Universidade Presbiteriana Mackenzie

PROJETO APLICADO 2: ECONOFORESIGHT – Análise de sentimento financeiro

Grupo 3:

DIONE LUCAS SOUZA DA COSTA - 22501150

DOUGLAS PEREIRA DE ARAUJO - 22021752

ISABEL DE FÁTIMA BATANETE RAMOS - 22520309

JULIA RODRIGUES DA CUNHA – 22502335

SAMUEL REGIS NASCIMENTO BARBOSA – 22012354

SÃO PAULO

2023

SUMÁRIO

Resumo

(Abstract)

Este trabalho se concentra na análise de sentimentos financeiros por meio de uma abordagem abrangente que envolve análise exploratória de dados, aquisição e processamento de dados, bem como mineração de dados, estatística preditiva, juntamente com a visualização dos resultados obtidos em cada fase. Sentimentos financeiros referem-se às emoções, opiniões e atitudes que os indivíduos e investidores têm em relação aos mercados financeiros, ativos e eventos econômicos. Compreender esses sentimentos é de grande importância, pois eles podem influenciar decisões de investimento, volatilidade de mercado e tomadas de decisão financeira em larga escala. O grupo 3, sob o nome da empresa fictícia “EconoForesight”, realizou a análise utilizando métodos estatísticos, aprendizado de máquina e técnicas de visualização para extrair insights significativos a partir de dados financeiros e sentimentos associados. O objetivo é identificar padrões, tendências e correlações que possam ajudar a prever movimentos de mercado e melhorar a tomada de decisões financeiras.

Abstract:

This paper focuses on the analysis of financial sentiments through a comprehensive approach that involves exploratory data analysis, data acquisition and processing, as well as data mining, predictive statistics, along with the visualization of results obtained at each stage. Financial sentiments refer to the emotions, opinions, and attitudes that individuals and investors hold towards financial markets, assets, and economic events. Understanding these sentiments is of great importance as they can influence investment decisions, market volatility, and large-scale financial decision-making. Group 3 under the fictitious company name "EconoForesight" conducted the analysis using statistical methods, machine learning, and visualization techniques to extract meaningful insights from financial data and associated sentiments. The objective is to identify patterns, trends, and correlations that may assist in predicting market movements and improving financial decision-making.

1. INTRODUÇÃO

Os sentimentos financeiros referem-se às emoções e percepções dos investidores em relação ao mercado financeiro e aos seus investimentos. É uma medida subjetiva que reflete a confiança e o otimismo dos investidores em relação ao mercado.

Eles desempenham um papel crucial, pois refletem as expectativas e decisões dos investidores, que podem ser influenciadas por uma série de fatores, como notícias econômicas, políticas, desempenhos passados e eventos globais. Eles podem afetar diretamente os preços das ações, títulos e outros ativos, gerando volatilidade nos mercados.

O nome escolhido para a Empresa foi EconoForesight, a razão do nome Econo, vem de economia e Foresight é a palavra inglesa para previsão, logo o grupo optou pela junção desses dois pontos relevantes do trabalho para criar o nome da empresa. Slogan: "Nós prevemos o seu sentimento financeiro!"

A EconoForesight é uma empresa fictícia de análise preditiva voltada aos sentimentos financeiros.

Neste projeto, visa-se prever o sentimento (positivo, negativo ou neutro) de artigos de notícias financeiras, publicações nas redes sociais relacionadas com mercados financeiros ou instrumentos financeiros específicos. Essa análise de sentimentos, permite que os investidores e analistas compreendam melhor as tendencias do mercado e tomem decisões mais informadas. Além de ajudar, também, na avaliação de riscos e na gestão de carteiras de investimentos.

O grupo encontrou, na plataforma Kaggle, uma base de dados sobre “Financial Sentiment Analysis”. O projeto aqui desenvolvido tem como modelo de classificação Naive Bayes.

O grupo, utilizou-se de ferramentas de análises de dados e algoritmos para captar o humor do mercado e fornecer insights uteis para auxiliar nas decisões financeiras. Com sua capacidade de compreender e interpretar padrões de linguagem, ele pode analisar com mais eficácia grandes quantidades de dados não estruturados, como artigos de notícias e manchetes, para fornecer uma compreensão abrangente do sentimento do mercado.

Por exemplo, o sentimento expresso numa discussão sobre uma nova política regulatória governamental pode diferir significativamente, dependendo se o contexto é um fórum de investidores

que discute potenciais impactos no mercado ou um fórum de consumidores, discutindo mudanças nas taxas de serviço (Poria et al., 2017).

A classificação das palavras utilizadas como positivas, como *solução, lucro, vendas*, pode nos fornecer resultados positivos. Esta análise de sentimento melhorada pode, por sua vez, informar a tomada de decisões em áreas como estratégias de investimento, gestão de risco e otimização de carteiras, levando a investidores melhor informação e potencialmente decisões mais elevadas (Tetlock, 2007; Chene outros, 2014).

2. OBJETIVOS E METAS

Desenvolver uma pesquisa sobre a aplicação da análise de sentimento financeiro na Empresa EconoForesight, afim de ajudar as instituições financeiras a entender o impacto das notícias nos mercados, prever tendências e tomar decisões de investimento informadas.

3. METODOLOGIA

3.1. Definição do grupo de trabalho

Haja vista o projeto aplicado 2 a ser entregue, a distribuição eficaz de tarefas é fundamental para garantir o sucesso da empresa fictícia criada pelo grupo 3. Nesse contexto, uma colaboração eficiente e uma alocação estratégica de responsabilidades entre os colegas busca uma alocação de alta qualidade do conhecimento adquirido. Em nosso projeto, optamos por uma divisão de tarefas que capitalizou as habilidades e competências específicas de nossa equipe.

Os membros do grupo, reconhecendo as suas respectivas forças e áreas de especialização, adotaram uma abordagem cuidadosamente planejada. A fim de otimizar a eficiência e a qualidade do projeto, optamos por dividir as responsabilidades com base em duas áreas principais: coleta, processamento e codificação de dados, e estruturação de texto e gerenciamento de aspectos metodológicos. Nossos colegas Douglas, Samuel e Dione desempenharam um papel fundamental no projeto. Eles se dedicaram à coleta abrangente de dados, que envolveu a seleção de fontes relevantes e a utilização de técnicas de web scraping para reunir as informações necessárias.

Além disso, eles foram responsáveis pelo processamento dos dados brutos, realizando a limpeza e a preparação dos dados para análise. Isso incluiu a detecção e tratamento de valores ausentes, a normalização, codificação de variáveis e os gráficos. A habilidade e o conhecimento técnico demonstrados pelos colegas foram fundamentais para garantir que a base de dados fosse

sólida e confiável, um elemento crucial em qualquer projeto de Ciência de Dados, bem como para ter os primeiros resultados dos testes. Enquanto isso, as colegas Julia e Isabel desempenharam um papel igualmente crucial. Elas lideraram a estruturação dos textos, garantindo que os relatórios e documentação do projeto estivessem bem-organizados e de fácil compreensão.

Também foram responsáveis por gerenciar o repositório do projeto no GitHub, garantindo que todos os arquivos, códigos e documentos estivessem devidamente armazenados e versionados. Além dessas responsabilidades, as integrantes se encarregaram de garantir que todas as exigências metodológicas do projeto fossem atendidas. Isso incluiu a revisão das diretrizes do projeto, a definição de critérios de avaliação e a garantia de que todos os aspectos do trabalho estivessem alinhados com as melhores práticas em Ciência de Dados. 10 Nossa abordagem colaborativa permitiu que cada membro da equipe desempenhasse um papel significativo e complementar no projeto, aproveitando suas habilidades e conhecimentos individuais.

Essa divisão de trabalho estratégica não apenas facilitou a conclusão eficiente do projeto, mas também resultou em um trabalho final de alta qualidade, que atendeu às expectativas e aos padrões rigorosos da disciplina de Projeto Aplicado 2. Este projeto é um exemplo do sucesso que pode ser alcançado quando uma equipe trabalha de forma coordenada, aproveitando a diversidade de talentos e habilidades que cada membro traz para a mesa.

3.2. Definição e descrição das bases teóricas dos métodos

Foram usadas as bases teóricas aprendidas no decorrer do curso de Tecnologia em Ciência de Dados, principalmente em relação às matérias de aprendizado de máquina, aquisição e preparação de dados e estatística preditiva. Também se seguiram as orientações da matéria de projeto aplicado e guias de trabalhos disponibilizados pela Universidade.

No intuito de esclarecer os pormenores das bases teóricas dos métodos utilizadas pelo grupo 3, apresentam-se abaixo os detalhes:

Tratamento de Dados: O tratamento de dados é uma etapa fundamental no processo de análise de dados e desempenha um papel crítico na obtenção de insights precisos e confiáveis a partir de conjuntos de dados brutos. Essa fase abrange uma série de atividades que visam preparar os dados para análise, modelagem ou visualização.

Coleta de Dados: A primeira etapa é a coleta de dados, onde foram obtidas informações das fontes já citadas e que podem ser verificadas em https://github.com/BelBatanete/Mackenzie/blob/main/DATA%20-CSV.zip, como bancos de dados e arquivos CSV. A qualidade e a integridade dos dados coletados são fundamentais para o sucesso do tratamento subsequente.

