UNIVERSIDADE DO OESTE DE SANTA CATARINA

MATEUS EDUARDO COLLING EIDT

IDENTIFICAÇÃO DOS ASSUNTOS MAIS DISCUTIDOS NO X (TWITTER) SOBRE INTELIGÊNCIA ARTIFICIAL ATRAVÉS DA MODELAGEM DE TÓPICOS

CHAPECÓ, SC

2024

MATEUS EDUARDO COLLING EIDT

IDENTIFICAÇÃO DOS ASSUNTOS MAIS DISCUTIDOS NO X (TWITTER) SOBRE INTELIGÊNCIA ARTIFICIAL ATRAVÉS DA MODELAGEM DE TÓPICOS

Projeto de Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação, Área das Ciências Exatas e da Terra, da Universidade do Oeste de Santa Catarina – Unoesc Campus de Chapecó, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Orientador: Prof. Jacson Luiz Matte

Chapecó, SC

2024

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BTM – *Biterm Topic Model*

IDE – *Integrated Development Environment*

CRISP-DM – *Cross-Industry Standard Process for Data Mining*

LDA – *Latent Dirichlet Allocation*

PLSA – *Probabilistic Latent Semantic Analysis*

SUMÁRIO

1 INTRODUÇÃO

Com o avanço das redes sociais e a proliferação de textos breves na internet, há uma crescente demanda por modelos de tópicos capazes de analisar esse tipo de conteúdo e lidar com o desafio dos dados dispersos. Em resposta a essa necessidade, (YAN et al., 2013) apresentaram um novo modelo simplificado, denominado *Biterm Topic Model* (BTM), projetado especificamente para a modelagem de textos curtos.

Segundo (YAN et al., 2013), o *Biterm Topic Model* (BTM) aborda a modelagem destacando a co-ocorrência de palavras comuns, o que aprimora a compreensão dos tópicos e resolve a questão das palavras esparsas nos documentos. Dessa forma, o processo de análise de documentos contendo textos breves se torna mais eficiente e preciso com a aplicação desse modelo.

Nos últimos anos, houve um aumento significativo no interesse por temas relacionados à inteligência artificial (IA). Avanços tecnológicos contínuos, juntamente com a integração cada vez maior da IA em várias esferas da vida cotidiana e dos negócios, têm impulsionado esse crescente interesse. Desde o desenvolvimento de algoritmos de aprendizado de máquina até a automação de processos industriais e a criação de assistentes virtuais inteligentes, a inteligência artificial tem demonstrado um potencial transformador em uma variedade de campos.

Assim, a descoberta de tópicos em textos relacionados à inteligência artificial é fundamental para diversas tarefas de análise de conteúdo nesse campo em constante evolução. Com base nesse pressuposto, este estudo propõe a extração e análise de um conjunto de dados textuais com o intuito de explorar, agrupar e classificar os principais temas abordados nessa área. Para isso, será empregado o modelo BTM com o objetivo de identificar e analisar os principais tópicos discutidos durante o segundo semestre de 2023 na rede social X (twitter).

1.1 DELIMITAÇÃO DO TEMA DE PESQUISA

Este estudo busca realizar uma análise exploratória no dataset construído a partir da ferramenta de extração de dados fornecida através da plataforma da APIFY sobre assuntos relacionados inteligência artificial. Os dados coletados estão publicamente acessíveis na rede social do X. Serão apenas tratados dados coletados no período do segundo semestre de 2023, com o intuito de classificar os tópicos mais frequentes no conjunto de dados, serão utilizadas palavras chaves como *ChatGPT, Inteligência Artificial, OpenAI, Gemini, Machine Learning, MidJourney, Bing, Copilot, Chat Bot, ia e Rede neural* para montar o dataset.

Diante do exposto, a metodologia de análise que melhor se encaixa para os documentos textuais coletados é a BTM, que conforme (YAN et al., 2013), busca aprimorar a aprendizagem de tópicos superando a dificuldade das palavras esparsas nos documentos maiores, resultando em uma análise mais eficaz, especialmente em textos curtos.

Os códigos utilizados para o pré-processamento do dataset, e para a modelagem de tópicos, serão desenvolvidos na linguagem Python, usufruindo dos recursos da *Integrated Development Environment* (IDE) do Visual Studio Code, e com o auxílio de bibliotecas como *pandas, numpy, sklearn, bitermplus, entre outras...* Essas mesmas bibliotecas se destacam por sua eficiência em manipular dados e apresentar resultados em forma gráfica e quantitativa.

1.2 PROBLEMA DE PESQUISA E JUSTIFICATIVA

As mídias sociais representam um canal fundamental para a interação global, permitindo a rápida disseminação de informações e opiniões, transmitindo informações até o usuário de forma muito prática e rápida (LEE et al., 2011). Plataformas como o X, por exemplo, registram diariamente milhões de posts, e que sem uma visão analítica se tornam dados desperdiçados, abrangendo uma ampla gama de tópicos, desde expressões de opinião até discussões sobre tecnologias e outras áreas de conhecimento.

De acordo com (FUKUYAMA; WAKABAYASHI, 2018), a extração de tópicos em microblogs emerge como uma abordagem promissora para identificar os temas populares em escala global. Através dessa análise, é possível compreender padrões comportamentais e tendências ao longo do tempo. Neste contexto, a presente pesquisa propõe-se a realizar uma análise detalhada dos dados da plataforma X, focando na extração e identificação dos principais tópicos discutidos, que serão obtidos e classificados como resultados desse estudo. O objetivo é preparar os dados, removendo elementos irrelevantes, e aplicar técnicas de análise de texto para identificar, delimitar e classificar os tópicos mais relevantes e seu comportamento ao longo do tempo.

Ao final deste estudo, espera-se obter resultados significativos sobre os temas mais debatidos na plataforma, contribuindo assim para uma melhor compreensão da dinâmica das discussões online e seu potencial impacto em áreas como comunicação, marketing e tomada de decisões.

1.3 OBJETIVOS

Nesta seção serão discutidos os objetivos a serem alcançados ao longo do desenvolvimento deste trabalho.

1.3.1 Objetivo geral

O objetivo do trabalho é criar um modelo para analisar e identificar os principais tópicos discutidos no X durante um período de crescente interesse em inteligência artificial. A abordagem utilizada foi a modelagem de tópicos com foco em textos curtos, em um dataset construído com os dados extraídos da rede social.

1.3.2 Objetivos específicos

- Extrair os textos curtos do X utilizando a plataforma da *APIFY*, no período do segundo semestre de 2023.
- Montar e preparar o dataset coletado, removendo elementos indesejados dos textos.
- Aplicar a modelagem de tópicos utilizando a metodologia BTM.
- Treinar o modelo e definir a parametrização ideal para o processamento de dados.
- Identificar e classificar os principais tópicos obtidos como resultado.
- Avaliar a coerência dos tópicos e relevância dos resultados alcançados.

2 REVISÃO DA LITERATURA

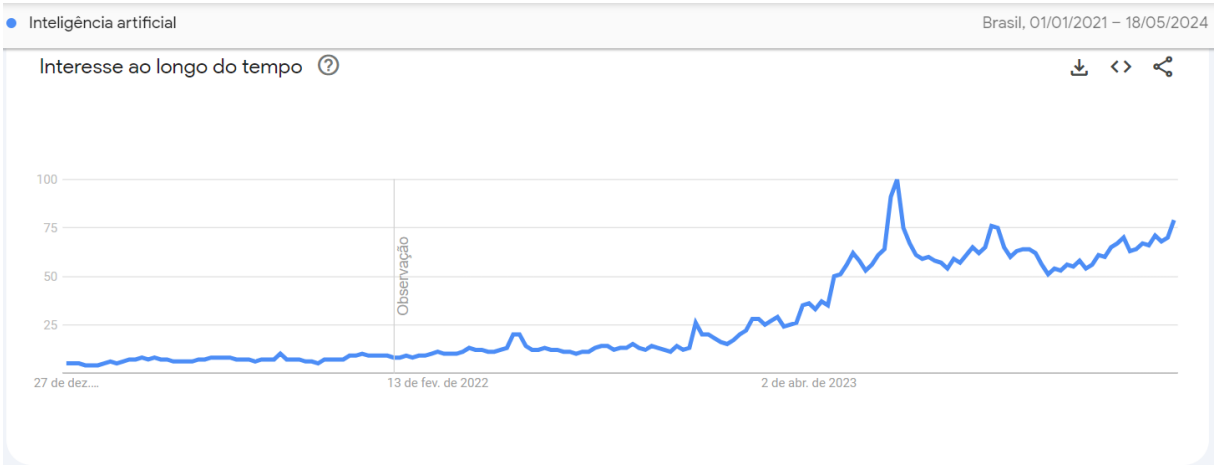
Nesta seção do trabalho, serão apresentadas as pesquisas teóricas que fundamentam este projeto, bem como as decisões relacionadas às tecnologias utilizadas no trabalho.

