

UFRRJ

**PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA,
TECNOLOGIA E INOVAÇÃO EM AGROPECUÁRIA**

TESE

**Desenvolvimento de Software e Aplicação de
Inteligência Artificial na Obtenção de Informações
Agropecuárias em Tempo Real a partir de Mídias
Sociais**

FÁBIO CARDOZO DA SILVA

2024



**UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO PROGRAMA DE
PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E INOVAÇÃO EM
AGROPECUÁRIA**

**DESENVOLVIMENTO DE SOFTWARE E APLICAÇÃO DE
INTELIGÊNCIA ARTIFICIAL NA OBTENÇÃO DE INFORMAÇÕES
AGROPECUÁRIAS EM TEMPO REAL A PARTIR DE MÍDIAS
SOCIAIS**

FÁBIO CARDOZO DA SILVA

Sob a Orientação da Professor
Mauro Antonio Homem Antunes

e Coorientação do Professor
José Airton Chaves Cavalcante Junior

Tese submetida como requisito parcial
para obtenção do grau de **Doutor**, no
Programa de Pós-graduação do Ciência,
Tecnologia e Inovação em
Agropecuária, Área de Concentração e
Área de concentração em Recursos
Naturais e Proteção Ambiental.

Seropédica, RJ
Dezembro de 2024

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central/Seção de Processamento Técnico

Ficha catalográfica elaborada
Com os dados fornecidos pelo(a) autor(a)

S586d

Silva, Fábio Cardozo da, 1980-

Desenvolvimento de software e aplicação de Inteligência Artificial na obtenção de informações agropecuárias em tempo real a partir de mídia sociais / Fábio Cardozo da Silva.

- Seropédica, 2024.

60 f: il.

Orientador: Mauro Antonio Homem Antunes.

Tese (Doutorado). – – Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação em Ciência, Tecnologia e Inovação em Agropecuária, 2024.

1. Inteligência Artificial. 2. Bots. 3. Big Data. 4. Mídias Sociais. 5. Agropecuária. I. Antunes, Mauro Antonio Homem, 1963-, orient. II. Universidade Federal Rural do Rio de Janeiro. III. Programa de Pós-Graduação em Ciência, Tecnologia e Inovação em Agropecuária. IV. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

É permitida a cópia parcial ou total desta Tese, desde que seja citada a fonte.



**MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E
INOVAÇÃO EM AGROPECUÁRIA**



HOMOLOGAÇÃO DE TESE DE DOUTORADO Nº 42 / 2024 - PPGCTIA (12.28.01.84)

Nº do Protocolo: 23083.071670/2024-19

Seropédica-RJ, 20 de dezembro de 2024.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E INOVAÇÃO EM AGROPECUÁRIA

FÁBIO CARDOZO DA SILVA

Tese submetida como requisito parcial para obtenção do grau de Doutor, no Programa de Pós-Graduação em Ciência, Tecnologia e Inovação em Agropecuária, área de concentração em Recursos Naturais e Proteção Ambiental.

TESE APROVADA EM 17/12/2024

Mauro Antonio Homem Antunes. Ph.D. UFRRJ

(Orientador)

Maria Fernanda Moura. Dra., Embrapa - Agricultura Digital

Alan Miranda Monteiro de Lima. Dra., UFRJ

Claudia Rosa Ledesma. Dra., UNRC

Marlucio Barbosa. Dr., UFRRJ

(Assinado digitalmente em 20/12/2024 11:20)
MARLUCIO BARBOSA
PROFESSOR DO MAGISTERIO SUPERIOR
DeptTL/IM (12.28.01.00.00.90)
Matrícula: 3749825

(Assinado digitalmente em 20/12/2024 11:17)
MAURO ANTONIO HOMEM ANTUNES
PROFESSOR DO MAGISTERIO SUPERIOR
DEPTOENG (12.28.01.00.00.00.44)
Matrícula: 1315209

(Assinado digitalmente em 13/01/2025 14:58)
CLAUDIA ROSA LEDESMA
ASSINANTE EXTERNO
Passaporte: 16991615

(Assinado digitalmente em 20/12/2024 12:28)
MARIA FERNANDA MOURA
ASSINANTE EXTERNO
CPF: 470.318.246-72

(Assinado digitalmente em 21/12/2024 21:42)
ALAN MIRANDA MONTEIRO DE LIMA
ASSINANTE EXTERNO
CPF: 004.119.917-01

Visualize o documento original em <https://sipac.ufrj.br/public/documentos/index.jsp>
informando seu número: **42**, ano: **2024**, tipo: **HOMOLOGAÇÃO DE TESE DE DOUTORADO**, data
de emissão: **20/12/2024** e o código de verificação: **15f07ee221**

DEDICATÓRIA

Primeiramente agradeço a Deus,

“Por isso, ó nosso Deus, agora vos damos graças e louvamos vosso Nome glorioso.”

I Crônicas 29,13

Em seguida minha mãe (Sandra Cardozo) e minha vó (Marli Casanova), que me educaram como pessoa e me ensinaram a necessidade de seguir seus próprios caminhos e lutar por seus ideais.

A minha linda e paciente esposa (Laura Cardozo) que em todas os desafios, que por mais complicados e longos que fossem, sempre permaneceu firme em me apoiar.

As minhas filhas Ana Julia e Ana Vitória, que espero estimular no mundo acadêmico.

AGRADECIMENTOS

Agradeço a Deus, por todas as oportunidades concedidas e desafios ultrapassado.

Agradeço minha colega de trabalho da UFRRJ/PROPLADI Alden Peres, que em muito me deu suporte durante esses anos.

A mais uma colega de trabalho, Renata Gomes de Souza (PPGCTIA), que sempre me ajudou em todas as necessidades ligadas ao curso.

A todos os professores do PPGCTIA, primeiramente por servirem como inspiração para o meu desenvolvimento acadêmico, e pela grande contribuição científica que eles representam.

Ao meu orientador Mauro Antonio Homem Antunes, que me apoiou de forma maravilhosa e animadora.

Ao meu coorientador José Airton Junior, um verdadeiro amigo com quem pude contar.

À UFRRJ, instituição que me acolheu como servidor e aluno.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo papel fundamental que desempenha no fortalecimento e na qualidade da pós-graduação brasileira, reconheço que a existência e o funcionamento do programa de doutorado que me acolheu só foram possíveis graças ao trabalho contínuo desta agência na avaliação, regulamentação e fomento dos programas de pós-graduação no país.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

RESUMO

SILVA, Fábio Cardozo da. **Desenvolvimento de software e aplicação de Inteligência Artificial na obtenção de informações agropecuárias em tempo real a partir de mídias sociais**. 2024. 60f. Tese (Doutorado em Ciência, Tecnologia e Inovação em Agropecuária) Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

Este trabalho apresenta o desenvolvimento de um método inovador que integra bots de coleta de dados e modelos de inteligência artificial generativa para automatizar o pré-processamento de informações agropecuárias em tempo real oriundas de mídias sociais, como Facebook, Instagram e X. A metodologia proposta abrange a varredura automatizada de postagens pelas bots, a conversão de dados não estruturados em formatos semiestruturados (exemplo: JSON), e a classificação criteriosa das informações mediante uma métrica gradual de pertinência, aplicada por motores de IA que avaliam e organizam os dados sem gerar conteúdo novo. Os resultados indicam que a automação do pré-processamento proporciona ganhos significativos em produtividade e economia de tempo, ao possibilitar uma triagem objetiva e eficiente da massa de dados disponível, cuja magnitude tornaria inviável a análise manual tradicional. Além disso, a categorização realizada pela IA atendeu a níveis aceitáveis de precisão e consistência, validando a confiabilidade do método mesmo diante dos desafios típicos como as alucinações de IA. A integração dos dados processados com ferramentas de Business Intelligence (BI) permite visualizações dinâmicas e filtragens personalizadas, fomentando a aplicação de técnicas avançadas de storytelling para a comunicação científica. Este método detém grande potencial para transformar a pesquisa agropecuária e a formulação de políticas públicas, ao estruturar dados dispersos em conhecimento acionável e de alta relevância, inaugurando um novo paradigma no uso de dados digitais. Recomenda-se sua adoção em estudos de larga escala e em conjunto com tecnologias de Big Data para maximizar o impacto e a abrangência das análises baseadas em mídias sociais.

Palavras-chave: Inteligência artificial. Bots. Big Data. Mídias sociais. Agropecuária. Análise de dados. Business intelligence. Storytelling científico.

ABSTRACT

SILVA, Fábio Cardozo da. **Development of software and application of Artificial Intelligence for obtaining real-time agricultural information from social media.** 2024. 60p. Thesis (Doctorate in Science, Technology, and Innovation in Agriculture) Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

This work presents the development of an innovative method that integrates data collection bots and generative artificial intelligence models to automate the real-time preprocessing of agricultural information sourced from social media platforms such as Facebook, Instagram, and X. The proposed methodology encompasses the automated scanning of posts by bots, the conversion of unstructured data into semi-structured formats (e.g., JSON), and the rigorous classification of information based on a gradual relevance metric applied by AI engines that evaluate and organize the data without generating new content. The results indicate that the automation of preprocessing provides significant gains in productivity and time savings by enabling an objective and efficient screening of the available data volume, whose magnitude would render traditional manual analysis unfeasible. Furthermore, the AI-driven categorization achieved acceptable levels of precision and consistency, validating the reliability of the method even in the face of typical challenges such as AI hallucinations. Integration of the processed data with Business Intelligence (BI) tools allows for dynamic visualizations and personalized filtering, fostering the application of advanced storytelling techniques for scientific communication. This method has great potential to transform agricultural research and public policy formulation by structuring dispersed data into actionable and highly relevant knowledge, inaugurating a new paradigm in the use of digital data. Its adoption is recommended in large-scale studies and in conjunction with Big Data technologies to maximize the impact and scope of analyses based on social media.

Keywords: Artificial intelligence. Bots. Big Data. Social media. Agriculture. Data Analysis. Business intelligence. Scientific storytelling.

RESUMEN AMPLIADO

SILVA, Fábio Cardozo da. **Obtención de información agropecuaria en tiempo real para el apoyo a políticas públicas, a partir de redes sociales mediante software y la aplicación de Inteligencia Artificial**. 2024. 60 p. Tesis (Doctorado en Ciencia, Tecnología e Innovación en Agropecuaria) Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

1. Introducción

La creciente demanda de alimentos, impulsada por una población mundial que debe alcanzar 9,8 mil millones de personas para 2050, impone desafíos significativos a la seguridad alimentaria y la sostenibilidad de la producción agropecuaria. En este escenario, la tecnología de la información y comunicación (TIC) surge como un pilar fundamental para optimizar procesos, aumentar la eficiencia y apoyar la toma de decisiones estratégicas. Las redes sociales, en particular, con más del 59% de la población global como usuaria activa, representan una fuente masiva e inexplorada de datos en tiempo real sobre eventos, tendencias y desafíos del sector agropecuario. Sin embargo, el volumen y la naturaleza no estructurada de estos datos hacen que el análisis manual sea impracticable.

Este trabajo presenta el desarrollo y validación de un método innovador que integra bots de recolección de datos, inteligencia artificial (IA) generativa y herramientas de Business Intelligence (BI) para automatizar la obtención, el preprocesamiento y el análisis de informaciones agropecuarias a partir de redes sociales como Facebook, Instagram y X (anteriormente Twitter). El objetivo central es transformar datos dispersos y no estructurados en conocimiento accionable, capaz de subsidiar investigaciones científicas y la formulación de políticas públicas más eficaces y responsivas.

La metodología propuesta aborda desde el rastreo automatizado de publicaciones hasta la clasificación criteriosa de las informaciones por medio de una métrica de pertinencia, aplicada por motores de IA. La integración con herramientas de BI permite la creación de visualizaciones dinámicas y la aplicación de técnicas de storytelling científico, facilitando la comunicación de los resultados para diferentes públicos. Este estudio, por tanto, no solo propone una solución tecnológica, sino que también explora un nuevo paradigma para la investigación agropecuaria en la era digital, donde el Big Data social se convierte en un activo estratégico para el avance del conocimiento.

2. Revisión de la Literatura

La investigación se fundamenta en una revisión interdisciplinaria que abarca desde la importancia histórica y económica de la agropecuaria hasta los avances más recientes en tecnología de la información. La literatura destaca la urgencia de ampliar la producción de alimentos de forma sostenible, lo que demanda innovaciones tecnológicas. Estudios sobre la adopción de TICs en el campo, como los de Kifli et al. (2021) y Fonseca, Silveira & Massruhá (2018), revelan un crecimiento en el uso de dispositivos móviles y acceso a internet en el medio rural, lo que potencializa la generación de datos.

La infraestructura logística es otro punto crítico abordado, con base en trabajos como el de Abbade (2020), que demuestra la correlación directa entre la calidad de la logística y factores como pérdidas de alimentos, precios y diversificación de la dieta. Las redes sociales son analizadas como fuentes de datos masivos (Big Data), con potencial para monitorear eventos en tiempo real, como problemas de infraestructura, brotes de enfermedades o tendencias de mercado, conforme explorado por Milusheva et al. (2021).

El marco teórico sobre inteligencia artificial se basa en los conceptos de redes neuronales, aprendizaje automático y, más específicamente, IA generativa. La capacidad de

estos modelos de procesar lenguaje natural y clasificar informaciones es central para la metodología propuesta. Sin embargo, desafíos como la "alucinación" de IA, discutidos por Alkaissi & McFarlane (2023), son considerados, reforzando la necesidad de validación y uso ético de la tecnología. Finalmente, la investigación se alinea con los conceptos de políticas públicas apoyadas por sistemas de información y workflows científicos, que garantizan la transparencia, la reproducibilidad y la eficiencia del método desarrollado.

3. Método Propuesto

El método desarrollado consiste en un workflow científico automatizado, implementado en Python, que organiza la recolección y el procesamiento de datos en etapas secuenciales e integradas. La solución combina scripts, bots, base de datos, archivos semiestructurados (JSON) y la aplicación de IA, sin la necesidad de un software monolítico.

El proceso se inicia con la parametrización de la investigación, donde el usuario define las palabras clave (tags) de interés (ej: "brusone", "vassoura de bruxa") y el alcance geográfico (región, estado o ciudad). Esta etapa permite dirigir la recolección y aumentar la relevancia de los datos. A continuación, un bot de web scraping es ejecutado para realizar búsquedas en las redes sociales (Facebook, X e Instagram), recolectando publicaciones (textos e imágenes) que correspondan a los términos definidos. Los datos brutos, no estructurados, son entonces convertidos y almacenados en archivos JSON, que preservan la información original y añaden una estructura flexible. Posteriormente, los datos son transferidos a una base de datos relacional (SQLite), donde son estructurados con códigos de geolocalización (IBGE) y otras informaciones relevantes, permitiendo consultas complejas y la eliminación de duplicidades.

La etapa más innovadora del método es la aplicación de IA generativa para la validación y clasificación de los datos. Utilizando las APIs del ChatGPT (OpenAI) y del Gemini (Google), cada publicación es evaluada en cuanto a su pertinencia al tema investigado. Un prompt estandarizado solicita que la IA atribuya una puntuación (score) de 1 a 5, donde 1 representa baja adherencia y 5, alta adherencia. Un F1-score adaptado, calculado por el promedio de las notas de las dos IAs, genera un índice de relevancia que permite filtrar y priorizar las informaciones más importantes.

Para demostrar la aplicabilidad y la robustez del método, fueron realizados tres experimentos principales:

1. Brusone: Enfocado en una importante enfermedad que afecta el cultivo del arroz, probando la capacidad de recolección y análisis de informaciones sobre fitosanidad.

2. Vassoura de Bruxa no Cacaueiro: Utilizado para validar la aplicación de la IA en la clasificación de relevancia de las publicaciones.

3. Brucelosis (en español): Una prueba con datos de Argentina para verificar la adaptabilidad del método a diferentes idiomas y contextos geográficos.

Finalmente, los datos procesados y clasificados son exportados a una herramienta de Business Intelligence (Power BI). En esta capa de visualización, los resultados son presentados en dashboards interactivos, con mapas, gráficos y tablas dinámicas, facilitando el análisis exploratorio y la comunicación de los hallazgos por medio de técnicas de storytelling científico.

4. Resultados y Discusión

La aplicación del método propuesto generó resultados significativos que validan su eficacia y potencial. La automatización del proceso de recolección y preprocesamiento de datos demostró ganancias expresivas de productividad y economía de tiempo, permitiendo el análisis de un volumen de informaciones (aproximadamente 1.740 registros en los experimentos) que sería inviable por métodos manuales.

La calidad de los datos recolectados fue alta, con 100% de las publicaciones presentando relación con las palabras clave investigadas. La diversidad de fuentes (productores,

instituciones, empresas) enriqueció la base de datos, pero también evidenció la necesidad de la etapa de clasificación para extraer los insights más relevantes.

La clasificación por IA alcanzó niveles satisfactorios de precisión y consistencia. La utilización de dos motores de IA (ChatGPT y Gemini) permitió una validación cruzada, con la diferencia promedio entre las puntuaciones manteniéndose en torno al 20% (un punto en la escala de 1 a 5). El F1-score adaptado se mostró como una métrica eficaz para crear un ranking de relevancia, permitiendo que el investigador se enfoque en los datos con mayor potencial informativo. Por ejemplo, textos más detallados y técnicos recibieron scores más altos (e.g., 0,9), mientras que publicaciones más genéricas o comerciales recibieron notas menores (e.g., 0,5), en concordancia con la evaluación humana.

La integración con la capa de exhibición de datos (Power BI) fue fundamental para transformar los datos brutos en conocimiento visual e interactivo. Los dashboards permitieron la identificación de patrones geográficos, el análisis de tendencias temporales y el filtrado dinámico por relevancia, facilitando la aplicación de técnicas de storytelling científico. Este enfoque separa el "cómo" (la complejidad técnica del workflow) del "qué" (los resultados y sus significados), haciendo la comunicación científica más accesible e impactante.

El estudio también reconoce las limitaciones, como la posibilidad de "alucinaciones" de la IA, que, aunque no invalidan el método, refuerzan la importancia de la supervisión de especialistas. Además, para aplicaciones a gran escala, la migración a una base de datos NoSQL puede ser necesaria para garantizar el rendimiento.

5. Conclusión

Este trabajo demostró que la integración de bots, inteligencia artificial generativa y herramientas de Business Intelligence constituye un método poderoso y escalable para la obtención y análisis de informaciones agropecuarias a partir de redes sociales. La automatización del preprocesamiento de datos no solo optimiza el trabajo de investigación, sino que también inaugura un nuevo paradigma en el uso de datos digitales, transformando el inmenso volumen de informaciones de las redes sociales en un activo estratégico para la ciencia y la gestión pública.

El método propuesto ofrece ganancias significativas en productividad, permitiendo una selección objetiva y eficiente de datos que, de otra forma, serían subutilizados. La clasificación por IA, validada por múltiples motores, probó ser un enfoque confiable para evaluar la pertinencia de las informaciones, incluso ante los desafíos inherentes a esta tecnología.

El potencial de aplicación de este método es vasto, abarcando desde el monitoreo de enfermedades y plagas en tiempo real hasta el análisis del impacto de políticas públicas y la identificación de tendencias de mercado. Al estructurar datos dispersos en conocimiento accionable y de alta relevancia, esta investigación contribuye a la construcción de estudios más innovadores y políticas públicas más informadas y eficaces.

Se recomienda la adopción de este método en estudios a gran escala, en conjunto con tecnologías de Big Data, para maximizar el impacto y el alcance de los análisis. La continua evolución de los modelos de IA y de las herramientas de análisis de datos promete perfeccionar aún más la capacidad de extraer valor de las redes sociales, consolidándolas como una fuente indispensable de información para el futuro de la agropecuaria.

Palabras clave: Inteligencia artificial. Robots. Grandes datos. Redes sociales. Agricultura. Análisis de datos. Inteligencia de negocios. Narración científica.