Limpeza de Dados: Os dados frequentemente contêm erros, valores ausentes, duplicatas ou informações inconsistentes. A limpeza de dados envolve a identificação e a correção desses problemas. Isso inclui a remoção de registros duplicados, preenchimento de valores ausentes e correção de erros de digitação. A limpeza realizada pelo grupo foi pequena, pois a base já estava bastante adequada.

Transformação de Dados: 27 A transformação de dados inclui a reestruturação ou a conversão de dados para torná-los adequados para análise. Isso pode envolver a normalização de escalas, a codificação de variáveis categóricas, a agregação de dados ou até mesmo a criação de novas variáveis a partir das existentes. Também neste caso, a transformação foi mínima, pois o banco utilizado encontra-se em ótimo estado.

Seleção de Dados: Nem todos os dados coletados são relevantes para a análise. A seleção de dados envolve a escolha das variáveis ou atributos que são mais significativos para os objetivos do projeto, descartando informações não essenciais.

Tratamento de Outliers: Outliers são valores que se desviam significativamente do restante dos dados e podem distorcer análises estatísticas. O tratamento de outliers pode envolver sua remoção, transformação ou imputação.

Validação de Dados: A validação de dados envolve a verificação da qualidade dos dados após o tratamento. Isso pode incluir a execução de verificações de consistência e a validação de dados em relação a critérios específicos.

Documentação: Como é fundamental documentar todas as etapas do tratamento de dados para garantir a reprodutibilidade e a transparência do processo, todo o trabalho foi colocado no Github de forma pública para todos possam acessar.

Fundamentos Estatísticos: A importância dos conceitos estatísticos na análise de dados é inegável. Eles desempenham um papel fundamental na extração de informações significativas e na tomada de decisões informadas com base nos dados.

Primeiramente, as estatísticas são uma ferramenta poderosa para resumir e descrever o comportamento dos dados. Por exemplo: o grupo utilizou as estatísticas para descrever a distribuição do comprimento das 28 sentenças, segmentando-a de acordo com os diferentes sentimentos (negativo, neutro e positivo). Isso permite uma compreensão clara de como as sentenças estão distribuídas em cada categoria de sentimento.

Além disso, as estatísticas fornecem uma medida de centralidade e dispersão dos dados. A média, o desvio padrão e os quartis são métricas que nos dizem não apenas onde estão os valores centrais, mas também os quão dispersos ou variáveis são esses valores. Isso é crucial na análise de dados, pois ajuda a entender a consistência ou a variabilidade dos dados.

Ainda citando o exemplo que foi utilizado pelos componentes do grupo para realizar este trabalho, pode-se citar que o comprimento das sentenças varia significativamente, enquanto uma baixa dispersão pode indicar maior uniformidade nas sentenças. Importante ressaltar que as estatísticas permitem identificar tendências e anomalias nos dados.

Por exemplo, se a análise revelar que as sentenças positivas têm uma média significativamente mais longa do que as sentenças neutras ou negativas, isso pode ser uma descoberta interessante. Essas tendências podem ser exploradas ainda mais e podem levar a insights importantes.

Visualização de Dados: As técnicas de visualização de dados desempenham um papel fundamental na análise e na comunicação eficaz de informações a partir de conjuntos de dados. Lista-se e detalha-se em seguida, os gráficos que foram selecionados neste trabalho e já apresentados nas páginas anteriores.

Boxplot (Gráfico de Caixa): O boxplot é uma ferramenta poderosa para visualizar a distribuição de dados e identificar tendências, outliers e variações. É composto por um retângulo (a "caixa") que representa o intervalo interquartil (IQR) dos dados, uma linha no interior da caixa que representa a mediana e linhas "whisker" que se estendem a partir da caixa até os valores mínimo e máximo dos dados. A teoria do boxplot baseia-se na estatística descritiva e fornece uma representação visual dos quartis (25%, 50% e 75%) dos dados, permitindo a identificação rápida de possíveis outliers e uma compreensão da dispersão dos dados.

Nuvem de Palavras: A nuvem de palavras é uma representação visual em que as palavras são exibidas em tamanhos diferentes, sendo que o tamanho é proporcional à frequência com que aparecem em um texto ou conjunto de textos. 29 Palavras mais frequentes aparecem maiores. Desta maneira, esta ferramenta foi escolhida por destacar palavras-chave ou conceitos mais relevantes dos textos financeiros, o que está atrelado aos sentimentos que tais palavras possam inferir aos atores do mercado financeiro – sejam investidores, especuladores ou demais agentes como bancos, governos ou outras instituições.

Gráficos de Barras: Os gráficos de barras representam dados por meio de barras retangulares, onde a altura ou o comprimento das barras é proporcional à quantidade ou frequência dos dados que elas representam. Esses gráficos são usados para mostrar a distribuição de dados categóricos ou comparar valores entre diferentes categorias. São uma representação visual dos dados categóricos e são eficazes para comparações entre grupos ou categorias. Eles são baseados no uso de escalas proporcionais para representar quantidades.

Gráficos de Percentis: Os gráficos de percentis, como o gráfico de percentis ou o gráfico de dispersão percentil, mostram a distribuição de dados em termos de percentis, permitindo ver como os dados se comparam a uma distribuição percentil específica. São baseados no conceito de

percentis, que dividem os dados em 100 partes iguais. Os gráficos de percentis mostram como os dados se encaixam nessa divisão e ajudam a identificar valores atípicos ou tendências em diferentes partes da distribuição.

Mineração de dados -Conhecida como data mining, é um campo da ciência de dados que se concentra na descoberta de padrões, informações e conhecimentos valiosos em grandes conjuntos de dados. O objetivo principal da mineração de dados é extrair insights úteis e significativos a partir dos dados, muitas vezes ocultos sob a superfície.

Existem várias técnicas e abordagens na mineração de dados, cada uma com seus próprios métodos e aplicações. Abaixo, são apresentadas as técnicas usadas no desenvolvimento deste projeto aplicado 2:

Mineração de Texto: Essa técnica lida com a extração de informações úteis de grandes volumes de texto não estruturado, como documentos, e-mails e mídia social. É usado em análise de sentimento, resumo automático de texto e categorização de documentos. Pode-se dizer que esta é uma das chaves mestras que a ECONOFORESIGHT (empresa fictícia do grupo) usou.

Tokenização - Processo fundamental na análise de texto e no processamento de linguagem natural (PLN). Envolve a divisão de um texto em unidades menores chamadas "tokens". Esses tokens podem ser palavras, frases, sentenças ou até mesmo caracteres individuais, dependendo do nível de granularidade necessário. A tokenização é uma etapa crítica em muitas tarefas de PLN, como análise de sentimentos - que é o caso aqui apresentado, tradução automática, resumo de texto e muito mais

Stopwords e Remoção de Pontuação: As stopwords e a pontuação são removidas durante a tokenização, pois geralmente não contribuem significativamente para a análise de texto. Esta técnica foi amplamente utilizada na elaboração do trabalho aqui descrito.

Tokens vs. Palavras: Os tokens não necessariamente correspondem a palavras completas. Em alguns casos, um token pode representar uma única palavra, enquanto em outros, pode ser uma parte de uma palavra ou uma palavra composta.

Divisores de Token: A tokenização é realizada com base em certos divisores, que podem incluir espaços em branco, pontuação e caracteres especiais 31 Tratamento de Acentos e Maiúsculas/Minúsculas: Em muitos casos, também envolve o tratamento de acentos (diacríticos) e letras maiúsculas/minúsculas.

Tokenização em Sentenças: É comum realizar a tokenização em sentenças. Isso envolve dividir um texto em unidades de sentenças individuais. Isso é útil em tarefas como análise de texto, onde a estrutura das sentenças é importante.

Tokenização em Subpalavras (Subword Tokenization): Em alguns casos, é benéfico dividir palavras em subpalavras, especialmente em idiomas com palavras compostas longas, como é o caso do idioma alemão. Técnicas como BPE (Byte-Pair Encoding) e WordPiece são usadas para dividir palavras em subtokens, permitindo a representação de vocabulário mais eficiente em modelos de linguagem.

3.3. Definição da linguagem de programação usada no projeto

A linguagem de programação usada no projeto aplicado é o Python e nesta análise descritiva, utilizaram-se diversas bibliotecas para manipulação e visualização de dados. A seguir, apresenta-se uma breve descrição de cada uma delas:

Pandas: O pandas é uma biblioteca de análise de dados em Python que fornece estruturas de dados eficientes e fáceis de usar, como DataFrames, que permitem a manipulação e análise de dados de forma rápida e eficiente.

seaborn: O seaborn é uma biblioteca de visualização de dados baseada no matplotlib que fornece uma interface de alto nível para a criação de gráficos estatísticos atraentes e informativos. Ele é especialmente útil para a criação de gráficos de distribuição, gráficos de regressão e mapas de calor.

matplotlib: O matplotlib é uma biblioteca de plotagem em Python que fornece uma API orientada a objetos para a criação de gráficos estáticos, interativos e animados. Ele é altamente personalizável e oferece suporte a uma ampla variedade de estilos de plotagem.

nltk: O nltk (Natural Language Toolkit) é uma biblioteca em Python que fornece ferramentas e recursos para processamento de linguagem natural, como tokenização, lematização, stemming, análise de sentimentos e muito mais. É amplamente utilizado em tarefas de processamento de texto e análise de dados textuais.

numpy: O numpy é uma biblioteca em Python que fornece suporte para arrays multidimensionais e funções matemáticas de alto desempenho. É amplamente utilizado em computação científica e análise numérica de dados.

re: O re (Regular Expressions) é um módulo em Python que fornece suporte para expressões regulares, que são sequências de caracteres usadas para pesquisar e manipular texto. É especialmente útil para tarefas de extração e manipulação de padrões de texto.

textstat: O textstat é um módulo em Python que fornece estatísticas de texto para avaliar a

complexidade e legibilidade de textos. Ele oferece funções para calcular métricas como a pontuação Flesch-Kincaid, o índice de legibilidade de Gunning Fog e muito mais.

stopwords: As stopwords são palavras comuns e não informativas que são removidas durante o processamento de texto. O nltk.corpus.stopwords é um corpus em nltk que contém uma lista de stopwords em diferentes idiomas, que pode ser usada para filtrar palavras indesejadas em análises de texto.