2.1 INTELIGÊNCIA ARTIFICIAL

Nos últimos anos, a inteligência artificial (IA) tem testemunhado um interesse crescente, impulsionado por avanços tecnológicos e pela sua aplicação em diversas áreas. A pesquisa de IA e robótica tem se acelerado significativamente, com um aumento exponencial no número de artigos publicados, refletindo o crescimento e a inovação na área (NATURE, 2022). Este fenômeno é impulsionado pela capacidade da IA de transformar setores como saúde, educação, transporte e muitos outros, tornando-se uma ferramenta indispensável para a análise de grandes volumes de dados e automação de tarefas complexas (CROMPTON; BURKE, 2023).

A Ilustração 1 abaixo mostra como a busca pelo termo *Inteligência Artificial* cresceu nos últimos anos:

Ilustração 1 – Busca pelo termo *Inteligência Artificial* no Google.



Fonte: O Autor.

O aumento do interesse na IA também se deve à sua capacidade de oferecer soluções inovadoras e eficientes em várias indústrias. No setor educacional, por exemplo, a IA está sendo utilizada para personalizar a aprendizagem, automatizar tarefas administrativas e desenvolver sistemas de tutoria inteligente. Estudos mostram que a pesquisa sobre a aplicação da IA na educação superior tem se expandido, refletindo um reconhecimento crescente de seu potencial

para revolucionar o ensino e a aprendizagem (CROMPTON; BURKE, 2023).

Além disso, a IA está cada vez mais presente no desenvolvimento de tecnologias emergentes, como veículos autônomos, assistentes virtuais e sistemas de recomendação, que estão remodelando a maneira como interagimos com a tecnologia no dia a dia. Este crescimento rápido é suportado por investimentos substanciais em pesquisa e desenvolvimento, tanto no setor público quanto no privado, e pela colaboração entre acadêmicos e a indústria (NATURE, 2022).

2.2 X (TWITTER)

O Twitter, atualmente renomeado como X, é uma rede social e plataforma de microblogging que permite aos usuários publicar e interagir com mensagens curtas chamadas ”tweets”. Desde sua criação em 2006 por Jack Dorsey, Noah Glass, Biz Stone e Evan Williams, a plataforma ganhou destaque por seu formato ágil de comunicação, limitando inicialmente os tweets a 140 caracteres, posteriormente ampliados para 280. A simplicidade e a velocidade de disseminação de informações tornaram o Twitter uma ferramenta essencial para notícias de última hora, debates públicos e campanhas de marketing (MURTHY, 2018).

Em 2023, o Twitter passou por uma rebranding significativo, mudando seu nome para X, sob a liderança de Elon Musk, que adquiriu a plataforma em outubro de 2022. A mudança de nome foi parte de uma estratégia mais ampla para transformar a rede social em uma ”super app”multifuncional, integrando serviços como pagamentos, compras e mais (SCRAGG, 2023).

A Ilustração 2 mostra um exemplo de um *tweet* retirado do dataset. O conteúdo começa com o nome do usuário em destaque, seguido pela imagem do perfil e o ID do usuário. Logo abaixo, aparece o texto, que pode ser acompanhado por uma imagem ou vídeo escolhidos pelo usuário. No final do texto, são incluídas hashtags que referenciam o tema do tweet e a data de publicação do texto.

Ilustração 2 – Exemplo de um *tweet* retirado do dataset.

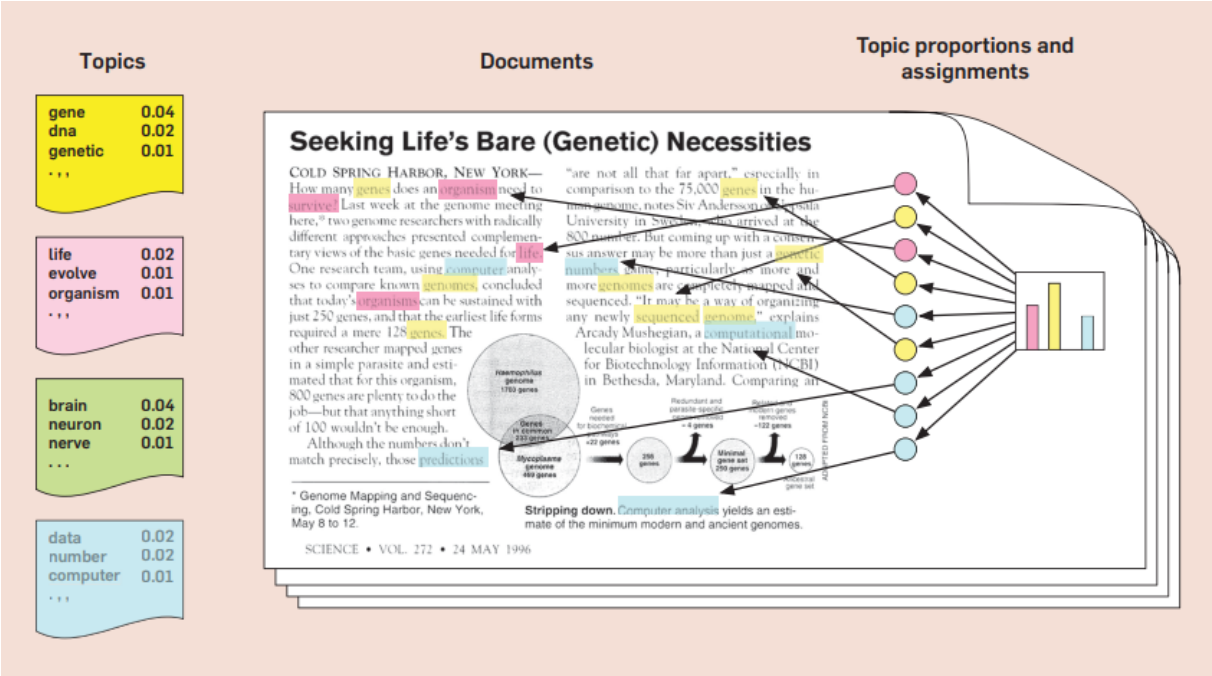
Fonte: O Autor.

Twitter, agora X, manteve uma base de usuários globalmente significativa, sendo especialmente popular para a disseminação de notícias e engajamento em tempo real. Em 2023, a plataforma contava com aproximadamente 368 milhões de usuários ativos mensais globalmente, com cerca de 18,97% do tráfego vindo dos Estados Unidos, seguido pelo Japão com 15,90% e a Índia com 4,80% (SINGH, 2024). Essa vasta quantidade de usuários e a natureza pública dos tweets fornecem uma riqueza de dados para análise, tornando a plataforma um recurso valioso para estudos de opinião pública, marketing digital. Esse intenso uso diário não apenas demonstra a popularidade e a relevância contínua da plataforma, mas também cria um fluxo constante de informações que pode ser analisado para entender tendências e comportamentos sociais.

2.3 MODELAGEM DE TÓPICOS

A modelagem de tópicos é uma técnica fundamental na análise de textos para identificar e extrair informações relevantes sobre os assuntos discutidos em um corpus de documentos. Essa abordagem visa descobrir padrões subjacentes nos dados textuais, agrupando-os em tópicos distintos que representam diferentes temas ou conceitos discutidos na coleção de documentos (BLEI; NG; JORDAN, 2003). Os modelos de tópicos são geralmente baseados em algoritmos de aprendizado não supervisionado, como o LDA, que identifica distribuições de palavras em tópicos latentes e a distribuição de tópicos em documentos individuais (BLEI; NG; JORDAN, 2003).

Ilustração 3 – Exemplo de uma aplicação de modelagem de tópicos.



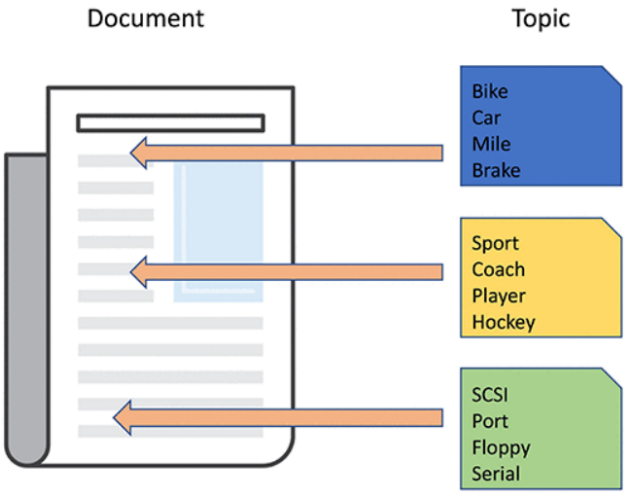
Fonte: (BLEI, 2012).