LISTA DE FIGURAS

Figura 1. Exemplo de notícias e postagens que podem compor uma base de dados em agropecuária. Fonte: Internet, 2023.	3
Figura 2. Caminhão atolado com grãos. Fonte: Post realizado em Facebook.....	5
Figura 3. Relação logística, preço com a presença de alimentos e seu preço: Fonte: Abbade (2020).....	6
Figura 4. Métricas e influência. Fonte: Abbade (2020).....	6
Figura 5. Relação de aspectos logísticos e a distribuição e preços de produtos. Fonte: Abbade (2020).....	6
Figura 6. Market share das mídias sociais. Fonte: “Global Social Media Statistics Research Summary”, 2022.	8
Figura 7. Presença da população nas mídias sociais. Fonte: Global social media statistics research summary, 2022.	9
Figura 8. Origem da Inteligência Artificial, neurônio representado matematicamente. Fonte: Silva, 2011.	10
Figura 9. Comparação entre IA, Aprendizado de máquina e aprendizado profundo. Fonte: Sarker, 2021.....	11
Figura 10. Mapa de Eventos. Fonte: Milusheva et al., 2021.	18
Figura 11. Identificação de Animal por IA - Banco de dados de espécies como o tamanduá-bandeira permitiu treinar para a realidade do país modelo utilizado em outros países. Fonte: imagem montagem de Gabriel Souto Ferrante (sobre foto de Miguel Rangel Jr/Creative Commons).....	19
Figura 12. Diagrama de execução e demonstração de ambiente da aplicação. Fonte: Elaborado pelo Autor.	20
Figura 13. Parametrização de termos de pesquisa. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	21
Figura 14. Parametrização de região. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	21
Figura 15. Resultado de pesquisa. Fonte: Aplicação desenvolvida na pesquisa pelo Autor. ...	21
Figura 16. Arquivo JSON com resultado de mineração. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	22
Figura 17. Arquivo JSON com resultado de mineração. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	22
Figura 18. Resultado da estruturação dos dados obtidos. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	23
Figura 19. Modelo de banco de dados. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	23
Figura 20. Exemplo de dados obtidos na aplicação, submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	24

Figura 21. Exemplo de dados obtidos na aplicação, submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	24
Figura 22. Nova parametrização. Fonte: Aplicação desenvolvida na pesquisa pelo autor.	25
Figura 23. Brusone em folha. Fonte: EMBRAPA, 2024.....	25
Figura 24. Nova parametrização. Fonte: Elaborado pelo Autor.....	25
Figura 25. Exemplo de conjunto de dados obtido na pesquisa. Fonte: Aplicação desenvolvida na pesquisa pelo autor.....	26
Figura 26. Dashboard em Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo autor.....	26
Figura 27. Parametrização. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.....	27
Figura 28. A distribuição de postagens por cidade. Fonte: Aplicação desenvolvida na pesquisa pelo autor.	27
Figura 29. Arquivo JSON resultado de processamento. Fonte: Aplicação desenvolvida na pesquisa pelo autor.	28
Figura 30. Detalhamento do campo score (Resultado de aplicação de IA). Fonte: Aplicação desenvolvida na pesquisa pelo autor.....	30
Figura 31. Integração OpenAI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.....	30
Figura 32. Tabela de dados, resultado de processamento. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	31
Figura 33. Resultado de processamento submetido ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	31
Figura 34. Parametrização para processamento d dados argentinos Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	32
Figura 35. Resultado de processamento de dados argentinos. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.....	32
Figura 36. Dados argentinos submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.	33
Figura 37. Prompt utilizado no ChatGPT e Gemini. Fonte: Autor	34

LISTA DE TABELAS

Tabela 1. Presença de computadores em estabelecimentos rurais.....	4
Tabela 2. Participação e tempo de uso da sociedade em mídias sociais.....	7
Tabela 3. Contagem de dados obtidos pelo Bot.	28
Tabela 4. A distribuição de postagens por cidade, e score de dados gerado por IA.....	32

SUMÁRIO

1. INTRODUÇÃO.....	1
2. REVISÃO DE LITERATURA	3
2.1 Agropecuária.....	3
2.2 Tecnologia da Informação (TI) e Atividade Agrícola	4
2.3 Infraestrutura Logística e seus Impactos na Agropecuária	4
2.4 As Mídias Sociais	7
2.4.1 Facebook	8
2.4.2 Instagram.....	8
2.4.3 Twitter (X)	8
2.4.4. Análise de dados em mídias sociais	8
2.5 Inteligência Artificial (IA)	9
2.5.1 Inteligência artificial gerativa	11
2.5.2 Modelos de avaliação de IA gerativa	12
2.5.3 Alucinação de IA gerativa.....	14
2.6 Políticas Públicas Apoiadas por Sistemas de Informação	14
2.7 Workflow Científico	15
2.8 Bot's (Robot)	15
2.9 Técnica de Web Scraping	15
2.10 Big Data	16
2.11 IoT.....	16
2.12 Computação em Nuvem.....	16
2.13 Bancos de Dados e Tipos de Dados	16
2.13.1 Dados estruturados	16
2.13.2 Dados semiestruturados	17
2.13.3 Dados não estruturados	17
2.14 Estudo de Caso.....	17
2.14.1 Mapeamento de acidentes automotivos em zonas urbanas	17
2.14.2 Identificação de animais selvagens em estradas	18
3. MÉTODO PROPOSTO.....	20
3.1 Descrição da Solução.....	20
3.2 Experimento - Brusone	25
3.3.1 Volume de dados.....	28

3.3.2 Da veracidade dos dados.....	28
3.3.3 Da produção de factoides	29
3.4 Aplicação de IA	29
3.4.1 Sistemas de categorização ou score	29
3.5 Experimento – Vassoura Bruxa no Cacaueiro – Com IA	30
3.6 Experimento em Espanhol - “Brucelosis Enfermedad”	32
4. RESULTADOS E DISCUSSÃO.....	34
4.1 Análise dos Dados.....	35
4.2 Camada de Exibição de Dados	35
5. CONCLUSÕES	37
6. REFERÊNCIAS BIBLIOGRÁFICAS	38
7. ANEXOS	41

1. INTRODUÇÃO

O presente trabalho visa analisar os resultados da obtenção de dados voltados à pesquisa científica em agropecuária, obtidos por meio de sistema de informação, capazes de extrair, classificar e organizar dados provenientes das seguintes mídias sociais X (Twitter), Instagram, e Facebook. O motivo deve-se ao fato de que essas plataformas são as que têm a maior influência e de participação de 59,4% da população global, de acordo com *Global Social Media Statistics Research Summary* (2022).

Deste modo, a análise a ser realizada, deve observar a relevância, qualidade e volume de dados extraídos, e sua forma de aproveitamento em pesquisas na agropecuária, nos eixos de agroecologia, políticas públicas e tecnologias aplicadas à agropecuária.

Naturalmente, os dados e as análises devem passar pelo crivo de pesquisadores e estudantes das áreas em questão, por meio de aferimento de dados gerados pelos sistemas, nos assuntos de interesses dos envolvidos, e validados por meio de pesquisa qualitativa sobre os resultados obtidos dos dados e seus aproveitamentos.

A análise e discussão de recursos tecnológicos são o resultado da observação de situações complexas que envolvem a nossa sociedade atual, em termos de desenvolvimento populacional e sua segurança alimentar; onde especialistas demonstram a necessidade iminente da ampliação dos recursos de produção agrícola em função do crescimento populacional (COLLOMB, 1999).

O crescente número da parcela mundial que utiliza mídias sociais, associado a evolução da presença de computadores e celulares nos domicílios rurais, nos permite pensar na existência de dados relativos a fatos cotidianos que serviriam como informações relevantes a pesquisadores e órgãos públicos.

Validar a relevância de dados obtidos por meio de mídias sociais e a consequente utilização desses dados para geração de estatísticas, rastreamento de eventos em assuntos relativos ao desenvolvimento agropecuário e de políticas públicas torna-se uma necessidade, tendo em vista o volume de dados existentes e a possibilidade de utilização desses dados na pesquisa científica.

Em análise das mídias sociais de relevância como Facebook, Twitter e Instagram, é possível constatar seu grande volume de dados, sendo essas postagens de cunho político, entretenimento, assuntos locais (regionalismo) e demais. Estes conteúdos, que podem ser textos, fotos ou vídeos se relacionam com assuntos de interesse da presente pesquisa, como produção agrícola e animal.

Essa relação ocorre por meio de postagens que narram fatos diários como acidentes de trânsito envolvendo a logística de produtos agrícolas, situação de doença de animais ou mesmo rebanhos. Além da narração de eventos, existem postagens que relacionam essas situações com responsabilidades governamentais, como a situação das estradas em áreas de grande produção, falta de apoio ao produtor no que tange a saúde animal.

Postagens sobre buracos em estradas Federais, que cortam estradas que fazem parte da logística de escoamento da produção agropecuária, transformam essas estradas em recursos importantes de infraestrutura, e em coeficientes no grande “cálculo” de valores e variáveis capazes de alterar preços, qualidade ou mesmo a disponibilidade de itens agropecuário, podem estar contidos não em um formulário de pesquisa, mas em diversas postagens feitas por vários agentes locais. Outros tópicos, que podem ser convertidos em informação e estatísticas de ocorrências da área, que se lidos por autoridades e ou entidades, podem gerar ações que reverterem problemas de infraestrutura em soluções logísticas.

Associada às questões ligadas ao volume de informações a serem obtidas por meio de postagens e mídias sociais, o que naturalmente geraria um grande volume de dados a serem filtrados pelos interessados, temos como aliada a aplicação de Inteligência Artificial (IA). A IA pode ser utilizada como automatizadora de parte desses dados, para a classificação de relevância de textos e ou imagens obtidas por mídias sociais, de forma a identificar sua relação com o assunto, classificação e categorização.

Em outras palavras, podemos dizer que existe um grande espaço, para uma solução em sistemas de informação, capaz de realizar a obtenção e organização de dados de forma automatizada, sobretudo se impulsionada pela IA generativa, capacitando essa aplicação na etapa de processamento de dados, e isso resulta em uma grande fonte de dados capazes de impulsionar pesquisas ligadas ao desenvolvimento agropecuário.

A pergunta norteadora desta pesquisa é: seriam as mídias sociais, uma fonte de dados, que, se organizados de maneira automatizada e de forma sistemática, se tornam capazes de gerar informações consistentes capazes de nortear pesquisas científicas ou orientar políticas públicas?

Essa pergunta baseia-se na possibilidade de utilizar os dados contidos nas mídias sociais, como campo de busca para alimentação de pesquisas científicas e ou mesmo na possibilidade de organização dessas a fim de gerar mapas do tipo “hotspot”, de acordo com perguntas de interesse. Como exemplo, imaginemos um banco de dados (BD), alimentado para responder perguntas que contenham “tag’s” como arroz, acidente e br. Um sistema de informação digital, construído com redes neurais artificiais realizando a categorização e validação dos textos e imagens relativos às pesquisas (tag’s), montaria uma tabela com o número de registros com essas tag’s, localização geográfica dos posts. Além disso, seria exibido um dashboard com mapa que relacionaria essas estatísticas, referentes ao número de registros das tags de interesse, com as regiões do Brasil.

Com base em todas as observações descritas nos capítulos anteriores, formula-se a seguinte hipótese: um sistema de informação digital que utiliza IA gerativa é eficaz na transformação de dados de mídias sociais em informações úteis para a pesquisa científica, a formação de políticas públicas e o desenvolvimento agropecuário.

O objetivo geral desta pesquisa consiste em desenvolver um método eficaz quanto à automação e redução de tempo na etapa de pré-processamentos em pesquisa científica com base em mídias sociais.

Considera-se para tal, o desenvolvimento de um sistema que realize a consulta (busca e mineração) de postagens nas mídias sociais (inicialmente Facebook), baseado no assunto de interesse (parametrização via tags). Ainda, quando identificado um post contendo o assunto o mesmo deve ser armazenado, catalogado juntamente com seus metadados (ex. informação geográfica (latitude + longitude)). Por fim, os metadados estão relacionados não somente ao texto, mas também com as fotos.

Assim, pretende-se realizar uma análise qualitativa e quantitativa para desvendar a pergunta norteadora, tendo os seguintes objetivos específicos:

a) Criação de software que utiliza IA gerativa realizando pré-processamento de dados: Cluster de assuntos, Validação de imagens e Scores de relevância de post (escala de cinza).

b) Criação de módulos, mineração de posts (Bots), filtragem e armazenamento de posts (BD), dashboard de cálculo e apresentação das estatísticas sobre os dados.

2. REVISÃO DE LITERATURA

2.1 Agropecuária

Como descrito por Feldens (2018), os sapiens iniciaram a manipulação da vida de plantas e animais há 10 mil anos, sendo esse um dos pontos de partida para a agricultura e pecuária. Sem dúvida, em condições técnicas e em formato muito diferentes do que conhecemos hoje, mas ainda assim com o objetivo de produzir e garantir a subsistência de nossa espécie. Hoje a agropecuária vai além da subsistência e permeia áreas como trabalho, economia e saúde apresentando-se como agronegócio, commodities e demais relações econômicas ligadas à área.

Ainda com visão sobre a importância da agropecuária, temos as previsões da necessidade de dobrar a produção vegetal, a fim de garantir, as condições adequadas de nutrição da população mundial (COLLOMB, 1999).

O fato determinante sobre a agropecuária é ser de suma importância na manutenção da vida. Partindo da base, ter levado a humanidade de uma condição nômade, para uma condição de construção social complexa e sendo um dos primeiros insumos comerciais, no decorrer do tempo essa importância, relevância e dinâmica se tornaram basais em toda sociedade humana.

Conforme Collomb (1999), com a expectativa de atingirmos uma população mundial na ordem de 9,8 bilhões de habitantes em 2050, torna-se um fator de risco para a produção mundial de alimentos, tendo em vista a necessidade de que esta cresça proporcionalmente ao percentual de desenvolvimento populacional. No entanto, existem preocupações nessa área, como a capacidade produtiva de alimentos de forma sustentável, e também as condições políticas e econômicas para que além da produção, seja respeitada a preservação dos recursos naturais e, seja de alcance das diversas faixas sociais da população mundial.

Nesse contexto, a necessidade de aplicação de recursos tecnológicos no auxílio às pesquisas de produção de alimentos mais especificamente TIC 's (Tecnologias de Informação e Comunicação), que são amplamente aplicadas nas diversas áreas da atuação humana, como por exemplo, na agricultura onde elas possuem um papel importante, permitindo não somente a melhoria na produção, mas também na capacidade logística e comercial de produtores, compartilhando a visão Kifli et al. (2021), com a relevante capacidade de além de exercer influência positiva nas áreas citadas, as TIC 's apresentam ainda um fator de melhoria na qualidade de vida dos produtores.

Na Figura 1, observam-se resultados de pesquisas realizadas no Twitter (twitter.com). Os assuntos pesquisados estão relacionados a infraestrutura (buracos em vias federais), produção (suinocultura), condições de trabalho no campo (ministério do trabalho e agricultura), e produção agrícola (segurança alimentar).



Figura 1. Exemplo de notícias e publicações que podem conformar uma base de dados agrícola.

Fonte: Elaborado pelo autor (2023), com base em publicações de G1 Goiás (2023) e DNIT (2023).

Considerando os aspectos descritos nos parágrafos anteriores, há argumentos mais do que suficientes para o desenvolvimento desta pesquisa na área agropecuária. Além disso, existem alertas no âmbito de segurança alimentar (a serem descritos a seguir), que trazem a necessidade de aplicação de políticas públicas, desenhadas através do emprego das novas tecnologias de obtenção e processamento de dados em agropecuária.

2.2 Tecnologia da Informação (TI) e Atividade Agrícola

Segundo Kifli et al. (2021), a TI torna-se no mundo o principal componente para uma atmosfera de disrupção. Essa atmosfera é construída em diversas áreas incluindo a agricultura.

A TI é um fator de mudança no gerenciamento de negócios, pois é um desenvolvedor dos negócios que a aplicam.

Ainda sobre o texto de Kifli et al. (2021), a observação sobre as propriedades produtivas se inicia por uma pesquisa, que relaciona o percentual populacional da Indonésia que possui smartphone, frente a população agrícola que possuem o mesmo dispositivo.

Nesta pesquisa observou-se que 64% da população da Indonésia já possui um smartphone e, desse grupo apenas 16% trabalham em atividades agrícolas.

Dessa forma o estudo realizado em *Disruption of information technology and farming empowerment strategies in Indonesia* (KIFLI et al., 2021), chama a atenção para o entendimento da informatização das atividades de agropecuária, a partir da disponibilidade ou não de smartphones da população envolvida na atividade. Esse resultado relaciona-se ao objetivo desta tese, dado que o seu objeto principal, é o levantamento de informações de apoio à agricultura coletados por meio de mídias sociais.

Se no estudo de Kifli et al. (2021), os indicadores utilizados são da Indonésia, no trabalho de Fonseca, Silveira e Massruhá (2018) são apontados dados do Censo agropecuário de 2006, sobre o número de estabelecimentos agropecuários que utilizam computador e acesso à internet, na data em questão, era 4,54% para estabelecimentos com computador e 1,87% com computador e acesso à Internet (Tabela 1).

Tabela 1. Presença de computadores em estabelecimentos rurais.

Tipo de eletrodoméstico utilizado	Número de estabelecimentos agropecuários	
	Unidades	Percentual
Computador	183.623	4,54
Acesso à internet	75.407	1,87

Fonte: Fonseca, Silveira e Massruhá, 2018.

2.3 Infraestrutura Logística e seus Impactos na Agropecuária

A relevância da infraestrutura para a produção agrícola, é um ponto fundamental para o seu desenvolvimento, redução de custos do alimento na mesa do consumidor, e também uma questão preocupante em termos de segurança alimentar.

Essa situação pode ser observada facilmente no Brasil e demonstrada de forma científica conforme Abbade (2020) em seu trabalho “*Food Price, Losses and Logistics Affecting Diet Diversification and Food Security*”. Nesse estudo é aplicado a análise de correlação e regressão linear múltipla, para evidenciar os impactos logísticos na perda de alimentos, elevação de preço, redução na diversificação dietética. Apesar do olhar global do estudo são apresentados números por região de acordo com o, Logistics Performance Index (LPI), índice de interesse para o presente estudo.

Para tornar compreensível a avaliação realizada no estudo de (ABBADE, 2020), é necessário citar os estudos da *Food and Agriculture Organization* (FAO) ligados ao *Food Lost*

and Waste (FLW), onde se pretende mensurar a perda de alimentos, seja na produção ou distribuição. Observando o FLW, o estudo de Abbade (2020) converte a estimativa dessa perda de alimentos, em calorias e os relaciona com os dados de déficit alimentar de uma população, então avalia essa perda em relação a logística de uma determinada região (país) criando uma classificação. O objetivo do trabalho do FLW é demonstrar a influência negativa promovida por condições ruins em termos de infraestrutura e a necessidade de geração de instrumentos que avaliem e apontem pontos importantes na infraestrutura nacional.

Conforme o trabalho de Abbade (2020), as questões logísticas (infraestrutura), são consideradas como principais variáveis capazes de influenciar negativamente a economia, e a quantidade de comida que chega à mesa das famílias, seja pela perda desses alimentos ou pelo acréscimo de preço gerado na redução de disponibilidade e redução da diversidade de alimentos. (Figura 2).

Na Figura 3, são ilustradas as hipóteses ligadas à influência da infraestrutura logística na melhoria ou piora das características de preços de produtos agropecuários, diversidade alimentar e déficit alimentar.

Na Figura 4, destacam-se as principais variáveis de influência de distribuição, diversidade e preços de produtos agrícolas utilizadas no estudo de Abbade (2020). E na Figura 5 apresentam-se os índices derivados que formam os pesos no sistema proposto, demonstrando a influência do coeficiente logístico nas hipóteses apresentadas na Figura 4.



Figura 2. Caminhão atolado com grãos. Fonte: Post realizado em Facebook.

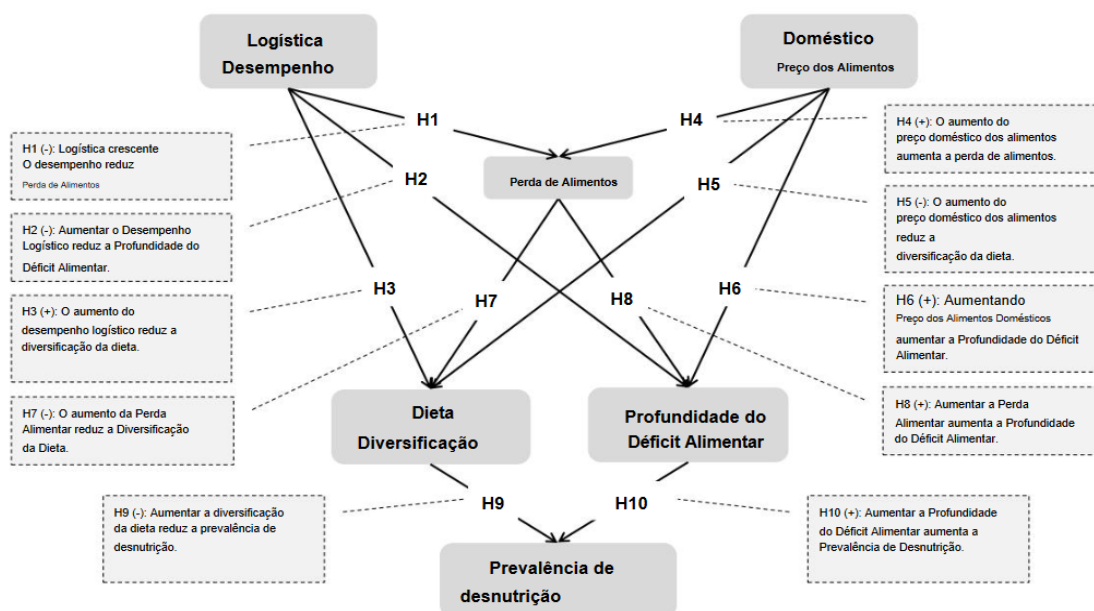


Figura 3. Relação logística, preço com a presença de alimentos e seu preço: Fonte: Abbade (2020).

Variável	Descrição	Fonte e Período
Logística Índice de Performance (LPI)	O LPI é a média ponderada das pontuações dos países nas seis dimensões principais, publicada pelo Grupo Banco Mundial (Banco Mundial, 2019).	Banco Mundial (2019)
Comida Doméstica Índice de Preços (DFPI)	O DFPI é dado pela razão entre a Paridade do Poder de Compra de Alimentos (Fppp) e a PPP Geral, resultando em um índice do preço dos alimentos no país relativo ao preço da cesta de consumo genérica (Capone et al., 2014; FAO, 2001, 2018a)	FAO (2011-2014)
Perda de Alimentos	Esta variável é uma medida que considera a perda de alimentos pós-colheita e pré-consumo como uma proporção da oferta agrícola doméstica (Gfsi, 2019).	GFSI (2019)
Diversificação da Dieta	Uma medida da participação de alimentos não amiláceos no consumo total de energia dietética. Maior é a participação de alimentos não amiláceos, maior é a diversidade de grupos de alimentos na dieta (Gfsi, 2019).	GFSI (2011-2013)
Profundidade do déficit alimentar (quilocalorias por pessoa por dia)	A profundidade do déficit alimentar indica quantas calorias seriam necessárias para elevar o subnutrido de seu estado, tudo o mais sendo constante (FAO, 2018a)	FAO (2011-2013)
Prevalência de Desnutrição (%)	A prevalência de desnutrição (PoU) representa a proporção da população cujo consumo alimentar habitual não é suficiente para fornecer a quantidade de energia dietética necessária para manter uma vida ativa e saudável normal (FAO, 2019)	FAO (2014)

Figura 4. Métricas e influência. Fonte: Abbade (2020).

	de Desempenho Logístico (LPI)	Doméstico Preço dos Alimentos Índice	Perda de Alimentos	Dieta Diversificação	Profundidade do déficit alimentar
Comida Doméstica Índice de Preço	-0,7740*** n = 122	1	-	-	-
Perda de Alimentos	-0,5006*** n = 103	+0,4696*** n = 102	1	-	-
Dieta Diversificação	+0,6977*** n = 103	-0,8288*** n = 102	-0,4441*** n = 113	1	-
Profundidade do déficit alimentar	-0,2907** n = 100	+0,4869*** n = 88	+0,0724 n = 75	-0,3618** n = 75	1
Prevalência de desnutrição	-0,5152*** n = 143	+0,6342*** n = 133	+0,3148** n = 106	-0,5891*** n = 106	+0,9099*** n = 114

Figura 5. Relação de aspectos logísticos e a distribuição e preços de produtos. Fonte: Abbade (2020).