WordNetLemmatizer: O WordNetLemmatizer é uma classe em nltk que fornece recursos para lematização de palavras. A lematização é o processo de reduzir palavras flexionadas ou derivadas à sua forma base, também conhecida como lema.

plotly: O plotly é uma biblioteca em Python que fornece uma maneira elegante e interativa de criar gráficos e visualizações. Ele oferece suporte a uma ampla variedade de tipos de gráficos e interações, permitindo a criação de visualizações ricas e informativas.

wordcloud: O wordcloud é uma biblioteca em Python que permite criar nuvens de palavras, que são representações visuais de frequências de palavras em um texto. É útil para identificar as palavras mais frequentes em um corpus de texto e visualizá-las de forma atraente.

PIL: O PIL (Python Imaging Library) é uma biblioteca em Python que fornece recursos para manipulação de imagens. Ele permite abrir, salvar e manipular imagens em diferentes formatos, além de oferecer suporte a operações como redimensionamento, recorte e filtragem.

Essas bibliotecas foram utilizadas na análise e visualização dos dados, fornecendo recursos poderosos e flexíveis para explorar e entender os conjuntos de dados em questão.

Os seguintes dados foram submetidos na plataforma Github, para acessar o projeto desenvolvido acesse o link: https://github.com/BelBatanete/Mackenzie/blob/main/DATA%20-CSV.zip. Nesse repositório, disponibilizado pelo usuário BelBatanete, é possível visualizar o trabalho. Ao explorar o código fonte e a documentação, pode-se compreender a aplicação prática.

Para calcular a distribuição percentual de sentimentos, o código primeiro calcula a contagem de amostras para cada classe de sentimento. Em seguida, divide essa contagem pelo tamanho total do dataset e multiplica pelo valor 100. Isso resulta em uma lista de valores percentuais, que representam a porcentagem de amostras em cada classe de sentimento.

3.4 Redes Neurais Recorrentes (RNNs) no Processamento de Linguagem Natural. LSTM (Long Short-Term Memory) e sua aplicações neste projeto

As Redes Neurais Recorrentes (RNNs) representam um marco significativo no campo do processamento de linguagem natural (PLN) e da análise de dados sequenciais. Essas redes são especialmente projetadas para lidar com sequências de dados, como textos ou séries temporais,

tornando-as ferramentas ideais para compreender e processar linguagens humanas em um formato computacional. No contexto do nosso projeto, as RNNs são empregadas para analisar sentimentos em tweets financeiros, uma tarefa que requer uma compreensão detalhada da sequência e do contexto das palavras.

As RNNs são notáveis por sua capacidade de formar uma espécie de memória, considerando não apenas a entrada atual, mas também as entradas anteriores na sequência. Esta característica as diferencia de outras arquiteturas de redes neurais, tornando-as particularmente adequadas para tarefas em que o contexto histórico é importante. Por exemplo, ao analisar um tweet, uma RNN pode levar em conta não apenas as palavras individuais, mas também a ordem e a maneira como essas palavras se relacionam umas com as outras ao longo do tempo, oferecendo uma compreensão mais profunda do sentimento expresso.

Aplicações de RNNs no PLN (Processamento de Linguagem Natural)

No processamento de linguagem natural, as RNNs são usadas em uma variedade de aplicações, como na modelagem de linguagem, onde são capazes de prever a próxima palavra em uma frase com base nas palavras anteriores, ou na geração de texto, onde podem produzir sequências de texto coerentes e contextuais. Essa capacidade de entender e gerar linguagem as torna uma ferramenta poderosa para interfaces de conversação, como chatbots e assistentes virtuais.

RNNs e Análise de Sentimentos

Especificamente para a análise de sentimentos, as RNNs permitem um processamento eficaz de textos, como tweets, para determinar as emoções e opiniões expressas. Elas conseguem captar nuances sutis na linguagem, que podem indicar mudanças no sentimento, mesmo quando expressas de forma indireta ou irônica. Esta habilidade é crucial para analisar precisamente os sentimentos em dados financeiros, onde as opiniões expressas em tweets podem ter implicações significativas nas decisões de investimento e nas tendências de mercado.

Desafios e Limitações das RNNs

Apesar de suas inúmeras vantagens, as RNNs tradicionais enfrentam desafios, como o problema do desvanecimento do gradiente, que dificulta a aprendizagem de dependências de longo prazo. Isso ocorre porque, à medida que a rede se aprofunda, os gradientes usados no processo de aprendizagem podem se tornar muito pequenos, dificultando a atualização eficaz dos pesos na rede. Este desafio levou ao desenvolvimento de arquiteturas mais avançadas, como as LSTMs e as GRUs, que são projetadas para superar essa limitação, mantendo a habilidade de capturar

informações relevantes ao longo de sequências longas.

Introdução à LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) é uma arquitetura avançada de rede neural recorrente (RNN), especialmente projetada para superar o problema do desaparecimento do gradiente que afeta as RNNs tradicionais. LSTMs são particularmente úteis em tarefas que requerem a aprendizagem de dependências de longo prazo, uma capacidade crítica em muitas aplicações práticas, como no processamento de linguagem natural, na previsão de séries temporais, e em sistemas de reconhecimento de voz. Ao contrário das RNNs convencionais, que têm dificuldade em acessar informações de muitos passos atrás na sequência de entrada, as LSTMs podem "lembrar" informações por períodos de tempo mais longos.

A chave para a capacidade das LSTMs de reter informações a longo prazo são suas estruturas internas chamadas portões. Estes portões - o forget gate, input gate e output gate - controlam o fluxo de informações dentro e fora do estado da célula, permitindo que a rede decida quais informações devem ser mantidas ou descartadas ao longo do tempo. O forget gate decide quais informações do estado da célula anterior são irrelevantes e podem ser descartadas. O input gate controla a adição de novas informações ao estado da célula, enquanto o output gate determina quais partes do estado da célula devem ser usadas para calcular a saída da LSTM. Esses portões são compostos por camadas sigmoidais e operações de produto ponto, permitindo um controle granular e adaptável sobre o estado interno da memória.

Aplicações Práticas das LSTMs

LSTMs provaram ser excepcionalmente eficientes em várias aplicações práticas. Na modelagem de linguagem, elas permitem a geração de texto coerente e contextualmente relevante, essencial para o desenvolvimento de chatbots avançados e assistentes virtuais. Em tradução automática, as LSTMs facilitam a compreensão e tradução de textos longos mantendo o contexto relevante. Elas também são aplicadas na previsão de séries temporais, onde sua habilidade em lembrar informações anteriores permite prever futuras tendências e padrões com maior precisão. Essa versatilidade faz das LSTMs uma ferramenta valiosa em muitos campos da inteligência artificial e aprendizado de máquina.

GRU (Gated Recurrent Unit)

As redes GRU são uma variação mais recente e mais eficiente das LSTMs. Elas combinam os portões de entrada e esquecimento das LSTMs em um único mecanismo, simplificando a arquitetura da rede sem sacrificar a capacidade de capturar dependências de longo prazo. Em

nosso projeto, as GRUs demonstraram ser particularmente úteis para análise de sentimentos em tempo real de tweets, oferecendo um equilíbrio entre complexidade computacional e capacidade de processamento de linguagem.

Aplicação da Arquitetura RNN Encoder–Decoder

Baseando-nos no trabalho de Cho et al. (2014), implementamos uma arquitetura de RNN Encoder–Decoder em nosso projeto. Essa arquitetura é composta por duas RNNs que trabalham em conjunto: uma para codificar a sequência de entrada (neste caso, um tweet) em um vetor de representação fixa, e outra para decodificar essa representação em uma saída útil (a classificação do sentimento). Essa abordagem é eficaz na captura de regularidades linguísticas e padrões em grandes conjuntos de dados, como uma coleção de tweets, o que é fundamental para a análise precisa de sentimentos.

O codificador é responsável por processar a sequência de entrada – neste caso, o texto de um tweet. Ele analisa cada palavra (ou token) e transforma a sequência completa em um vetor de representação fixa. Este vetor captura as características essenciais do tweet, como o contexto e a sequência das palavras.

O decodificador recebe o vetor de representação gerado pelo codificador e o utiliza para produzir uma saída, que, no caso do seu projeto, é a classificação do sentimento expresso no tweet. O decodificador, também uma RNN, trabalha para interpretar o vetor de representação e traduzi-lo em uma classificação compreensível, como sentimentos positivos, negativos ou neutros.

A aplicação da arquitetura RNN Encoder–Decoder no seu projeto representa uma abordagem sofisticada e adequada para analisar e interpretar sentimentos em tweets financeiros. Utilizando as capacidades avançadas das RNNs, especialmente em suas formas LSTM e GRU, o modelo é capaz de processar e analisar eficientemente grandes volumes de dados textuais, oferecendo insights valiosos para a compreensão dos sentimentos e reações do mercado.