A Ilustração 3 acima exemplifica a modelagem de tópicos aplicada à um artigo intitulado como *Seeking Life's Bare Genetics Necessities*, que discute o uso da análise de dados para identificar quantos genes um organismo precisa para sobreviver. Conforme descrito por (BLEI, 2012) foram destacadas algumas palavras do artigo que se identificam dentro de um tópico específico, palavras em amarelo pertencem ao tópico de *genética*, em rosa *biologia evolucionária*, palavras em azul se encaixam no tópico de *análise de dados*. Os dados dessa figura são meramente ilustrativos e não representam dados reais.

2.3.1 Documentos

Em modelagem de tópicos, os documentos são unidades básicas de texto que são analisadas para identificar os temas subjacentes presentes em uma coleção de textos, conforme ilustra a Ilustração 4. Esses documentos podem variar em tamanho e conteúdo, desde pequenos trechos, como tweets e postagens em redes sociais, até textos longos, como artigos científicos e livros. A variação no tamanho dos documentos apresenta desafios e oportunidades diferentes para a modelagem de tópicos. Por exemplo, textos curtos, como os encontrados no Twitter, podem exigir técnicas especiais para capturar contextos e significados adequados, enquanto documentos longos podem fornecer uma riqueza de informações que facilita a identificação de tópicos (STEYVERS; GRIFFITHS, 2007).

Ilustração 4 – Documento textual.



Fonte: (PING et al., 2018).

Modelagens como PLSA e LDA são amplamente utilizadas para a descoberta de tópicos em documentos. O PLSA, introduzido por (HOFMANN, 1999), é uma técnica que modela cada documento como uma mistura de tópicos, onde cada tópico é uma distribuição probabilística sobre palavras. Já o LDA, aprimora essa abordagem ao introduzir distribuições de Dirichlet como uma forma de regularização, permitindo uma inferência mais robusta dos tópicos. Ambas as técnicas são eficazes, mas LDA é particularmente preferida por sua capacidade de lidar com grandes corpora e fornecer resultados mais estáveis em diferentes tamanhos de documentos (BLEI; NG; JORDAN, 2003).

2.3.2 Tópicos

Os tópicos podem ser definidos como a distribuição de palavras dentro de um documento textual, as palavras que definem um tópico seguem um padrão probabilístico de ocorrência, de acordo com (BLEI, 2012) um tópico é definido através da distribuição de um vocabulário fixo de palavras. A Ilustração 5 aponta 4 tópicos e as quinze palavras mais frequentes que os compõem, as palavras *humano* e *genoma* tem uma alta probabilidade de serem encontradas dentro do tópico *Genética*, assim como no tópico *Computadores* as palavras *computador* e *modelos* são as mais frequentes (BLEI, 2012). Elas estão ordenadas conforme sua probabilidade de aparecer no tópico.

Ilustração 5 – Exemplo de tópicos e das suas palavras mais frequentes.

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Fonte: (BLEI, 2012).

2.3.3 BTM

A modelagem de tópicos BTM é uma técnica avançada desenvolvida para lidar especificamente com textos curtos, como tweets, que são caracterizados pela limitação de contexto e pela informalidade da linguagem. A principal inovação do BTM é o uso de bi-termos (pares de palavras) ao invés de palavras individuais, o que permite capturar melhor as relações semânticas

e co-ocorrências frequentes de palavras em documentos curtos (YAN et al., 2013).

Para ilustrar o funcionamento do BTM, considere um exemplo de documento curto, como um tweet:

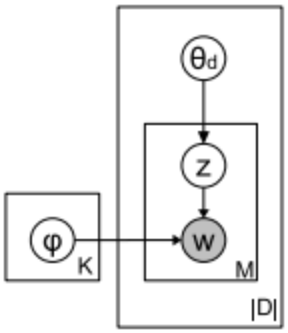
”Inteligência artificial está revolucionando a tecnologia”.

Na abordagem BTM, esse documento seria processado para identificar todos os pares possíveis de palavras (bi-termos). Os bi-termos extraídos desse tweet seriam:

(”Inteligência”, ”artificial”), (”Inteligência”, ”está”), (”Inteligência”, ”revolucionando”), (”Inteligência”, ”tecnologia”), (”artificial”, ”está”), (”artificial”, ”revolucionando”)...

Ao analisar esses bi-termos, o BTM pode identificar que certas palavras frequentemente ocorrem juntas e, portanto, são prováveis de pertencer ao mesmo tópico.

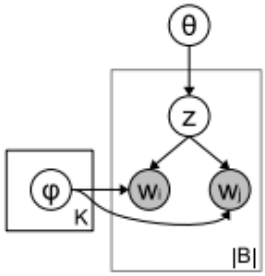
Ilustração 6 – Ilustração gráfica modelo LDA.



Fonte: (YAN et al., 2013).

No modelo LDA, mostrado na Ilustração 6, cada documento é gerado inicialmente desenhando uma distribuição de tópicos em nível de documento θ_d, e então iterativamente amostrando uma atribuição de tópico z para cada palavra w no documento. O LDA captura implicitamente os padrões de co-ocorrência de palavras em nível de documento, uma vez que a variável de atribuição de tópicos z de cada palavra depende de outras palavras no mesmo documento através do compartilhamento da mesma distribuição de tópicos θ_d. No entanto, quando os documentos são curtos, o LDA sofre do problema de escassez de dados devido à sua excessiva dependência das observações locais para a inferência da atribuição de tópicos ϕ.

Ilustração 7 – Ilustração gráfica modelo BTM.



Fonte: (YAN et al., 2013).

Por outro lado, o BTM, mostrado na Ilustração 7, supera o problema de escassez de dados do LDA ao desenhar a atribuição de tópicos *z* a partir da distribuição de tópicos em nível de corpus θ. Dessa forma, o BTM não apenas mantém a correlação entre palavras, mas também captura múltiplos gradientes de tópicos em um documento, já que as atribuições de tópicos de diferentes bitermos em um documento são independentes. Isso permite que o BTM modele de maneira mais eficaz os tópicos em textos curtos(YAN et al., 2013).

2.4 APIFY

A Apify fundada por (CURN; BALADA, 2015) é uma plataforma projetada para atender necessidades de *web-scrapping* e automação de dados em larga escala e alto desempenho. Ela oferece uma interface de fácil acesso para instâncias de mineração de dados (conhecidas como ”atores”), armazenamento conveniente de requisições e resultados, proxies, agendamento e webhooks, tudo acessível através da interface web Console, API do Apify, ou pelos clientes API em JavaScript e Python.

Os atores do Apify são componentes centrais da plataforma, cada um projetado para executar tarefas específicas de automação e *web-scrapping*. Eles podem ser usados para extrair dados de diversas fontes na internet, desde sites de e-commerce até redes sociais. Por exemplo, a Apify oferece atores prontos para uso que podem coletar informações do X, facilitando a extração de tweets, perfis e outras interações relevantes na plataforma. Estes atores permitem configurar e personalizar a coleta de dados conforme as necessidades específicas do usuário, ajudando a superar limitações comuns como a proteção anti-raspagem e a gestão de grandes volumes de dados.

Ilustração 8 – Ator para data mining disponibilizado pela plataforma.

Fonte: (CURN; BALADA, 2015).

A Ilustração 8 mostra a interface de um desses atores da plataforma, a monetização da plataforma adota um modelo de pagamento por resultado, cobrando $0,30 por cada 1.000 tweets extraídos. O ator em questão é desenvolvido e mantido pela própria comunidade. De acordo com o que representa a Ilustração 9 podemos personalizar alguns parâmetros para facilitar e objetificar melhor a atuação do ator para coletar os dados com mais precisão e volumetria. Os principais parâmetros, destacados em vermelho, são os termos de busca, a linguagem dos tweets e o período de busca.

Ilustração 9 – Parâmetros personalizáveis do ator.

```
{
    "customMapFunction": "(object) => { return {...object} }",
    "end": "2024-01-01",
    "includeSearchTerms": true,
    "maxItems": 10000,
    "maxTweetsPerQuery": 10000,
    "onlyImage": false,
    "onlyQuote": false,
    "onlyTwitterBlue": false,
    "onlyVerifiedUsers": false,
    "onlyVideo": false,
    "searchTerms": [
        "ChatGPT",
        "Inteligência Artificial",
        "OpenAI",
        "Gemini",
        "Machine Learning",
        "MidJourney",
        "Bing",
        "Copilot",
        "Chat Bot",
        "ia",
        "Rede neural"
    ],
    "start": "2023-12-31",
    "tweetLanguage": "pt"
}
```

Fonte: (CURN; BALADA, 2015).

2.5 TRABALHOS RELACIONADOS

Esta seção discute alguns trabalhos que são pertinentes a esta proposta. Foram escolhidos dois estudos específicos, (PEREIRA, 2019) e (PIERRE, 2022), que utilizam técnicas de extração de tópicos em documentos textuais e que também forneceram fundamentos importantes para esta pesquisa.

