UNIVERSIDADE PRESBITERIANA MACKENZIE

Faculdade de Computação e Informática

Prof. Dr. Ivan Carlos Alcântara de Oliveira

Inteligência Artificial – 7º Semestre CC - Noite

Análise Automatizada de Relatórios Financeiros Utilizando Modelos de Linguagem Natural

Integrantes

- Ian Merlini: 10402831
- Lucas Farias: 10402521

Resumo. *Este projeto propõe o desenvolvimento de um sistema de análise automatizada de relatórios financeiros utilizando técnicas de Processamento de Linguagem Natural (PLN) e a API do ChatGPT. O objetivo é extrair informações relevantes de demonstrações financeiras, identificar padrões e tendências, e gerar resumos executivos que auxiliem investidores e analistas na tomada de decisões. A solução empregará um modelo pré-treinado de linguagem natural combinado com técnicas específicas de análise financeira para produzir insights valiosos a partir de documentos financeiros complexos. A implementação incluirá a coleta e estruturação de dados financeiros, pré-processamento de textos, análise exploratória dos dados e validação dos resumos gerados. Espera-se que o sistema desenvolvido aumente a eficiência na análise de relatórios financeiros, proporcionando resumos precisos e de fácil compreensão.*

4. Introdução

a. Contextualização

O mercado financeiro gera uma quantidade massiva de documentos textuais, incluindo relatórios trimestrais, demonstrações financeiras anuais e comunicados a investidores. Analisar manualmente esses documentos é uma tarefa demorada e sujeita a erros humanos. Conforme destacado por Fisher et al. (2016), os relatórios financeiros contêm informações críticas sobre a saúde financeira e as perspectivas futuras das empresas, mas sua extensão e complexidade representam um desafio significativo para investidores e analistas.

b. Justificativa

Com o avanço das técnicas de Processamento de Linguagem Natural (PLN) e o desenvolvimento de grandes modelos de linguagem, como o GPT-4, tornou-se possível automatizar a análise de documentos financeiros com alto grau de precisão. Loughran e McDonald (2020) demonstraram que algoritmos de PLN podem identificar aspectos sutis da linguagem usada em relatórios financeiros que têm impacto direto no desempenho das ações. Além disso, a automatização da análise permite processar um volume muito maior de documentos, possibilitando análises comparativas entre diferentes empresas e setores.

c. Objetivo

Desenvolver um sistema que utilize a API do ChatGPT para:

- Extrair informações-chave de relatórios financeiros, como indicadores de desempenho, projeções e fatores de risco;
- Identificar padrões linguísticos que possam indicar tendências futuras;
- Gerar resumos executivos personalizados conforme as necessidades do usuário;
- Comparar relatórios de diferentes períodos para identificar mudanças significativas.

d. Opção do Projeto

Este projeto seguirá a Opção API ChatGPT, empregando um grande modelo de linguagem para implementar uma solução de análise de dados de negócio, especificamente focada em relatórios financeiros e documentos de relação com investidores.

5) Descrição do Problema

Analisar manualmente relatórios financeiros é uma tarefa demorada e complexa para investidores e analistas. Esses documentos, frequentemente extensos e repletos de termos técnicos, dificultam a rápida compreensão das informações essenciais. Além disso, a variabilidade na estrutura dos relatórios entre diferentes empresas torna comparações eficientes um desafio, limitando a identificação de tendências e mudanças importantes. A interpretação humana também pode introduzir vieses subjetivos, afetando a objetividade das análises. Segundo Kearney e Liu (2014), a análise textual desses relatórios pode revelar insights valiosos que os números sozinhos não capturam. Portanto, há uma necessidade clara de automatizar esse processo utilizando técnicas avançadas de Processamento de Linguagem Natural (PLN) para extrair, analisar e sintetizar informações de maneira eficiente e precisa, facilitando a tomada de decisões informadas.

6) Aspectos Éticos do Uso da IA e sua Responsabilidade

A utilização de inteligência artificial na análise de relatórios financeiros traz importantes considerações éticas que devem ser rigorosamente abordadas. Primeiramente, é fundamental garantir a transparência dos métodos utilizados, informando claramente aos usuários como os resumos são gerados e quais são as limitações da ferramenta. Além disso, é essencial proteger a confidencialidade dos dados processados, assegurando que informações sensíveis sejam tratadas com rigorosos protocolos de segurança, mesmo quando se trata de dados públicos.

Outro ponto crítico é a mitigação de vieses nos modelos de IA, para evitar interpretações tendenciosas que possam levar a decisões injustas ou equivocadas. Isso requer uma calibração cuidadosa dos algoritmos e uma revisão contínua de seus resultados. Também é importante implementar mecanismos de validação para garantir a precisão e a confiabilidade dos resumos gerados, prevenindo omissões ou distorções de informações essenciais.