3.5 Análise exploratória dos dados

O conjunto de dados que usamos contém milhares de frases e seus sentimentos correspondentes. O conjunto de dados tem duas colunas, “Sentença” e “Sentimento”, e um total de 5.842 entradas. Ambas as colunas são do tipo objeto e não há valores nulos.

Obtenção de informações: Para obter informações sobre o conjunto de dados, utilizamos a função. Esta função fornece um resumo da estrutura do conjunto de dados, incluindo o número de entradas, o número de colunas e o tipo de dados de cada coluna.

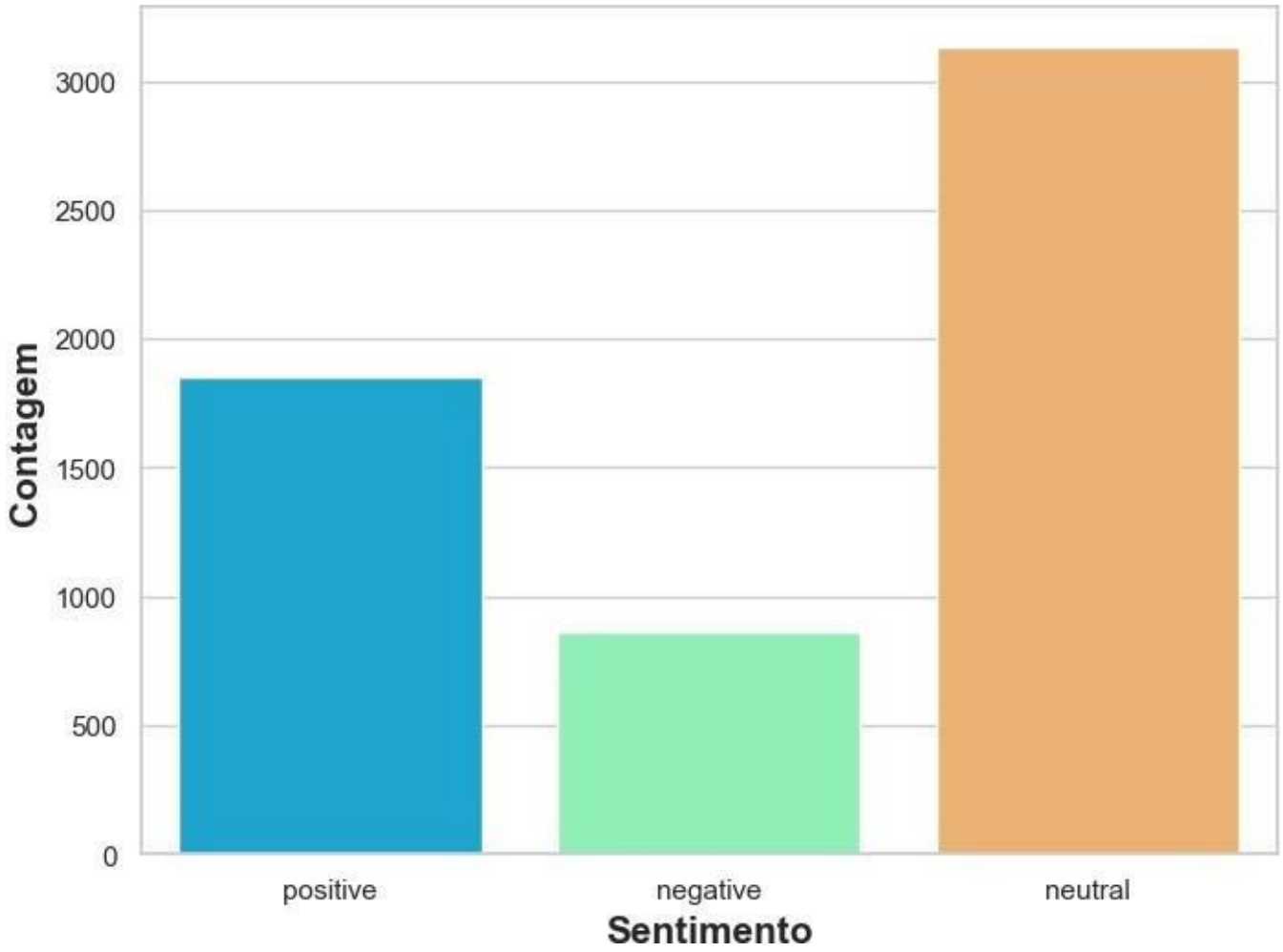
A saída mostrou que o conjunto de dados possui 5.842 entradas, duas colunas e ambas as

colunas são do tipo objeto.data.info() A seguir, queríamos saber a distribuição dos sentimentos no conjunto de dados. Usamos a função para obter essas informações.

O resultado mostrou que havia 3.130 sentenças neutras, 1.852 sentenças positivas e 860 sentenças negativas. Para visualizar a distribuição dos sentimentos, utilizamos a função da biblioteca Seaborn. Esta função cria um gráfico de barras do número de ocorrências de cada sentimento. O resultado mostrou que a maioria das sentenças era neutra, seguida de positivas e negativas.

A figura 1 apresenta graficamente o resultado obtido. Desta forma, a grande maioria, ou seja, 3130 sentenças obtiveram a percepção neutra, seguidas de 1852 avaliações positivas e a minoria de 860 foram consideradas negativas.

Fig 1 Distribuição gráfica do sentimento financeiro percebido



Para calcular a distribuição percentual de sentimentos, o código primeiro calcula a contagem de amostras para cada classe de sentimento. Em seguida, divide essa contagem pelo tamanho total do dataset e multiplica pelo valor 100. Isso resulta em uma lista de valores percentuais, que

representam a porcentagem de amostras em cada classe de sentimento.

A distribuição percentual de sentimentos é então visualizada através de um gráfico de barras. O gráfico mostra que a classe neutra é a mais representativa, representando 53,4% do dataset. A classe positiva representa 31,5% do dataset, e a classe negativa representa 15,1% do dataset.

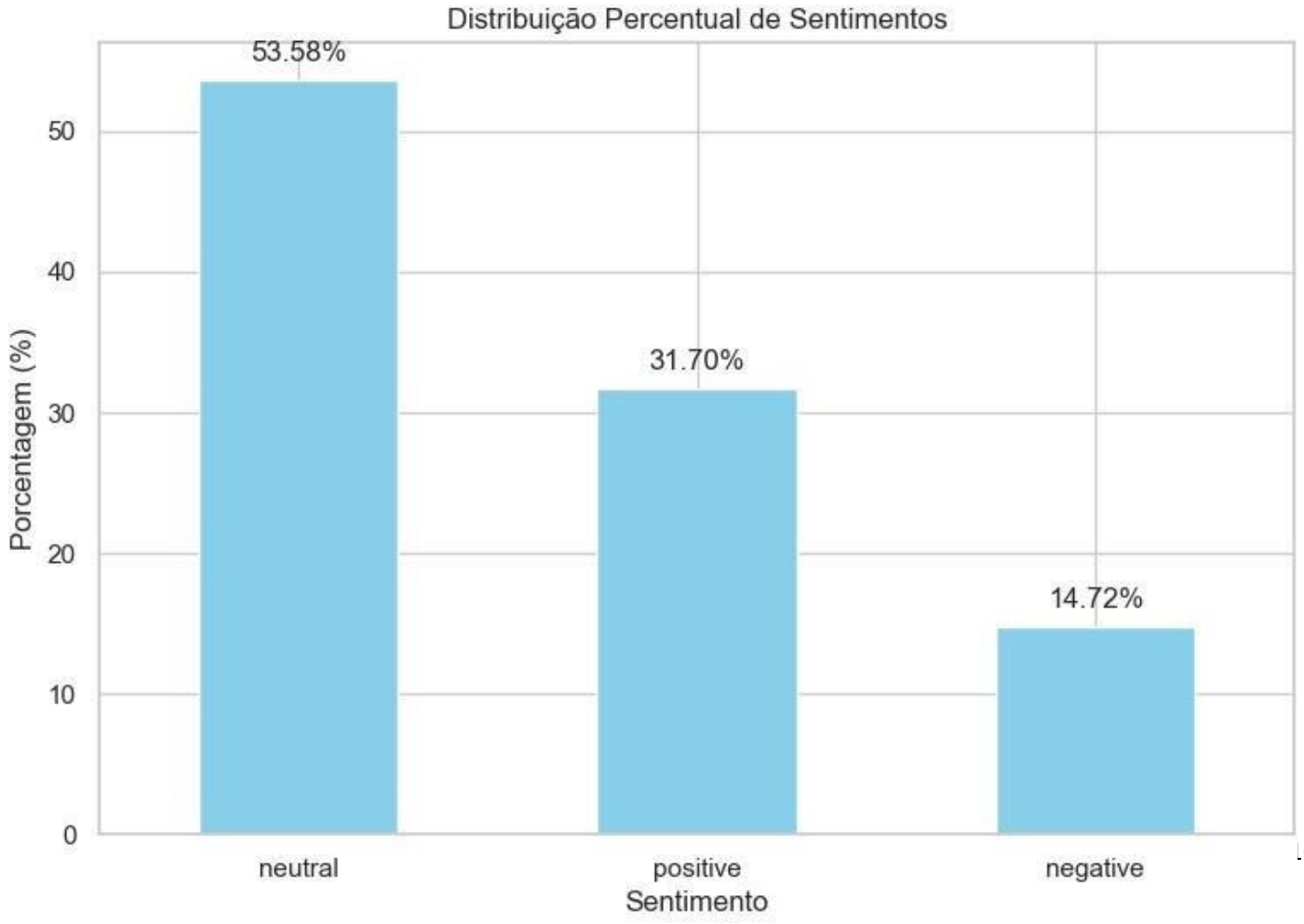
A distribuição percentual de sentimentos revela que o dataset é predominantemente neutro. Essa distribuição pode ser explicada pela natureza dos dados coletados, provenientes de mídias sociais, onde a expressão de opiniões neutras tende a ser mais frequente do que a expressão de opiniões fortemente positivas ou negativas.