O trabalho de (PIERRE, 2022) tem seu foco em aplicar a modelagem de tópicos para assuntos relacionados ao Haiti no twitter. Utilizando-se da metodologia LDA, o trabalho faz uso de um dataset com os dados coletados no período de Janeiro de 2018 a Maio de 2022, foram extraídos tweets na linguagem portuguesa com o objetivo de classifica-los em políticos e não políticos. A Ilustração 10 abaixo mostra de forma visual as palavras mais frequentes nos textos, os gráficos em forma de nuvem mostram as principais palavras relacionadas à política no período eleitoral.

Ilustração 10 – Parâmetros personalizáveis do ator.

Fonte: (PIERRE, 2022). Legenda: tweets políticos (esquerda) e não políticos (direita).

No dataset utilizado por (PIERRE, 2022) obteve-se um total de 185.348 tweets e os mesmos passaram por um processo de pré-processamento, a Tabela 1 mostra o antes e depois do processo de limpeza dos dados. Após o pré-processamento, a quantidade de dados totalizou em 150.174.

Tabela 1 – Exemplo dados antes e depois do pré-processamento.

Antes do pré-processamento
FUGA DO BOZO: ”Presidente, o senhor é racista?”, pergunta repórter a Trump sobre polêmica envolvendo Haiti e países africanos ele saiu sem responder. https://t.co/AUprDvBNTO
Depois do pré-processamento
fuga bozo presidente senhor racista perguntar reporter trump sobre polemica envolver Haiti pais africano sair responder

Fonte: (PIERRE, 2022).

Para treinamento do modelo foram definidos aleatoriamente o número de tópicos, e os dados foram repartidos em 53 arquivos, sendo assim, um arquivo para cada mês durante o período especificado para a extração dos dados. Após a modelagem inicial foram identificadas várias palavras que não faziam sentido no dataset, logo foram descartadas na etapa do pré-processamento. Após a extração dos tópicos, foram calculados os tópicos mais frequentes de cada período, em seguida foi realizada a rotulação manual dos tópicos, de acordo com a interpretação das palavras contidas em cada tópico.

A Tabela 2 a seguir apresenta o tópico mais frequente para cada mês do ano de 2018, junto de suas 10 palavras mais frequentes e da rotulação manual do assunto. A ordenação das palavras segue de acordo com a sua probabilidade de aparecer dentro do tópico, sendo a primeira com maior probabilidade e a última com menor chance de aparecer. O mês de janeiro apresentou temas relacionados a *imigração* e *racismo*, enquanto o ultimo mês apresentou temas de *imigração* e *educação*.

Tabela 2 – Tópicos predominantes no ano de 2018.

Mês	Palavras	Assunto
jan	haiti, trump, brasil, eua, africa, pessoa, brasileiro, chamar, onu, terremoto	imigração e racismo
fev	exercito, militar, general, brasileiro, povo, rio_janeiro, trabalho, usar, coisa, sexual	intervenção militar
mar	haiti, brasil, rio, gente, exercito, brasileiro, militar, sirio, ajudar, venezuela	intervenção militar
abr	haiti, brasil, exercito, usar, militar, vida, missao_paz, terremoto, mandar, missao	intervenção militar
maio	contra, messi, selecao, jogo, grande, cuba, argentino, amistoso, fazer_gol, hat_trick	jogo de futebol
jun	haiti, brasil, brasileiro, selecao, povo, ficar, ajudar, futebol, pessoa, venezuela	jogo de futebol
jul	haiti, gente, brasileiro, jamaica, ficar, selecao, chorar, pessoa, mundo, pais	jogo de futebol
ago	haiti, brasil, militar, brasileiro, ficar, chegar, bolsonaro, missao, mundo, grande	intervenção militar
set	haiti, brasileiro, passar, pessoa, missao_paz, contra, governo, mundo, missao, militar	intervenção militar
out	haiti, povo, cuba, brasileiro, passar, militar, presidente, paz, conhecer, guerra	comparação econômica
nov	haiti, brasil, grande, ajudar, mundo, contra, eua, usar, coisa, deixar	imigração
dez	haiti, brasil, militar, brasileiro, imigrante, bolsonaro, haitiana_cega, governo, povo, aprovar_oab	imigração e educação

Fonte: (PIERRE, 2022).

O trabalho de (PEREIRA, 2019) se propõe a fazer uma análise exploratória de tweets coletados no período da olimpíadas de 2016. Ao contrário dos modelos mais convencionais de modelagem de tópicos (LDA, PLSA) que foram desenvolvidos com base em documentos textuais de grande porte, o foco dessa análise gira em torno da modelagem de tópicos BTM, que se encaixa melhor com textos curtos, como tweets e posts de redes sociais.
Alguns dos objetivos específicos do trabalho são destacados abaixo:

- Definir o período de datas para obtenção dos tweets.
- Identificar a melhor forma para obter os dados (tweets) de acordo com as métricas pré estabelecidas.
- Realizar o pré-processamento e eliminar stop words da coleção de textos obtidos.
- Comparar e classificar os resultados obtidos.

Para extração dos dados foi definido o período de 02 de agosto de 2016 a 24 de agosto de 2016, os dados foram extraídos através da ferramenta *Get Old Tweets*, como parâmetro para extração foram utilizadas as seguintes palavras chaves: ”rio2016”, ”olimpiadas”, ”olimpiada”, ”cerimoniadeabertura”, ”cerimoniadeencerramento”, ”jogosolimpicos”, entre outros…Foram coletados ao total 2.806.785 tweets, os dados foram separados em 9 arquivos diferentes cada um com um período de data, conforme mostra a Tabela 3.

Tabela 3 – Quantidade de tweets por arquivo.

Arquivo	Dias	Tamanho	Quantidade de Tweets
1	02.08.2016 a 04.08.2016	27,2MB	128.527
2	05.08.2016	65,9MB	334.125
3	06.08.2016 a 08.08.2016	127MB	624.444
4	09.08.2016 a 11.08.2016	51MB	239.912
5	12.08.2016 a 14.08.2016	52,1MB	245.149
6	15.08.2016 a 17.08.2016	64,7MB	311.581
7	18.08.2016 a 20.08.2016	59,2MB	279.432
8	21.08.2016	53,9MB	260.063
9	22.08.2016 a 24.08.2016	74MB	350.763

Fonte: (PEREIRA, 2019).

Para o pré-processamento dos dados, foram realizadas várias etapas, entre elas se destacam a remoção de URLs, menções e hashtags, além da remoção de stop-words, palavras inde-

sejadas e acentuação. Importante destacar que processos como *lemmatization* e *stemming*, que consistem em reduzir palavras flexionadas ao seu radical, não foram utilizados. Dando sequencia ao pré-processamento dos dados, foi executado um script para a aplicação da modelagem de tópicos, tendo como entrada os dados ja limpos.

A Tabela 4 a seguir representa 5 tópicos e suas palavras mais frequentes obtidas como resultado, os resultados são referentes ao arquivo 2 contido no dataset apresentado da Tabela 3. A coluna P(z) representa a probabilidade de ocorrência do tópico na coleção de documentos, no tópico 3, P(z) é igual a 0.094771, ou seja, o tópico tem a probabilidade de 9,44% de aparecer na coleção de documentos.

Tabela 4 – Tópicos obitos a partir do arquivo 2.

Tópico	**P(z)**	**Palavras**
1	0.159879	pokemon, vendo, hoje, cerimonia, mundo, casa, assistir, assistindo, estar, tv.
2	0.124873	pais, lindo, copa, mundo, brasileiro, povo, dinheiro, bonito, mal, festa.
3	0.094771	lindo, deus, gisele, demais, cerimonia, orgulho, maravilhosa, mundo, amo, parabens.

Fonte: (PEREIRA, 2019).

Para melhorar a visualização dos tópicos gerados, foi realizado um processo de rotulação manual dos dados. A definição dos rótulos envolveu a seleção da modalidade e/ou assunto que correspondia à maioria das palavras em cada tópico. Os tópicos que não se encaixaram em nenhuma categoria descrita nos sites foram rotulados de maneira empírica. Na Tabela 5 abaixo é possível visualizar a rotulação dos tópicos obtidos com base no arquivo 4 da da Tabela 3.

Tabela 5 – Tópicos rotulados obitos a partir do arquivo 4.

Rótulo	**Palavras**
Tópico 01: Vôlei	jogo, esporte, volei, melhor, futebol, mulher, assistir, galvao, globo, vendo.
Tópico 02: Transmissão	jogo, hoje, assistir, vendo, casa, assistindo, ficar, vamos, tv, tempo.
Tópico 03: Transmissão das Olimpíadas	atleta, brasileiro, esporte, medalha, pais, mundo, futebol, jogo, torcida, copa.
Tópico 04: Futebol	gol, selecao, neymar, jogo, jesus, galvao, gabriel, hoje, dinamarca, time.
Tópico 05: Futebol	masculino, feminino, jogo, selecao, futebol, volei, hoje, argentina, basquete, brasileira.