Como resultado, pode-se observar a relação negativa entre o desempenho logístico (infraestrutura) onde: o preço doméstico dos alimentos ($\hat{y} = -0,7744$; $p < 0,0001$), nível de perda de alimentos ($\hat{y} = -0,5006$; $p < 0,0001$), profundidade da déficit alimentar ($\hat{y} = -0,2907$; $p = 0,003$) e prevalência de desnutrição ($\hat{y} = -0,5152$; $p < 0,0001$), sugerindo que países com melhor desempenho logístico tendem a ter menor nível de perda de alimentos e consequentemente preços mais baixos nos alimentos fornecidos internamente, menor profundidade do déficit alimentar e menor gravidade na prevalência da desnutrição da população. Por outro lado, o desempenho logístico tem correlação positiva e significativa com a diversificação da dieta ($\hat{y} = +0,6977$; $p < 0,0001$), sugere que países com logística mais aprimorada tendem a oferecer maior diversidade alimentar as suas populações.

Dessa forma o estudo de Abbade (2020), colabora com o presente trabalho, como base para uma das vertentes de assuntos a serem pesquisados e respondidos. Especialmente sobre a importância de aspectos logísticos na produção agropecuária, sendo essa uma das bases de informação a serem construídas como produto deste trabalho.

O trabalho de Abbade (2020), ganha destaque no presente texto, devido ao interesse do autor em destacar a importância das variáveis de infraestrutura (logística) na produção e distribuição agropecuária. Sendo esse requisito um dos alvos de pesquisa da solução desenvolvida, permitindo a observação dos pontos críticos para distribuição dos produtos agropecuários e, consequentemente, um dos principais pontos de influência para sua política de preços, disponibilidade e diversificação na mesa do consumidor final.

2.4 As Mídias Sociais

As mídias sociais, tornaram-se parte da vida das pessoas na última década, e sem dúvida, o volume de pessoas associadas a essas mídias cresce a cada dia. Neste capítulo vamos tratar, as plataformas, estatísticas e capacidade de influência dessas ferramentas.

Arora et al., (2019), no estudo *Measuring social media influencer index - insights from Facebook, Twitter and Instagram*, classifica as plataformas do título do trabalho como as mais populares, sendo essas as principais bases de estudo quando as fontes de dados são as plataformas de mídia sociais.

Observando a pesquisa apresentada pela empresa norte-americana Smart Insights “Global social media statistics research summary (2022)”, em junho de 2022, o número de usuários das mídias sociais chegou a 4.76 bilhões, representando 59%, da população mundial, conforme mostrado na Tabela 2.

Os números em si demonstram o alcance dessas plataformas, e tornam-se ainda mais impressionantes quando observamos a média de tempo de utilização por usuário, é de 2 horas e 31 minutos por dia.

Tabela 2. Participação e tempo de uso da sociedade em mídias sociais.

Utilização de mídias sociais – visão geral – jan. 2023				
Nº. de usuários	Evolução trimestral	Evolução anual	Tempo médio gasto / dia	Nº. médio de plataformas usadas ao mês / pessoa
4,76 bilhões	+ 23 milhões +0,5 %	+ 137 milhões +3,0 %	2h e 31m	7,2
Usuários x população	Usuários x população +18	Usuários das plataformas x usuário de internet	Usuárias femininas	Usuários masculinos
59,4%	77,8%	92,3%	46,3%	53,7%

Fonte: “Global social media statistics research summary, 2022”.

2.4.1 Facebook

O Facebook é controlado pela empresa Meta de Mark Zuckerberg. Apesar da postura recente da Meta em idealizar um início de uso em grande escala, de tecnologias imersivas como o Metaverso, a plataforma Facebook, é definida pela empresa, como uma plataforma de “conexão” entre pessoas, que devem ocorrer orientadas por grupos familiares ou de interesses comum por assuntos. Esta plataforma conta ainda com um Marketplace, onde é possível realizar compra e venda de itens.

Segundo o relatório Digital 2023: Global Overview Report, o Facebook alcançou a marca de 2,95 bilhões de usuários, consolidando-se como a maior rede social do mundo (KEPIOS, 2023).

2.4.2 Instagram

Também controlado pela Meta, o Instagram possui uma apresentação diferente do Facebook, onde é priorizado o registro de momentos específicos por meio de fotografias. Essa plataforma se destacou pela possibilidade de postagem de vídeos curtos associados a temas musicais (Reels), ainda segundo a Smart Insights o Instagram possui 2 bilhões de usuário sendo a quarta maior mídia social.

2.4.3 Twitter (X)

Diferentemente das anteriores, essa mídia é controlada por Elon Musk. Se caracteriza pelas originais postagens de até 140 caracteres, hoje ampliados para 280. Permite postagens com fotos e vídeos curtos. Dentre as três plataformas, o Twitter se destaca por ser a plataforma com maior número de postagens sobre política. O Twitter possui 556 milhões de usuários.

2.4.4. Análise de dados em mídias sociais

Sobre a análise de dados contidos nas mídias sociais, ilustrado na Figura 6, no estudo de Milusheva et al. (2021) é levantada a hipótese das potencialidades das mídias sociais como arcabouço de dados para o planejamento de gestão urbana, onde o autor identifica a possibilidade de uso desses dados, na identificação de locais com grande incidência de acidentes (o estudo foca na cidade de Nairobi no Kenya).

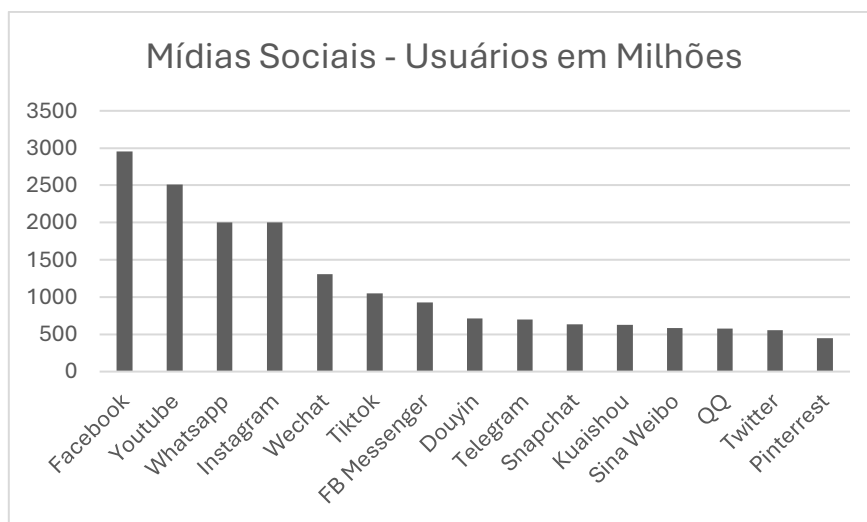


Figura 6. Market share das mídias sociais. Fonte: “Global Social Media Statistics Research Summary”, 2022.

Outro fator relevante sobre as mídias sociais, é o fenômeno Big Data, onde, sem dúvida, o grande volume de dados dessas plataformas pode ser visto como potenciais fontes estatísticas para o planejamento de infraestrutura, além de métricas de rastreamento de pobreza, doenças e outros índices sociais (MILUSHEVA et al., 2021).

Com o objetivo de localizar (regionalizar) o estudo estatístico sobre as mídias sociais são observadas também a representação do percentual de usuários ativos (no mundo) das mídias. Na América do Sul encontra-se em 7,7%, superando os 6,9% da América do Norte. No confronto entre população e usuários, a América do Sul possui 71,3% de sua população ativa nas mídias sociais. Fazendo uma transferência direta entre o percentual estimado e a população brasileira, teremos o número de 151 milhões de usuários ativos nas redes sociais como ilustrado na Figura 7.

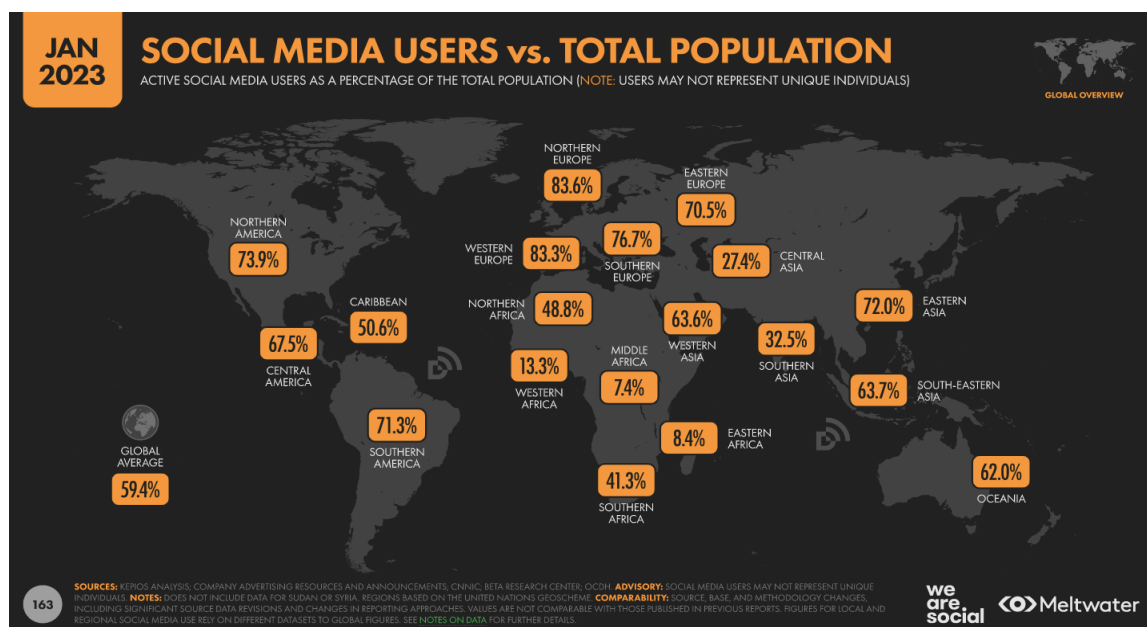


Figura 7. Presença da população nas mídias sociais. Fonte: *Global social media statistics research summary*, 2022.

2.5 Inteligência Artificial (IA)

Apesar das aplicações atuais como ChatGPT, que tem chamado muita atenção e criado inúmeras discussões sobre o futuro das IA's e a possibilidade de regulação do uso de suas ferramentas, os estudos científicos na área datam de 1950. No iconico trabalho de McCulloch e Pitts (1943), os autores tratam da neurofisiologia, com o objetivo de desenvolver uma teoria matemática da descrição do funcionamento dos neurônios e consequentemente, das redes neurais artificiais. As redes neurais artificiais, baseiam-se em um sistema computacional inspirado no funcionamento do cérebro humano.

Os sistemas baseados em rede neurais são aplicados ao processamento de linguagem natural, fala, reconhecimento de imagens e classificação em geral. A capacidade desse processamento é explorada em áreas de grande complexidade, devido à capacidade de aprendizado dessas redes, com grande capacidade de reconhecimento de padrões.

A capacidade das redes neurais de “aprenderem”, sem que necessariamente tenham sido programadas para determinado reconhecimento, é conhecido como Aprendizado de Máquina (machine learning). Este termo foi utilizado primeiramente pelo Prof. Arthur Samuel (WIEDERHOLD & MCCARTHY, 1992).

Trazendo a discussão para termos de aplicação, o trabalho de Zhang et al. (2021), trata da utilização de reconhecimento de imagens por meio de aplicação de IA, baseadas em redes neurais. As IA's, combinadas às técnicas de reconhecimento de imagem, permitem a extração de dados, processamento e classificação aprimorados. Hoje essas tecnologias estão bastante difundidas e fazem parte de sistemas como o de QR code, ou mesmo sistemas mais complexos, a exemplo das cidades inteligentes, onde é feito o reconhecimento de placas de veículos (via expressa) e de pessoas por meio da leitura facial (aeroportos). Outra aplicação importante, são as tecnologias de direção assistida, que contam com os algoritmos de reconhecimento de padrões para identificar objetos e obstáculos em ambientes dinâmicos, como o redor de um veículo em movimento. Essa capacidade deve-se a aplicação das redes neurais em reconhecimento de padrões (Figura 8).

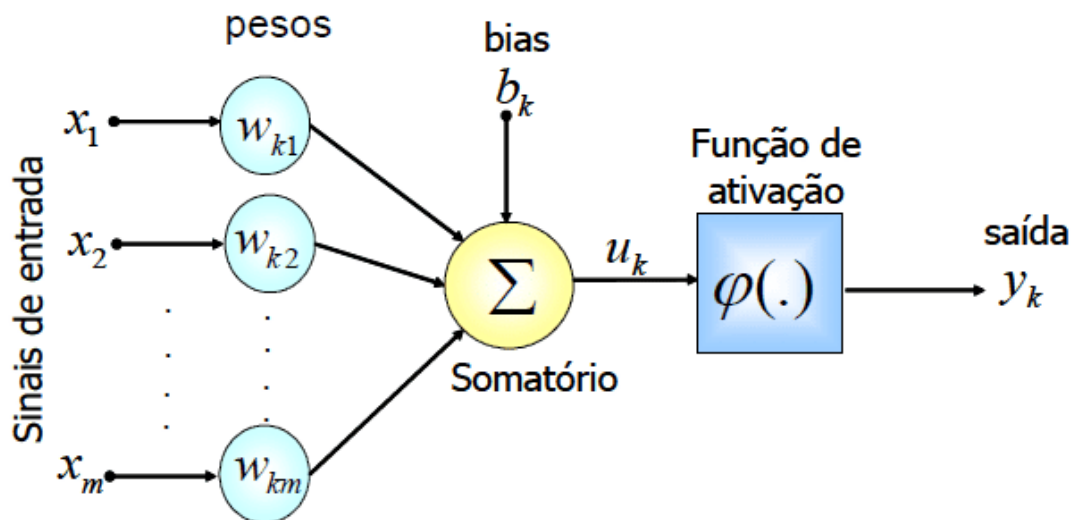


Figura 8 – Modelo de um neurônio artificial. Fonte: Adaptado de Haykin (2001).

Sobre a IA, ainda se faz necessário a descrição de aprendizado profundo (Deep Learning) que se caracteriza pela capacidade de processamento de alto volume de dados, devido ao alto número de camadas na rede (SARKER, 2021). Naturalmente essa abordagem traz custos e tempo de processamento bem mais altos, tendo em vista que objetiva a categorização de grandes volumes de dados.

Na Figura 9, pode-se ter uma ideia de como as subáreas da IA se relacionam.

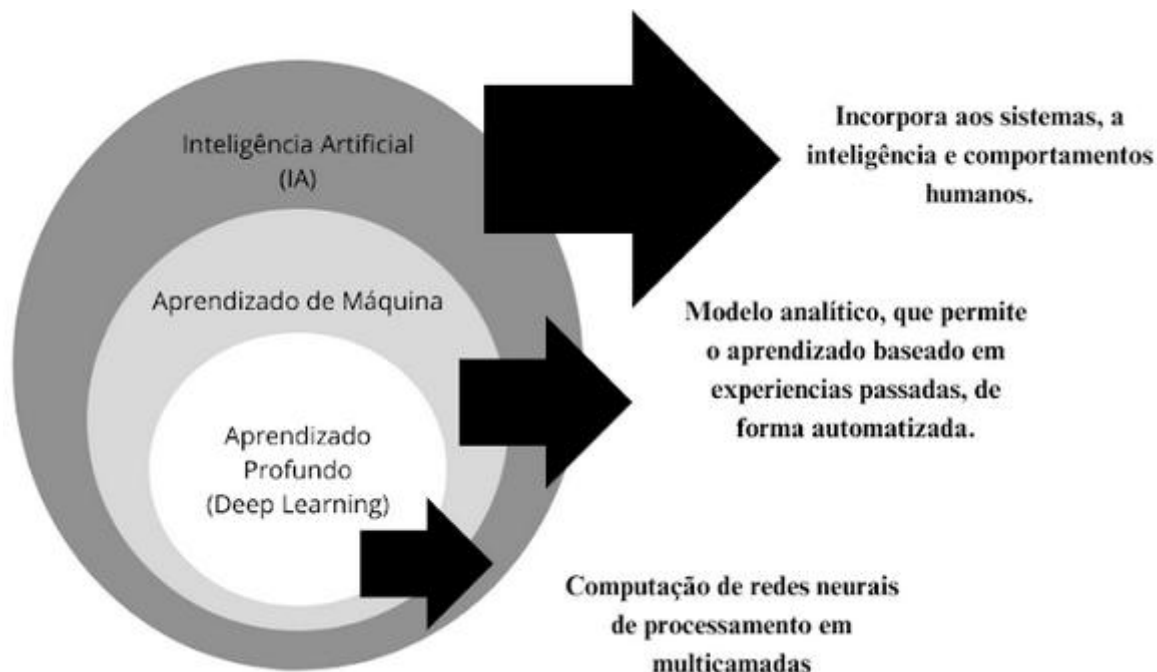


Figura 9. Comparação entre IA, Aprendizado de máquina e aprendizado profundo. Fonte: Sarker, 2021.

2.5.1 Inteligência artificial gerativa

Com o grande destaque dado nos últimos anos, ao ChatGpt da OpenAI, o termo Inteligência Artificial gerativa ganhou grande destaque. Segundo Singh e Sengar(2024), são os sistemas baseados em inteligência artificial, capazes de criar textos, imagens e demais formas de mídia. Esses sistemas baseiam-se em modelos de aprendizado capazes de compreender padrões existentes em seus dados de treinamento e a partir desses criar novos dados.

A seguir os principais tipos de IA's gerativas, e suas aplicações, quanto à produção de texto e ou imagem:

- 1 Redes Adversariais Generativas (GANs):
 - a. Função: Gerar novos dados que imitam as características dos dados de treinamento;
 - b. Tipo de Aplicação: Principalmente imagem (ex: geração de imagens, face swapping).
- 2 Modelos Baseados em Transformadores (Transformers):
 - a. Função: Realizar tarefas de geração de linguagem, como tradução automática e geração de texto coerente;
 - b. Tipo de Aplicação: Texto (ex: geração de texto, tradução de idiomas).
- 3 Autoencoders Variacionais (VAEs):
 - a. Função: Reduzir a dimensionalidade dos dados e gerar novas amostras a partir de uma representação latente;

- b. Tipo de Aplicação: Pode ser aplicado tanto em texto quanto em imagem, mas frequentemente utilizado em imagens (ex: geração de imagens a partir de representações latentes).

4 Modelos de Difusão (*Diffusion Models*):

- a. Função: Aprender distribuições de probabilidade complexas para gerar novos dados a partir de uma distribuição base simples;
- b. Tipo de Aplicação: Principalmente imagem (ex: geração de imagens de alta qualidade).

2.5.2 Modelos de avaliação de IA gerativa

Baseado no trabalho de Celikyilmaz et al. (2006), os métodos de avaliação de IA's generativas, são o conjunto de técnicas utilizadas na medição de eficácia e precisão das IA's com o objetivo de conferir o funcionamento desejado.

Esses métodos podem ser classificados em 3 categorias:

- Métrica automáticas: que não requerem intervenção humana, baseadas em precisão, recall e F1-score;
- Métrica centrada na avaliação humana: utilizam especialistas que analisam e classificam o resultado de processamento da IA;
- Métrica de aprendizado de máquina: utilizam modelos de aprendizado, treinados para avaliar saída de outras IA's.

Como forma de exemplificar os métodos mais utilizados quanto à avaliação gerativa de textos, serão considerados os modelos a seguir:

- a) Precisão (*Precision*): utiliza a proporção de verdadeiros positivos em relação ao total de classificações positivas feitas pelo modelo. A equação é:

$$\text{Precision} = \frac{Tp}{TP+FP}$$

Onde TP representa os verdadeiros positivos e FP, os falsos positivos.

- b) Sensibilidade (*Recall*): Avalia a proporção de verdadeiros positivos em relação ao total de elementos que realmente pertencem à classe positiva. A equação é:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Onde FN corresponde aos falsos negativos.

- c) **F1-score**: Representa a média harmonica entre precisão (*Precision*) e abrangência (*Recall*), balanceando ambos os critérios. A equação é:

$$F1\ 2x = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Exemplo:

A) Calculando precision:

- Palavras da resposta gerada pela IA: "Exercício", "melhora", "a", "saúde", "mental", "aumenta", "a", "força", "muscular", "e", "promove", "um", "coração", "saudável" (14 palavras no total).
- Palavras da resposta de referência (correta): "O", "exercício", "físico", "regular", "melhora", "a", "saúde", "mental", "promove", "um", "coração", "saudável", "e", "ajuda", "no", "controle", "de", "peso" (18 palavras no total).
- Dentre as 14 palavras geradas pela IA, as seguintes estão corretas em relação à referência: "Exercício", "melhora", "a", "saúde", "mental", "promove", "um", "coração", "saudável" (8 palavras corretas).

Calculando a precisão: $\text{Precision} = \frac{8}{14} = 0,57$.

B) Aplicação do modelo de avaliação centrada no humano:

Exemplo de Resposta Gerada pela IA:

Pergunta: “*Quais são os benefícios da leitura regular?*”

Resposta da IA: “*A leitura é importante porque ajuda a aprender novas coisas, relaxa a mente e melhora a concentração.*”

Para mensurar a eficácia da resposta gerada, implementam-se as seguintes métricas de avaliação por especialistas humanos:

1. Fluência

Critério: O texto gerado é gramaticalmente correto e de fácil leitura? O fluxo das ideias é natural?