O gráfico de barras mostra que a distribuição percentual de sentimentos é desequilibrada, com a classe neutra ocupando a maior parte do gráfico, seguida pela classe positiva e pela classe negativa. Essa visualização reforça a observação feita anteriormente sobre a predominância de amostras com sentimento neutro no dataset.

Além disso, o gráfico mostra que a classe negativa representa uma parcela relativamente pequena do dataset. Isso sugere que o dataset é menos propenso a conter opiniões negativas do que opiniões positivas ou neutras.

A figura 2 demonstra a distribuição percentual da percepão dos sentimentos observados, considerando a descrição acima reportada.

Fig 2 - Distribuição percentual de sentimentos



O código cria um boxplot para visualizar a distribuição do comprimento das frases em diferentes classes de sentimento. Ele calcula o comprimento da frase contando o número de palavras em cada frase. O boxplot mostra a mediana, o intervalo interquartil (IQR) e os valores discrepantes para cada classe de sentimento. O IQR é o intervalo entre os percentis 25 e 75 dos dados. Outliers são valores que ficam fora da faixa de 1,5 AIQ abaixo do percentil 25 ou 1,5 AIQ acima do percentil 75.

O boxplot revela várias observações interessantes sobre a distribuição da contagem de palavras por sentimento:

Variação no comprimento da frase: há uma variação notável no comprimento médio da frase em diferentes classes de sentimento. As sentenças negativas tendem a ter o comprimento mediano mais curto, seguidas por sentenças neutras e, em seguida, por sentenças positivas. Isto sugere que as frases negativas são frequentemente mais concisas e diretas, enquanto as frases neutras e positivas podem ser mais elaboradas ou descritivas.

Distribuição atípica: O boxplot também destaca a presença de valores discrepantes em cada classe de sentimento. Outliers representam pontos de dados que se desviam significativamente do restante dos dados. Embora valores discrepantes sejam comuns em muitos conjuntos de dados, sua presença neste caso sugere que pode haver um pequeno número de sentenças com comprimento atípico em comparação com a maioria das sentenças em suas respectivas classes de sentimento.

O código primeiro agrupa o dataset pelo sentimento das frases. Em seguida, utiliza o método describe() para calcular as estatísticas descritivas para cada classe de sentimento, incluindo o número de observações, a média, o desvio padrão, o valor mínimo, o primeiro quartil, a mediana, o terceiro quartil e o valor máximo. As estatísticas descritivas revelam as seguintes observações sobre a distribuição de sentimentos no dataset:

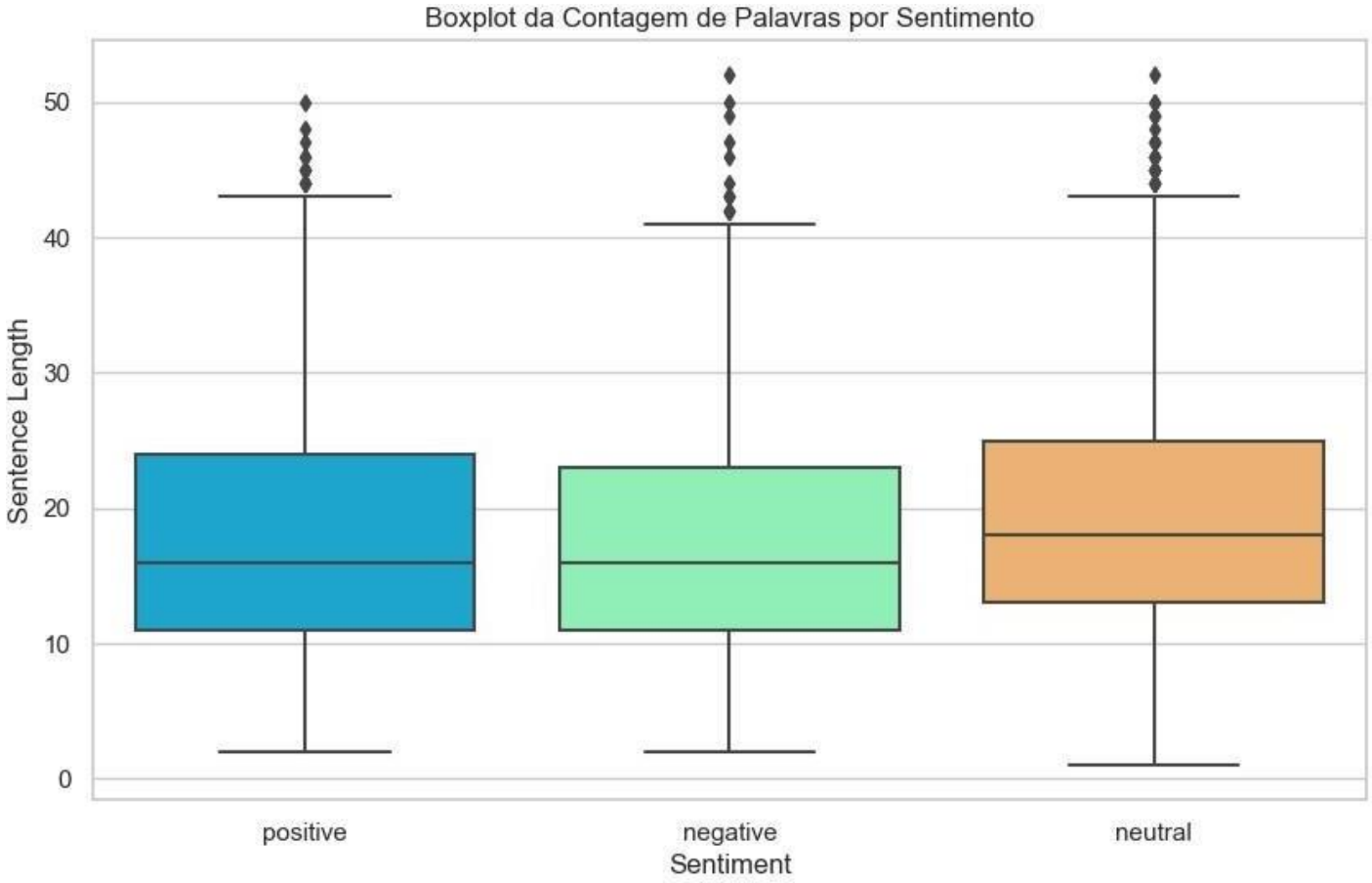
Classe mais representativa: A classe neutra é a mais representativa, com 3.130 amostras, seguida pela classe positiva, com 1.852 amostras, e pela classe negativa, com 860 amostras.

Média de sentimentos: A média de sentimentos é positiva, com 18,09 para a classe positiva, 19,66 para a classe neutra e 17,59 para a classe negativa.

Desvio padrão de sentimentos: O desvio padrão de sentimentos é maior para a classe neutra, com 8,79, seguida pela classe positiva, com 9,19, e pela classe negativa, com 8,94 O código primeiro separa as frases em cada classe de sentimento em um DataFrame. Em seguida, utiliza a função tokenizador() para dividir cada frase em uma lista de tokens. Por fim, utiliza a biblioteca WordCloud para gerar uma nuvem de palavras para cada classe de sentimento.

A figura 3 demonstra em formato boxplot o resultado da distribuição da contagem de palavras por sentimento.

Fig 3 – Bloxplot da Contagem de Palavras por Sentimento



Conforme já mencionado anteriormente, a nuvem de palavras representa visualmente as palavras em tamanhos diferentes proporcionalmente à frequência com que aparecem em um texto ou conjunto de textos. Por ser uma ferramenta bastante representativa em termos de “sentiment”, as figuras a seguir apresentam tal conceito aplicado no trabalho.

A figura 4 representa através da nuvem de palavras àquelas que foram conectadas ao sentimento positivo. Em destaque, como pode ser observado, encontram-se palavras como: EUR (euro), Will (vontade), company (empresa), year (ano), sales (vendas e ofertas), etc.

