AVALIAÇÃO E TESTES DE VETORIZAÇÃO E EBR NO ECOSSISTEMA DE MODELO DE AGRONEGÓCIOS: CASO CITROS.

CASSITAS, Júlia dos Santos julia.cassitas@fatec.sp.gov.br Fatec Cotia

SILVA JUNIOR, Braz Izaias braz.silva@fatec.sp.gov.br Fatec Cotia

1. INTRODUÇÃO

Com o desenvolvimento de IA generativas e a progressão nos últimos anos tornou-se possível modelos que foquem nos significados semânticos para cada ramo específico do conhecimento e aplicações. O projeto visa a vetorização e *Embedding* de Dados Estruturados (EBR) no ecossistema de modelos de agronegócios utilizando o modelo de *Large Language Model (LLM)* [3], [4]na base de dados de agronegócios e utilizar a arquitetura *Transformer* no capítulo de Citros do "Boletim100".O foco deste projeto é a avaliação e testes de métodos de vetorização de dados.

2. METODOLOGIA

O estudo inicial é o estudo preliminar dos modelos de Redes Neurais Recorrentes (RNNs) e da arquitetura *Trasnformer* na Base de dados de agronegócio, manejo de lavouras e nutrição De plantas[1]. Após a escolha do melhor modelo o projeto será. dividido em 5 fases: fase de recebimento das informações, fase de seleção dos dados, fase de vetorização, consulta por similaridade e mineração de dados. O projeto tem como foco a vetorização e *Embedding* fazendo a transformação de dados estruturados em representações vetoriais, que capturam a essência semântica de dados de Agronegócios. O desenvolvimento envolveu as seguintes atividades:

- Identificação das vetorizações e padronizações mais adequadas para as atividades do projeto.
- Implementação de duas vetorizações candidatas para comparação.
- Seleção de uma base de dados para o teste dos algoritmos.
- Execução bem-sucedida de um grupo de mensagens e arquivos para testar a modelagem.
- Destacar abordagens mais eficazes e suas aplicações práticas

3. RESULTADOS E DISCUSSÕES

Algoritmos de aprendizado de máquina operam em um espaço de atributos numéricos, ou seja, eles esperam que uma matriz bidimensional; para executar aprendizado de máquina no texto, é necessário converter os documentos de origem em representações vetoriais. Através da revisão bibliográfica foi escolhido que o modelo LLM a ser adotado para o SmartBoletim100(SB100) será baseado na arquitetura *Transformer,* que utiliza mecanismos de atenção autorregressivas, permitindo a captura de relações entre palavras e frases em um texto, bem como contexto e significado do mesmo. Foi avaliado e testados 3 tipos de vetorização, sendo elas: TF-IDF, ELMO e BERT. Os três tipos apresentaram um bom desempenho, entretanto há uma técnica que se destacou, sendo ela a BERT. O uso da TF-IDF foi descartado pelo fato de ser um método que aceita documentos individualmente, e apesar de funcionar muito bem para o capítulo citros o foco é o desenvolvimento de um método que seja capaz de aceitar diversos documentos. Já o ELMO e BERT são modelos de *deep learning* para dados extensos, ambos são *multi-layered*, a escolha do uso do BERT dar-se-á pelo utilização da arquitetura *Transformer* que apresenta um desempenho superior para esse tipo de dado [2].

Houve desafios como por exemplo as "*stopwords*", pois poucas bibliotecas apresentam português no seu repertório, além disso os dados de agricultura apresentam muitos números associados ao contexto, como por exemplo: "no máximo em 2 dias", "3g de K", na qual são essenciais para o entendimento semântico e para ao treinamento do nosso algoritmo, o desafio se dá pelo fato de que números muitas vezes não são considerados no processamento dos dados para a vetorização e treinamento de máquina, por isso surgiu a ideia de transformá-los em texto corrido antes da vetorização. Entretanto, tanto a conversão quanto a vetorização estão sendo estudadas em busca das melhores soluções.

4. CONCLUSÕES

Com base em uma revisão bibliográfica e nos testes realizados, a arquitetura *Transformer* demonstra um desempenho significativamente superior no processamento de dados voltados ao agronegócio, especialmente em termos de precisão. Além disso, as etapas de pré-processamento do texto, como limpeza e normalização, anteriores à vetorização, revelaram-se essenciais para garantir um treinamento eficiente e preciso dos modelos de machine learning, melhorando tanto a qualidade dos dados quanto a performance dos algoritmos.

REFERÊNCIAS

[1] KOZHEVNIKOV, Vadim Andreevich; PANKRATOVA, Evgeniya Sergeevna. Research of the text data vectorization and classification algorithms of machine learning. ISJ Theoretical & Applied Science, v. 5, n. 85, p. 574-585, 2020.

[2] SHUTSKE, John M. **Harnessing the power of large language models in agricultural safety & health**. *Journal of Agricultural Safety and Health*, p. 0, 2023.

[3] LI, Jiajia et al. Foundation models in smart agriculture: basics, opportunities, and challenges. Computers and Electronics in Agriculture, v. 222, 2024, p. 109032.

[4] TZACHOR, Asaf et al. Large language models and agricultural extension services. Nature food, p. 1-8, 2023.