Avaliação: A frase é gramaticalmente correta e fluente, sem erros aparentes. Não há interrupções ou problemas que prejudiquem a leitura.

Nota (escala de 1 a 5): 5 (Fluente);

2. Coerência

Critério: As ideias da resposta fazem sentido e estão ligadas de forma lógica?

Avaliação: A resposta é coerente. Menciona benefícios como "aprender", "relaxar" e "melhorar a concentração", todos relacionados ao tema da leitura.

Nota (escala de 1 a 5): 4 (Coerente, mas poderia ser mais detalhada);

3. Relevância

Critério: A resposta está diretamente relacionada à pergunta feita? Fornece informações úteis e precisas?

Avaliação: A resposta é relevante e responde diretamente à pergunta sobre os benefícios da leitura. No entanto, poderia incluir mais detalhes como "aumento do vocabulário" ou "melhoria nas habilidades cognitivas".

Nota (escala de 1 a 5): 4 (Relevante, mas não completa);

4. Criatividade

Critério: A resposta é inovadora ou traz alguma perspectiva criativa sobre o tema?

Avaliação: A resposta é bastante genérica e não traz uma visão criativa ou uma nova perspectiva sobre a leitura. Poderia ter citado exemplos mais criativos, como "leitura como uma forma de viajar por diferentes mundos sem sair do lugar".

Nota (escala de 1 a 5): 2 (Pouco Criativa);

5. Utilidade

Critério: A resposta fornece informações úteis e aplicáveis ao contexto?

Avaliação: A resposta fornece informações úteis, mas um pouco superficiais. Embora útil, faltam mais detalhes ou exemplos práticos que o usuário possa aplicar em sua rotina.

Nota (escala de 1 a 5): 3 (Útil, mas limitada).

Resumo da Avaliação:

Fluência: 5/5

Coerência: 4/5

Relevância: 4/5

Criatividade: 2/5

Utilidade: 3/5

2.5.3 Alucinação de IA gerativa

Hoje o uso da IA gerativa tornou-se uma ferramenta oportuna e natural ao usuário de internet. No entanto, situações como a criação da informação por esses IA's podem incorrer em respostas factualmente incorretas, que possuem uma aparência de credibilidade. Segundo Alkaissi; McFarlane (2023), esse fenômeno é uma preocupação significativa, na utilização de modelos de IA gerativas. Dessa forma essas alucinações são um ponto de atenção em seu uso, que envolvem principalmente:

A – Integridade Científica;

B – Confiança do usuário;

C – Necessidade de detecção e monitoramento no uso de IA's.

Essa possibilidade não invalida o uso da IA, mas, permite o reconhecimento de suas limitações, a necessidade de aperfeiçoamento contínuo e revalida a importância do desenvolvimento de diretrizes e comprometimento ético no uso dessas tecnologias.

2.6 Políticas Públicas Apoiadas por Sistemas de Informação

Um dos autores pioneiros na escrita sobre políticas públicas apoiadas por sistemas de informação, foi Kling (1978), em seu artigo *Information systems in policy making: Computer*

technology and organizational arrangements. Este texto traz diversos apontamentos sobre a necessidade de melhorias nos métodos de processamento e forma de interpretação dos dados.

Já em um recente trabalho de Vahedi; Arvand (2020), os sistemas de informação são ferramentas indispensáveis para a formação de políticas públicas. Essa visão deve-se principalmente as seguintes características: acesso em tempo real da informação, capacidade de análise e verificação de tendências, monitoramento de avaliação de políticas públicas, capacidade de transparência na prestação de contas.

Dessa forma os sistemas de informação apresentam-se como uma ferramenta confiável na orientação e tomada de decisão em políticas públicas, além da eficácia e eficiência no desenvolvimento dessas.

Finalizando os estudos ligados a políticas públicas, a base central sobre o entendimento do assunto fica com o texto de Almeida; Gomes (2018), onde o conceito geral de política pública é como um plano de ação ou decisão tomada por um governo ou organização pública para abordar questões específicas e alcançar objetivos de interesse público.

Essas políticas são desenvolvidas para lidar com problemas sociais, economicos, ambientais ou políticos, e geralmente, envolvem a alocação de recursos e a implementação de medidas para resolver ou mitigar esses problemas. As políticas públicas podem abranger uma ampla gama de áreas, como saúde, educação, segurança, meio ambiente, economia, entre outras, e são fundamentais para a governança e o funcionamento de uma sociedade.

2.7 Workflow Científico

De acordo com Aalst e Basten (2002), um Fluxo de Trabalho (*workflow*) científico é a organização sistemática das etapas de atividades necessárias à execução de uma pesquisa ou experimento.

Nesse contexto, os workflows científicos permitem a execução de tarefas (*in virtuo*) de um plano de trabalho, de acordo com o planejamento das atividades necessárias ao alcance de um objetivo. Em específico, neste trabalho, sendo o objetivo a obtenção, o tratamento e a análise de dados, inclui-se ainda a comunicação dos dados obtidos por meio de sua execução.

O uso de workflows científicos permite a transparência, reprodutibilidade e eficiência na execução das atividades de pesquisa.

2.8 Bot's (Robot)

O que é um Bot? A definição geral de Bot, é um programa de computador capaz de realizar tarefas repetitivas automatizadas, previamente programadas. Um termo comumente utilizado, quando tratamos de incidentes de privacidade ou obtenção de dados via Internet.

Uma de suas principais finalidades é a interação por meio de mídias sociais, para a coleta de dados (KOVAČIĆ et al, 2022).

2.9 Técnica de Web Scraping

O *web scraping*, é a técnica que utiliza a Internet como fonte de coleta de dados, esse processo deve ser automatizado preferencialmente por meio de Bots. É uma técnica que se relaciona diretamente com o Big Data, pelo seu acesso a grande volumes de dados, que combinado com IA, torna-se uma ferramenta valiosa no reconhecimento de padrões (LOTFI et al., 2023).

2.10 Big Data

O Big Data é o ativo do processamento do grande volume de dados produzidos devido à popularização da Internet e a utilização em massa de dispositivos móveis, que conferiu aos seus usuários a possibilidade de produzir dados na forma de texto, áudio, vídeos. Além disso os dispositivos, por si só, geram constantemente dados (IoT), sobre localização, configuração e acesso dos diversos sistemas utilizados por seus usuários.

Esses ativos, são o resultado da aplicação de métodos analíticos, que contam com grande infraestrutura para sua execução (MAURO et al, 2016).

2.11 IoT

Internet das Coisas (Internet of Things) refere-se a conexão dos mais diversos dispositivos, que vão desde nossos smartphones, smartwatches, leitores de livros digitais, geladeiras, cafeteiras, fones de ouvido, lâmpadas, interruptores, câmeras. Até mesmo equipamentos industriais, entre os quais sensores, motores, filtros e qualquer outro equipamento que seja acessível pela Internet (BERTE, 2018).

Essa conectividade de “todas as coisas”, é parte integrante das plataformas de Big Data, tendo em vista que as coisas enviam dados para as nuvens computacionais de seus fabricantes, como forma de monitoramento e acesso de seus usuários por meio de APP’s ou sites web.

2.12 Computação em Nuvem

A Computação em nuvem (Cloud Computing), pode ser definida como o modelo de acesso aos serviços de TiC, por meio da Internet. Estes serviços vão desde softwares como uma simples suíte office (em seu navegador, como um serviço pago mensalmente) até uma sofisticada rede de computadores, com servidores de grande porte e sistemas de segurança como firewalls e sistemas de detecção de invasores AHMAD et al., (2017).

Esses serviços permitem aos seus usuários, recursos de TI, sem a necessidade de gerenciamento de infraestrutura. O que confere ao serviço flexibilidade, escalabilidade, além da eficiência no atendimento das demandas.

2.13 Bancos de Dados e Tipos de Dados

A definição geral de banco de dados é do trabalho de Codd (1970), que o descreve como um conjunto organizado de dados, gerenciado de forma a permitir acesso e, recuperação dos dados armazenados.

2.13.1 Dados estruturados

Os dados estruturados, são os que respeitam uma estrutura de organização, onde os dados são claramente definidos utilizando chaves de identificação primária e estrangeira para relação entre eles. Usando exemplo comum de fonte de dados estruturados são os bancos de dados relacionais.

2.13.2 Dados semiestruturados

São os dados, que não cumprem completamente as regras de dados estruturados, onde possuem alguma estrutura, porém esta é flexível. Sendo um ponto divergente do dado estruturado, onde o banco de dados possui uma estrutura rígida.

Os dados semiestruturados, são frequentemente encontrados nos formatos como XML (Extensible Markup Language) e JSON (JavaScript Object Notation).

2.13.3 Dados não estruturados

Dados não estruturados são as informações que não possuem formato definido, ou organização estabelecida como no modelo de banco de dados. Estes dados podem ser textos, imagens, vídeos, áudios, como os comumente encontrados em posts de redes sociais, publicações de sites, etc.

Por não possuírem estrutura definida, torna-se um desafio seu processamento, quanto análise, qualificação e organização.

2.14 Estudo de Caso

Como base de estudo é utilizado o trabalho de Milusheva et al. (2021), que aplica técnicas similares as propostas no presente trabalho, com o objetivo de criar uma mapa de eventos. Onde serão destacados os acidentes de uma determinada região.

2.14.1 Mapeamento de acidentes automotivos em zonas urbanas

Na busca por pesquisas correlatas, que utilizem dados públicos (Internet), para mapeamento de zonas ligadas a acidentes, encontra-se o trabalho de Milusheva et al. (2021). Nesse trabalho, uma coleção de dados públicos, contidos nas mídias sociais, é utilizado com o objetivo de estratificar dados relacionados a colisões de trânsito em rodovias (Road Traffic Crash - RTC).

No trabalho Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning, de Milusheva et al. (2021), a partir de postagens realizadas no Twitter, sobre colisões de trânsito, são coletadas as informações relativas à localização geográfica dos acidentes. Na sequência, aplicou-se Machine Learning com objetivo de gerar estatística de pontos críticos, onde os acidentes ocorrem com maior frequência, e é levantada a hipótese do uso desses dados para o planejamento urbano. Ou seja, a criação de políticas públicas com o interesse na redução da violência no trânsito e questões relacionadas à infraestrutura.

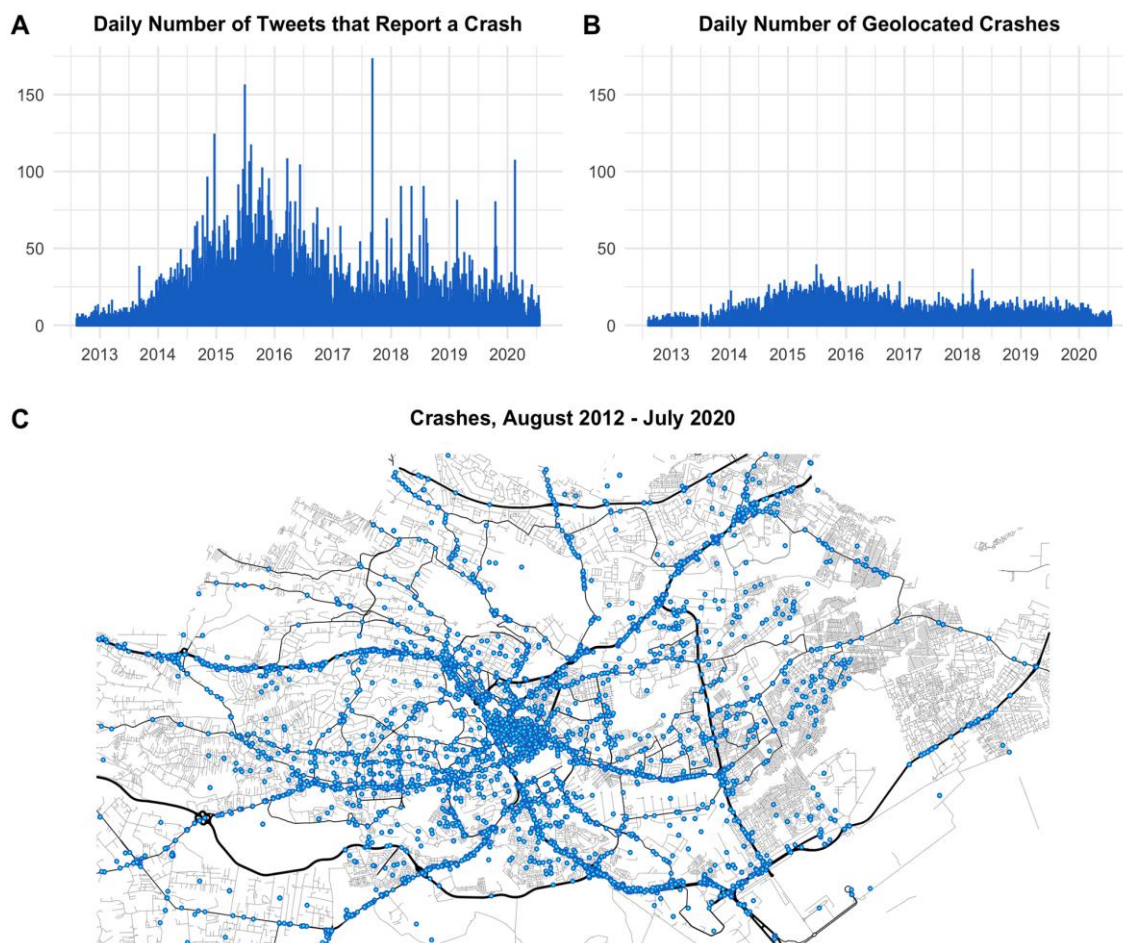


Figura 10. Mapa de Eventos. Fonte: Milusheva et al., 2021.

Como resultado, foi possível realizar o levantamento de dados e com o auxílio de sistemas de informação, realizar a criação de mapas, nos quais são identificados os “hotspot” da concentração de acidentes na cidade de Nairobi.

Dessa forma, o trabalho de Milusheva et al. (2021) evidência a capacidade de observação de fenômenos de nossa sociedade, por meio de dados publicados da Internet pela população civil, isto é, de forma não profissional ou ligada a qualquer tipo de entidade governamental.

2.14.2 Identificação de animais selvagens em estradas

Em destaque temos o trabalho de Ferrante et al. (2024), onde um sistema que combina câmeras em estradas e aprendizado de máquina, pretende identificar animais em extinção em estradas, com o objetivo de mitigar os riscos e entender as espécies que são atingidas por esse tipo de acidente (Figura 11).

Naturalmente esse tipo de aplicação torna-se um agregador de informações sobre os impactos do desenvolvimento na fauna brasileira.

Neste trabalho, a ação dos faróis e ruídos destacam-se como principais causas de acidentes com animais, pois geram confusão nos animais que estão em travessia.



Figura 11. Identificação de Animal por IA - Banco de dados de espécies como o tamanduá-bandeira permitiu treinar para a realidade do país modelo utilizado em outros países. Fonte: imagem montagem de Gabriel Souto Ferrante (sobre foto de Miguel Rangel Jr/Creative Commons).

3. MÉTODO PROPOSTO

Como forma de organizar a execução das atividades, os eventos foram transformados em processos e a descrição realizada por meio de diagrama de fluxo, conforme mostrado na Figura 12:

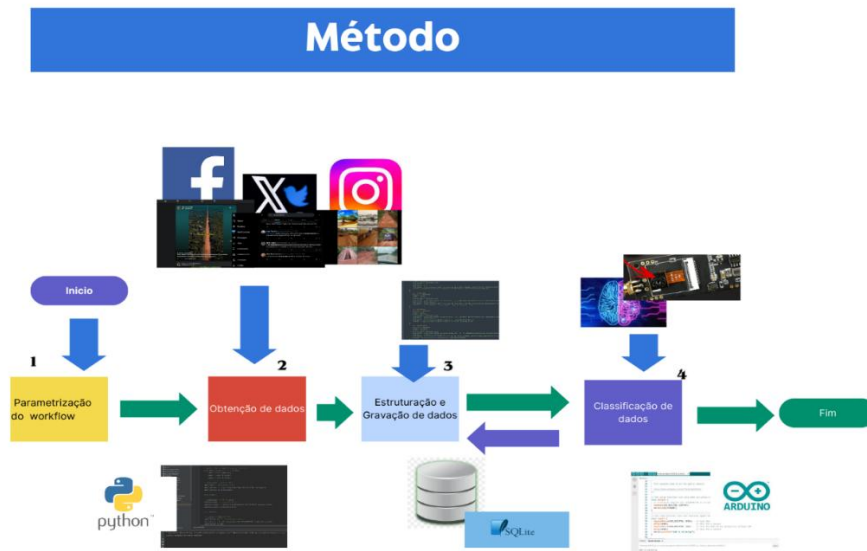


Figura 12. Diagrama de execução e demonstração de ambiente da aplicação. Fonte: Elaborado pelo Autor.

Etapa 1: início da execução de workflow escrito em linguagem de programação Python, onde serão definidas as palavras-chaves (tags) a serem pesquisadas.

Etapa 2: execução de um Bot, que por meio da técnica de web scraping realizará a busca dos termos parametrizados nas mídias sociais (Facebook, X e Instagram).

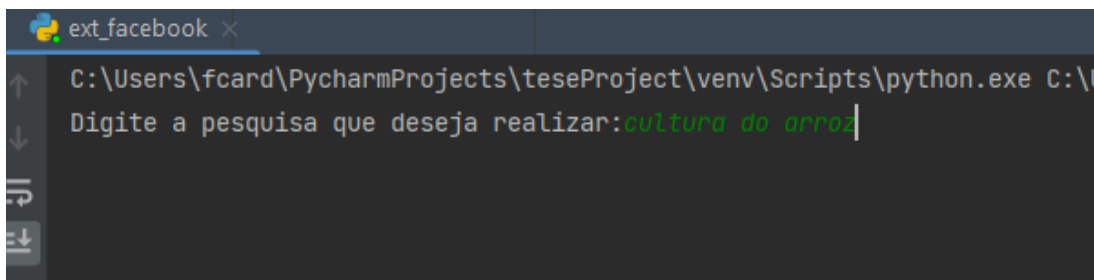
Etapa 3: todos os posts (textos) e imagens, relacionados aos termos definidos na pesquisa, serão gravados em arquivos semiestruturados (no formato JSON), que ao fim serão transferidos para uma base dados em SQLite.

Etapa 4: utilização de engine (motor) de IA gerativa (chat GPT e Gemini) para validação de textos e imagens (em desenvolvimento) obtidas na etapa 2.

3.1 Descrição da Solução

A construção da solução, consiste no formato de workflow científico, onde não é desenvolvido produto de software em si, mas neste caso a combinação de scripts, Bot's, banco de dados, arquivos de dados semiestruturados (JSON), treinamento e aplicação de IA.

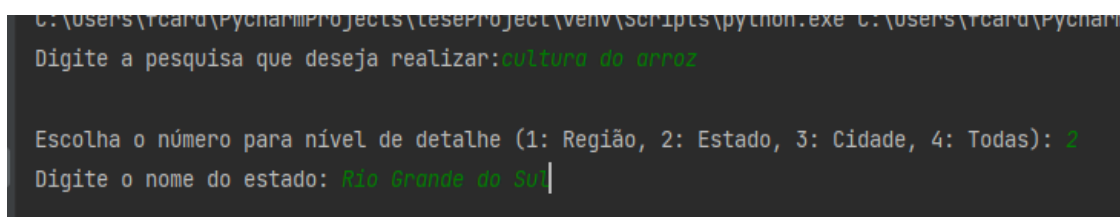
O início da aplicação, solicita os termos a serem pesquisados (ex. cultura do arroz), em seguida é iniciada a execução do Bot, que realiza o login na rede social, e busca os diversos posts contidos com o termo desejado (Figura 13).



```
ext_facebook x
C:\Users\fcarr\PycharmProjects\teseProject\venv\Scripts\python.exe C:\U
Digite a pesquisa que deseja realizar:cultura do arroz|
```

Figura 13. Parametrização de termos de pesquisa. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Em seguida a aplicação permite a definição da região geográfica, onde é permitido selecionar por estado, região ou cidade (Figura 14).



```
C:\Users\fcarr\PycharmProjects\teseProject\venv\Scripts\python.exe C:\Users\fcarr\Pycharm
Digite a pesquisa que deseja realizar:cultura do arroz

Escolha o número para nível de detalhe (1: Região, 2: Estado, 3: Cidade, 4: Todas): 2
Digite o nome do estado: Rio Grande do Sul|
```

Figura 14. Parametrização de região. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Dessa forma, a pesquisa é direcionada, esse filtro tem o objetivo de aumentar a assertividade do conteúdo a ser processado, que vai gerar o conteúdo de pesquisa na área selecionada.

Sobre a abrangência geográfica da pesquisa, é realizado o controle por meio de um arquivo do tipo JSON, que define a região, estados e cidades. Tanto no nível de regiões e estados, é possível inserir ou excluir cidades de acordo com a importância desses em uma pesquisa. Essa opção deve-se ao fato da existência de 5.568 municípios segundo dados do IBGE.

Para a organização, os dados de pesquisa são separados por diretório, onde ficam concentrados os textos e imagens obtidos (Figura 15).