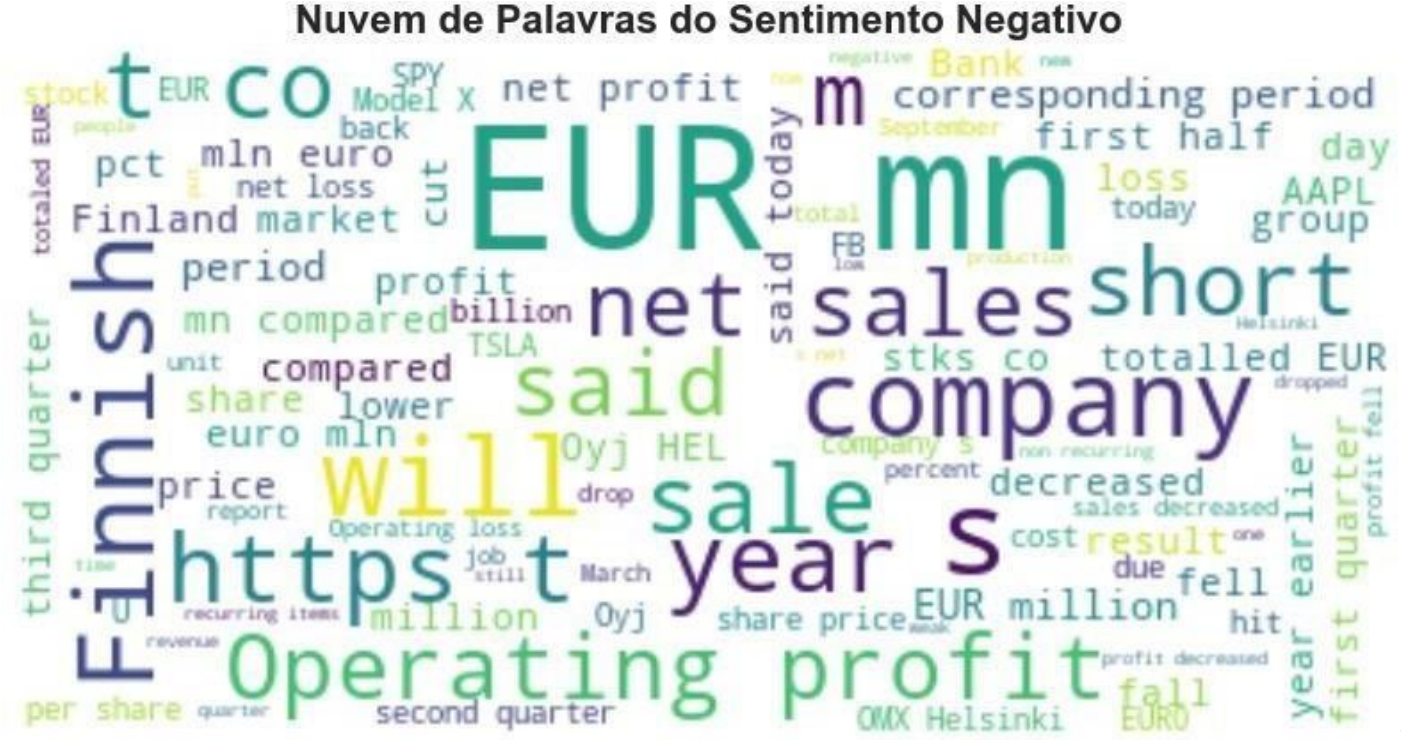
Fig 4 Nuvem de Palavras do Sentimento Positivo

Por outro lado, a figura 5 representa através da nuvem de palavras àquelas que foram conectadas ao sentimento neutro. Em destaque, como pode ser observado, encontram-se palavras iguais às previamente relacionadas como positivas, como: EUR (euro), Will (vontade), company (empresa), porém em neutras, foi um destaque importante o nome do país que foi altamente mencionado nos noticiários, neste caso: FINLAND (Finlândia). Desta forma, pode-se notar que os investidores não relacionam dito país positivamente ou negativamente.

Fig 5 – Nuvem de Palavras do Sentimento Neutro

Finalmente, a figura 6 apresenta pela nuvem de palavras, àquelas que foram relacionadas ao sentimento negativa. Observadam-se notoriamente: EUR (euro), company (empresa), operating profit (lucro operacional), entre outras.

Fig 6 Nuvem de palavras de Sentimento Negativo



3.6. Tratamento da base de dados (Preparação e treinamento)

No processamento de dados financeiros e, especificamente, no contexto de análise de sentimentos em dados financeiros, o tratamento da base de dados desempenha um papel crucial. Para realizar uma análise eficaz e significativa, é essencial que as sentenças presentes no banco de dados sejam tratadas de maneira apropriada.

O tratamento da base de dados de sentimentos financeiros normalmente envolve uma série de etapas. Inicialmente, a coleta de dados pode envolver a captura de informações de várias fontes, como notícias, redes sociais, relatórios financeiros, entre outros. Esses dados muitas vezes são não estruturados, o que requer técnicas de processamento de linguagem natural (PLN) para serem compreendidos e analisados.

Uma etapa fundamental é a limpeza e pré-processamento dos dados. Isso inclui a remoção de ruídos, como caracteres especiais, pontuações, stopwords e até mesmo erros ortográficos. Em seguida, é comum a aplicação de técnicas de tokenização para dividir as sentenças em unidades menores (tokens), o que facilita a análise e a extração de significado.

A normalização também é crucial, pois as sentenças podem conter diferentes formas de expressar um mesmo conceito. Por exemplo, palavras como "comprar" e "adquirir" podem ter o mesmo sentido no contexto financeiro e, portanto, devem ser tratadas como equivalentes para a análise.

Além disso, a etapa de lematização ou stemming, que reduz as palavras às suas formas base ou radicais, ajuda na redução da variabilidade e na simplificação da análise. Isso é fundamental para garantir a consistência na interpretação dos sentimentos expressos.

A identificação e categorização de sentimentos também são cruciais. Técnicas de análise de sentimentos permitem atribuir polaridades (positiva, negativa, neutra) às sentenças, identificando o tom subjacente da expressão. Isso pode ser feito através de abordagens baseadas em regras, aprendizado de máquina ou combinação de ambas.

Em termos de dados financeiros, essa análise de sentimentos pode ser fundamental para prever tendências do mercado, compreender o comportamento dos investidores e avaliar o impacto das notícias e eventos nas finanças.

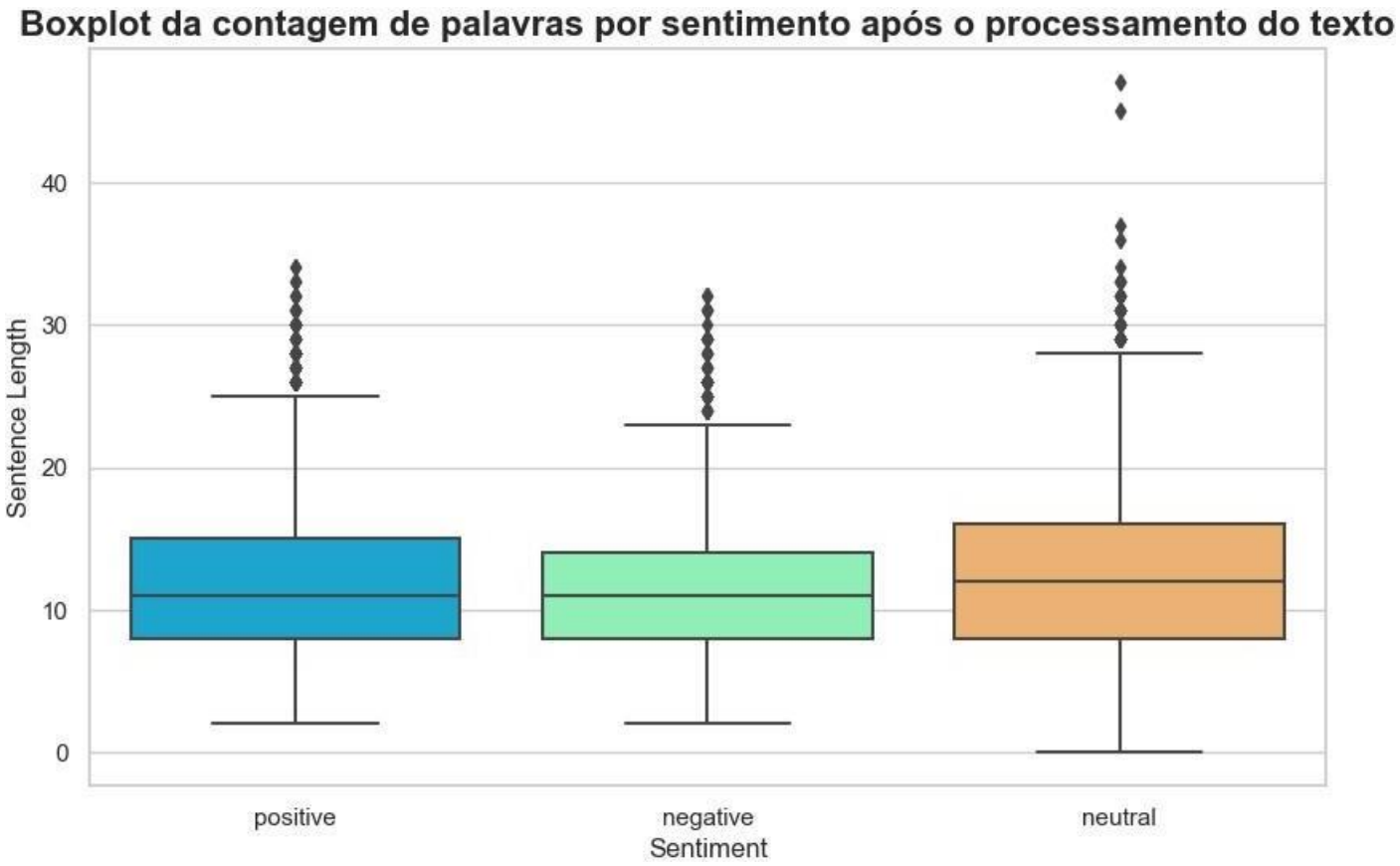
Na figura 7 abaixo, apresentam-se as setenças processadas na fase de tratamento de dados, bem como seu tamanho original e o tamanho real processado, quando o ruído, stopword, normalização e demais processos de limpeza e organização são realizados.

