *CoRR*, abs/2201.01337, 2022. Disponível em: <https://arxiv.org/abs/2201.01337>.

[17] ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. *Computational linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 34, n. 4, p. 555–596, 2008.

[18] SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

[19] BENDER, J. et al. Foundations of factor investing. *Available at SSRN 2543990*, 2013.

[20] CHAN, L. K.; JEGADEESH, N.; LAKONISHOK, J. Momentum strategies. *The Journal of Finance*, Wiley Online Library, v. 51, n. 5, p. 1681–1713, 1996.

[21] HURST, B.; OOI, Y. H.; PEDERSEN, L. H. A century of evidence on trend-following investing. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 44, n. 1, p. 15–29, 2017.

[22] MOSKOWITZ, T. J.; OOI, Y. H.; PEDERSEN, L. H. Time series momentum. *Journal of financial economics*, Elsevier, v. 104, n. 2, p. 228–250, 2012.

[23] JENSEN, M. C. The performance of mutual funds in the period 1945-1964. *The Journal of finance*, JSTOR, v. 23, n. 2, p. 389–416, 1968.

[24] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 403–417.

[25] HIEW, J. Z. G. et al. Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*, 2019.

[26] KILGARRIFF, A.; GREFENSTETTE, G. Introduction to the special issue on the web as corpus. *Computational linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 29, n. 3, p. 333–347, 2003.

[27] AMINI, M. R.; USUNIER, N.; GOUTTE, C. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in neural information processing systems*, v. 22, 2009.

[28] SUN, C. et al. How to fine-tune bert for text classification? In: SPRINGER. *China national conference on Chinese computational linguistics*. [S.l.], 2019. p. 194–206.

[29] PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

[30] FAN, A. et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, v. 22, n. 107, p. 1–48, 2021.

[31] EFRON, B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 569–593.