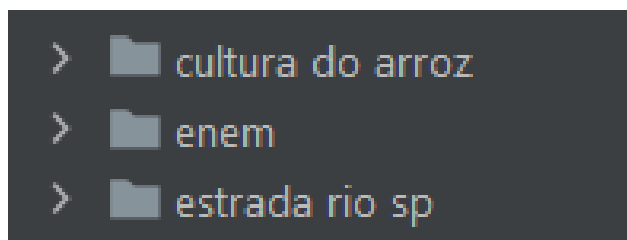


Figura 15. Resultado de pesquisa. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Como forma primária de organização dos dados, a aplicação cria um arquivo de dados semiestruturado do tipo JSON, com todos os textos e referência de imagens encontradas sobre o assunto da pesquisa (Figura 16 e Figura 17).

```

C: > Users > fcard > PycharmProjects > teseProject > cultura do arroz > {} cultura_do_arroz.json > {} 0
2      {
3          "ID": 458811665981,
4          "Data post": "27 de março de 2017",
5          "Cidade": "PORTO ALEGRE",
6          "Estado": "RIO GRANDE DO SUL",
7          "Hora extração": "23/03/2024 10:52",
8          "Link perfil": null,
9          "Nome perfil": null,
10         "Link imagem": "https://scontent-gig4-1.xx.fbcdn.net/v/t31.18172-8/17434680_17
11         "Conteúdo": "Sugestão de almoço delicioso! Apenas 18,00 reais almoço + suco na
12     },
13     {
14         "ID": 286474007485,
15         "Data post": "26 de março de 2017",
16         "Cidade": "PORTO ALEGRE",
17         "Estado": "RIO GRANDE DO SUL",

```

Figura 16. Arquivo JSON com resultado de mineração. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

```

{
    "ID": 443708700251,
    "Data post": "17/06/2024",
    "Cidade": "Pelotas",
    "Estado": "Rio grande do sul",
    "Hora extração": "06/04/2024 12:11",
    "Link perfil": null,
    "Nome perfil": null,
    "Link imagem": "https://scontent.fsdu12-1.fna.fbcdn.net/v/t1.6435-9/199157653_2990378257862905_4991276245506360826_n.jpg?stp
    "Conteúdo": "#NOTÍCIA - Técnica é capaz de aumentar produtividade de grãos em 26 sacas por hectare no Sul Uma técnica de pla
},

```

Figura 17. Arquivo JSON com resultado de mineração. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Em seguida os dados são transferidos para um banco de dados, onde características dos dados como localização da postagem, recebem códigos de estado e município visando uma futura aplicação de geolocalização de postagens e assuntos, são utilizados códigos baseados no mapa do IBGE.

Além disso, a estruturação dos dados, permite o tratamento de questões como duplicação de dados, e permite a consulta e geração de relatórios baseados em linguagem SQL.

Nas Figuras 18 e 19, é possível observar o modelo de banco de dados utilizado no projeto.

Grande	dataPost	conteudo	linkPerfil	img
1	21/07/2023	Manejo integrado da mancha parda em arr	https://www.facebook.com/novohorizor	y0yà JFIF yi 6Photosho...
2	21/07/2023	Manejo integrado da mancha parda em arr	https://www.facebook.com/novohorizor	y0yà JFIF yi Photosho...
3	03/03/2023	Alô, alô produtores de Arroz de todo mund	https://www.facebook.com/equipfieldc	y0yà JFIF yi Photosho...
4	03/03/2023	Alô, alô produtores de Arroz de todo mund	https://www.facebook.com/equipfieldc	PNG IHDR öy... [59
5	27/01/2023	Com a valorização do arroz, é importante ni	https://www.facebook.com/dressleni?_c	y0yà JFIF yi Photosho...
6	27/01/2023	Com a valorização do arroz, é importante ni	https://www.facebook.com/dressleni?_c	y0yà JFIF yi Photosho...
7	27/01/2023	Com a valorização do arroz, é importante ni	https://www.facebook.com/dressleni?_c	y0yà JFIF yi Photosho...
8	27/01/2023	Com a valorização do arroz, é importante ni	https://www.facebook.com/dressleni?_c	y0yà JFIF yi Photosho...
9	08/09/2023	Você sabia que algumas doenças podem ch	https://www.facebook.com/massarifertil	y0yà JFIF yi Photosho...

Figura 18. Resultado da estruturação dos dados obtidos. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

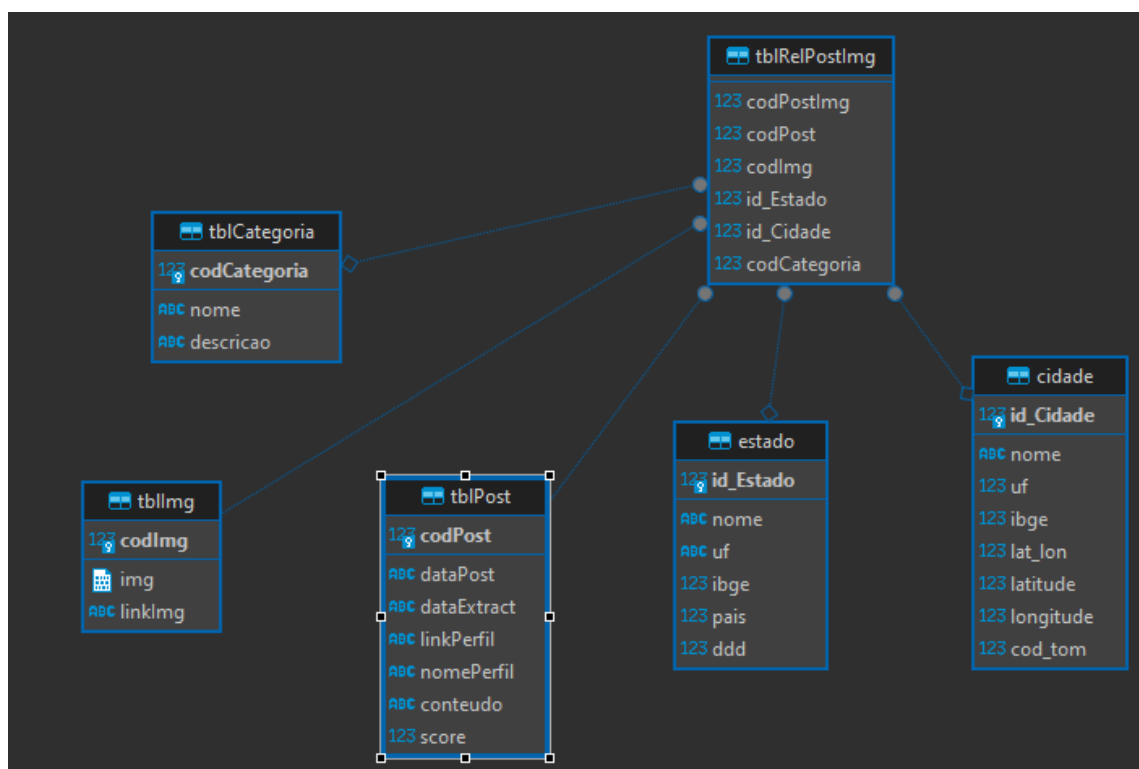


Figura 19. Modelo de banco de dados. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Após a obtenção e tratamento dos dados, estes são submetidos a uma ferramenta de Business Intelligence (BI), plataforma analítica que integra, processa e visualiza dados para gerar insights e apoiar processos decisórios organizacionais. Neste exemplo, foi utilizado o Power BI da Microsoft. Como resultado, obtêm-se os destaques de localização, número e percentual de ocorrências por região de um determinado termo pesquisado. Essa disponibilidade dos dados permite o rastreo de medida de ocorrências de uma determinada questão citada pela população, por meio das mídias sociais.

Após a exposição do funcionamento do workflow, será exemplificado um caso completo de pesquisa. No entanto é importante citar as características do projeto até o presente ponto, onde a obtenção de dados realizada pelo workflow do presente projeto, se apresenta como uma ferramenta de obtenção e análise de dados, capaz de obter dados de interesse por meio de mídias sociais. Em condições de serem submetidos a técnicas de BI, se tornando uma possível ferramenta de apoio a pesquisa e tomada de decisão, por permitir rastrear assuntos diversos, com rastreabilidade de dados (fonte, origem e data) (Figura 20 e Figura 21).

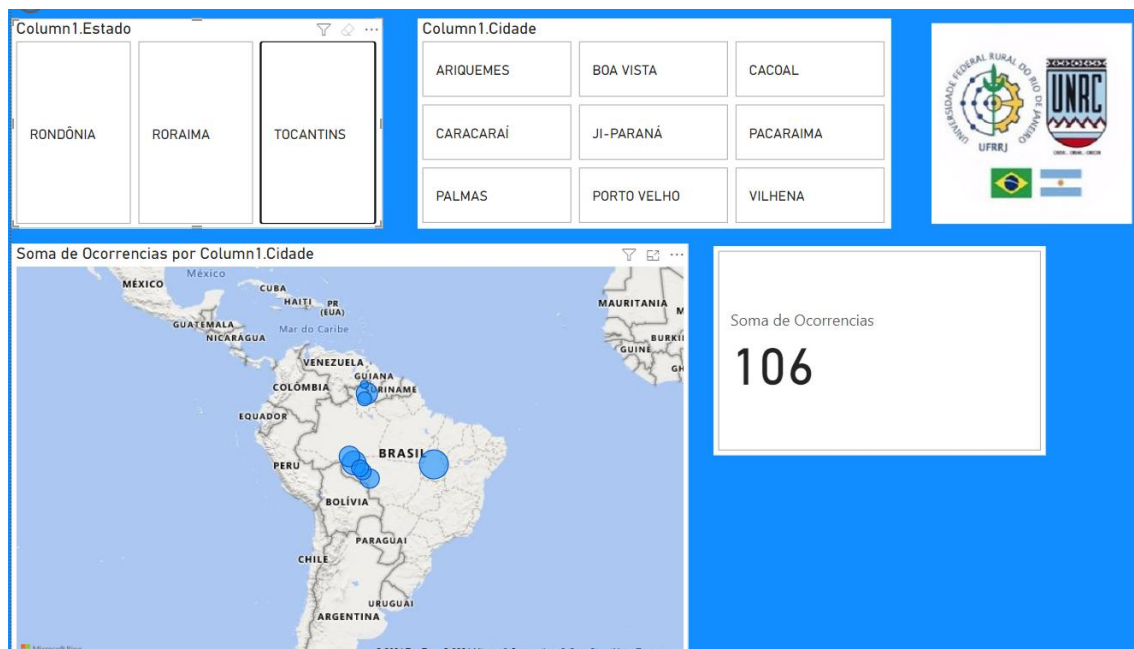


Figura 20. Exemplo de dados obtidos na aplicação, submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

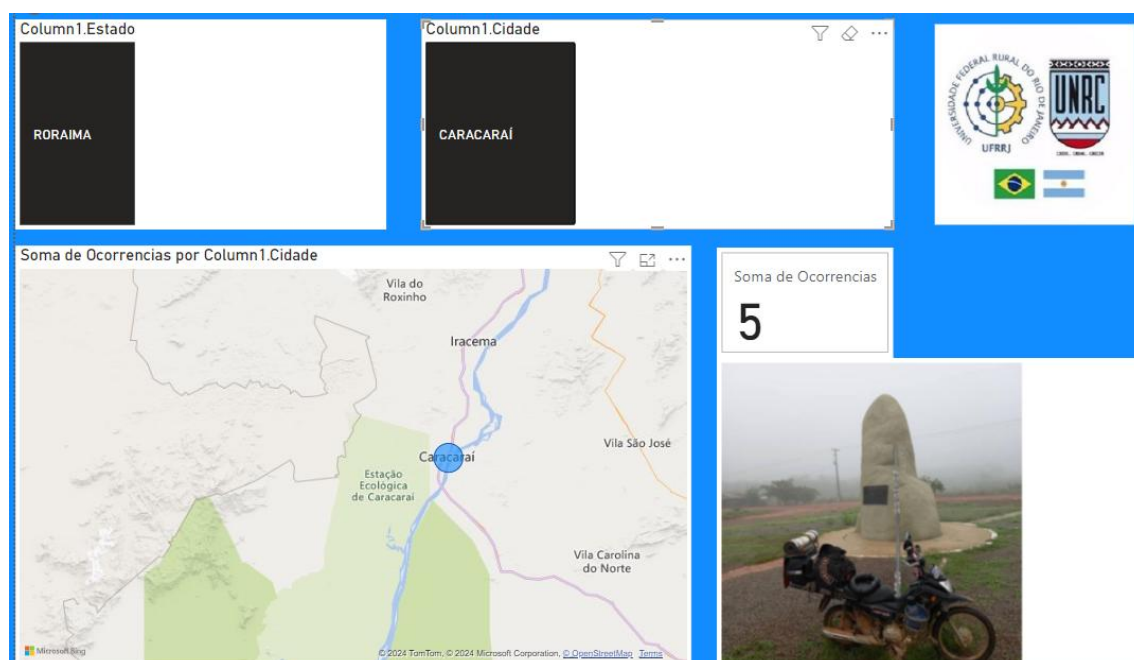


Figura 21. Exemplo de dados obtidos na aplicação, submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Ainda em tempo, é necessário destacar a forma de apresentação dos dados após submissão ao Power BI, onde torna-se possível aplicar filtros de forma dinâmica, onde podemos ver o conjunto de dados por estado, município e visualizar a incidência de informação no mapa além da pré-visualização do conteúdo de mídia da pesquisa.

3.2 Experimento - Brusone

Tendo em vista o fundamento de tese, que trata da importância de dados de mídias sociais, no apoio a pesquisa e formação de política públicas, por meio de sistema capaz de obter, tratar e categorizar dados dessas fontes.

Nesse experimento será o workflow será parametrizado, para resgatar postagens que possuam o termo Brusone (Figura 22).

```
ext_facebook
C:\Users\fcad\PycharmProjects\teseProject\venv\Scripts\python.exe C:\Users\fcad\PycharmProjects\teseProject\ext_facebook.py
Digite a pesquisa que deseja realizar:brusone
```

Figura 22. Nova parametrização. Fonte: Aplicação desenvolvida na pesquisa pelo autor.

Brusone é o nome de um dos principais patógenos da lavoura de arroz (Figura 23) e trigo, e possuem diversos artigos no repositório da Embrapa – Manejo de Doenças, esses artigos constam na bibliografia deste trabalho.



Figura 23 – Brusone em folha de trigo. Fonte: Lau, Maciel e Zoldan (2021).

A região predominante para a busca, será a região Sul, tendo em vista a relevância na produção de trigo e arroz (veja Figura 24).

```
C:\Users\fcad\PycharmProjects\teseProject\venv\Scripts\python.exe C:\Users\fcad\PycharmProjects\teseProject\ext_facebook.py
Digite a pesquisa que deseja realizar:brusone

Escolha o número para nível de detalhe (1: Região, 2: Estado, 3: Cidade, 4: Todas): 1
Digite o nome da região: Sul
```

Figura 24. Nova parametrização. Fonte: Elaborado pelo Autor.

O Bot, realizará as buscas com filtragem de posts envolvendo a região sul, dessa forma procuramos direcionar a pesquisa com o que deve ser a região com a maior incidência do termo Brusone (Figura 25).

específico, foi decido a anotação de um segundo experimento com termos mais amplos, o que pode ajudar no volume de dados e maior pluralidade na natureza das fontes, uma vez que a maioria das fontes do primeiro experimento vem de universidades, empresas e pesquisadores. Com isso procuramos um parâmetro de pesquisa que tenha maior participação da população em geral.

3.3 Experimento – Vassoura de Bruxa no Cacaueiro.

O segundo experimento será parametrizado, com os termos vassoura, bruxa e cacau. Deste modo procuramos encontrar dados sobre a produção de cacau, vamos especificar a região como o estado da Bahia (Figura 27).

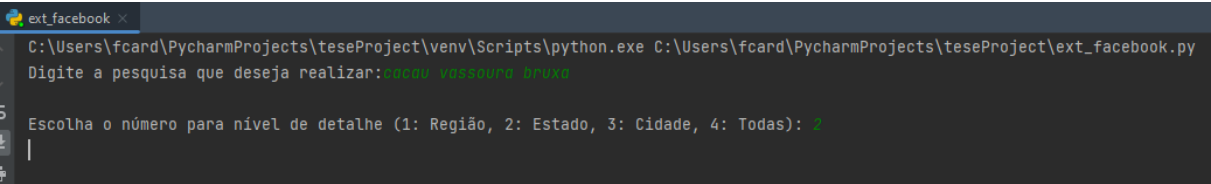


Figura 27. Parametrização. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Nesse experimento os municípios filtrados pelo sistema foram respectivamente "ILHEUS", "PORTO SEGURO", "VALENÇA", "JEQUIE", "SALVADOR", "FEIRA DE SANTANA", "VITÓRIA DA CONQUISTA", "CAMACARI", "ITABUNA", onde foram obtidas 80 entradas de dados, com um tempo de execução de 15 min. Das 80 linhas, apenas 1 possuía erro e não foi processada nos métodos de BI (veja Figura 28, Tabela 3 e Figura 29).

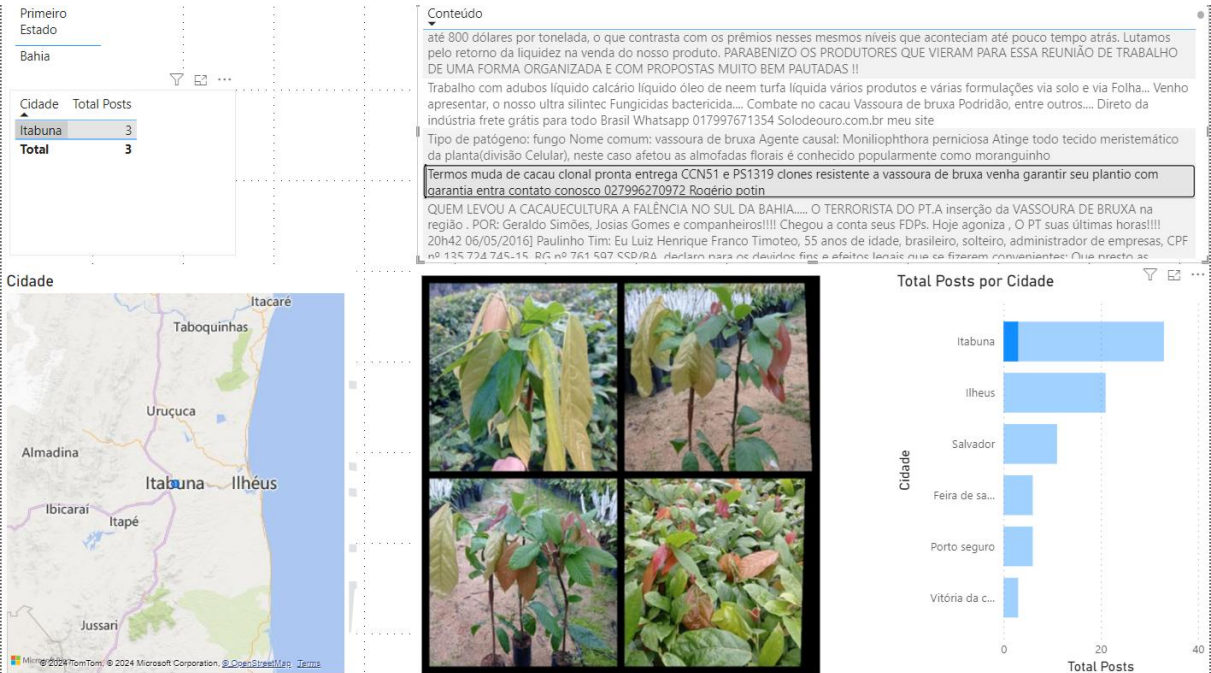


Figura 28. A distribuição de postagens por cidade. Fonte: Aplicação desenvolvida na pesquisa pelo autor.

Tabela 3. Contagem de dados obtidos pelo Bot.

Município	Número de Posts
Feira de Santana	6
Ilhéus	21
Itabuna	33
Porto Seguro	6
Salvador	11
Vitória da Conquista	3
Total	80

Fonte: O Autor.

```
1  [
2    {
3      "ID": 286966644547,
4      "Data post": "17/03/2024",
5      "Cidade": "Ilheus",
6      "Estado": "Bahia",
7      "Hora extração": "18/04/2024 16:33",
8      "Link perfil": "https://www.facebook.com",
9      "Nome perfil": null,
10     "Link imagem": "https://scontent.fsd12-1.fna.fbcdn.net/v/t39.30808-6/276032459_3268064...
11     "Conteúdo": "IHÉUS É DESTAQUE NO JORNAL ESTADO DE MINAS! A nossa cidade levada aos qua
12   },
13   {
14     "ID": 646605840810,
15     "Data post": "17/03/2024",
16     "Cidade": "Ilheus",
17     "Estado": "Bahia",
18     "Hora extração": "18/04/2024 16:33",
19     "Link perfil": "https://www.facebook.com",
20     "Nome perfil": null,
21     "Link imagem": "https://static.xx.fbcdn.net/images/emoji.php/v9/t58/1/16/1f929.png",
22     "Conteúdo": "IHÉUS É DESTAQUE NO JORNAL ESTADO DE MINAS! A nossa cidade levada aos qua
23   },
24   {
25     "ID": 380362252789,
26     "Data post": "01/07/2024",
27     "Cidade": "Ilheus",
28     "Estado": "Bahia",
29     "Hora extração": "18/04/2024 16:33",
30     "Link perfil": "https://www.facebook.com"
```

Figura 29. Arquivo JSON resultado de processamento. Fonte: Aplicação desenvolvida na pesquisa pelo autor.

3.3.1 Volume de dados

O volume de dados obtidos (80 registros), refere-se ao total filtrado e selecionado pelo Bot, naturalmente esse número não reflete o número total de posts da plataforma de mídia social. Enfatiza-se que é utilizado um Bot automatizador, e que em caso de uso de API o número poderá se alterar. Levando em consideração que a API dá acesso diretamente à base de dados da plataforma e o Bot acessa os posts oferecidos pela plataforma por meio de site web e seu acesso é feito por meio do html da página apresentada.

3.3.2 Da veracidade dos dados

Neste momento é importante destacar, que os dados utilizados na pesquisa, são postagens públicas, realizada nas mídias sociais, sem nenhum tipo de checagem quanto à sua veracidade. Porém os assuntos pesquisados não são politicamente sensíveis, tratando essencialmente de situações cotidianas.