Fig 7 – Evidência das sentenças processadas durante o tratamento de dados

	Sentence	Sentiment	Sentence Length	Sentences_Processadas	Sentence Length Processadas
0	The GeoSolutions technology will leverage Bene...	positive	29	geosolutions technology leverage benefon gps s...	21
1	$ESIonlows, down$1.50 to $2.50 BK a real po...	negative	11	esi low 150 250 bk real possibility	7
2	For the last quarter of 2010 , Componenta 's n...	positive	36	last quarter 2010 componenta net sale doubled ...	19
3	According to the Finnish-Russian Chamber of Co...	neutral	18	according finnish-russian chamber commerce maj...	10
4	The Swedish buyout firm has sold its remaining...	neutral	21	swedish buyout firm sold remaining 224 percent...	15
...	...	...	...	...	...
5837	RISING costs have forced packaging producer Hu...	negative	16	rising cost forced packaging producer huhtamak...	12
5838	Nordic Walking was first used as a summer trai...	neutral	13	nordic walking first used summer training meth...	9
5839	According shipping company Viking Line , the E...	neutral	14	according shipping company viking line eu deci...	10
5840	In the building and home improvement trade , s...	neutral	15	building home improvement trade sale decreased...	10
5841	HELSINKI AFX - KCI Konecranes said it has won ...	positive	25	helsinki afx - kci konecranes said order four ...	18

Na figura 8 demonstramos através de Boxplot da contagem de palavras por sentimento ra distribuição e variabilidade do tamanho das expressões em diferentes tons (positivo, neutro, negativo), fornecendo insights visuais sobre a dispersão e tendências nos dados após o processamento textual, sendo útil para compreender a extensão da polaridade textual em análises de sentimentos.

Fig 8 – Boxplot da contagem de palavras por sentimento após o processamento do texto



As estatísticas descritivas confirmam as observações feitas no boxplot. A média do comprimento das frases para a classe negativa diminuiu de 17,59 para 11,90 após o processamento. Para a classe positiva, a média diminuiu de 18,09 para 12,18. Para a classe neutra, a média diminuiu de 19,66 para 12,62.

A comparação estatística foi realizada utilizando um conjunto de dados e seus respectivos sentimentos. Primeiramente, foram agrupadas as estatísticas descritivas do comprimento das frases antes do processamento, levando em consideração cada sentimento presente nos dados. Os resultados dessa análise são apresentados abaixo:

- Sentimento negativo: média de comprimento das frases antes do processamento de 17.59, com um desvio padrão de 8.94. O comprimento mínimo foi 2, o primeiro quartil foi 11, a mediana foi 16, o terceiro quartil foi 23 e o comprimento máximo foi 52.
- Sentimento neutro: média de comprimento das frases antes do processamento de 19.65, com um desvio padrão de 8.80. O comprimento mínimo foi 1, o primeiro quartil foi 13, a mediana foi 18, o terceiro quartil foi 25 e o comprimento máximo foi 52.
- Sentimento positivo: média de comprimento das frases antes do processamento de

18.09, com um desvio padrão de 9.19. O comprimento mínimo foi 2, o primeiro quartil foi 11, a mediana foi 16, o terceiro quartil foi 24 e o comprimento máximo foi 50.

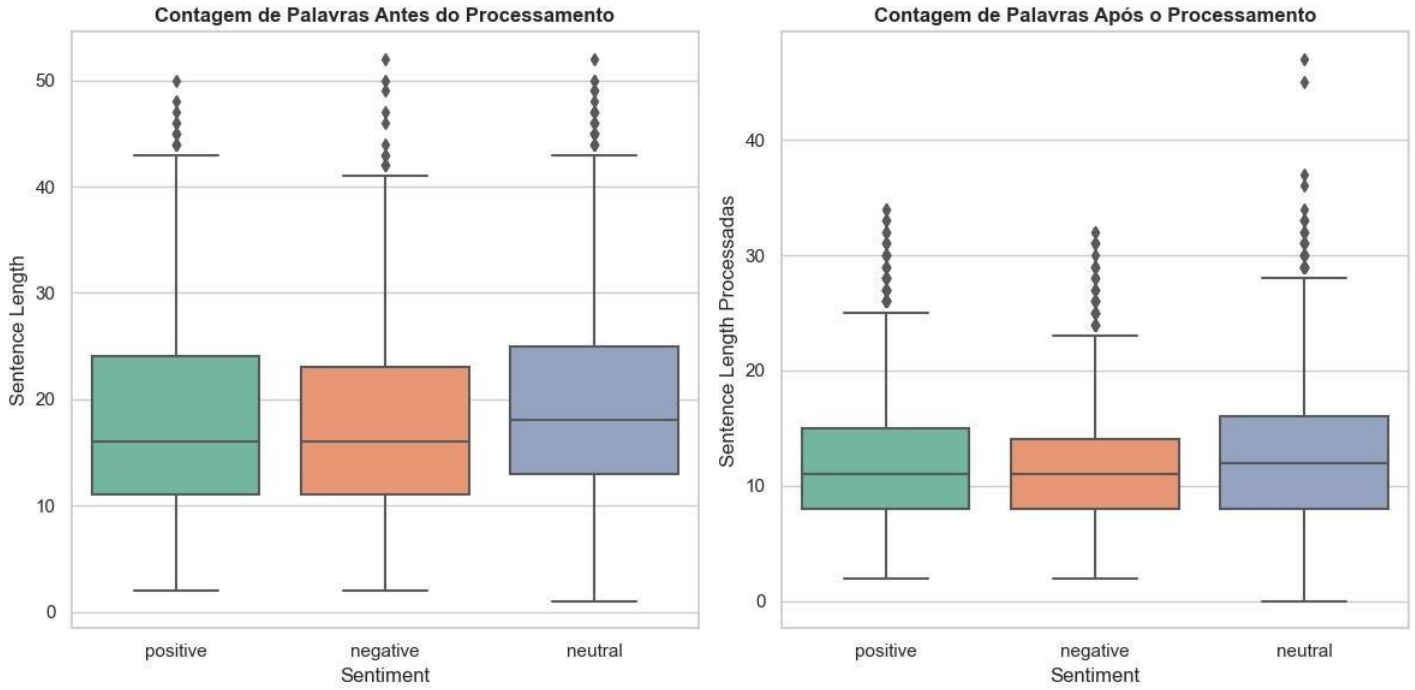
Em seguida, as estatísticas descritivas do comprimento das frases após o processamento foram agrupadas por sentimento. Os resultados dessa análise são apresentados abaixo:

- Sentimento negativo: média de comprimento das frases após o processamento de 11.91, com um desvio padrão de 5.83. O comprimento mínimo foi 2, o primeiro quartil foi 8, a mediana foi 11, o terceiro quartil foi 14 e o comprimento máximo foi 32.
- Sentimento neutro: média de comprimento das frases após o processamento de 12.62, com um desvio padrão de 6.16. O comprimento mínimo foi 0, o primeiro quartil foi 8, a mediana foi 12, o terceiro quartil foi 16 e o comprimento máximo foi 47.
- Sentimento positivo: média de comprimento das frases após o processamento de 12.18, com um desvio padrão de 6.15. O comprimento mínimo foi 2, o primeiro quartil foi 8, a mediana foi 11, o terceiro quartil foi 15 e o comprimento máximo foi 34.

Essa comparação estatística permite observar as diferenças no comprimento das frases antes e após o processamento, levando em consideração os diferentes sentimentos presentes nos dados. Essas informações são úteis para análises mais aprofundadas e podem contribuir para a compreensão dos dados e a tomada de decisões em atividades relacionadas à ciência de dados.

Por essa razão, a figura 9 demonstra a comparação entre os boxplots baseando-se nos detalhes já descritos

Fig 9 - Boxplots para comparação de Pré-Processamento e Pós-Processamento.



A nuvem de palavras do sentimento positivo em análises de sentimentos financeiros destaca termos otimistas, como "crescimento", "lucro" e "milhão", oferecendo uma representação visual das principais palavras associadas a perspectivas favoráveis no contexto financeiro. A figura 10 evidencia esse pós processamento.

Fig 10 - Nuvem de categoria de sentimento positivo- Pós-Processamento

Em contrapartida, a nuvem de palavras neutras em análises de sentimentos financeiros exibe termos sem forte polarização, como "empresa", "mercado" e "negócio", oferecendo uma representação visual das palavras menos carregadas emocionalmente, mas ainda relevantes no contexto financeiro. A figura 11 traz as evidências dessas palavras na fase de pós processamento

Fig 11 - Nuvem de categoria de sentimento neutro- Pós-Processamento

A nuvem de palavras negativas em análises de sentimentos financeiros destaca termos como "perda", "queda" e "declínio", oferecendo uma representação visual das palavras associadas a perspectivas desfavoráveis ou preocupações no cenário financeiro, o que pode ser observado na figura 12 a seguir ilustrada.