Fonte: (PEREIRA, 2019).

3 PROCEDIMENTOS METODOLÓGICOS E TÉCNICOS

Neste capítulo, serão apresentados os procedimentos metodológicos e técnicos adotados para a execução deste projeto. Esses procedimentos abrangem a delimitação do estudo, a caracterização da metodologia empregada, as questões de pesquisa, a aplicação dos métodos e as considerações sobre as limitações do estudo.

3.1 CARACTERIZAÇÃO DA METODOLOGIA DE PESQUISA

A presente pesquisa assume uma abordagem aplicada, alinhada à concepção de (KERLINGER, 2009), que a define como resolução de problemas específicos por meio da utilização de ferramentas e métodos práticos. Nesse contexto, a investigação visa identificar os temas mais discutidos no Twitter sobre inteligência artificial, empregando a modelagem de tópicos como principal instrumento.

A abordagem adotada neste estudo será quantitativa, pautada na avaliação da eficácia do modelo de modelagem de tópicos por meio de métricas objetivas. Entre essas métricas, destacam-se a coerência dos tópicos identificados e a perplexidade do modelo, fornecendo uma base sólida para avaliar a qualidade e a relevância dos resultados obtidos. Ao empregar essa metodologia quantitativa, busca-se não apenas descrever os temas discutidos, mas também mensurar sua consistência e representatividade.

Para embasar a presente pesquisa, será realizada uma extensa revisão bibliográfica, englobando tanto os fundamentos teóricos da inteligência artificial quanto as técnicas e abordagens utilizadas na modelagem de tópicos. A pesquisa bibliográfica será essencial para contextualizar o estudo dentro do panorama acadêmico, além de fornecer subsídios teóricos sólidos para a análise e interpretação dos resultados obtidos.

3.2 DELIMITAÇÃO DO ESTUDO

Este projeto delimita-se como sendo um estudo que fará a identificação dos tópicos e assuntos mais discutidos no X sobre inteligência artificial, servindo como uma ferramenta para buscar informações sobre termos relacionados a área, para posteriormente analisar e modelar os dados encontrados.

Tendo isso em vista, o projeto será realizado em Chapecó/SC ao longo do ano de 2024. Durante o primeiro semestre de 2024 será conduzido a pesquisa bibliográfica, visando aprofundar o conhecimento sobre as áreas de interesse para realização do projeto. Em seguida, no segundo semestre de 2024 será dedicado para a extração e processamento dos dados, assim como o desenvolvimento prático da modelagem de tópicos.

O projeto tem como responsável o acadêmico Mateus Eduardo Colling Eidt, sob orientação do professor Jacson Luiz Matte.

3.2.1 População e amostra

A população da pesquisa compreende todo o conjunto de dados que será utilizado no desenvolvimento da aplicação. Este conjunto de dados foi construído especificamente para esse estudo, contendo 1 arquivo para cada mês em questão, tendo seus dados obtidos por meio da ferramenta de extração fornecida pela *APIFY* (CURN; BALADA, 2015), resultando em uma média de 559 tweets por dia dentro de um mês, e totalizando 100.647 posts durante o período especificado em 2023, conforme indica a tabela abaixo.

Tabela 6 – Quantitativo de dados coletados por mês pela ferramenta da APIFY.

Mês	**Quantidade**
Julho	17.412
Agosto	15.967
Setembro	15.276
Outubro	17.051
Novembro	17.853
Dezembro	17.088
TOTAL	**100.647**

Fonte: O autor.

A seleção dos dados envolveu apenas tweets em português e que continham uma das palavras-chave especificadas para a pesquisa (*ChatGPT*, *Inteligência Artificial*, *OpenAI*, *Gemini*, *Machine Learning*, *MidJourney*, *Bing*, *Copilot*, *Chat Bot*, *ia*, *Rede neural*).

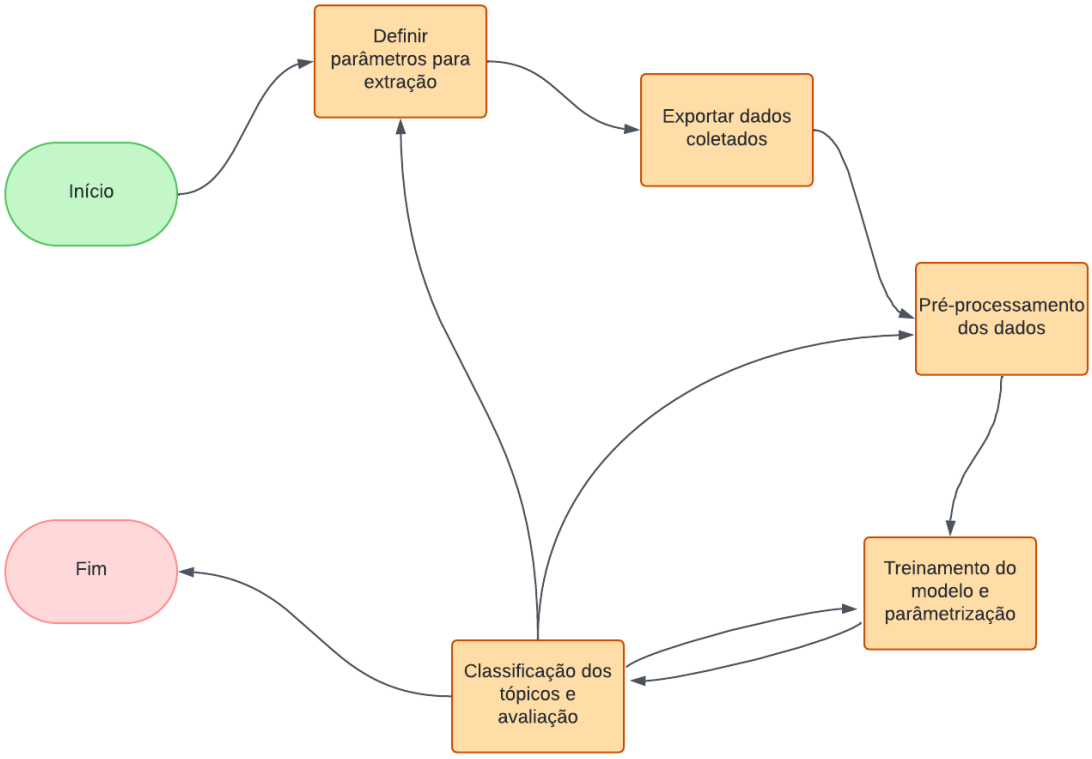
3.3 QUESTÕES DE PESQUISA

O trabalho proposto é capaz de coletar, processar, aplicar a modelagem de tópicos e apresentar resultados que facilitem o entendimento da categorização dos assuntos encontrados, utilizando-se dos dados que foram extraídos para construção do dataset?

3.4 ESBOÇO DO PROJETO NA PRÁTICA

Diante do contexto delineado, este projeto visa à identificação e classificação de tópicos discutidos no Twitter sobre inteligência artificial, através da metodologia de modelagem de tópicos *Biterm Topic Model* (BTM). Para alcançar este propósito, a Ilustração 11 abaixo descreve as etapas essenciais para alcançar os objetivos delineados.

Ilustração 11 – Fluxograma visão geral do projeto.



Fonte: O Autor.

Conforme apresentado na Ilustração 11, o projeto está organizado em 5 etapas fundamentais:

- Definir parâmetros para extração: Na fase inicial, foi estabelecido os critérios para a coleta

dos dados, como palavras-chave e o período de tempo desejado, garantindo que as informações obtidas sejam abrangentes o suficiente para o seu estudo.

- Exportar dados coletados: Após a definição dos parâmetros, a ferramenta APIFY é utilizada para coletar os dados do X conforme as especificações estabelecidas, trazendo as informações de forma organizada para a análise posterior.
- Pré-processamento dos dados: Nesta fase, os dados coletados passam por uma série de etapas de limpeza e preparação, incluindo a remoção de emojis e hastags, stopwords e a normalização do texto, visando deixar apenas termos relevantes nos documentos textuais, para garantir uma análise mais precisa.
- Treinamento do modelo e parametrização: Aqui, é aplicada a modelagem de tópicos com base no dataset construído, ajustando os parâmetros conforme necessário, como a quantidade de tópicos e iterações. As iterações durante o treinamento do modelo permitem refinar, melhorar a precisão e a dispersão dos documentos textuais.
- Classificação dos tópicos e avalização: Na etapa final, os tópicos identificados pelo modelo são classificados e avaliados quanto à sua perplexidade e coerência. Isso permite uma compreensão mais profunda dos assuntos discutidos sobre inteligência artificial.