Por fim, as recomendações fornecidas pelo sistema de IA devem ser apresentadas como apoio à decisão, mantendo a responsabilidade humana sobre as escolhas finais. Dessa forma, asseguramos que a tecnologia complementa, e não substitui, a análise profissional. Seguindo as diretrizes éticas propostas por Floridi et al. (2018), buscamos desenvolver uma solução que seja justa, transparente e confiável, promovendo o uso responsável da inteligência artificial no contexto financeiro.

7) Dataset, Análise Exploratória e Preparação dos Dados

a. Descrição do Dataset

O projeto utilizará o conjunto de dados EDGAR (Electronic Data Gathering, Analysis, and Retrieval) da SEC (Securities and Exchange Commission), que contém relatórios financeiros de empresas de capital aberto nos Estados Unidos. Esses dados são públicos e contêm uma vasta quantidade de informações sobre o desempenho financeiro das empresas.

Conforme descrito por Loughran e McDonald (2016), os documentos do EDGAR seguem formatos padronizados que facilitam a extração e análise automatizada. Nosso foco inicial será nos relatórios 10-K (anuais) e 10-Q (trimestrais) de empresas do setor tecnológico nos últimos 5 anos, totalizando aproximadamente 100 documentos.

b: Análise Exploratória e Preparação dos Dados

A análise exploratória inicial consistirá em:

Coleta e Organização dos Relatórios Financeiros Selecionados: Utilização de ferramentas como BeautifulSoup e Scrapy para extrair dados do site da SEC. Armazenamento dos documentos em formato PDF e posterior conversão para texto usando bibliotecas como PyPDF2 e PDFMiner.

Análise Estatística do Vocabulário e Estrutura dos Documentos: Utilização de Pandas para manipulação de dados e Matplotlib e Seaborn para visualizações. Cálculo de métricas como frequência de termos, diversidade lexical e distribuição de tópicos.

Identificação de Padrões Linguísticos Recorrentes: Aplicação de técnicas de Tokenização e Lematização usando a biblioteca NLTK. Análise de sentenças e identificação de frases que indicam desempenho financeiro, riscos e projeções futuras.

Pré-processamento dos Textos para Análise: Tokenização: Divisão do texto em palavras e frases para facilitar a análise. Remoção de stopwords: Eliminação de palavras comuns que não agregam significado relevante. Normalização: Padronização de termos financeiros para melhorar a consistência das análises.

Para a preparação dos dados, seguiremos as técnicas descritas por Devlin et al. (2019), adaptando o processamento textual para o contexto financeiro. Será utilizada a biblioteca NLTK para as tarefas básicas de PLN e implementações específicas para análise financeira com o objetivo de otimizar a extração e interpretação das informações contidas nos relatórios.

8. Metodologia e Resultados Esperados

a. Metodologia

O projeto será desenvolvido seguindo estas etapas:
Coleta e Preparação de Dados:
Coleta de relatórios financeiros do EDGAR.
Conversão de PDFs para texto estruturalizado.
Limpeza e normalização dos textos para padronização dos dados.

b. Desenvolvimento do Modelo

O projeto será desenvolvido seguindo as seguintes etapas:

1. Planejamento: Definir objetivos, funcionalidades e cronograma do projeto.
2. Coleta e Preparação de Dados: Obter e limpar os relatórios financeiros do EDGAR.
3. Análise Exploratória: Avaliar a estrutura e identificar padrões nos dados.
4. Desenvolvimento do Modelo: Integrar a API do ChatGPT e criar prompts para extração e sumarização.
5. Implementação da Sumarização: Desenvolver algoritmos para gerar resumos precisos.
6. Validação e Testes: Comparar os resumos gerados com análises humanas e ajustar o modelo.
7. Otimização: Refinar os algoritmos e melhorar a eficiência do sistema.
8. Documentação: Elaborar o relatório final detalhando todas as etapas e resultados.

c. Resultados Esperados

Esperamos que o sistema desenvolvido seja capaz de:

- Reduzir em 70% o tempo necessário para analisar um relatório financeiro: Automatizando a extração e a síntese das informações, tornando o processo significativamente mais eficiente.
- Identificar com precisão superior a 80% os principais indicadores financeiros: Utilizando técnicas avançadas de PLN para capturar informações críticas que impactam a saúde financeira das empresas.
- Gerar resumos executivos que capturem adequadamente as informações mais relevantes: Produzindo resumos claros e concisos que facilitem a compreensão rápida dos relatórios.
- Fornecer análises comparativas entre diferentes períodos e empresas: Permitindo uma visão abrangente das mudanças financeiras ao longo do tempo e das diferenças entre empresas.

9. Referências

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2020). Deep learning for event-driven stock prediction. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (pp. 2327-2333).

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, 23(3), 157-214.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, 33, 171-185.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187-1230.

Loughran, T., & McDonald, B. (2020). Measuring firm complexity. The Journal of Corporate Finance, 65, 101812.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

OpenAI. (2023). ChatGPT API Documentation. Disponível em: https://platform.openai.com/docs/guides/chat ocs/guides/chat