3.3.3 Da produção de factoides

Dentro da produção de dados por IA's gerativas, observamos o fenómeno chamado de “Artificial Hallucinations in ChatGPT” (Alucinações Artificiais no ChatGPT), que é a geração de informações aparentemente verdadeiras, mas que não são fatos reais (por Alkaissi H, McFarlane S I).

Desta forma, na presente pesquisa não foi solicitado ao “Engine” da IA a criação de textos, mas sim a categorização do conjunto de dados obtidos pelo Bot.

3.4 Aplicação de IA

Naturalmente buscas realizadas com maiores volumes de dados, terão necessidade de validação campos que permitam filtrar os textos por meio de alguma métrica que indique o quanto um determinado texto (post) encontra-se dentro do assunto e ou categoria desejada.

Hoje com a evolução da IA, torna-se preponderante a aplicação dessa técnica em sistemas voltados a obtenção e tratamento de dados em pesquisas científicas ou de apoio a políticas públicas, sendo essa a técnica escolhida para o presente projeto.

Como método será utilizada a biblioteca “OpenAI”, que permite acesso à engine de IA “gpt-3.5-turbo”. Com isso tem-se a aplicação de um modelo de IA gerativa para avaliação dos textos encontrados pelo Bot, implicando no preenchimento da lacuna de ausência de validação de conteúdo.

Sobre a escolha do método baseado em OpenAI, vale destacar que é possível a utilização de outras bibliotecas como “numpy” e “pandas”, para execução de rede neural local para a classificação do texto. Porém devido ao grande número de possibilidades de assuntos e infinidade de textos de interesse nas áreas de agropecuária e políticas públicas, a necessidade de treinamento de redes para a classificação dos textos se tornaria inviável em termos técnicos, quanto a espaço, poder de processamento e tempo de treinamento.

3.4.1 Sistemas de categorização ou score

Observando as formas de realizar a filtragem dos dados dentro do interesse do pesquisador utilizador do sistema, nos deparamos com pelos menos duas formas, sendo a categorização, que visa definir um assunto entre uma base de opções, e o score, onde procura-se definir um valor de assertividade para o objeto em texto, em nosso caso o texto de um post encontrado pelo Bot (Figura 30).

Sendo assim, será utilizado o campo “score”, como uma espécie de escala de cinza, com número de 1 até 5, onde 1 significa texto sem aderência ao assunto e 5, onde o texto considerado completamente aderente.

A partir desse ponto vamos “Zerar” todo o banco de dados, tendo em vista que as pesquisas anteriores não realizavam o Score de texto via IA. Em seguida repete-se as pesquisas da sessão anterior com a inclusão desse recurso (veja Figura 31).

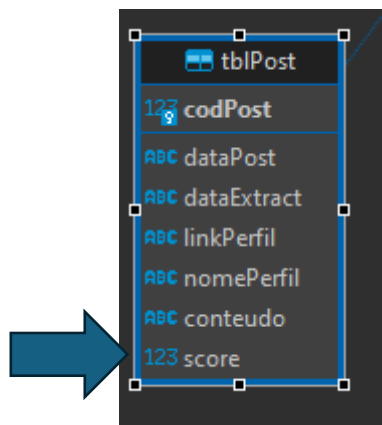


Figura 30. Detalhamento do campo score (Resultado de aplicação de IA). Fonte: Aplicação desenvolvida na pesquisa pelo autor.

```
import openai

# Defina sua chave de API
openai.api_key = 'sk-proj-5g230YFbhs15w0dt23w0T3B1bkFJ4D8pTu4R2vGRc20XIeXZ'

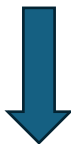
# Função para gerar texto
1 usage new *
def classificatexto(prompt):
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": prompt}
        ]
    )
    return response['choices'][0]['message']['content']

2 usages new *
def retornavalorAI(promptX, promptE):
    # Exemplo de uso
    respostaX = classificatexto(promptX+promptE)
    print(respostaX)
    return int((respostaX))
```

Figura 31. Integração OpenAI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

3.5 Experimento – Vassoura Bruxa no Cacaueiro – Com IA

A repetição do experimento, consiste na execução com todos os parâmetros da rodada descrita anteriormente. Naturalmente nesta rodada contaremos com a presença do campo “score”, que possui a leitura de IA, para determinação da taxa de pertinência do assunto com o texto (veja Figura 32).



	ABC estado	ABC cidade	123 codPost	ABC data	ABC conteudo	123 score	ABC link
1	Bahia	Salvador	440.503.143.257	17/09/2024	VAMOS A BRASÍLIA I	4	https://scon
2	Bahia	Salvador	156.052.191.335	17/09/2024	VAMOS A BRASÍLIA I	5	https://scon
3	Bahia	Salvador	813.637.154.470	17/09/2024	VAMOS A BRASÍLIA I	4	https://scon
4	Bahia	Salvador	406.599.149.329	27/05/2024	https://www.youtu	1	https://extei
5	Bahia	Salvador	563.405.049.349	01/02/2024	Itacaré - Ba Antigo re	3	https://scon
6	Bahia	Salvador	816.471.192.062	01/02/2024	Itacaré - Ba Antigo re	3	https://scon
7	Bahia	Salvador	289.571.257.445	01/02/2024	Itacaré - Ba Antigo re	3	https://scon
8	Bahia	Salvador	1.015.104.962.021	01/02/2024	Itacaré - Ba Antigo re	2	https://scon
9	Bahia	Salvador	310.277.065.179	01/02/2024	Itacaré - Ba Antigo re	2	https://scon

Figura 32. Tabela de dados, resultado de processamento. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Será repetida também a submissão do resultado de dados ao Power BI, agora contando com a possibilidade de filtragem por score, permitindo reduzir o número de resultados de acordo com o critério de assertividade selecionada pelo pesquisador.

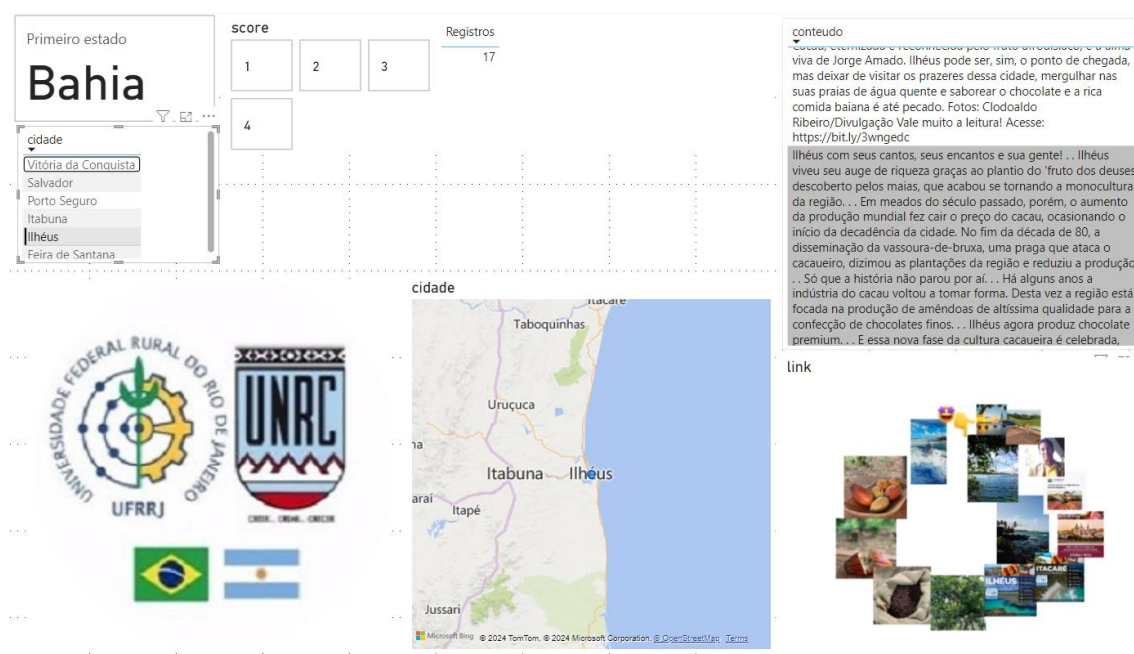


Figura 33. Resultado de processamento submetido ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Em destaque, na Figura 33, o campo score, que traz a avaliação da IA, quanto a assertividade do texto, possibilitando a filtragem ou orientação quanto a importância dos textos obtidos pelo Bot (ilustrado na Tabela 4).

Tabela 4. A distribuição de postagens por cidade, e score de dados gerado por IA.

Cidade	Número Posts	Scores das postagens
Feira de Santana	10	4 e 5
Ilhéus	17	1, 2, 3, e 4
Itabuna	33	1, 2, 3, 4, e 5
Porto Seguro	6	1 e 4
Salvador	18	1, 2, 3, 4, e 5
Vitória da Conquista	6	1, 2, 3 e 5
Total		90

Fonte: Elaborado pelo Autor.

3.6 Experimento em Espanhol - “Brucelosis Enfermedad”

Tendo em vista a identidade do PPGCTIA, que o ocorre no formato binacional (Brasil-Argentina), em parceria com a UNRC (Universidad Nacional de Río Cuarto), foram realizadas adaptações no sistema a fim de possibilitar a utilização do workflow, também em espanhol sobretudo com as características da agropecuária Argentina.

Dessa forma incluímos nesse estudo a parametrização “Brucelosis enfermedad”, com o objetivo de obter dados ligados à doença que é um fator relevante na produção argentina (veja Figura 34).

```
C:\Users\fcard\PycharmProjects\teseProject\venv\Scripts\python.exe C:\Users\fcard\PycharmPro
Digite a pesquisa que deseja realizar:Brucelosis enfermedad

Escolha o número para nível de detalhe (1: Região, 2: Estado, 3: Cidade, 4: Todas): 1
Digite o nome da região: ARGENTINA
```

Figura 34. Parametrização para processamento de dados argentinos Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Houve sucesso na coleta e estruturação de dados, além da criação dos Score dos dados, apontado a possibilidade uso do formato proposto que pode ser adaptado para diversas línguas. Foi utilizado como região geográfica de busca as seguintes regiões: Buenos Aires, Córdoba, Santa Fé, "Entre Ríos e La Pampa (Figura 35).



	estado	cidade	codPost	data	conteudo	score	
1	Argentina	Buenos Aires	169.299.264.034	11/07/2024	Las Zoonosis son enfermedades infecciosas	4	h
2	Argentina	Buenos Aires	484.026.563.172	04/12/2024	Nos reunimos con el vicepresidente de Sena	5	h
3	Argentina	Buenos Aires	724.468.349.045	04/12/2024	Nos reunimos con el vicepresidente de Sena	5	h
4	Argentina	Buenos Aires	882.101.813.530	16/12/2024	Hola amigos queridos, Hoy no son buenas l	1	h
5	Argentina	Buenos Aires	193.870.632.537	16/12/2024	Hola amigos queridos, Hoy no son buenas l	1	h
6	Argentina	Buenos Aires	851.977.842.607	22/03/2024	Campaña de concientización sobre la impo	1	h
7	Argentina	Buenos Aires	636.739.301.843	20/05/2024	COMPRAR: https://conradolibros.mercados	5	h
8	Argentina	Buenos Aires	842.812.270.458	01/04/2024	Campaña de concientización sobre la impo	3	h
9	Argentina	Buenos Aires	604.365.877.427	01/04/2024	Campaña de concientización sobre la impo	2	h

Figura 35. Resultado de processamento de dados argentinos. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

Como os dados geográficos utilizados no sistema, são baseados nas informações do IBGE, foi necessário cadastrar a Argentina como um estado brasileiro, para que fosse possível realizar a filtragem de região.

Quanto a utilização dos dados no Power BI, o comportamento foi o mesmo dos dados em português, o que permitiu a pesquisa e criação de dashboards interativos (vide Figura 36).

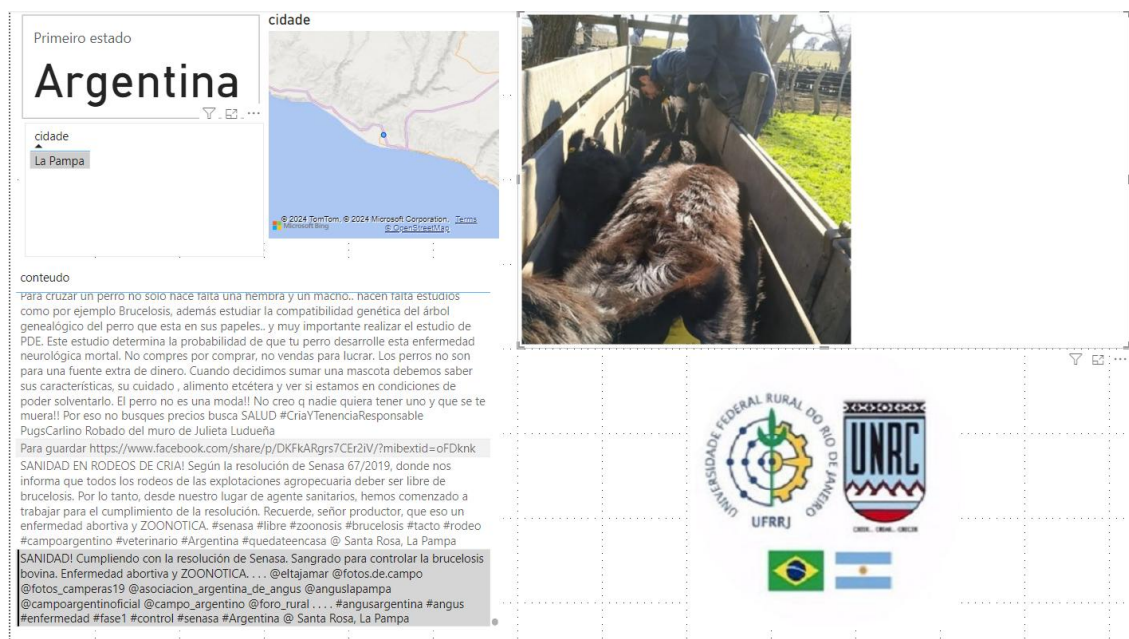


Figura 36. Dados argentinos submetidos ao Power BI. Fonte: Aplicação desenvolvida na pesquisa pelo Autor.

4. RESULTADOS E DISCUSSÃO

Tendo vista a utilização de IA's sendo um dos assuntos centrais do presente trabalho, a validação de dados foi realizada por meio da submissão de conteúdos, a dois principais motores de IA, ChatGPT da OpenIA e Gemini do Google (Figura 37).

Os dados foram submetidos utilizando o mesmo prompt, que retorna um inteiro de 1 até 5, onde 1 significa fraca aderência entre o texto e o assunto pesquisado, e 5 forte aderência entre o assunto pesquisado e o texto obtido pelo Bot.

Em uma escala de 1 até 5 (responder exclusivamente com numeral), onde 1 significa muito pouco e 5 significa de acordo, classifique o texto a seguir quanto a aderência sobre o assunto, na área de agropecuária:”, conteúdo

Figura 37. Prompt utilizado no ChatGPT e Gemini. Fonte: Autor.

Para o cálculo do “Acerto” foi utilizado o modelo de F1-score adaptado (SOKOLOVA; LAPALME, 2009), onde foi considerada a seguinte equação:

$$\text{Acerto} = \frac{N_{\text{ChatGpt}} + N_{\text{Gemini}}}{10}$$

Caso 1:

Mudas cacau pronta entrega PS1319 e CCN51 clone resistente a vassoura de bruxa e mudas de ótima qualidade.

Nota ChatGPT = 3

Nota Gemini = 2

$$\text{Avaliação} = \frac{3+2}{10} = 0,5.$$

O resultado neste caso 50% de chance de que o dado seja de interesse de pesquisa, ou seja deve ser observado pelo pesquisador.

Caso 2:

Hoje é dia de falar sobre a brusone do trigo! Essa doença é causada por Magnaporthe oryzae, fungo que ataca mais de 100 espécies de gramíneas - dentre elas, arroz, cevada, triticale, centeio, azevém, milheto e milho. Seus sintomas podem ser observados em folhas, colmos, bainhas, sendo que na espiga é onde causa maior dano. Na ráquis da espiga infectada, identifica-se facilmente o ponto de penetração do fungo, caracterizado por uma lesão de coloração negra brilhante e de formato irregular. No ponto de penetração, ocorre o estrangulamento da ráquis, que impede o transporte de nutrientes para a parte superior da espiga, resultando na descoloração ou branqueamento da espiga, sendo esse o sintoma mais conhecido. A doença é mais favorecida por temperaturas entre 24 a 28 °C e molhamento foliar, que pode ser orvalho, de no mínimo 15 horas. As medidas de controle da doença ainda não

apresentaram o êxito almejado, pois sob condições favoráveis, a eficiência dos fungicidas registrados é reduzida e devido à alta variabilidade genética do fungo, poucos cultivares resistentes estão disponíveis.

$$Avaliação = \frac{4+5}{10} = 0,9.$$

Nos Casos 1 e 2, foram apresentados textos, obtidos pela aplicação, que tiveram diferentes avaliações por cada IA, porém apesar de valores diferentes ambas concordam com o enquadramento gradual de cada um dos textos, pois no caso 1 o GPT enquadra do texto como 3 e, o Gemini como 2, porém estão próximos dentro dessa escala. Ao observar o texto é possível compreender que sua extensão e profundidade são limitadas, o que leva à concordância do autor com a classificação.

No Caso 2, submetemos um texto mais completo e profundo, o que levou à elevação da classificação, ainda que com diferença entre as IA's, com o GPT atribuindo o valor 5 e o Gemini o valor 4. Pela métrica aplicada, o indicativo de 90% de chance de se tratar de um dado relevante para uma pesquisa científica/técnica.

Os casos citados acima, servem como norteadores para o leitor deste trabalho, pois o método foi aplicado a aproximadamente 1740 linhas (tuplas), com resultados dentro do esperado onde a diferença dos dados entre IA's se manteve em uma média de 20%, ou seja, um ponto dentro da escala proposta.

Um ponto a ser testado (possibilidade futura), é a utilização de banco de dados NoSQL, pois o método encontra-se no contexto de BigData, em caso de aplicação em larga escala. O que poderia ter seu desempenho comprometido pela utilização de banco de dados relacional, conforme a versão em que se encontra.

4.1 Análise dos Dados

Sobre a pertinência e qualidade dos dados, 100% dos posts possuem relação com o conjunto de tags pesquisadas, porém eles são de categorias diferentes, ou seja, tem-se posts de produtores rurais, de instituições de ensino, empresas de agrotóxicos, profissionais liberais, comerciais de novelas e pontos turísticos.

Essa situação demonstra a necessidade da implementação de função de categorização ou classificação de textos admitidos da busca, a fim de melhorar a precisão de acesso aos registros. Tendo em vista que uma das vantagens de se utilizar um sistema no formato proposto é a redução no tempo de acesso a dados de interesse de pesquisa.

4.2 Camada de Exibição de Dados

Por definição o Storytelling, é a técnica de organização e exibição de dados de forma que o público-alvo se sinta envolvido com o assunto, permitindo ao autor a condução da exposição dos dados por meio de narrativas que tragam significado ao seu público (SERRAT, 2017).

Fundamentado nas técnicas de Storytelling, este trabalho identificou que a orquestração entre IA, Bots e BI cria uma jornada narrativa única: dados brutos se transformam em insights, que por sua vez se convertem em histórias científicas cativantes. Esse processo emergente posiciona essas tecnologias como protagonistas de uma nova era da comunicação científica, onde cada resultado de pesquisa pode ser contado como uma história envolvente.

Observa-se no sistema desenvolvido, a oportunidade de separar a massa bruta de dados e jargões (dados estruturados, semiestruturado, JSON, xml, etc), o que na perspectiva da

divulgação científica, estas questões técnicas são dificultadores da exposição de resultados (CORTINA, 2020).

Dessa maneira, é possível separar a pesquisa, primeiramente em um arcabouço técnico, onde encontramos o workflow contendo a execução de parametrização do sistema, a execução em série dos Bot's de obtenção de dados, estruturação e filtragem, e a classificação dos dados por IA. Em seguida a exportação dos dados para uma aplicação de Power Bi, que trata exclusivamente da visualização dos dados e da filtragem de acordo com a sua classificação, origem geográfica.

Portanto, o pesquisador pode criar uma experiência narrativa onde o público se conecta primeiro com o 'porquê' e o 'o que significa' dos dados, deixando o 'como foi obtido' como um capítulo opcional da história. Essa abordagem de storytelling científico prioriza o engajamento e a compreensão, mantendo a transparência metodológica sempre ao alcance daqueles que desejam explorar os bastidores da descoberta.

5. CONCLUSÕES

Neste trabalho foram utilizadas técnicas de inteligência artificial para a obtenção de informações agropecuárias através de Bots. A partir do observado durante as etapas de desenvolvimento e validação do método proposto no presente trabalho, foi possível constatar que as atuais Inteligências Artificiais (IAs) gerativas possuem grande capacidade para automatizar o pré-processamento de dados. Esse potencial é particularmente evidente quando o critério de avaliação dos dados envolve a definição de uma métrica gradual de pertinência, que estabelece o grau de relevância de determinado conteúdo textual em relação ao tema pesquisado. Esse tipo de análise representa um avanço significativo, pois permite uma triagem inicial eficiente e baseada em parâmetros objetivos, reduzindo a necessidade de intervenção humana em fases iniciais do processo de análise.

Outro aspecto central analisado durante a pesquisa refere-se ao uso de dados provenientes de mídias sociais no contexto acadêmico e científico. Essas plataformas apresentam um imenso potencial como fonte de dados, especialmente devido ao volume massivo de informações geradas continuamente por usuários ao redor do mundo. Contudo, a utilização eficiente desse tipo de dado em estudos científicos apresenta desafios consideráveis, sobretudo relacionados à coleta, triagem e organização. Por métodos convencionais, como buscas manuais, avaliação qualitativa e catalogação individualizada, o trabalho exigiria um esforço hercúleo, tornando-se frequentemente inviável diante da magnitude de dados disponíveis. Nesse sentido, a integração de métodos automatizados, baseados em IA, para coleta e avaliação inicial desses dados mostra-se como uma solução altamente eficaz e escalável.

Adicionalmente, o uso de métodos automatizados apoiados por IA não apenas facilita o pré-processamento, mas também amplia as possibilidades de análise, uma vez que as ferramentas baseadas em aprendizado de máquina e processamento de linguagem natural conseguem identificar padrões, tendências e associações que, muitas vezes, escapam ao olhar humano. A aplicação dessas tecnologias não só otimiza o trabalho de pesquisadores, mas também proporciona insumos robustos para a tomada de decisão em políticas públicas, especialmente em contextos que demandam agilidade e precisão no processamento de grandes volumes de dados.

Dessa forma, pode-se afirmar que métodos automatizados de coleta e pré-processamento de dados, sustentados por Inteligência Artificial, representam ferramentas de enorme potencial para a pesquisa científica e para a formulação de estratégias e políticas públicas. A produtividade, a velocidade e a escalabilidade obtidas com essas abordagens permitem o aproveitamento pleno dos dados que circulam nas mídias sociais, transformando informações dispersas em conhecimento estruturado, relevante e acionável. Além disso, a aplicação de tais métodos contribui para um novo paradigma no uso de dados digitais, estabelecendo bases mais sólidas para a construção de estudos inovadores e políticas mais informadas e efetivas.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- AALST, W. M. P. van der; BASTEN, T. Inheritance of workflows: an approach to tackling problems related to change. **Journal of Software Maintenance and Evolution: Research and Practice**, v. 14, n. 1, p. 55-81, 2002. DOI: 10.1002/smr.241.
- ABBADE, E. B. Food price, losses and logistics affecting diet diversification and food security. **Revista de Gestão Social e Ambiental**, v. 14, n. 3, p. 57-74, 2002. DOI: 10.24857/rgsa.v14i3.2456.
- AHMAD, I.; BAKHT, H.; MOHAN, U. Cloud Computing – A Comprehensive Definition. **International Journal of Computer Science and Information Technologies**, v. 8, n. 1, p. 1-4, 2017.
- ALKAISSI, H.; MCFARLANE, S. I. Artificial hallucinations in ChatGPT: implications in scientific writing. **Cureus**, v. 15, n. 2, e35179, 2023. DOI: 10.7759/cureus.35179.
- ALMEIDA, L. A.; GOMES, R. C. Processo das políticas públicas: revisão de literatura, reflexões teóricas e apontamentos para futuras pesquisas. **Cadernos EBAPE.BR**, Rio de Janeiro, v. 16, n. 3, p. 444-455, jul./set. 2018. DOI: 10.1590/1679-395164550.
- ARORA, A. *et al.* Measuring social media influencer index-insights from Facebook, Twitter and Instagram. **Journal of Retailing and Consumer Services**, v. 49, p. 86-101, 2019. DOI: 10.1016/j.jretconser.2019.03.012.
- BERTE, D. Defining the IoT. **Proceedings of the International Conference on Business Excellence**, v. 12, n. 1, p. 118–128, 2018. DOI: 10.2478/picbe-2018-0012.
- CELIKYILMAZ, A.; CLARK, E.; GAO, J. **Evaluation of Text Generation: A Survey**. 2020. arXiv:2006.14799. Disponível em: <https://arxiv.org/abs/2006.14799>. Acesso em: 22 set. 2024.
- CODD, E. F. A relational model of data for large shared data banks. **Communications of the ACM**, v. 13, n. 6, p. 377-387, 1970. DOI: 10.1145/362384.362685.
- COLLOMB, P. **Une voie étroite pour la sécurité alimentaire d'ici à 2050**. Roma: FAO, 1999.
- CORTINA, A. **Textos de divulgação científica: análise de duas reportagens sobre agrotóxicos**. 2020. Trabalho de Conclusão de Curso (Graduação em Letras) – Universidade Federal de Santa Maria, Santa Maria, 2020.
- DNIT [@DNIToficial]. Alô, Ceará! O DNIT realiza a manutenção da BR-020/CE... **X**, 6 mar. 2023. Disponível em: <https://twitter.com/DNIToficial/status/1632833488438140930>. Acesso em: 28 jul. 2024.
- EMBRAPA. **Manejo de pragas e doenças no cultivo de arroz em terras altas**. [S. l.]: Embrapa, [s. d.]. Disponível em: <https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/arroz/producao/sistema-de-cultivo/arroz-de-terras-altas/manejo-de-pragas/manejo-de-doencas>. Acesso em: 15 abr. 2024.
- FELDENS, L. **O homem, a agricultura e a história**. 1. ed. Lajeado: Univates, 2018.

FERRANTE, G. S. *et al.* Evaluating YOLO architectures for detecting road killed endangered Brazilian animals. **Scientific Reports**, v. 14, n. 1, 2024. DOI: 10.1038/s41598-024-51699-y.

FONSECA, F.; SILVEIRA, F. F.; MASSRUHÁ, S. M. F. S. **Tecnologia da informação e comunicação e suas relações com a agricultura**. Brasília, DF: Embrapa, 2018. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1093083/tecnologia-da-informacao-e-comunicacao-e-suas-relacoes-com-a-agricultura>. Acesso em: 7 dez. 2024.

G1 GOIÁS. Internado em hospital, trabalhador rural conta que foi atacado por onça em fazenda de Goiás. **G1**, 7 mar. 2023. Disponível em: <https://g1.globo.com/go/goias/noticia/2023/03/07/internado-em-hospital-trabalhador-rural-counta-que-foi-atacado-por-onca-em-fazenda-de-goias.ghtml>. Acesso em: 28 jul. 2024.

HAYKIN, S. **Redes Neurais: princípios e prática**. 2. ed. Porto Alegre: Bookman, 2001.

JULIÃO, A. Sistema utiliza inteligência artificial para detectar animais selvagens na pista e evitar acidentes. **Agência Brasil**, 11 mar. 2024. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2024-03/sistema-utiliza-inteligencia-artificial-para-detectar-animais-selvagens-na-pista-e-evitar-acidentes>. Acesso em: 23 mar. 2024.

KEPIOS. **Digital 2023: Global Overview Report**. Datareportal, 26 jan. 2023. Produzido em parceria com We Are Social e Meltwater. Disponível em: <https://datareportal.com/reports/digital-2023-global-overview-report>. Acesso em: 28 jul. 2024.

KIFLI, G. *et al.* Disruption of information technology and farming empowerment strategies in Indonesia. **IOP Conference Series: Materials Science and Engineering**, v. 1098, n. 5, 052087, 2021. DOI: 10.1088/1757-899X/1098/5/052087.

KLING, R. Information systems in policy making: computer technology and organizational arrangements. **Telecommunications Policy**, v. 2, n. 1, p. 22-32, 1978. DOI: 10.1016/0308-5961(78)90048-X.

KOVAČIĆ, M. *et al.* Using Artificial Intelligence for Creating and Managing Organizational Knowledge. **Tehnički vjesnik**, v. 29, n. 4, p. 1413-1418, 2022. DOI: 10.17559/TV-20211222120653.

LAU, D.; MACIEL, J. L. N.; ZOLDAN, S. M. **Brusone em trigo: como identificar, monitorar e controlar**. Passo Fundo, RS: Embrapa Trigo, 2021. (Documentos Online, n. 220). Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1136355/brusone-em-trigo-como-identificar-monitorar-e-controlar>. Acesso em: 28 jul. 2024.

LOTFI, C.; SRINIVASAN, S.; ERTZ, M.; LATROUS, I. Web Scraping Techniques and Applications: A Literature Review. *In: PROCEEDINGS OF THE 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS, AND INFORMATION ENGINEERING (ICEIE 2021)*. [S. l.]: Francis Academic Press, 2021. p. 38-44. DOI: 10.52458/978-93-91842-08-6-38.

MAURO, A. D.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its essential features. **Library Review**, v. 65, n. 3, p. 122-135, 2016. DOI: 10.1108/LR-06-2015-0061.

MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943. DOI: 10.1007/BF02478259.

MILUSHEVA, S. *et al.* **Using Big Data and AI to Identify High Risk Traffic Accident Hotspots in Nairobi**. Washington, D.C.: World Bank, 2021. (Policy Research Working Paper, n. 9732).

SARKER, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. **SN Computer Science**, v. 2, n. 6, 420, 2021. DOI: 10.1007/s42979-021-00815-1.

SENGAR, S. S. *et al.* **Generative Artificial Intelligence: A Systematic Review and Applications**. 2024. arXiv:2405.11029. Disponível em: <https://arxiv.org/abs/2405.11029>. Acesso em: 28 jul. 2024.

SERRAT, O. Storytelling. *In*: SERRAT, O. **Knowledge Solutions**. Singapore: Springer, 2017. p. 91-95. DOI: 10.1007/978-981-10-0983-9_91.

VAHEDI, N.; ARVAND, F. Machine Learning Applications in Food Safety and Quality Control: A Systematic Review. **Food Control**, v. 119, 107431, 2021. DOI: 10.1016/j.foodcont.2020.107431.

WIEDERHOLD, G.; MCCARTHY, J. Artificial Intelligence and its Applications. **Communications of the ACM**, v. 35, n. 7, p. 77–87, 1992.

ZHANG, W.; ZHAO, M.; YANG, L. Advances in Smart Agriculture Using Artificial Intelligence: Trends and Future Directions. **Computers and Electronics in Agriculture**, v. 187, 106274, 2021. DOI: 10.1016/j.compag.2021.106274.

7. ANEXOS

7.1 Anexo 1 – Dados do Experimento Brusone (Fração dos Dados)

estado	cidade	codPost	data	conteudo	link
Rio Grande do Sul	Santa Maria	560.518.564.179	27/01/2023	Com a valorização do arroz, é importante não perder o timing para o manejo de doenças. As condições estão propícias para doenças foliares, principalmente brusone e mancha parda.	https://scontent.fsd37-1.fna.fbcdn.net/vh1.6435-9/1433957194_3643194225763293_221137213261688538_n.jpg?stp=dst-jpg_p280x280&nc_cat=102&ccb=1-7&nc_sid=dd63ad8&nc_ohc=pQy14ipYkLAAx8YgFp-8&nc_ht=scontent.fsd37-1.fna&oh=00_AiBFFTKGQAG1wHYhPLTGx9DSqcm86uVnCFVjU-U-NvkE-Rw&oe=655273D6
Rio Grande do Sul	Santa Maria	71.980.717.665	27/01/2023	Com a valorização do arroz, é importante não perder o timing para o manejo de doenças. As condições estão propícias para doenças foliares, principalmente brusone e mancha parda.	https://scontent.fsd37-1.fna.fbcdn.net/vh1.6435-9/143394286_3643194324293956_6565416214246229584_n.jpg?stp=dst-jpg_p280x280&nc_cat=105&ccb=1-7&nc_sid=dd63ad8&nc_ohc=QipKl3YIOEAX3Ysb2o8&nc_ht=scontent.fsd37-1.fna&oh=00_AiCH12lmn3yDuoayqQsipkoPsdMf7HMY1DnyU5Z8Vztsg&oe=65527778
Rio Grande do Sul	Santa Maria	328.327.977.381	27/01/2023	Com a valorização do arroz, é importante não perder o timing para o manejo de doenças. As condições estão propícias para doenças foliares, principalmente brusone e mancha parda.	https://scontent.fsd37-1.fna.fbcdn.net/vh1.6435-9/143297544_3643194379096617_1753719175429368611_n.jpg?stp=dst-jpg_s600x600&nc_cat=107&ccb=1-7&nc_sid=dd63ad8&nc_ohc=kmi21CT_HQsAX9FHVQJ&nc_ht=scontent.fsd37-1.fna&oh=00_AiAhLzU0hG3n-UekizAi3kITJsz4tCAcJGADuvstHUKQ&oe=65525CB7
Rio Grande do Sul	Santa Maria	347.634.007.534	27/01/2023	Com a valorização do arroz, é importante não perder o timing para o manejo de doenças. As condições estão propícias para doenças foliares, principalmente brusone e mancha parda.	https://scontent.fsd37-1.fna.fbcdn.net/vh1.6435-9/142584843_3643194485763273_3429411720225541004_n.jpg?stp=dst-jpg_s600x600&nc_cat=108&ccb=1-7&nc_sid=dd63ad8&nc_ohc=7LIQCCwhKJAX9_nBq&nc_ht=scontent.fsd37-1.fna&oh=00_AiA3peI5gTRfaDkD7FuvpiSyDjIGiZht6iXVoJQTXzI5w&oe=655284C2
Rio Grande do Sul	Santa Maria	523.882.370.716	03/00/2023	O Zynergy é a melhor escolha quando falamos no manejo da Brusone e Mancha Parda na cultura do arroz. Mantenha a sua plantação saudável e protegida para obter uma colheita ainda mais produtiva! #omevbrasil #plantacaodearroz #produtivadadenocampo #arroz #agrobrasil #tecnologiainocampo #arrozprodutor, a BRS CATIANA é uma excelente opção para o cultivo de arroz no Tocantins. Com ampla adaptação e alto teto produtivo, ela tem moderada resistência à Brusone nas folhas e nas panículas e mancha parda e de grãos. Feita em parceria com a Embrapa, essa cultivar pode transformar sua produtividade nessa safra. Conte com a qualidade da Unigigel Sementes para sua lavoura de arroz irrigado. Faça uma cotação no site: www.unigigelsementes.com.br .	https://scontent.fsd37-1.fna.fbcdn.net/vh139.30808-6/387833823_828626139265750_8274248743248038597_n.jpg?stp=dst-jpg_p526x296&nc_cat=109&ccb=1-7&nc_sid=5f2048&nc_ohc=syllAnGnjA2gAX8V329f&nc_ht=scontent.fsd37-1.fna&oh=00_AiAHEspvyGNW6raOvdshgigPrlQ2Tsovd2TBwWwXPmLIQ&oe=6530AD79
Rio Grande do Sul	Santa Maria	114.861.537.431	24/08/2023	Produtor, a BRS CATIANA é uma excelente opção para o cultivo de arroz no Tocantins. Com ampla adaptação e alto teto produtivo, ela tem moderada resistência à Brusone nas folhas e nas panículas e nas panículas e	https://scontent.fsd37-1.fna.fbcdn.net/vh139.30808-6/239648915_3998245290298722_8369345582952721011_n.jpg?stp=dst-jpg_p600x600&nc_cat=109&ccb=1-7&nc_sid=5f2048&nc_ohc=s3DjwPNE3h8AX-DdsJB&nc_ht=scontent.fsd37-1.fna&oh=00_AiBjAtcPv53C12NDU6JnNSClkF5aIBvKPN6X8qdHrFQ&oe=6530A34A
				Produtor, a BRS CATIANA é uma excelente opção para o cultivo de arroz no Tocantins. Com ampla adaptação e alto teto produtivo, ela tem moderada resistência à Brusone nas folhas e nas panículas e nas panículas e	https://scontent.fsd37-1.fna.fbcdn.net/vh139.30808-6/239877153_3998245246965393_2817408216342844158_n.png?stp=dst-png_p600x600&nc_cat=110&ccb=1-7

7.2 Anexo 2 – Código do Bot

```
import os
import sys
import requests
import time
import re
import json
from datetime import datetime, timedelta
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup

def login_facebook(driver):
    """Função para fazer o login no Facebook"""
    driver.get('https://www.facebook.com')
    time.sleep(2)

    # Lendo o email e senha do arquivo "secret.txt"
    with open("secret.txt", "r") as file:
        lines = file.readlines()
        email = lines[0].strip()
        senha = lines[1].strip()

    # Preenchendo o campo de email
    email_path = 'input[name="email"]'
    email_element = driver.find_element(By.CSS_SELECTOR, email_path)
    email_element.send_keys(email)

    time.sleep(2)

    # Preenchendo o campo de senha
    password_path = 'input[name="pass"]'
    password_element = driver.find_element(By.CSS_SELECTOR, password_path)
    password_element.send_keys(senha)

    time.sleep(2)

    # Clicando no botão de login
```

```

login_button_path = 'button[name="login"]'
login_button = driver.find_element(By.CSS_SELECTOR, login_button_path)
login_button.click()

time.sleep(4)

def search_facebook_posts(driver, search_term):
    """ Função para realizar a busca no Facebook por posts """
    query_params = {'q': search_term}
    search_url = 'https://www.facebook.com/search/posts/?='
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/94.0.4606.61 Safari/537.36'
    }

    time.sleep(3)

    response = requests.get(search_url, params=query_params, headers=headers)
    driver.get(response.url)

    time.sleep(2)
    mouse = ActionChains(driver)
    x_coord = 401
    y_coord = 88
    mouse.move_by_offset(x_coord, y_coord).click().perform()
    time.sleep(2)

def check_fim_resultados(driver, timeout=10):
    """ Função para verificar se chegou ao final dos resultados """
    try:
        WebDriverWait(driver, timeout).until(
            EC.presence_of_element_located((By.XPATH, "//span[@class='x193iq5w xeuugli
x13faqbe x1vvkbs x1xmvt09 x1lliihq x1s928wv xhkezso x1gmr53x x1cpjm7i x1fgarty
x1943h6x xudqn12 x3x7a5m x6prxxf xvq8zen xo1l8bm xi81zsa x2b8uid' and contains(text(),
'Fim dos resultados')]))
        )
        return True
    except:
        return True

def scroll_to_end(driver):

```

```

"""Função para rolar até o final da página e carregar todos os posts"""
soup = BeautifulSoup(driver.page_source, 'html.parser')
divs = soup.find_all("div", attrs={"data-ad-comet-preview": "message"})

if not divs or any("Não encontramos nenhum resultado" in div.get_text() for div in divs):
    print("A página está vazia ou contém o conteúdo de erro. Não há dados para coletar.")
    return

time.sleep(3)

while True:
    last_height = driver.execute_script("return document.body.scrollHeight")
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    time.sleep(7) # Aguarde o carregamento dos novos resultados
    new_height = driver.execute_script("return document.body.scrollHeight")
    if new_height == last_height and check_fim_resultados(driver):
        break

def coletar_dados(search_term, cidade, estado, driver):
    """Coletar dados da div"""

    # Click no botão "ver mais."
    ver_mais_buttons = driver.find_elements(By.XPATH, "//*[contains(text(), 'Ver mais')]")

    for button in ver_mais_buttons:
        try:
            driver.execute_script("arguments[0].click();", button)
            WebDriverWait(driver, 20).until(EC.staleness_of(button))
        except:
            print("Não foi possível clicar no botão 'Ver mais' usando JavaScript.")

    time.sleep(5)

    soup = BeautifulSoup(driver.page_source, 'html.parser')
    divs = soup.find_all("div", attrs={"data-ad-comet-preview": "message"})

    # Função para gerar um ID único em hexadecimal
    def generate_unique_id(length=10):
        return hex(int.from_bytes(os.urandom(length), byteorder='big'))[2:length + 2]

    # Função para verificar e garantir a unicidade do ID
    def generate_unique_id_unique(existing_ids, length=10):
        while True:

```

```

id_cont = int(generate_unique_id(length), 16)
if id_cont not in existing_ids:
    existing_ids.add(id_cont)
    return id_cont

hora_atual = datetime.now().strftime("%d/%m/%Y %H:%M")
tudo = None
id_cont = None
perfil_link = None
nome = None
imagem_src = ""
conteudo = ""
horario = ""

# Construir o caminho para o arquivo JSON
json_path = os.path.join(search_term, f'{search_term.replace(" ", "_")}.json')

# Carregar dados existentes se o arquivo JSON já existir
if os.path.exists(json_path):
    with open(json_path, 'r', encoding="utf-8") as f:
        dados = json.load(f)
else:
    dados = []

# Inicializar conjunto para IDs existentes
existing_ids = set()

for div in divs:
    tudo = div.parent.parent.parent.find_all("a", attrs={"role": "link", "tabindex": "0"})

    conteudo_post = div.find_all("div", attrs={"style": "text-align: start;"})

    conteudo = []

    for div_element in conteudo_post:
        conteudo_text = div_element.get_text(separator=" ")
        conteudo_text_cleaned = " ".join(conteudo_text.split()) # Remove quebras de linha e
        junta o texto

        # Aplica a modificação para remover aspas e colchetes
        conteudo_text_cleaned = re.sub(r'\\', "", conteudo_text_cleaned) # Modificação na
        expressão regular

        # Adiciona ao conteúdo

```

```

        conteudo.append(conteudo_text_cleaned.strip())

# Junta todo o conteúdo em uma única string
conteudo_total = ''.join(conteudo)

for t in tudo:
    if t.find("span"):
        q = t.find("span")
        if "min" in q.text:
            horario = q.text

    procurar = t.find("span", attrs={"aria-labelledby": True})
    liberado = False

    if procurar and not procurar.parent.parent.has_attr("target") and
    procurar.parent.parent.name == "a":
        liberado = True

    if liberado:
        lo = procurar.find("use")
        hora = soup.find("span", attrs={"id": procurar["aria-labelledby"]})

        if hora:
            horario = hora.text

        elif lo and soup.find("text", attrs={"id": lo["xlink:href"].replace("#", "")}):
            texto = soup.find("text", attrs={"id": lo["xlink:href"].replace("#", "")})
            horario = texto.text

    if t.find("strong"):
        perfil_link = t['href']
        nome = t.strong.span.text

# Restante do código

elif t.find("img", attrs={"referrerpolicy": "origin-when-cross-origin"}):
    imagem_src = t.find("img", attrs={"referrerpolicy": "origin-when-cross-
origin"})["src"]

# Gerar ID único
id_cont = generate_unique_id_unique(existing_ids)

nome_arquivo = f'{id_cont}.jpg'

```

```

pasta_destino = search_term # Salvar na pasta do termo de busca

# Verificar id_cont
id_exists = any(post["ID"] == id_cont for post in dados)

if not id_exists:
    dados.append({
        "ID": id_cont,
        "Data post": horario,
        "Cidade": cidade,
        "Estado": estado,
        "Hora extração": hora_atual,
        "Link perfil": perfil_link,
        "Nome perfil": nome,
        "Link imagem": imagem_src,
        "Conteúdo": conteudo_total
    })

# Função para baixar a imagem
def baixar_imagem(url, pasta_destino, nome_arquivo):
    response = requests.get(url)
    if response.status_code == 200:
        caminho_salvar = os.path.join(pasta_destino, nome_arquivo)
        with open(caminho_salvar, 'wb') as arquivo:
            arquivo.write(response.content)
        print(f'Imagem baixada e salva em {caminho_salvar}')
    else:
        print(f'Falha ao baixar a imagem. Status code: {response.status_code}')

baixar_imagem(imagem_src, pasta_destino, nome_arquivo)

elif t.find("text", attrs={"y": "13"}):
    texto = t.find("text", attrs={"id": "gid4"}).text
else:
    texto = None

with open(json_path, 'w+', encoding="utf-8") as f:
    json.dump(dados, f, indent=2, ensure_ascii=False)

print(f'Pesquisa efetuada com sucesso na pasta: {search_term}')

# Tratamento do Json
def converter_data_para_dd_mm_aaaa(data_original):

```

```

# Verificar se a expressão é "Agora Mesmo"
if data_original.lower() == "Agora mesmo":
    data_atual = datetime.now()
    return data_atual.strftime("%d/%m/%Y %H:%M")

# Expressão regular para capturar horas, minutos ou segundos
regex_tempo = r"(\d+)\s*(h|min|seg)"

match = re.match(regex_tempo, data_original, re.I) # Ignorar maiúsculas/minúsculas
if match:
    valor, unidade = match.groups()
    valor = int(valor)

    # Converter para segundos
    if unidade == "h":
        valor *= 3600 # 1 hora = 3600 segundos
    elif unidade == "min":
        valor *= 60 # 1 minuto = 60 segundos

    # Calcular a nova data baseada no valor em segundos
    data_atual = datetime.now()
    nova_data = data_atual + timedelta(seconds=valor)
    return nova_data.strftime("%d/%m/%Y %H:%M") # Formato desejado para a data

# Lógica para converter datas no formato "dd de mês" ou "dd d"
meses = {
    'janeiro': '01', 'fevereiro': '02', 'março': '03', 'abril': '04',
    'maio': '05', 'junho': '06', 'julho': '07', 'agosto': '08',
    'setembro': '09', 'outubro': '10', 'novembro': '11', 'dezembro': '12'
}

padroes = [r"(\d{1,2}) de (\w+)", r"(\d{1,2}) d"]

for padrao in padroes:
    match = re.match(padrao, data_original)
    if match:
        grupos = match.groups()
        if len(grupos) == 1:
            dia = grupos[0]
            mes = ""
        elif len(grupos) == 2:
            dia, mes = grupos
        else:
            return None

```

```

        mes_numero = meses.get(mes.lower(), '00')
        ano_atual = datetime.now().year
        return f"{dia.zfill(2)}/{mes_numero}/{ano_atual}"

    return None

def limpar_conteudo(conteudo):
    # Remove as sequências \"
    conteudo_limpo = conteudo.replace("\"", "")

    # Remove os padrões \"
    conteudo_limpo = re.sub(r'\"(.*)\"|\\', r'\\1', conteudo_limpo)

    return conteudo_limpo

def limpar_json(json_data):
    for data in json_data:
        if "Conteúdo" in data:
            data["Conteúdo"] = limpar_conteudo(data["Conteúdo"])
    return json_data

def atualizar_data_no_json(caminho_arquivo):
    with open(caminho_arquivo, "r", encoding="utf-8") as arquivo:
        doc_json = json.load(arquivo)

    doc_json = limpar_json(doc_json)

    for data in doc_json:
        # Converter a cidade e estado para minúsculas e capitalizar a primeira letra
        data["Cidade"] = data.get("Cidade", "").capitalize()
        data["Estado"] = data.get("Estado", "").capitalize()

        data_post = data.get("Data post", "")
        data_formatada = converter_data_para_dd_mm_aaaa(data_post)

        if data_formatada:
            data["Data post"] = data_formatada

        # Verificar se o link do perfil está vazio e substituí-lo pelo link do Facebook
        if data["Link perfil"] is None or data["Link perfil"] == "":

```

```

data["Link perfil"] = "https://www.facebook.com"

# Verificar se Data esta vazia e substituí-lo por data atual
if data["Data post"] is None or data["Data post"] == "":
    data["Data post"] = datetime.now().strftime("%d/%m/%Y")

with open(caminho_arquivo, "w", encoding="utf-8") as arquivo:
    json.dump(doc_json, arquivo, ensure_ascii=False, indent=4)

print("Datas, Cidades e Estados atualizados no arquivo JSON com sucesso.")

def search_on_face(driver, query, cidade, estado):
    """Função para fazer a busca por cidades e estados"""

    search_box = driver.find_element(By.CSS_SELECTOR, 'input[placeholder="Localização marcada"]')
    time.sleep(3)

    for char in query:
        search_box.send_keys(char)
        time.sleep(0.1)

    time.sleep(3)
    wait = WebDriverWait(driver, 20)

    # Localizar a lista de opções usando o seletor CSS
    lista_opcoes = wait.until(EC.visibility_of_element_located((By.CSS_SELECTOR, 'ul[role="listbox"]'))))

    # Localizar o primeiro elemento da lista de opções usando o seletor CSS
    primeira_opcao = lista_opcoes.find_element(By.CSS_SELECTOR, 'li[role="option"]')

    time.sleep(1)
    primeira_opcao.click()
    time.sleep(5)

    scroll_to_end(driver)
    coletar_dados(search_term, cidade, estado, driver)

    time.sleep(2)
    # Limpar o local
    element = driver.find_element("xpath", "///div[contains(@aria-label, 'Limpar filtro Localização marcada')]")

```

```
element.click()
```

```
time.sleep(9)
```

```
def busca_generica(cidade, estado):
```

```
    scroll_to_end(driver)
```

```
    coletar_dados(search_term, cidade, estado, driver)
```

```
    time.sleep(10)
```

```
def caso_1(driver):
```

```
    regioao_escolhida = input("Digite o nome da região: ").upper()
```

```
    if regioao_escolhida in cidades_brasil:
```

```
        for estado, cidades in cidades_brasil[regiao_escolhida].items():
```

```
            for cidade in cidades:
```

```
                print("-" * 40)
```

```
                print(f"Campo de pesquisa: {cidade}, {estado}")
```

```
                search_on_face(driver, f"{cidade}, {estado}", cidade, estado)
```

```
            cidade = cidade # Chamando para a busca generica
```

```
            estado = estado
```

```
    else:
```

```
        print("Região não encontrada.")
```

```
    busca_generica(cidade, estado)
```

```
def caso_2(driver):
```

```
    estado_escolhido = input("Digite o nome do estado: ").upper()
```

```
    cidade_encontrada = None
```

```
    for regioao, estados in cidades_brasil.items():
```

```
        for estado, cidades in estados.items():
```

```
            if estado.upper() == estado_escolhido:
```

```
                cidade_encontrada = cidades
```

```
                break
```

```
    if cidade_encontrada:
```

```
        print(f"Cidades do estado {estado_escolhido}: {' '.join(cidade_encontrada)}")
```

```
        for cidade in cidade_encontrada:
```

```
            search_on_face(driver, f"{cidade}, {estado_escolhido}", cidade, estado_escolhido)
```

```
            cidade = cidade # Para a busca generica
```

```

else:
    print("Estado não encontrado.")

estado = estado_escolhido

busca_generica(cidade, estado)

def caso_3(driver):
    cidade_escolhida = input("Digite o nome da cidade: ").upper()
    cidade_encontrada = None

    for regioao, estados in cidades_brasil.items():
        for estado, cidades in estados.items():
            if cidade_escolhida in cidades:
                cidade_encontrada = cidade_escolhida
                break

    if cidade_encontrada:
        print(f"Cidade encontrada: {cidade_encontrada}")
        search_on_face(driver, f"{cidade_encontrada}", cidade_encontrada, None)
        estado = estado
    else:
        print("Cidade não encontrada.")

    cidade = cidade_escolhida
    busca_generica(cidade, estado)

def caso_4(driver):
    for regioao, estados in cidades_brasil.items():
        for estado, cidades in estados.items():
            for cidade in cidades:
                print("-" * 40)
                print(f"Campo de pesquisa: {cidade}, {estado}")
                search_on_face(driver, f"{cidade}, {estado}", cidade, estado)

    busca_generica(cidade, estado)

def default():
    print("Nível inválido.")

```

```

def funcionamento(variavel, driver):
    switch_case = {
        1: caso_1,
        2: caso_2,
        3: caso_3,
        4: caso_4,
    }
    switch_case.get(variavel, default)(driver)

if __name__ == "__main__":

    with open('regioes.json', 'r', encoding="utf-8") as json_file:
        cidades_brasil = json.load(json_file)

    search_term = input("Digite a pesquisa que deseja realizar:")
    nivel = int(input("\nEscolha o número para nível de detalhe (1: Região, 2: Estado, 3:
Cidade, 4: Todas): "))

    if not os.path.exists(search_term):
        os.makedirs(search_term)

    driver = webdriver.Chrome()
    driver.maximize_window()

    try:
        inicio = time.time()
        login_facebook(driver)
        search_facebook_posts(driver, search_term)
        funcionamento(nivel, driver)

        # Tratar Json
        # Substituir espaços por underscores e adicionar a extensão .json
        data_json = f'{search_term.replace(' ', '_')}.json'

        # Criar o caminho completo para o arquivo JSON
        caminho_json = os.path.join(search_term, data_json)

        # Chamar a função para atualizar o JSON
        atualizar_data_no_json(caminho_json)
        fim = time.time()
        print("Tempo de execução: ")
        print(fim - inicio)

```

```
except Exception as e:
    print("Ocorreu um erro:", e)
```

7.3 Anexo 3 - Código para Uso das AI's

```
import openai
import google.generativeai as genai

# Defina sua chave de API
openai.api_key = "
genai.configure(api_key)

#Função criar texto Gemini
def classificatextoGemini(prompt):
    model = genai.GenerativeModel('gemini-1.5-flash')
    chat = model.start_chat(history=[])
    response = chat.send_message(prompt)
    return(response.text)

# Função para gerar texto Chat GPT
def classificatextoGpt(prompt):
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": prompt}
        ]
    )
    return response['choices'][0]['message']['content']

def retornavalorAIGemini(promptX, promptE):
    # Exemplo de uso
    respostaY = classificatextoGemini(promptX+promptE)
    #print(respostaY)
    return int((respostaY))

def retornavalorAIGpt(promptX, promptE):
    # Exemplo de uso
    respostaX = classificatextoGpt(promptX+promptE)
    #print(respostaX)
    return int((respostaX))

valorE = "Cana de açúcar produz combutivel."
```

```

print(retornavalorAIGpt("Por gentileza, em uma escala de 1 até 5 (responder exclusivamente com numeral), onde 1 significa muito pouco e 5 significa de acordo, classifique o texto a seguir quanto a aderencia sobre o assunto central na área de agropecuária:",valorE))
print(retornavalorAIGemini("Por gentileza, em uma escala de 1 até 5 (responder exclusivamente com numeral), onde 1 significa muito pouco e 5 significa de acordo, classifique o texto a seguir quanto a aderencia sobre o assunto central na área de agropecuária:",valorE))

```

7.4 Anexo 4 – Estrutura de Banco de Dados

```

import sqlite3 as conector

```

```

conexao = conector.connect("teseBD.db")
cursor = conexao.cursor()

```

```

comandos = {
    "comando1": """
        CREATE TABLE tblCategoria (
            codCategoria BIGINT NOT NULL,
            nome CHARACTER VARYING(120) NOT NULL,
            descricao CHARACTER VARYING(255),
            PRIMARY KEY (codCategoria)
        );
    """,

    "comando2": """
        CREATE TABLE tblPost (
            codPost BIGINT NOT NULL,
            dataPost DATE NOT NULL,
            dataExtract DATE NOT NULL,
            linkPerfil TEXT NOT NULL,
            nomePerfil TEXT NOT NULL,
            conteudo TEXT NOT NULL,
            xscore INT,
            yscore INT,
            acerto INT,
            PRIMARY KEY (codPost)
        );
    """,

    "comando3": """
        CREATE TABLE tblImg (
            codImg BIGINT NOT NULL,
            img BLOB,

```

```

        linkImg TEXT,
        PRIMARY KEY (codImg)
    );
    """,

```

```

"comando7": ""
CREATE TABLE pais (
    id BIGSERIAL NOT NULL,
    nome VARCHAR(60),
    nome_pt VARCHAR(60),
    sigla VARCHAR(2),
    bacen INTEGER,
    PRIMARY KEY (id)
);
""",

```

```

"comando8": ""
CREATE TABLE estado (
    id_Estado BIGSERIAL NOT NULL,
    nome VARCHAR(60),
    uf VARCHAR(2),
    ibge INTEGER,
    pais INTEGER,
    ddd JSON,
    PRIMARY KEY (id_Estado)
);
""",

```

```

"comando9": ""
CREATE TABLE cidade (
    id_Cidade BIGINT NOT NULL,
    nome CHARACTER VARYING(120),
    uf INTEGER,
    ibge INTEGER,
    lat_lon POINT,
    latitude DOUBLE PRECISION,
    longitude DOUBLE PRECISION,
    cod_tom SMALLINT DEFAULT 0,
    PRIMARY KEY (id_Cidade)
);
""",

```

```

"comando10": ""
CREATE TABLE "tblRelPostImg" ("codPostImg" INTEGER NOT NULL, "codPost"

```

```

BIGINT NOT NULL, "codImg" BIGINT NOT NULL, "id_Estado" BIGINT, "id_Cidade"
BIGINT, "codCategoria" BIGINT NOT NULL,
    PRIMARY KEY("codPostImg" AUTOINCREMENT),
    FOREIGN KEY("codCategoria") REFERENCES "tblCategoria"("codCategoria"),
    FOREIGN KEY("codPost") REFERENCES "tblPost"("codPost"),
    FOREIGN KEY("id_Cidade") REFERENCES "cidade"("id_Cidade"),
    FOREIGN KEY("id_Estado") REFERENCES "estado"("id_Estado"),
    FOREIGN KEY("codImg") REFERENCES "tblImg"("codImg")
);
"""
}

```

```

for comando_nome, comando_sql in comandos.items():
    cursor.execute(comando_sql)
    conexao.commit()

cursor.close()
conexao.close()

```

7.5 Anexo 5 – Estruturação de Dados e Aplicação de Score de Validação

```

import json
import os
import sqlite3 as conector
from retornavalorAIGpt import retornavalorAIGpt
from retornavalorAIGemini import retornavalorAIGemini

conexao = conector.connect("teseBD.db")
cursor = conexao.cursor()

os.chdir('./vassoura bruxa cacau doce')
def carregarJson(jNome):
    with open(jNome, encoding='utf-8') as meu_json:
        dic_bd = json.load(meu_json)
        return (dic_bd)

def prepImg(filename):
    # Convert digital data to binary format
    with open(filename, 'rb') as file:
        blobData = file.read()
    return blobData

def buscaIdEstado(nomeEstado):
    print(nomeEstado)

```

```

    comando1: str = f'SELECT id_Estado FROM estado WHERE (nome) = "{nomeEstado}"
COLLATE nocase '
    print("comando1", comando1)
    cursor.execute(comando1)
    #print("Resultado cursor.fetchone()", cursor.fetchone())
    #xResultEstado = cursor.fetchone()
    #print("xResultEstado", xResultEstado)
    #xAuxResultEstado = xResultEstado[0]
    #print(xAuxResultEstado, "id_Estado")
    xAuxResultEstado = cursor.fetchone()
    print(xAuxResultEstado[0])
    return xAuxResultEstado[0]

def buscaIdCidade(nomeCidade,idEstado):
    print("Buscacidade cidade, idestado", nomeCidade,idEstado)
    comando1: str = f'SELECT id_Cidade FROM cidade WHERE (nome) = "{nomeCidade}"
COLLATE nocase AND uf = "{idEstado}" '
    cursor.execute(comando1)
    #xResult2 = cursor.fetchone()
    #print(xResult2, "Resultad0 query cidade")
    #xResult2 = int(xResult2[0])
    #print(xResult2,"id_Cidade")
    xResult2 = cursor.fetchone()
    return xResult2[0]

def ecreveBD(dicData):
    for registro in dicData:
        ID = registro["ID"]
        xID = int(ID)
        dataPost = registro["Data post"]
        cidade = registro["Cidade"]
        estado = registro["Estado"]
        dataExtracao = registro["Hora extração"]
        linkPerfil = registro["Link perfil"]
        nomePerfil = registro["Nome perfil"]
        linkImagem = registro["Link imagem"]
        conteudo = registro["Conteúdo"]
        buscaIdEstado(estado)
        idXestado = buscaIdEstado(str(estado))
        idXcidade = buscaIdCidade(cidade,idXestado)
        xImg = preplmg(str(xID)+ ".jpg")
        xTupleIMG = ([xImg])
        print(conteudo)
        xScore = retornavalorAIGpt("Por gentileza, em uma escala de 1 até 5 (responder
exclusivamente com numeral), onde 1 significa muito pouco e 5 significa de acordo,

```

classifique o texto a seguir quanto a aderencia sobre o assunto, na área de agropecuária:",conteudo)

yScore = retornavalorAIGemini("Por gentileza, em uma escala de 1 até 5 (responder exclusivamente com numeral), onde 1 significa muito pouco e 5 significa de acordo, classifique o texto a seguir quanto a aderencia sobre o assunto, na área de agropecuária:",conteudo)

acerto = ((int(xScore) + int(yScore)) / 10)

print("Teste acerto")

print(acerto)

comando2: str = f'INSERT INTO tblPost

(codPost,dataPost,dataExtract,linkPerfil,nomePerfil,conteudo,xscore,yscore, acerto) VALUES

({xID}, "{dataPost}", "{dataExtracao}", "{linkPerfil}", "{nomePerfil}",

"{conteudo}","{xScore}","{yScore}","{acerto}");'

comando3= "INSERT INTO tblImg (codImg, linkImg, img) VALUES (?,?,?)"

comando4= "INSERT INTO tblRelPostImg (codPost, codImg, id_Estado, id_Cidade)

VALUES (?,?,?,?)"

xRelPostImg = (xID,xID, idXestado, idXcidade)

xAuxComando3 = (xID,linkImagem, xImg)

cursor.execute(comando2)

cursor.execute(comando3,xAuxComando3)

cursor.execute(comando4,xRelPostImg)

conexao.commit()

xjNome = input("Escreva seu nome do arquivo .json:")

ecrevedB(carregarJson(xjNome))

7.6 Anexo 6 – Principais Query's

SELECT p.conteudo, i.img from tblPost as p INNER JOIN tblImg as i ON P.codPost = I.codImg WHERE p.dataExtract = "17/10/2023 08:28" AND p.conteudo LIKE "%milho%" collate nocase;

SELECT P.dataPost, p.conteudo, i.img from tblPost as p INNER JOIN tblImg as i ON P.codPost = I.codImg WHERE p.dataExtract LIKE "24/10/2023" AND p.conteudo LIKE "%inpi%" collate nocase;

SELECT P.dataPost, p.conteudo, i.linkImg, c.nome from tblPost as p INNER JOIN tblImg as i ON P.codPost = I.codImg WHERE p.conteudo LIKE "%inpi%" collate nocase AND tblRelPostImg as r ON cidade as c c.nome = r.id_Cidade;

```
SELECT tblRelPostImg.codPostImg, tblPost.codPost , tblpost.dataPost, tblpost.conteudo,
tblimg.img from tblPost INNER JOIN tblImg ON tblpost.codPost = tblimg.codImg inner join
tblRelPostImg ON tblPost.codPost = tblRelPostImg.codPost WHERE tblpost.conteudo LIKE
"%arroz%" collate nocase;
```

```
SELECT estado.nome, cidade.nome, tblRelPostImg.codPostImg, tblPost.codPost,
tblPost.dataPost, tblPost.conteudo, tblImg.img
FROM tblPost
INNER JOIN tblImg ON tblPost.codPost = tblImg.codImg
INNER JOIN tblRelPostImg ON tblPost.codPost = tblRelPostImg.codPost
INNER JOIN estado ON tblRelPostImg.id_Estado = estado.id_Estado
INNER JOIN cidade ON tblRelPostImg.id_Cidade = cidade.id_Cidade
WHERE tblPost.conteudo LIKE "%arroz%" COLLATE NOCASE;
```

```
SELECT estado.nome as estado, cidade.nome as cidade, tblPost.codPost, tblPost.dataPost as
data, tblPost.conteudo, tblPost.score, tblImg.linkImg as link
FROM tblPost
INNER JOIN tblImg ON tblPost.codPost = tblImg.codImg
INNER JOIN tblRelPostImg ON tblPost.codPost = tblRelPostImg.codPost
INNER JOIN estado ON tblRelPostImg.id_Estado = estado.id_Estado
INNER JOIN cidade ON tblRelPostImg.id_Cidade = cidade.id_Cidade
WHERE tblPost.conteudo LIKE "%brusone%" COLLATE NOCASE;
```

```
SELECT estado.nome AS estado, cidade.nome AS cidade, tblPost.codPost, tblPost.dataPost
AS data, tblPost.conteudo, tblPost.Score, tblImg.Img AS link
FROM tblPost
INNER JOIN tblRelPostImg ON tblPost.codPost = tblRelPostImg.codPost
INNER JOIN estado ON tblRelPostImg.id_Estado = estado.id_Estado
INNER JOIN cidade ON tblRelPostImg.id_Cidade = cidade.id_Cidade
INNER JOIN tblImg ON tblPost.codPost = tblImg.codImg
WHERE tblPost.conteudo LIKE "%arroz%" AND tblPost.conteudo LIKE "%acidente%"
COLLATE NOCASE;
```