  Nessa fase a identificação de pontos de melhoria é fundamental, pois o modelo de desenvolvimento do projeto permite retroceder a etapas anteriores, sendo possível adicionar novas etapas de limpeza de texto, adicionar novos termos na extração dos dados, ou alterar os parâmetros na modelagem dos posts.

3.5 APLICAÇÃO DA METODOLOGIA

Neste capítulo do projeto será abordada a construção do referencial teórico e as metodologias aplicadas a pesquisa prática.

3.5.1 Construção do referencial teórico

Para a construção do referencial teórico deste projeto, foi conduzida uma pesquisa a partir de fontes acadêmicas através do levantamento bibliográfico. Os termos utilizados foram selecionados de acordo com área de conhecimento específica da pesquisa, incluindo modelagem de tópicos, *Twitter*, extração e análise de dados textual. A pesquisa bibliográfica foi conduzida em três principais bases de dados: Google Acadêmico, Scopus e Scielo. Essas bases foram

escolhidas devido à sua abrangência e diversidade de fontes, incluindo livros, artigos, sites e documentos, que se apresentaram mais relevantes no escopo do estudo.

A Tabela 7 exibe o número total de artigos encontrados em cada uma das bases, organizados de acordo com o ano de publicação:

Tabela 7 – Quantitativo de resultados encontrados nas fontes acadêmicas.

Base	Termo pesquisado	2018	2019	2020	2021	2022	2023	Total
Google Acadêmico	Topic modeling	6.430	6.990	7.730	9.060	8.790	8.100	47.100
Google Acadêmico	Twitter	30.600	30.500	283.00	30.800	28.300	24.500	173.000
Google Acadêmico	Data extraction	14.300	15.100	14.500	16.500	15.600	24.500	100.500
Google Acadêmico	Textual data analysis	14.400	15.400	16000	18.400	18.500	18.300	101.000
Scopus	Topic modeling	7	6	10	6	12	13	54
Scopus	Twitter	13	23	19	32	23	25	189
Scopus	Data extraction	29	44	51	54	46	71	295
Scopus	Textual data analysis	19	24	31	20	27	19	140
Scielo	Topic modeling	7	9	4	4	15	10	49
Scielo	Twitter	8	11	17	12	13	17	78
Scielo	Data extraction	11	15	32	28	26	22	134
Scielo	Textual data analysis	21	14	27	29	27	17	135
TOTAL		**65.845**	**68.136**	**66.721**	**74.945**	**71.379**	**75.594**	**422.620**

Fonte: O autor.

É perceptível na Tabela 7 que a base acadêmica do Google foi a que mais contribuiu para os resultados. Outro fator importante ressaltar é que foram utilizados os termos em inglês, e que a busca ocorreu através dos títulos e palavras-chaves das publicações.

Os dados da Tabela 8 revelam que ao longo do período analisado, de 2018 a 2023, um total de 422.620 artigos foram encontrados. O ano de 2023 se destacou como aquele com o maior número de artigos relacionados aos termos pesquisados durante esse intervalo. Além disso, a base de dados do Google Acadêmico se destacou como a fonte primária utilizada para as buscas, contabilizando um total de 421.600 artigos.

Tabela 8 – Quantitativo de resultados encontrados por base.

Base	2018	2019	2020	2021	2022	2023	Total
Google Acadêmico	65.730	67.990	66.530	74.760	71.190	75.400	421.600
Scopus	68	97	111	112	108	128	624
Scielo	47	49	80	73	81	66	396
TOTAL	**65.845**	**68.136**	**66.721**	**74.945**	**71.379**	**75.594**	**422.620**

Fonte: O autor.

Na Tabela 9 abaixo, é demonstrada a quantidade de livros, artigos, entre outros documentos, utilizados para realizar a pesquisa de fato.

Tabela 9 – Quantitativo total de obras utilizadas.

Tipo	Quantidade
Livros	13
Artigos	3
Teses	2
Sites	3
TOTAL	**7**

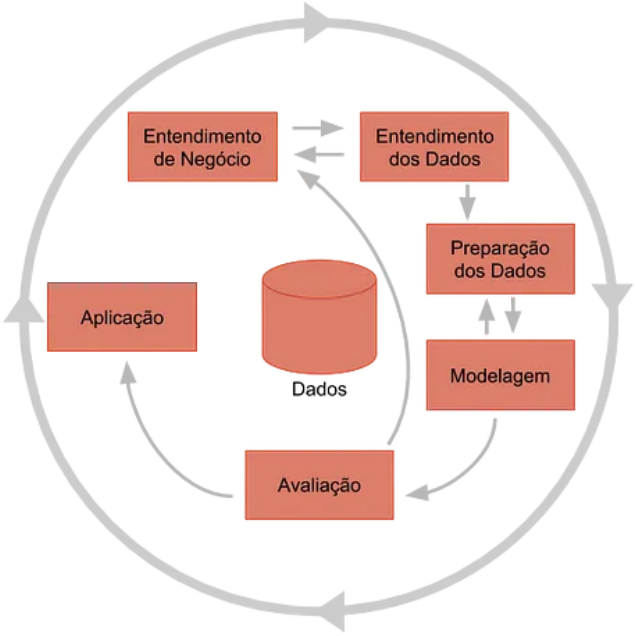
Fonte: O autor.

3.5.2 Desenvolvimento do sistema computacional

O modelo CRISP-DM é uma metodologia estabelecida e amplamente utilizada na indústria para orientar projetos de mineração de dados (CHAPMAN et al., 2000). Esta estrutura oferece uma abordagem sistemática e iterativa para o desenvolvimento de soluções analíticas, dividindo o processo em seis fases inter-relacionadas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação (CHAPMAN et al., 2000).

A Ilustração 12 demonstra de forma visual as seis fases do modelo CRISP-DM, destacando a sequência lógica e interdependência entre elas.

Ilustração 12 – Etapas desenvolvimento do projeto.



Fonte: Moura (2019).

3.5.2.1 Entendimento do negócio

A primeira etapa do modelo CRISP-DM é o ”Entendimento do Negócio”. Essa fase é crucial para definir os objetivos do projeto e entender o contexto em que os dados serão analisados. De acordo com (HAN; PEI; TONG, 2022), durante essa etapa, a pesquisa deve colaborar para identificar os requisitos do negócio, os objetivos do projeto e os critérios de sucesso. Além disso, é essencial realizar uma análise detalhada do domínio do problema e investigar as questões que precisam ser respondidas com a análise de dados (HAN; PEI; TONG, 2022). Essa compreensão inicial é fundamental para orientar todas as etapas subsequentes do processo de mineração de dados.

3.5.2.2 Entendimento dos dados

Na segunda etapa do modelo CCRISP-DM, denominada ”Compreensão dos dados”, o intuito é coletar e explorar os dados relevantes para o projeto. Isso envolve a obtenção de textos curtos do X que contenham discussões sobre inteligência artificial, bem como a realização de uma análise inicial para entender a estrutura e a qualidade desses dados.

Segundo (WITTEN et al., 2005), durante essa fase, é importante identificar quais variá-

veis e características dos dados são pertinentes para a análise de tópicos, bem como avaliar a integridade e a consistência dos dados coletados. Essa compreensão inicial dos dados é essencial para a fase de pré-processamento, pois a qualidade dos dados vai ditar o que deve ser removido e ajustado na etapa seguinte.

3.5.2.3 Preparação dos dados

A terceira etapa do modelo é a fase de ”Preparação de Dados”. Nesta etapa, os dados coletados anteriormente são preparados para a análise. Isso envolve atividades como limpeza de dados, integração de diferentes fontes de dados, seleção de variáveis relevantes e transformação dos dados em um formato adequado para a modelagem. Segundo (HAN; PEI; TONG, 2022), esta fase é crucial para garantir a qualidade e a consistência dos dados, pois dados sujos ou mal formatados podem levar a resultados imprecisos ou enviesados na análise de dados.

3.5.2.4 Modelagem

Nesta fase, a metodologia abordada no processo nos permite construir um modelo utilizando o dataset preparado, a fim de identificar tópicos subjacentes aos dados. Um exemplo notável dessa abordagem é a Modelagem BTM, que se destaca por sua capacidade de descobrir padrões sem a necessidade de rotulagem prévia dos dados. O BTM é uma técnica de aprendizado não supervisionado que extrai automaticamente tópicos em grandes conjuntos de documentos textuais, atribuindo palavras a esses tópicos com base em sua coocorrência em documentos.

Como mencionado por (HAN; PEI; TONG, 2022), essa etapa desempenha um papel crucial na garantia da qualidade dos dados utilizados para análise, pois a precisão e a integridade dos dados têm um impacto direto nos resultados finais dos modelos de mineração de dados.

3.5.2.5 Avaliação

Na quinta etapa do CRISP-DM, os modelos desenvolvidos são avaliados para determinar sua eficácia e adequação aos objetivos do projeto. Isso envolve a aplicação de métricas de desempenho e técnicas de validação cruzada para verificar a precisão e a generalização dos modelos. Conforme destacado por (WIRTH; HIPP, 2000), a avaliação é uma etapa crítica do processo de mineração de dados, pois é onde se verifica se os modelos desenvolvidos atendem aos critérios

de sucesso estabelecidos no início do projeto.

Diferente do modelo cascata, conhecido por seguir uma abordagem linear e sequencial, o CRISP-DM se destaca por sua flexibilidade. Essa flexibilidade é fundamental, pois permite corrigir erros, refinar as técnicas utilizadas e incorporar novas perspectivas de negócio que podem surgir durante o desenvolvimento do projeto. Utilizando a modelagem de tópicos BTM pode-se descobrir que certos parâmetros de configuração precisam ser ajustados para obter resultados mais precisos. Retroceder para a etapa de Preparação de Dados permitiria refinar o dataset, enquanto retroceder para a etapa de Modelagem permitiria ajustar os parâmetros do modelo. Essa capacidade de retroceder entre as etapas garante que o processo de mineração de dados seja iterativo e adaptável às necessidades e descobertas do projeto.

3.5.2.6 Implementação

A última fase do CRISP-DM, é a Avaliação. Tendo em vista que não é mais necessário retroceder etapas para realizar novos ajustes na limpeza dos dados por exemplo, os resultados do modelo são avaliados para determinar sua eficácia e relevância em relação aos objetivos de negócios estabelecidos na etapa inicial. Conforme (SHEARER, 2000) define, esse passo envolve uma análise detalhada da precisão do modelo, sua capacidade de generalização para novos dados e sua adequação às necessidades específicas do problema em questão. Além disso, considerações éticas e práticas são levadas em conta para garantir que as descobertas do modelo sejam interpretadas e utilizadas de maneira responsável e eficaz.

4 APRESENTAÇÃO DA SOLUÇÃO

Nesta seção, será exposta a solução de desenvolvimento do sistema computacional, considerando as informações abordadas no item 3.5.2, e seguindo o modelo previamente mencionado de mineração de dados CRISP-DM.

Logo abaixo, é apresentado o Tabela 10 com os requisitos funcionais do sistema computacional, seguido do Tabela 11 com os requisitos não funcionais do mesmo.

Tabela 10 – Requisitos funcionais do sistema computacional.

Número Requisito	Requisito Funcional
01	A coleta de dados deve ser parametrizável ao ponto de extrair tweets em um período de tempo específico, com base em palavras-chave diferentes e na linguagem especificada.
02	O sistema deve armazenar os tweets coletados em arquivos CSV para facilitar o acesso e análise.
03	O sistema deve limpar os tweets coletados, removendo stop words, URLs e caracteres especiais.
04	O sistema deve implementar o Biterm Topic Model (BTM) para a extração de tópicos de textos curtos.
05	O sistema deve permitir a configuração dos parâmetros do BTM, como o número de tópicos a serem gerados.
06	O sistema deve gerar uma lista de tópicos com suas respectivas palavras-chave.
07	O sistema deve apresentar uma visualização gráfica dos tópicos e suas distribuições.

Fonte: O Autor.

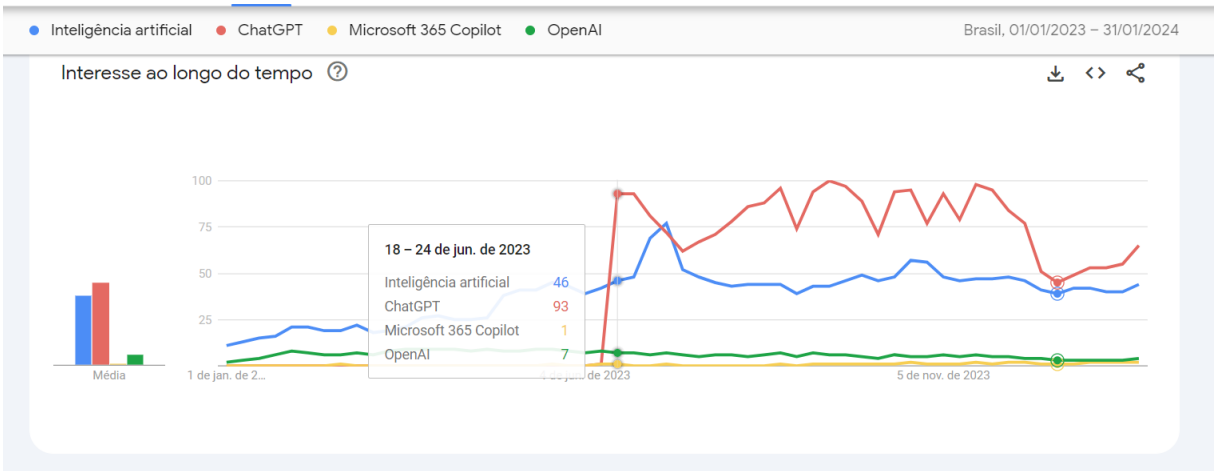
Tabela 11 – Requisitos não funcionais do sistema computacional.

Número Requisito	Requisito Não Funcional
01	O sistema computacional deve funcionar no Windows.
02	A coleta dos dados deve ser feita através da plataforma da APIFY.
03	Os dados devem ser armazenados em arquivos CSV.
04	Os dados devem ser armazenados em arquivos CSV.
05	O repositório do projeto deve estar no Github.

Fonte: O Autor.

No começo do estudo é necessário entender as motivações do projeto. O foco inicial era identificar qual período os assuntos estavam mais em alta para ter como foco na extração dos dados. A Ilustração 13 abaixo mostra que o interesse em inteligência artificial cresceu consideravelmente no período do segundo semestre de 2023, levando esse fato em consideração, podemos definir essa fatia de tempo como relevante para o projeto, pois é o período com o maior quantidade de informações circulando na rede social.

Ilustração 13 – Gráfico de interesse do Google Trends.



Fonte: O Autor.

Após definido o período alvo, na próxima fase é onde ocorre a extração dos dados, devido a questões de monetização da API oficial do X, seria inviável para o projeto realizar a coleta de dados dessa maneira. Tendo isso em vista, foi necessário buscar uma alternativa para buscar os dados, para facilitar esse processo foi utilizada a ferramenta da APIFY, que conforme a Ilustração 14, coleta os dados através de palavras-chave e outros parâmetros personalizáveis no JSON.

Ilustração 14 – Palavras-chave para extração dos dados.

```
{
    "customMapFunction": "(object) => { return {...object} }",
    "end": "2024-01-01",
    "includeSearchTerms": true,
    "maxItems": 10000,
    "maxTweetsPerQuery": 10000,
    "onlyImage": false,
    "onlyQuote": false,
    "onlyTwitterBlue": false,
    "onlyVerifiedUsers": false,
    "onlyVideo": false,
    "searchTerms": [
        "ChatGPT",
        "Inteligência Artificial",
        "OpenAI",
        "Gemini",
        "Machine Learning",
        "MidJourney",
        "Bing",
        "Copilot",
        "Chat Bot",
        "ia",
        "Rede neural"
    ],
    "start": "2023-12-31",
    "tweetLanguage": "pt"
}
```

Fonte: O Autor.

Outro objetivo nesse momento é classificar os tipos de dados que podemos encontrar na rede social, além de identificar os elementos textuais que compõe um post de dentro da plataforma. Isso é importante pois precisamos entender os elementos que deverão ser processados posteriormente. A Ilustração 17 em seguida exemplifica um documento textual e como ele é estruturado, tendo em destaque na cor azul a palavra-chave para obtenção da postagem. Esse exemplo foi retirado aleatoriamente do dataset, e contém elementos textuais como Emojis, *Hashtags*, e Menções respectivamente.

Ilustração 15 – Post exemplificando os elementos textuais.

Fonte: O Autor.

O diagrama abaixo representa a estrutura dos arquivos CSV exportados após a coleta de dados, nele temos várias informações úteis, de momento a informação mais importante se encontra na coluna *"text"*, onde contém o texto das postagens.

Ilustração 16 – Estrutura dos arquivos CSV.

Tweets	
url	VARCHAR
twitterUrl	VARHCAR
id	INT
text	VARCHAR
retweetCount	INT
replyCount	INT
likeCount	INT
quoteCount	INT
createdAt	TIMESTAMP
bookmarkCount	INT
isRetweet	BOOLEAN
isQuote	BOOLEAN

Fonte: O Autor.

Para o pré-processamento dos dados, foi desenvolvido um código em python que vai tratar o dataset para a modelagem de tópicos BTM. Essa etapa é especialmente importante devido à sensibilidade desses modelos quanto a qualidade e limpeza dos dados de entrada. O script realiza várias processos de limpeza e normalização do texto, como remoção de menções, links, hashtags, emojis, conversão para minúsculas, remoção de acentos, pontuações e stopwords.

Todos esses elementos foram identificados na etapa anterior, em seguida, duplicatas baseadas em texto são removidas e as linhas com texto vazio são eliminadas. Por fim, o DataFrame pré-processado é salvo em um novo arquivo CSV, a ilustração a seguir exemplifica um post antes e após o procedimento de limpeza.

Ilustração 17 – Comparação de texto antes e depois de processado.

Antes	Depois
Inteligência artificial consegue traduzir idioma mais antigo do mundo - Olhar Digital https://t.co/QbAKcZGDD9 #arqueologia #civilizações antigas #idiomas #Inteligência Artificial via @OlharDigital	inteligencia artificial consegue traduzir idioma antigo mundo olhar digital coes antigas encia artificial via
🤔 GitHub Copilot é confiável? Essa é uma dúvida muito comum nas empresas quando vamos começar a usar o GitHub Copilot no desenvolvimento dos projetos. Pensando nisso o GitHub Lançou ou GitHub Copilot Trust Center, vamos ver isso hoje https://t.co/WF7tcKoddi	github copilot confiavel duvida comum empresas vamos comecar usar github copilot desenvolvimento projetos pensando nisso github lancou github copilot trust center vamos hoje

Fonte: O Autor.

Após a aplicação da modelagem de tópicos espera-se obter resultados no formato da Tabela 12 (dados ilustrativos), nela temos a rotulação do tópico seguido de sua frequência no documento textual e das palavras mais frequentes. O processamento dos arquivos será particionado mensalmente, sendo assim, serão utilizados 6 arquivos que correspondem ao segundo semestre de 2023. Além de resultados como o do exemplo abaixo, pretende-se apresentar mapas de calor e outros visuais gráficos que auxiliem na interpretação dos resultados obtidos.

Tabela 12 – Formato esperado como resultado do processamento dos arquivos.

Tópico	Probabilidade	Palavras
01	0.96700000	palavra1, palavra2, palavra3, palavra4, palavra5
02	0.82500000	palavra1, palavra2, palavra3, palavra4, palavra5

Fonte: O Autor.

5 CRONOGRAMA

Nesta seção, será delineada a maneira pela qual o trabalho e o projeto de pesquisa serão conduzidos, levando em consideração o tempo disponível e o cumprimento das atividades programadas. O quadro a seguir mostra o cronograma utilizado para melhor organização das tarefas no primeiro semestre do Trabalho de Conclusão de Curso.

Tabela 13 – Cronograma do Projeto do Trabalho de Conclusão de Curso (Quinzenal).

Atividade		Fev		Mar		Abr		Mai		Jun	
Definir Orientador	P										
	R										
Definir Tema	P										
	R										
Delimitação do Estudo	P										
	R										
Elaborar Questões/Orçamento/ Problematização/Justificativa	P										
	R										
Elaborar Esboço/ Procedimentos Metodológicos	P										
	R										
Elaborar Solução/Cronograma	P										
	R										
Entregar Revisão Bibliográfica/ Apresentação Pré-banca	P										
	R										

Fonte: O autor. Legenda: P=Previsto, R=Realizado.

No próximo quadro, será exibido o cronograma da segunda parte do Trabalho de Conclusão de Curso. É importante lembrar que, a metodologia usada para execução dessas etapas do trabalho não segue uma sequência linear de passos, ou seja, ela permite voltar nas etapas conforme necessário.

Tabela 14 – Cronograma do Trabalho de Conclusão de Curso (Quinzenal).

Atividade		**Jul**		**Ago**		**Set**		**Out**		**Nov**	
Realizar correções apontadas pela banca	P										
Complementar referencial teórico	P										
Explorar diferente palavras-chave para extração dos dados	P										
Executar diferentes métodos e etapas para limpeza e processamento dos dados	P										
Desenvolvimento do modelo e definição dos parâmetros da modelagem	P										
Avaliação dos resultados e métricas obtidas	P										
Elaborar apresentação do projeto	P										
Entrega da monografia para banca	P										

Fonte: O autor. Legenda: P=Previsto.

6 ORÇAMENTO

O projeto será inteiramente desenvolvido pelo acadêmico, materiais como notebook ou softwares para o desenvolvimento da aplicação, já estão disponíveis ao acadêmico antes no início do projeto. O trabalho em questão não teve nenhum custo com hardware físico, outros custos no desenvolvimento do estudo podemos visualizar na Tabela 15 abaixo:

Tabela 15 – Orçamento do projeto.

Descrição	Valor
Cursos necessários sobre as áreas de conhecimento do projeto	R$ 110,00
Mensalidade Créditos da matéria TCC 1	R$ 125,38
Licença da ferramenta de extração de dados APIFY	R$ 255,92
TOTAL	**R$ 491,30**

Fonte: O autor.

REFERÊNCIAS

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, Jan, p. 993–1022, 2003.

CHAPMAN, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, n. 13, p. 1–73, 2000.

CROMPTON, Helen; BURKE, Diane. Artificial Intelligence in higher education: The state of the field. **International Journal of Educational Technology in Higher Education**, v. 20, n. 1, abr. 2023. DOI: 10.1186/s41239-023-00392-8.

CURN, Jan; BALADA, Jakub. **Apify**. Acesso em 9 de abril de 2024. 2015. Disponível em: <https://apify.com/about>.

FUKUYAMA, Satoshi; WAKABAYASHI, Kei. Extracting time series variation of topic popularity in microblogs. In: PROCEEDINGS of the 20th International Conference on Information Integration and Web-based Applications & Services. [S.l.: s.n.], 2018. P. 365–369.

HAN, Jiawei; PEI, Jian; TONG, Hanghang. **Data mining: concepts and techniques**. [S.l.]: Morgan kaufmann, 2022.

HOFMANN, Thomas. Probabilistic latent semantic indexing. In: PROCEEDINGS of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. [S.l.: s.n.], 1999. P. 50–57.

KERLINGER, Fred N. Metodologia da pesquisa em ciências sociais: um tratamento conceitual. In: METODOLOGIA da pesquisa em ciências sociais: um tratamento conceitual. [S.l.: s.n.], 2009. P. xv–378.

LEE, Kathy et al. Twitter trending topic classification. In: IEEE. 2011 IEEE 11th international conference on data mining workshops. [S.l.: s.n.], 2011. P. 251–258.

MOURA, Karina. **Ciclo de vida dos dados**. [S.l.: s.n.], jan. 2019. Disponível em: <https://medium.com/%5C@kvmoura/crisp-dm-79580b0d3ac4>.

MURTHY, Dhiraj. **Twitter**. [S.l.]: Polity Press Cambridge, 2018.

NATURE. **Nature**, v. 610, n. 7931, out. 2022. DOI: 10.1038/d41586-022-03210-9.

PEREIRA, Mariana. **Análise exploratória de tweets utilizando modelagem de tópicos para textos curtos: caso Olimpíadas Rio 2016**. 2019. Universidade Federal da Fronteira Sul.

PIERRE, Jod. **Utilização de modelagem de tópicos para identificar os assuntos mais discutidos sobre o Haiti no Twitter**. 2022. Universidade Federal da Fronteira Sul.

PING, David et al. **Introduction to the Amazon SageMaker Neural Topic Model**. [S.l.: s.n.], 2018. https://aws.amazon.com/pt/blogs/machine-learning/introduction-to-the-amazon-sagemaker-neural-topic-model/. Accessed: 2024-05-23.

SCRAGG, Gracie. **The Twitter to "X"case study: The ip implications of rebranding**. [S.l.: s.n.], set. 2023. Disponível em: <https://www.henryhughes.com/Site/news/the-twitter-to-x-case-study-the-ip-implications-of-rebrandig.aspx>.

SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

SINGH, Shubham. **Twitter User Statistics 2024 | DAU & MAU**. [S.l.: s.n.], abr. 2024. https://www.demandsage.com/twitter-user-statistics-2024.

STEYVERS, Mark; GRIFFITHS, Tom. Probabilistic topic models. In: HANDBOOK of latent semantic analysis. [S.l.]: Psychology Press, 2007. P. 439–460.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: MANCHESTER. PROCEEDINGS of the 4th international conference on the practical applications of knowledge discovery and data mining. [S.l.: s.n.], 2000. v. 1, p. 29–39.

WITTEN, Ian H et al. Practical machine learning tools and techniques. In: ELSEVIER AMSTERDAM, THE NETHERLANDS, 4. DATA mining. [S.l.: s.n.], 2005. v. 2, p. 403–413.

YAN, Xiaohui et al. A biterm topic model for short texts. In: PROCEEDINGS of the 22nd international conference on World Wide Web. [S.l.: s.n.], 2013. P. 1445–1456.