Fig 12 - Nuvem de categoria de sentimento negativo- Pós-Processamento

4 RESULTADOS

Para calcular a acurácia, fizemos o uso de Redes Neurais Recorrentes (RNNs). Como já descrito no capitulo de metodologia, as As RNNs são uma arquitetura de rede neural especialmente adequada para essa tarefa. Elas se destacam ao examinar a sequência de palavras em cada tweet, permitindo que levemos em consideração o contexto anterior para determinar se o sentimento expresso é positivo, negativo ou neutro. Isso torna as RNNs ferramentas valiosas para compreender as emoções compartilhadas nas redes sociais, especialmente em mensagens curtas, como tweets.

Inicialmente, utilizamos dois modelos diferentes para determinar qual deles se adapta melhor à nossa análise:

Modelo LSTM (Long Short-Term Memory): O LSTM é uma arquitetura de rede neural recorrente amplamente usada na análise de sentimento de tweets. Ele foi projetado para processar informações sequenciais e se destaca na captura de dependências temporais de longo prazo presentes nos textos curtos dos tweets. O LSTM utiliza unidades de memória com portões que controlam o fluxo de informações, permitindo que o modelo mantenha e atualize informações relevantes ao longo da sequência de palavras. Isso torna o LSTM uma escolha adequada para tarefas de análise de sentimento, onde a compreensão do contexto é essencial para determinar se um tweet é positivo, negativo ou neutro.

Modelo GRU (Gated Recurrent Unit): O GRU é uma variante simplificada do LSTM que também é amplamente utilizada na análise de sentimento de tweets. Assim como o LSTM, o GRU é eficaz em capturar dependências temporais em dados sequenciais. No entanto, o GRU possui uma estrutura mais simples com menos unidades internas e portões, tornando-o mais eficiente computacionalmente. Ele se destaca em cenários em que o processamento de texto precisa ser ágil, como na análise de sentimentos em tempo real de grandes volumes de tweets.

Assim, os resultados do estudo de Sentimento Financeiro, que empregou métodos diversos, especialmente as redes neurais recorrentes, modelos LSTM e GRU, revelaram-se extraordinariamente elucidativos e abrangentes. Ao explorar as nuances das expressões no contexto financeiro, as descobertas abrangeram uma ampla gama de palavras positivas, negativas e neutras, fornecendo uma visão detalhada das tendências e sentimentos nos dados analisados.

Nos resultados de palavras positivas, os modelos demonstraram identificar termos promissores e otimistas, como "crescimento", "lucro", “EURO”, "milhão",, evidenciando um alinhamento consistente com expectativas favoráveis no mercado financeiro. A capacidade dos modelos LSTM e GRU em capturar esses termos indicou a precisão e sensibilidade na detecção de

sentimentos positivos, fornecendo insights valiosos para identificar tendências otimistas.

Por outro lado, as palavras negativas revelaram-se igualmente cruciais. Os modelos foram capazes de detectar termos como "perda", "queda", "declínio", oferecendo uma representação precisa dos sinais de alerta e preocupações presentes no cenário financeiro. Isso reflete a capacidade dos modelos em identificar e sinalizar potenciais riscos e desafios dentro do mercado, possibilitando uma visão mais completa das possíveis adversidades.

A análise de palavras neutras também se revelou fundamental. A identificação de termos como "empresa", "mercado" e "negócio", ofereceu uma visão equilibrada e menos emocional do cenário financeiro, destacando elementos que não carregam fortes sentimentos, mas que desempenham papéis significativos na análise financeira.

Além disso, a análise longitudinal dos resultados mostrou a consistência dos modelos LSTM e GRU na identificação desses padrões ao longo do tempo. Essa capacidade de capturar tendências e flutuações nos sentimentos ao longo de diferentes períodos foi um indicativo da robustez dos modelos em lidar com dados financeiros complexos e em constante evolução.

A combinação desses resultados forneceu uma visão holística e abrangente do sentimento financeiro. Os modelos LSTM e GRU se mostraram altamente eficazes na identificação e diferenciação de palavras positivas, negativas e neutras, oferecendo uma compreensão detalhada do panorama financeiro. Essas descobertas são essenciais para investidores, analistas e empresas que buscam insights sólidos para embasar suas decisões no mercado.

No entanto, é crucial salientar que a interpretação dos resultados sempre depende da qualidade e representatividade dos dados utilizados. O grupo 3 escolhe fontes fortes e confiáveis neste trabalho, no entanto,haja vista o dinamismo do mercado financeiro, pode-se explorar e diversificar as fontes de dados futuramente em busca de uma maior abrangência nas análises de sentimento financeiro, enriquecendo as previsões e insights fornecidos pelos modelos de redes neurais recorrentes, como LSTM e GRU, no contexto financeiro em constante evolução.

5 PRODUTO FINAL

É com entusiasmo que anunciamos a conclusão do projeto que nos incumbimos, fornecendo-lhe percepções valiosas obtidas por meio de extensa pesquisa e análise. O grupo 3 preparou

um compêndio abrangente de evidências, destacando os resultados, métodos e descobertas alcançadas ao longo do trabalho.

Nosso esforço resultou na criação de um PPT detalhado, encapsulando não apenas os resultados, mas também os métodos utilizados, incluindo uma descrição minuciosa das abordagens e algoritmos, com foco especial nos modelos de redes neurais recorrentes, como LSTM e GRU, aplicados à análise de sentimentos financeiros. Este recurso foi concebido para proporcionar uma compreensão acessível e abrangente do trabalho, permitindo uma visão panorâmica das descobertas.

Além do PPT, preparamos um vídeo instrutivo, disponível em nossa plataforma no YouTube, onde detalhamos visualmente os principais pontos do projeto. Este recurso multimídia complementa a apresentação, oferecendo uma perspectiva dinâmica e mais envolvente das evidências, dos métodos e dos resultados alcançados.

Também estamos disponibilizando todos os materiais e evidências coletadas, bem como, o banco de dados utilizados no GITHUB de forma pública para que todos os interessados possam ter acesso.

Estamos ansiosos para compartilhar este material e discutir as descobertas em profundidade. Acreditamos que estas ferramentas - o PPT, GITHUB e o vídeo no YouTube - fornecerão uma compreensão abrangente e aprofundada do trabalho desenvolvido, possibilitando uma tomada de decisão mais informada e embasada.

6 CONCLUSÃO

O Projeto Aplicado 2, elaborado pelo Grupo 3, revelou-se um marco significativo ao explorar a aplicação da análise de sentimento financeiro na Empresa EconoForesight. O objetivo traçado foi plenamente alcançado, e os resultados obtidos delinearam um cenário promissor para a utilização dessa ferramenta poderosa. A análise de sentimento financeiro não apenas se mostrou eficaz na compreensão do impacto das notícias nos mercados pelos investidores e analistas financeiros, mas também poderão utilizar a ferramenta para realizar investimentos e análises futuras.

A equipe demonstrou com sucesso a eficácia da análise de sentimento na antecipação do

clima do mercado financeiro. Esta pesquisa reforça a ideia de que essa abordagem é um instrumento valioso para a antecipação de movimentos no mercado, possibilitando uma compreensão mais profunda dos fatores que influenciam as decisões de investimento.

Ao longo deste projeto, também foi evidenciada a importância das redes neurais como uma ferramenta vital na resolução de problemas complexos e na análise de grandes volumes de dados. As redes neurais artificiais, espelhando o funcionamento do cérebro humano, apresentaram-se como ferramentas incrivelmente versáteis, impulsionando avanços significativos em várias esferas da tecnologia.

Contudo, é importante ressaltar que a base de dados utilizada nessa pesquisa foi composta por fontes específicas. Em pesquisas futuras, explorar outras fontes de dados pode gerar resultados distintos e ampliar ainda mais a compreensão do sentimento financeiro. A diversificação das fontes pode enriquecer a análise e oferecer perspectivas mais abrangentes sobre as tendências e movimentos do mercado.

Este projeto não apenas consolidou a eficácia da análise de sentimento financeiro, mas também destacou o potencial de aprimoramento contínuo por meio da exploração de diversas fontes de dados. Essa abordagem, combinada com o poder das redes neurais, representa um campo robusto para futuras pesquisas e aplicações no universo financeiro.

O arquivo do projeto está disponível no GITHUB no endereço: https://github.com/BelBatanete/Mackenzie

O vídeo da apresentação está disponível no YouTube através do link: https://youtu.be/JBSHTUfuE18

7 REFERÊNCIAS BIBLIOGRÁFICAS

FERREIRA, R. G. C.; MIRANDA, L. B. A. D.; PINTO, R. A. et al. Preparação e Análise Exploratória de Dados. Disponível em: Minha Biblioteca, Grupo A, 2021.

LAROSE, C. D.; LAROSE, D. T. Data Science Using Python and R. Hoboken: Wiley, 2019 | Biblioteca do Mackenzie.

MARIANO, D. C. B.; MARQUES, L. T.; SILVA, M. S. et al. Data Mining. Porto Alegre: Grupo A, 2021.

DEVORE, Jay L. Probabilidade e estatística para engenharia e ciências. Tradução da 9ª edição norteamericana. São Paulo: Cengage Learning Brasil, 2018.

KAGGLE. Financial Sentiment Analysis. Disponível em: https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis/data

PDF.TRANSFORMING SENTIMENT ANALYSIS IN THE FINANCIAL DOMAIN WITH CHATGPT. Disponível em: https://browse.arxiv.org/pdf/2308.07935.pdf

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, 10(1), 1-309;
Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780;
Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10), 2451-2471

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation;
Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167;
Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112);
Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate.