UNIVERSIDADE FEDERAL DE SANTA CATARINA

FERRAMENTA DE COMPARAÇÃO DE SENTIMENTOS ENTRE USUÁRIOS DO TWITTER E APLICAÇÃO A FIGURAS PUBLICAS

DENYS PARAGUAY GASPAR

Florianópolis– SC

2024/1

UNIVERSIDADE FEDERAL DE SANTA CATARINA

DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

CURSO DE SISTEMAS DE INFORMAÇÃO

FERRAMENTA DE COMPARAÇÃO DE SENTIMENTOS ENTRE USUÁRIOS DO TWITTER E APLICAÇÃO A FIGURAS PÚBLICAS

DENYS PARAGUAY GASPAR

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de informação.

Florianópolis – SC

2024/1

DENYS PARAGUAY GASPAR

FERRAMENTA DE COMPARAÇÃO DE SENTIMENTOS ENTRE USUÁRIOS DO TWITTER E APLICAÇÃO A FIGURAS PUBLICAS

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de informação.

Orientador:

_____

Prof. José Eduardo De Lucca

Banca Examinadora:

_____

Prof. Elder Rizzon Santos

_____

Prof. Jônata Tyska Carvalho

SUMÁRIO

LISTA DE FIGURAS

RESUMO

O Twitter se destaca como um canal crucial para a comunicação entre eleitores e políticos devido à sua capacidade de facilitar interações e discussões políticas com alto potencial de “viralização”. Utilizando técnicas de Processamento de Linguagem Natural (PLN), o estudo analisa os sentimentos expressos em tweets relacionados a um tópico específico ao longo do tempo.

O objetivo principal é desenvolver uma ferramenta que permita aos usuários comparar e verificar a evolução do sentimento expresso em tweets por meio de uma curva temporal. A metodologia incluiu a coleta de dados de tweets dos ex-presidentes Lula e Bolsonaro durante dois meses, seguida de análise de polaridade e lematização. Os resultados revelaram diferenças na atividade e no foco das comunicações dos dois políticos, com Lula sendo mais ativo e abordando frequentemente temas nacionais e governamentais e Bolsonaro buscando uma comunicação mais direta e redirecionando o público da rede social para os seus canais de comunicação. Observou-se também que a polaridade positiva nem sempre indica uma celebração do tema principal, mas pode estar relacionada a eventos ou figuras específicas.

Um dos desafios enfrentados foi a obtenção de dados devido ao acesso pago à API do Twitter. A análise evidenciou que tweets frequentemente gerenciados por equipes de marketing podem não refletir os pensamentos autênticos dos autores. A capacidade de interpretar corretamente a polaridade dos tweets é crucial para entender o impacto das redes sociais na política e na opinião pública.

1. INTRODUÇÃO

As mídias sociais têm desempenhado um papel cada vez mais significativo na formação do discurso político em todo o mundo. No Brasil, em particular, as mídias sociais têm uma ampla adesão, com aproximadamente 60% da população utilizando essas plataformas (Pacete, 2023). Dentre as várias redes sociais disponíveis, o Twitter tem se destacado como um canal de comunicação entre eleitores e políticos. O Twitter permite que os usuários compartilhem mensagens curtas de até 280 caracteres, chamadas de tweets, e interajam por meio de retweets e menções (Conover et al., 2011). A importância das mídias sociais, especialmente do Twitter, como um espaço para o debate político é evidente há alguns anos. Desde 2011, pesquisadores têm analisado a utilização do Twitter para fins políticos, levando em consideração sua capacidade de conectar pessoas e promover interações entre usuários (Lassen & Brown, 2011). Essas interações no Twitter possuem um alto potencial de viralização, o que, combinado com a formação de bolhas políticas, torna essa rede social um ambiente relevante para a atuação dos homens públicos (Recuero & Gruzd, 2019).

O processamento de linguagem natural (PLN) é um subcampo da inteligência artificial que se dedica à análise e representação computacional de idiomas humanos (Chowdhary, 2020). Essa abordagem envolve a utilização de algoritmos e métodos para processar, compreender e extrair informações de textos escritos em linguagem natural. No contexto deste trabalho, o PLN será aplicado para realizar a análise de sentimento dos tweets relacionados a um tópico selecionada pelos usuários. A ferramenta proposta neste trabalho é uma plataforma que permite aos usuários

acompanhar e analisar os sentimentos expressos pelos usuários de redes sociais, especificamente no Twitter, em relação a uma hashtag específica ao longo do tempo.

1.1. Objetivo geral

O objetivo geral deste projeto é a implementação de uma ferramenta que através do uso de processamento de linguagem natural (PLN) permita aos usuários realizarem a comparação do sentimento expresso por usuários do Twitter e verificarem sua alteração ao longo do tempo por meio de uma curva temporal.

1.2. Objetivos específicos

O trabalho a ser desenvolvido procurará atender os seguintes objetivos:

- Compreensão do ambiente acadêmico literário, técnico e prático de processamento de linguagem natural;
- Exercitar a implementação de algoritmos de análise de sentimentos;
- Fornecer insights sobre a utilização de mídias sociais, contribuindo para uma análise mais aprofundada do debate público e das interações entre os atores políticos e a opinião pública;
- Propor possíveis aplicações e usos da ferramenta, no auxílio à tomada de decisões políticas.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma análise da literatura e referências relevantes ao trabalho, fornecendo conceitos essenciais para uma compreensão do escopo do trabalho e permitir a modelagem da proposta.

2.1. Mídias sociais

As mídias sociais desempenham um papel importante na formação do discurso político em todo o mundo (Conover et al, 2011), e segundo Pacete (2023) publicado na revista Forbes “o levantamento da Comscore mostra que o país (Brasil) é o primeiro da América Latina em acesso às plataformas, o equivalente a 131,5 milhões de pessoas”. Este indicativo de ampla adesão das redes sociais no Brasil, aproximadamente 60% da população, reforça o impacto das mídias no discurso público. E atualmente dentre as redes sociais o Twitter possui o papel de conectar eleitores e políticos. Diferentemente de outras plataformas de comunicação no Twitter existe uma via de mão dupla entre os usuários, cada perfil é locutor e interlocutor e as mensagens possuem um alto potencial de viralização (Lassen; Brown. 2011) que agregado com a construção de bolhas políticas tornam a rede social o habitat dos homens públicos). Portanto, nota-se a relevância das mídias sociais e em especial o Twitter que vem sendo analisado desde 2011 frente sua ampla utilização para fins políticos.

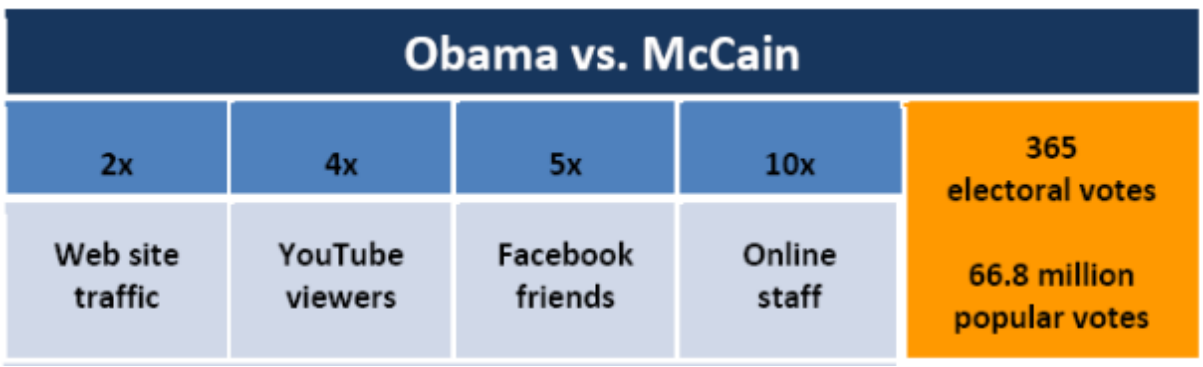
2.2. Utilização das mídias sociais em campanhas políticas

O impacto das mídias sociais na internet começou a ser notado nos Estados Unidos nas eleições de 2008, onde especialistas creditam a eleição do presidente

Barak Hussein Obama ao trabalho da sua equipe de marketing. Naquele momento, a rede social mais utilizada pelos americanos era o Facebook, com 21 milhões de contas criadas. Além da divulgação dos objetivos da campanha de Obama, foi realizada uma campanha de vaquinha online para arrecadação de dinheiro para a campanha.

O texto “Obama and the Power of Social Media and Technology”, escrito por Jennifer Aaker e Victoria Chang em 2009, analisa a estratégia de campanha de Barack Obama nas eleições presidenciais de 2008. Durante essa campanha, a equipe de Obama adotou uma abordagem inovadora, utilizando mídias sociais e tecnologia para arrecadar fundos e mobilizar voluntários.

De acordo com Aaker e Chang (2009), a campanha online de Obama foi fundamental para a eleição. Destacou-se que “Obama utilizou as mídias sociais de maneira eficaz para se conectar com os eleitores, mobilizar voluntários e arrecadar fundos”. Essa estratégia permitiu que a campanha alcançasse um amplo público e criasse um movimento de apoio. Em termos de seguidores, Obama superou seu concorrente, John McCain, nas redes sociais. Ele tinha milhões de seguidores no Twitter, Facebook e Youtube, enquanto McCain tinha uma presença online menos expressiva. Segundo os autores essa diferença de alcance e engajamento foi crucial para a vitória de Obama.



Fonte: Edelman Research, “The Social Pulpit,” Edelman, 2009, p. 2

Na imagem acima podemos perceber a diferença no alcance entre Obama e McCain nas diversas plataformas de mídia da época. Em números relativos o número de visualizações no Youtube de Obama passou 5x os números de McCain que não contava com uma equipe especializada nas novas mídias.

2.3. A rede social Twitter

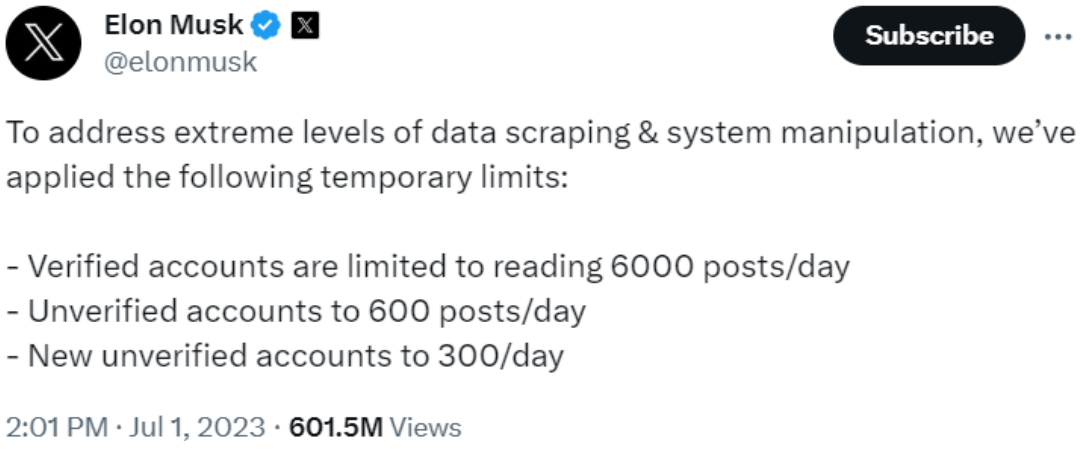
De acordo com Conover et al. (2011), o Twitter é uma plataforma de microblogging que permite aos usuários publicarem mensagens de até 280 caracteres, chamadas de tweets. Além de compartilhar tweets com seus seguidores, os usuários do Twitter podem interagir de duas maneiras principais: retweets e menções. Os retweets são utilizados como uma forma de endosso, permitindo que os usuários republiquem o conteúdo de outros usuários, aumentando sua visibilidade. Por outro lado, as menções possibilitam que os usuários se dirijam diretamente a outros usuários por meio do feed público. As hashtags também desempenham um papel importante no Twitter, permitindo que os usuários associem metadados aos tweets para especificar o tópico ou o público-alvo da comunicação Conover et al. (2011). Além das hashtags e retweets o Twitter também conta com a funcionalidade de “like” que atua como uma ferramenta de aceitação de conteúdo por parte dos usuários, uma validação da comunidade perante a informação, opinião ou manifestação, seja dos políticos ou das demais personalidades que estão presentes no cotidiano do Twitter.

Elon Musk, sul-africano de 52 anos, que por muito tempo era apenas mais um assíduo usuário do Twitter realizou a compra da ferramenta em abril de 2022, alegando ser defensor da liberdade de expressão e querer garantir que a rede social seria uma praça virtual para todos, além de reconhecer o potencial financeiro da rede social (Editorial g1, 2022). As principais mudanças para os usuários da ferramenta após a compra da rede social foram:

- Redução de funcionários: Musk demitiu cerca de 80% dos funcionários da empresa, alegando que a empresa estava superdimensionada (Editorial g1, 2023);
- Mudanças no algoritmo: Musk mudou o algoritmo do feed, tornando-o mais aleatório e menos baseado no que os usuários seguem.

Além das mudanças para o público geral da ferramenta também houveram alterações para os desenvolvedores que conectam seus sistemas a API do Twitter, o serviço de integração com os tweets que era gratuito desde sua disponibilização passou a ter um custo mensal individual de $100.00 (cem dólares) em 2 de fevereiro de 2023 (Figueiredo, 2023) e visando combater a utilização massiva de usuários robôs para extração de dados e por sua vez um pulo a barreira imposta pela cobrança de utilização do serviço o Twitter também passou a limitar o acesso aos tweets ao público geral, alegando que 600 tweets seriam o ideal para uma pessoa consumir no dia. Abaixo imagem do tweet publicado em 2023 pelo Elon Musk com algumas das alterações realizadas na rede social após sua compra.

Figura 1 – Tweet Elon Musk



Fonte: Musk, 2023

2.4. Raspagem de dados

De acordo com o autor Bo Zhao (2017), a raspagem de dados ou web scrapping é uma técnica amplamente reconhecida por sua eficiência e poder na coleta de dados da web. Essa prática, também conhecida como web scrapping ou coleta de dados, envolve a extração de informações da internet e seu armazenamento em um sistema de arquivos ou banco de dados para fins de análise ou recuperação. Essa técnica pode ser executada manualmente por um usuário ou automaticamente por meio de robôs ou web crawlers, e a grande quantidade de dados heterogêneos constantemente gerados na internet torna o web scrapping uma ferramenta essencial para a obtenção de informações.

O processo de coleta de dados da Internet envolve uma abordagem sequencial, dividida em duas etapas fundamentais (Zhao, 2017), solicitação HTTP, que pode ser chamada de aquisição dos dados, e o processamento dos dados adquiridos.

Neste trabalho, utilizou-se a raspagem de dados para leitura de dados do Twitter. Optou-se por essa abordagem devido à natureza pública dos dados e inviabilidade do pagamento da tarifa para consumo de dados da API, conforme descrito na seção 2.1.2.

2.5. Processamento de linguagem natural

Segundo Chowdhary (2020), o processamento de linguagem natural (PLN) é um subcampo da inteligência artificial, localizado nos estudos que engloba um conjunto de técnicas computacionais para análise automática e representação de idiomas humanos. A história da PLN data da década de 40, no período pós-Segunda Guerra Mundial, Weaver e Booth foram pioneiros em um dos primeiros projetos de Tradução Automática (TA) em 1946, tendo suas contribuições amplamente reconhecidas como inspiradoras para o campo (Liddy, 2001). O PLN envolve o uso de algoritmos e métodos para processar, compreender e extrair informações de textos escritos em linguagem natural. Para o autor, acesso e aquisição de características lexicais, semânticas e episódicas, identificação de construções básicas de linguagem (por exemplo, objetos e ações), representação de conceitos abstratos são alguns dos recursos obrigatórios para uma PLN de alto nível. Quanto à sua utilização, esta pode variar em diversos sentidos:

- Indexação e pesquisa de textos grandes;
- Recuperação de informações (IR);
- Classificação de texto em categorias;
- Extração de informações (IE);
- Tradução automática de idiomas;

- Resumo automático de textos;
- Resposta a perguntas (QA);
- Aquisição de conhecimento;
- Geração de textos/diálogos.

No contexto deste trabalho o PLN será utilizado para extração de informações através de técnicas de análises sintática e semânticas, conforme descrito a seguir.

2.5.1. Análises sintáticas e semânticas

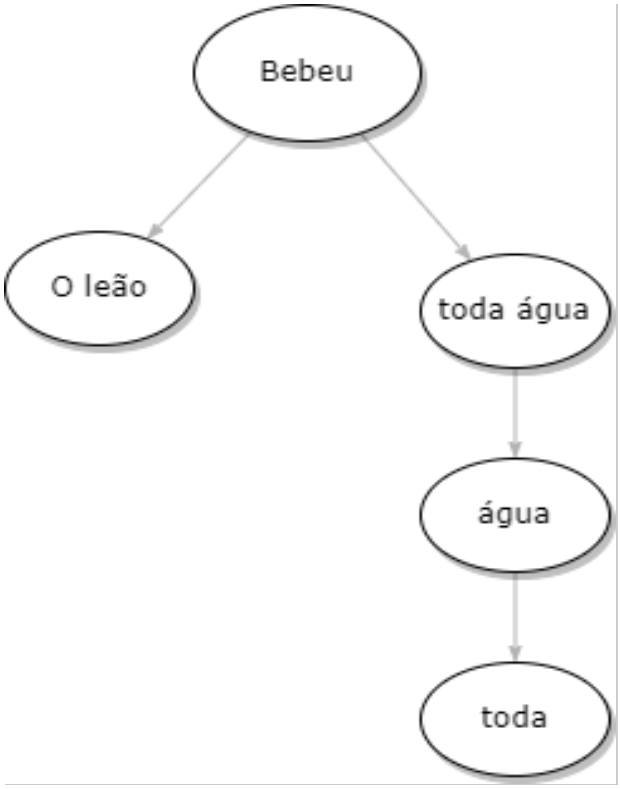
Se tratando de processamento de linguagem natural a análise sintática é responsável, tal qual no estudo linguístico de idiomas, das aplicações de regras gramaticais para análise de estrutura das frases. E esta, segundo Chowdhary (2020), pode ser dividida em duas funções: 1) Determinação da estrutura; 2) Regularização da estrutura de sintaxe;

A primeira função é a determinação da estrutura, na qual a análise de sintaxe identifica o sujeito, o objeto, as palavras modificadoras e a frase que cada palavra modifica, atribuindo uma estrutura de árvore durante o processo de análise.

Exemplo de análise sintática para a frase “O leão bebeu toda água.”:

- Sujeito da frase: "O leão", onde, "O" é artigo definido e "leão" exerce a função de substantivo.
- Verbo: "bebeu";
- Objeto direto: "toda água".

Figura 2 - Análise sintática



Fonte: Autoria própria

A segunda função é a regularização da estrutura de sintaxe, que simplifica o processamento subsequente, especialmente a análise semântica, ao mapear possíveis sentenças de entrada em um número reduzido de estruturas. Chowdhary exemplifica que certos elementos nas frases podem ser omitidos, mantendo o significado das frases intacto.

Para o autor o objetivo da análise semântica é determinar o significado de uma frase, buscando as condições em que ela é verdadeira. Além disso, a análise semântica caracteriza as regras de inferência para as sentenças da linguagem. No entanto, a caracterização da semântica das sentenças interrogativas e imperativas

apresenta maior complexidade. O autor destaca a importância do significado de uma frase, mesmo quando esta está correta ou sintaticamente correta. Ao criar um sistema de linguagem natural, o objetivo é que o sistema execute alguma ação em resposta à entrada, como recuperar dados ou realizar comandos em um robô. Nesse sentido, a tradução da linguagem natural para linguagens formais, como sistemas de recuperação de banco de dados ou sistemas de comando de robô, é necessária. Essas linguagens formais possuem propriedades não ambíguas, regras simples de interpretação e inferência, além de uma estrutura lógica determinada pela forma da sentença.

2.5.2. Análise de sentimentos

A análise de sentimentos (SA, Sentiment Analysis) ou mineração de opinião (OM, Opinion Mining) é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade. A entidade pode variar entre indivíduos, eventos ou tópicos, sendo mais comum encontrá-la em resenhas. A mineração de opinião extrai e analisa as opiniões das pessoas sobre uma entidade, enquanto a análise de sentimento identifica o sentimento expresso em um texto e, em seguida, o analisa. Assim, o objetivo da SA é encontrar opiniões, identificar os sentimentos expressos e classificar sua polaridade (MEDHAT et al. 2014)

Polaridade que segundo Benevenuto et al. (2013) “representa o grau de positividade e negatividade de um texto”, pode ser expressa de forma binária ou ternária, variando do método de medição e a tecnologia aplicada.

A polaridade pode ser calculada de diversas formas, seja de forma humanizada através de classificação manual até RNNs, redes neurais recorrentes, neste trabalho serão utilizadas para o cálculo da polaridade dos tweets as bibliotecas NLTK e spaCy.

O Natural Language Toolkit (NLTK) é uma biblioteca de código aberto em Python [1]amplamente reconhecida e utilizada para o processamento de linguagem natural (PLN). Com um conjunto de ferramentas e recursos. Entre as funcionalidades oferecidas pelo NLTK estão a tokenização, conforme será definida abaixo; a marcação POS (part-of-speech), que atribui a cada palavra uma classe gramatical, como substantivo, verbo ou adjetivo; a análise de sentimentos, que avalia a polaridade emocional do texto; a lematização, que reduz as palavras à sua forma base ou lema; e o stemming, que remove os sufixos das palavras para simplificá-las NLTK (2029).

VADER (Valence Aware Dictionary and sEntiment Reasoner) é um modelo baseado em regras desenvolvido para análise de sentimentos em textos de mídias sociais. Criado por C.J. Hutto e Eric Gilbert em 2014, o VADER combina uma lista padrão de características lexicais com cinco regras gerais que consideram convenções gramaticais e sintáticas que influenciam a intensidade do sentimento. Este modelo se destaca por sua capacidade de lidar eficientemente com textos curtos e informais, comuns em plataformas como o Twitter, e por ser capaz de interpretar

[1] Python é uma linguagem de programação voltada para ciência de dados e machine learning, mais informações podem ser obtidas em: https://www.python.org/

gírias, pontuação e uso de maiúsculas para transmitir intensidade emocional (Hutto & Gilbert, 2014).

Utilizou-se o VADER para realizar a análise de sentimentos dos tweets relacionados aos usuários selecionados. Esta escolha se deu devido à alta precisão do modelo em contextos de mídias sociais, onde ele supera avaliadores humanos individuais em termos de precisão de classificação, com uma acurácia F1 de 0.96 em comparação com 0.84 dos humanos. A implementação do VADER permitiu uma análise robusta dos sentimentos expressos nos tweets, fornecendo insights sobre as percepções e estratégias de comunicação dos atores políticos no Twitter (Hutto & Gilbert, 2014).

2.5.3. Tokenização e lematização

De acordo com Webster e Kit (1992), essa etapa envolve a fragmentação do texto em unidades chamadas de tokens, que podem corresponder a palavras individuais. Ao mesmo tempo, é desejável remover caracteres especiais, como pontuações e números, a fim de simplificar o texto para as etapas subsequentes. Em um algoritmo a aplicação da tokenização em uma frase resulta em uma matriz de palavras, conforme demonstrado abaixo.

Frase original:

"Os dias ensolarados são os melhores dias do ano"

Objeto resultante da tokenização: ["Os", "dias", "ensolarados", "são", "os", "melhores", "dias", "do", "ano"].

Por outro lado, a lematização leva em conta a análise da estrutura morfológica das palavras, buscando identificar e agrupar diferentes formas verbais no infinitivo, bem como substantivos em uma única forma. Geralmente, esse processo requer o uso de um vocabulário específico (Schütze, Manning, Raghavan, 2008). A seguir o exemplo de aplicação da lematização para uma frase em português.

Frase original:” Os dias ensolarados são os melhores dias do ano”

Resultado da lematização: “O dia ensolarado ser o melhor dia do ano”

2.5.4. Stopwords

Stopwords são termos comuns que normalmente são excluídos durante o pré-processamento de texto em tarefas de Processamento de Linguagem Natural (PLN). Estas palavras, tais como "o", "a", "de" e "para", têm pouco impacto semântico no texto e, portanto, são filtradas para aprimorar a eficácia e a precisão das análises textuais. Essa prática de remoção de stopwords visa otimizar a interpretação dos dados, destacando palavras mais relevantes para a análise.

2.5.5. Estudo de caso da aplicação de PNL e tweets

O artigo “Análise de sentimento dos usuários do Twitter em relação ao COVID19”. contém uma demonstração prática de aplicação de inteligência artificial para análise de dados, além da correlação do número de casos com o volume de tweets relacionados à pandemia. Sendo que esta correlação pode ser comparada com pesquisas similares realizadas em países com proporções demográficas parecidas com às do Brasil. Para realização do estudo utilizou-se a seguinte metodologia:

- Divisão do trabalho em etapas:
- Definição das Hashtags e extração de dados;
- Pré-processamento e construção do algoritmo;
- Análise de sentimentos e classificação dos tweets.
- Execução do trabalho;
- Análise de resultados.

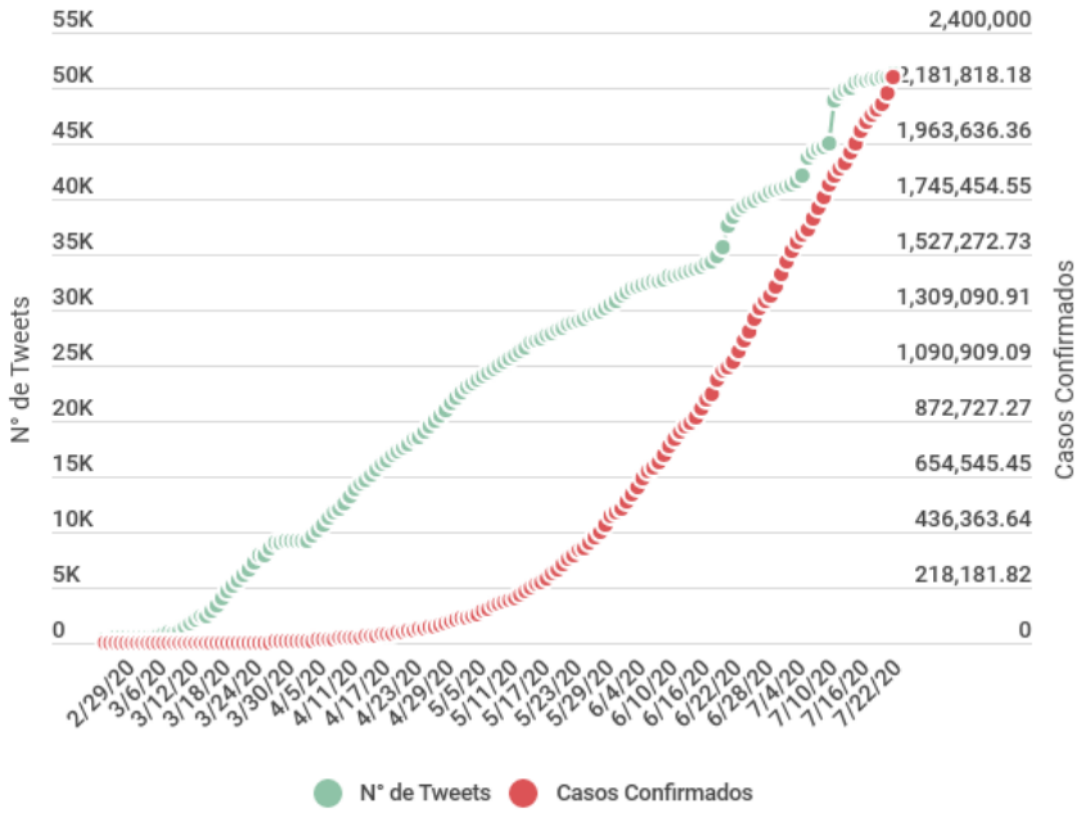
Etapa 1 – Recolheram-se os dados a partir do Twitter utilizando uma API que permite a recolha e geração de uma base de dados de tweets sobre um tópico específico, selecionou-se o período de janeiro a julho de 2020, totalizando 7 meses de tweets selecionados, utilizando hashtags como COVID-19 e FiqueEmCasa para pesquisar e selecionar tweets. Inicialmente a base consistia em 56.090 tweets.

Etapa 2 - A segunda fase da investigação envolveu a seleção apenas de tweets em localizados no Brasil, a remoção de caracteres especiais, números, stopwords, conteúdo visual, e a conversão de caracteres maiúsculos para minúsculas. O foco desta fase é otimizar o processo de análise de sentimentos.

Etapa 3 – Classificaram-se os dados (sentimentos) nas classes: a) positivos; b) neutros; c) negativos.

Após a coleta e saneamento da base de dados a técnica Bag of Words para a conversão dos textos em vetores, subdividindo a base em duas classes: 1) classe das frases positivas; 2) classe das frases negativas.

Figura 3 - Gráfico tweets por dia



Fonte: Pessanha et al. (2020)

O gráfico acima demonstra o crescimento da quantidade de citações a pandemia na rede social e o relaciona com o número de casos confirmados. Os autores também citam demais pesquisas realizadas com tweets de língua inglesa e a correlação encontrada nas demais fontes não difere do gráfico brasileiro. Em PESSANHA et al. (2020, p10):

> Características de tendência (linear e exponencial) das curvas de número de tweets e casos de confirmados de COVID-19 no Brasil são semelhantes aos padrões encontrados por Medford et al. (2020) em tweets em língua inglesa e os casos confirmados do vírus a nível mundial.

Abaixo as nuvens de palavras referentes aos tweets, a esquerda temos tweets classificados como positivos no Bag of Words e a direita temos os tweets classificados como negativos.

Fonte: (Pessanha et al. 2020)

Os autores apontam a presença de palavras como: “morte”, “casos”, “vida”, “sus”, “pandemia”, “quarentena”, e menciona-se a presença de um tom de preocupação por parte dos usuários da rede social. A maior parte dos tweets foram negativos, ao final do levantamento obtém-se os seguintes números:

Classe	Percentual
Negativo	63,27%
Positivo	27,55%
Neutro	9,18%

Neste ponto é realizado um paralelo com pesquisas internacionais e novamente a pesquisa brasileira coincide com os números externos: “[...] estes resultados estão em sintonia com os estudos recentes realizados por Pastor (2020) nas Filipinas, Xue

et al. (2020) na China, Medford et al. (2020) em tweets de língua inglesa e Lwin (2020) em 170 países.” (Pessanha et al. 2020).

Ao concluir analisa-se a correlação do aumento do número de *tweets* com o aumento da contaminação das pessoas pela COVID19. Retoma-se a importante mensagem de possuirmos, enquanto país, sistemas de processamento de linguagem natural integrados com mídias sociais para o apoio à decisão e a importância do monitoramento das redes sociais para o direcionamento de ações públicas, tanto para a pandemia que se passou quanto para as demais crises que possam vir.

3. TRABALHOS CORRELATOS

A área de estudos e análises relacionados às mídias sociais e seu impacto na sociedade é ampla e conta com a contribuição de diversos autores e grupos. Nesta seção, serão apresentadas breves descrições de trabalhos que se dedicam a promover a transparência do poder público, explorar a integração com as mídias sociais e realizar análise de sentimentos. Esses estudos têm como objetivo aprofundar nosso entendimento sobre o papel das mídias sociais na sociedade e suas implicações, contribuindo para o avanço do conhecimento nessa área.

3.1. Quem me representa

A página "Quem me representa?" é uma ferramenta online que permite aos usuários descobrir quais políticos e representantes estão atuando de acordo com suas opiniões (Marques; Andrade; Arthur; Freire, 2023). O objetivo do site é facilitar o acesso à informação sobre os representantes políticos eleitos em diferentes esferas

governamentais, como vereadores, deputados estaduais, deputados federais e senadores.

Ao visitar o site "Quem me representa?", os usuários podem inserir seu endereço para obter uma lista dos políticos que os representam em cada nível de governo. Além disso, a plataforma também fornece informações sobre o histórico político dos representantes, incluindo: 1) seu partido político; 2) mandatos anteriores; 3) projetos de lei de sua autoria; 4) posicionamentos públicos.

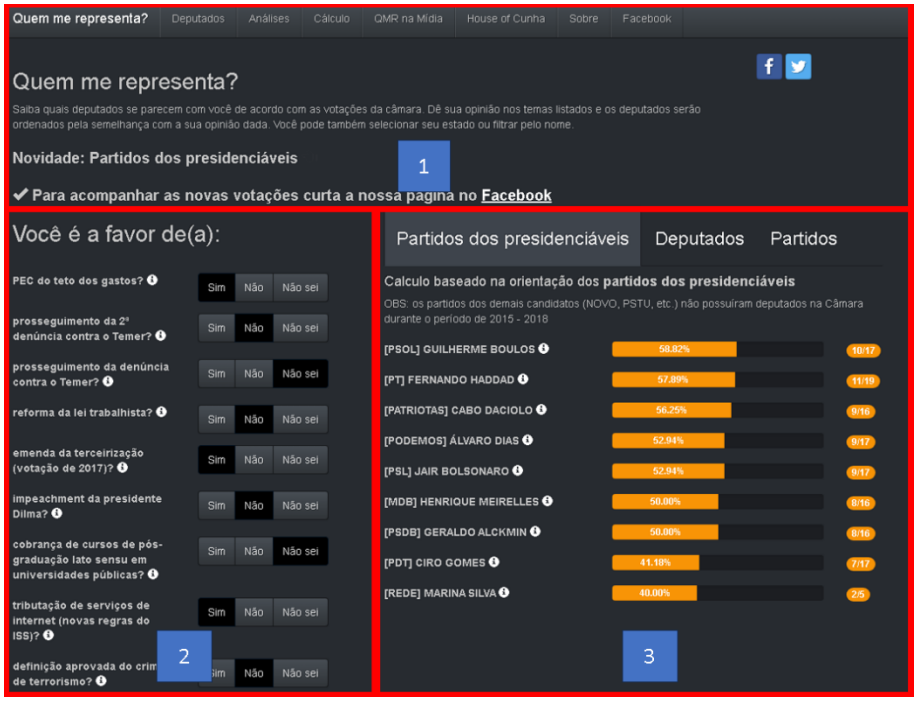
Abaixo uma imagem da página inicial do site, as seções são distribuídas da seguinte forma:

1) cabeçalho da página;

2) campos para a seleção das preferências do usuário sobre os temas pré-cadastrados na plataforma;

3) Os resultados da plataforma que inclui: os presidenciáveis, os deputados e os partidos que possuem maior afinidade com as preferências do usuário.

Figura 4 - Quem me representa



Fonte: (Marques; Andrade; Arthur; Freire, 2015)

De acordo com os desenvolvedores da plataforma (Marques; Andrade; Arthur; Freire, 2023) o cálculo é realizado da seguinte forma:

- O eleitor preenche o formulário (seção 2);
- A plataforma realiza um filtro na listagem de resultados (seção 3) com base na votação dos candidatos, excluindo abstenções e obstruções;
- Se o candidato não votou sobre o tema o tema é desconsiderado;
- Após os passos anteriores a plataforma exibe de forma proporcional qual o candidato melhor se adequa ao perfil do eleitor.

A página visa promover a transparência e a participação cidadã, permitindo que os eleitores conheçam seus representantes e acompanhem suas atividades. Dessa forma, os cidadãos podem ter maior conhecimento sobre quem está tomando

decisões políticas em seu nome e buscar uma participação mais informada na política. Pode ser acessado através do endereço: qmrepresenta.com.br

3.2. Ranking dos políticos

A página "Ranking dos Políticos" é uma plataforma online que tem como objetivo avaliar e classificar o desempenho dos políticos brasileiros (Ranking dos Políticos, 2024). Segundo os autores do site, a página busca fornecer informações aos cidadãos sobre o trabalho dos políticos eleitos, promovendo a transparência e a prestação de contas.

Abaixo a tela inicial da página:

Figura 5 - Pagina inicial "Ranking dos políticos"

Fonte: Ranking dos políticos

Ao clicar sobre um dos políticos exibidos o usuário é direcionado para os detalhes da legislatura do indivíduo:

Figura 6 - Detalhes dos políticos

Fonte: Ranking dos políticos, 2023

No site do Ranking dos Políticos, disponível em politicos.org.br, os usuários podem encontrar rankings atualizados dos políticos brasileiros, classificados de acordo com critérios como a presença nas sessões legislativas, a qualidade de seus projetos de lei e o uso consciente de recursos públicos. A página também disponibiliza informações sobre processos judiciais e condenações de políticos.

Abaixo a fórmula, disponível no site do portal Ranking dos Políticos (2023), de como é calculada a nota do político no site.

Para calcular a pontuação final dos políticos, aplicar os seguintes pesos:

Votações: 3x

Gastos (Presenças e Economia de Verbas): 2÷

Processos Judiciais (Ficha Limpa): quando há processos, subtrai pontos da nota final

Outros: soma ou subtrai pontos da nota final

Fórmula:

[(V x 3)+(G / 3)]/4 + P + OT = Nota Final

V = Votações

G = Gastos

P = Processos Judiciais

OT = Outros

Sendo que:

- Votações: O posicionamento dos parlamentares nas principais votações do Congresso é avaliado de acordo com a orientação do nosso Conselho. O aproveitamento de acerto nos votos reflete a nota do político neste critério. O Ranking avalia apenas parlamentares que ficam no mínimo 06 meses em seus respectivos mandatos.

- Gastos: Políticos que não faltam ao trabalho e que economizam na Cota Parlamentar e Verba de Gabinete ganham pontos no ranking.
- Presenças nas sessões deliberativas: A nota varia de acordo com a taxa de comparecimento do político durante o período em que ele esteve em exercício. Faltas justificadas contam como presença.
- Economia de Verbas: O desempenho neste quesito é definido pelo percentual economizado pelo político em relação à sua cota parlamentar e verba de gabinete disponíveis enquanto esteve em exercício.
- Cota Parlamentar: Valor mensal para custear as despesas do mandato. Seu valor varia de acordo com o estado de origem do parlamentar, por conta dos diferentes custos de passagens aéreas para Brasília.
- Verba de Gabinete: É o valor mensal destinado para pagar o salário dos assessores dos parlamentares.
- Processos Judiciais: Os parlamentares condenados em crimes, principalmente contra a administração pública, ou que respondem a inquéritos do STF, perdem pontos no Ranking dos Políticos.
- Outros: Iniciativas relevantes não previstas nos demais critérios podem gerar o acúmulo ou a perda de pontuações extras para os parlamentares. Os pontos são subtraídos da nota final apurada nos demais critérios.

Exemplo prático para o deputado fictício João da Silva.

Nome: Deputado João da Silva

Votações: Fez 60 pontos em 100 possíveis nas votações

Nota: 6,0

Gastos: Teve 90 presenças em 100 sessões deliberativas possíveis = nota 9

Economizou R$ 70 mil de R$ 100 mil disponíveis para Cota Parlamentar + Verba de Gabinete= nota 7

Nota: 8,0 (média das duas notas)

Processos Judiciais: Condenado em inquérito da Operação Lava-Jato = -1,0 ponto

Outros: Aprovou o projeto de sua autoria que cria o Marco das Ferrovias = +0,5 ponto

O deputado João da Silva totalizou os seguintes pontos: v) 6,0; g) 8,0; p) -1,0; o)0,5;

Cálculo da nota final: [(6,0 x 3) + 8,0 /3)]/4 - 1,0 + 0,5 = 4,66

3.3. Cascatas de fake news políticas: um estudo de caso no Twitter

O estudo “Cascatas de Fake News Políticas: um estudo de caso no Twitter” (Recuero, 2019), teve como objetivo analisar a estrutura e difusão de “cascatas de fake news” relacionadas ao caso do julgamento e prisão do ex-presidente Lula. Para alcançar esse objetivo, a seguinte abordagem foi adotada:

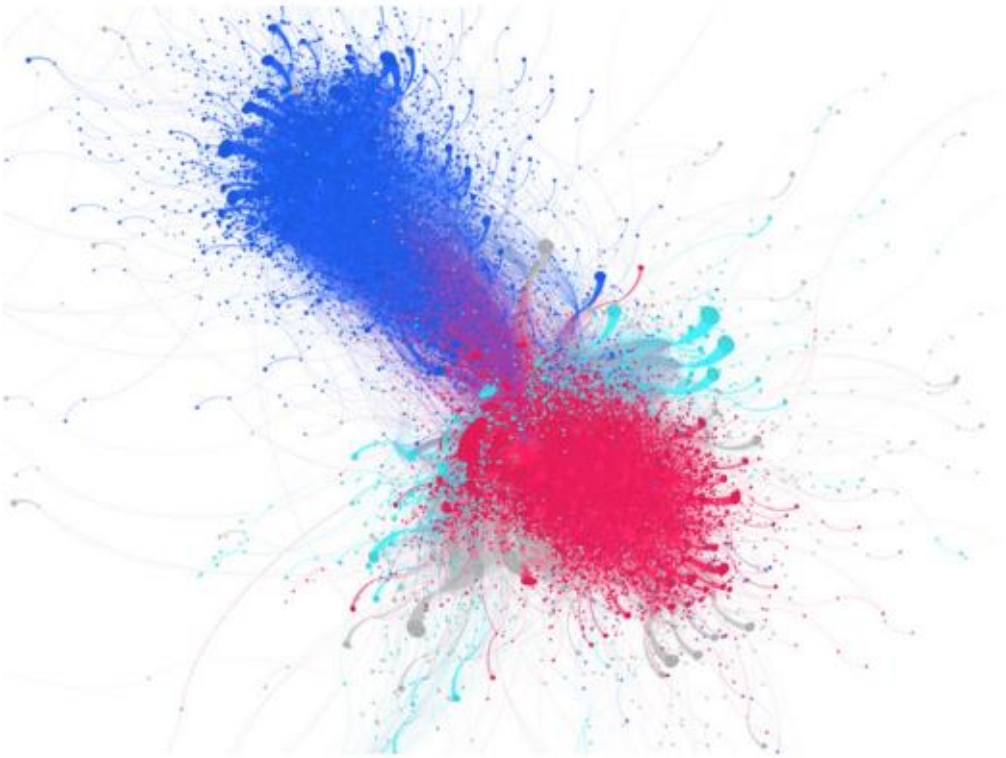
Para a coleta de dados utilizou-se um *crawler* com acesso à streaming API do Twitter. Este foi configurado para coletar tweets que estivessem relacionados à palavra-chave "lula". Ao final do processo de coleta, obteve-se um dataset inicial contendo 2.430.468 tweets, que foram coletados no período compreendido entre 01 de fevereiro e 26 de abril de 2018.

A etapa de preparação dos dados foi composta pela filtragem utilizando palavras-chave descritivas, que permitiram identificar as discussões pertinentes ao caso do julgamento e prisão do ex-presidente Lula. Com o dataset filtrado, iniciou-se a análise da estrutura das redes de disseminação das fake news, utilizaram-se métricas da análise de redes sociais, baseadas nos estudos de Degenne e Forsé (1999) e Wasserman e Faust (1994). E como ferramenta de análise utilizou-se o software Gephi.

Na representação das redes, os nós foram associados às contas no Twitter e as conexões foram representadas pelos retweets e menções entre essas contas. Essa abordagem permitiu visualizar e estudar a estrutura das redes de disseminação das fake news.

Os resultados obtidos sugerem que as "cascatas de fake news" políticas tendem a ficar restritas aos seus próprios clusters ideológicos, conforme a imagem abaixo:

Figura 7 - Cluster fake news



Fonte: Recuero, 2019

No cluster vermelho, encontram-se os nós que demonstram maior afinidade com partidos de esquerda e apoiam o ex-presidente (Lula), enquanto no cluster azul, estão os que demonstram posição contrária e são mais alinhados com partidos de direita. Os nós em azul claro representam grupos não vinculados a nenhum partido.

4. PROPOSTA

O presente trabalho propõe o desenvolvimento de uma ferramenta que permita aos usuários visualizarem através de gráficos os sentimentos expressos pelos usuários de redes sociais em relação a um tópico específico no Twitter através de uma linha do tempo. Além disso, a ferramenta possibilitará aos usuários selecionarem figuras políticas para obter o sentimento expresso em seus tweets relacionados aos tópicos. A análise de sentimento será realizada utilizando algoritmos de Processamento de Linguagem Natural (PLN) aplicados aos tweets. É importante destacar que tanto os tweets quanto os tópicos e as figuras políticas disponíveis para seleção serão previamente cadastrados na ferramenta, restringindo a capacidade dos usuários de processar novos dados, apesar desta restrição, será disponibilizado aos usuários a opção de cadastramento de perfil no Twitter o qual este deseja ter acesso aos sentimentos, este cadastramento permitirá uma análise do time do projeto e estudo da viabilidade de inclusão ou não deste novo perfil.

4.1. Definição dos tópicos

Para definir os tópicos, inicialmente considerou-se a seguinte abordagem:

- Incluir manualmente as hashtags obtidas por meio da integração com o Twitter.

No entanto, ao realizar uma análise manual do conteúdo publicado, notou-se que nem todos os políticos aderem às hashtags. Portanto, optou-se por uma segunda abordagem:

- Listar as palavras utilizadas nos Tweets. Nesta primeira versão, a lista será criada por meio da lematização das palavras presentes nos tweets, permitindo

aos usuários buscarem as palavras e visualizarem os sentimentos relacionados a cada uma delas.

4.2. Políticos selecionados

Para a amostragem inicial selecionou-se os políticos que atingiram o segundo turno das eleições presidenciais de 2022, que então, seriam os mais relevantes da política nacional. Portanto, as contas de Twitter utilizadas para análise foram de Jair Messias Bolsonaro (@jairbolsonaro) e Luiz Inácio Lula da Silva (@lulaoficial).

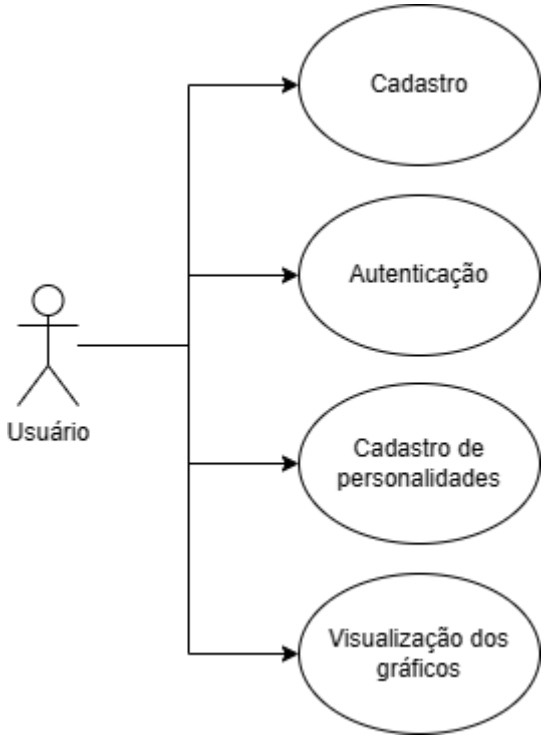
Além das contas pré-cadastradas também será possível a inclusão de novos perfis, mediante a análise do time do projeto. Haverá um serviço disponibilizado na nuvem para o cadastramento e atualização dos perfis em tempo real, utilizando tecnologias de raspagem de dados.

A política de atualização das informações será através de rotinas de execução de serviços, estas rotinas serão responsáveis pelo acionamento de um ou mais serviços Python capazes de identificar novas contas e realizar a raspagem destas informações e atualizar as contas dos usuários. Atualmente a previsão de execução dessa rotina é de forma diária para atualização e quinzenal para inclusão de novos usuários.

4.3. Funcionalidades adjacentes

Para melhor operacionalização da ferramenta as seguintes funcionalidades extras serão disponibilizadas: Cadastro, login e cadastro de pessoas.

Figura 8 - Casos de uso



Fonte: Autoria própria

4.3.1. Cadastro e login

A ferramenta apesar de poder ser clonada para realização de analise em ambiente local, também poderá ser disponibilizada em plataformas de nuvem, exemplos: Amazon, SAP Cloud Platform, Google Cloud e etc.

Para favorecer o hospedeiro e monitorar quais usuários estão operando a ferramenta, a funcionalidade de cadastro tem o intuito de restringir o número de pessoas inseridas para adesão e criação de gráficos e lematização em massa de atores não tão relevantes.

A funcionalidade de login, atualmente, apenas restringe o cadastro de novos políticos.

4.3.2. *Cadastro de pessoas*

Propõe-se a liberação de cadastro de novas pessoas e suas respectivas redes sociais para execução do PNL. Este cadastro poderá ser realizado pelos usuários após a realização do login na aplicação.

5. DESENVOLVIMENTO

Neste capítulo abordaremos as etapas que compreenderam o processo de desenvolvimento do projeto, destacando os principais aspectos metodológicos e técnicos adotados para a concretização da proposta.

5.1. Metodologia

No início do projeto optou-se por objetivos mensais e a realização de encontros semanais de acompanhamento para a avaliação contínua do avanço das atividades. Contudo, ao longo do desenvolvimento, observou-se que determinadas fases da implementação demandaram um período de tempo superior à estimativa inicial. Portanto, optou-se por ajustar a abordagem adotada, introduzindo sessões de planejamento durante as reuniões semanais, as reuniões foram realizadas de forma

remota através do Google Meet[2]. Nesse contexto, desenhou-se objetivos mais específicos e realizáveis, mediante uma colaboração efetiva entre a equipe de desenvolvimento e o orientador. Na prática realizaram-se reuniões semanais às segundas-feiras definindo os objetivos para a entrega da próxima semana tendo como guia o macro cronograma inicial. As entregas eram marcadas por demonstrações do funcionamento do sistema através do compartilhamento de tela e as falhas, dúvidas e próximos passos eram anotadas informalmente e discutidas no próximo encontro semanal.

5.2. Arquitetura

A arquitetura do sistema é composta por várias camadas que interagem de forma coordenada para realizar a análise de sentimentos no Twitter, interface de usuário, camada de serviço NodeJS[3], serviço Python de raspagem de dados e serviço Python de processamento de dados.

A interação do usuário com o sistema inicia-se ao pressionar um botão no navegador, o que dispara uma chamada HTTPS. Esta ação invoca uma função no back-end, implementada utilizando a tecnologia NodeJS em conjunto com o framework Express[4]. As rotas definidas no back-end estabelecem a conexão com o

[2] Google Meet é um software que permite a realização de chamadas de vídeo/ áudio da empresa Google, pode ser acessado em: https://meet.google.com/

[3] NodeJS é um ambiente de execução do JavaScript no lado do servidor, pode ser acesso em: https://nodejs.org/en

[4] Express é utilizado como um framework para auxiliar nas atividades de programação que envolvem o NodeJS, pode ser acessado em: https://expressjs.com/pt-br/.

banco de dados através de TCP/IP. O banco de dados é alimentado por uma função Python responsável pelo processamento dos dados cadastrados. Esses dados têm origem em outro serviço Python que processa os novos perfis de redes sociais cadastrados pelos usuários na interface do front-end.

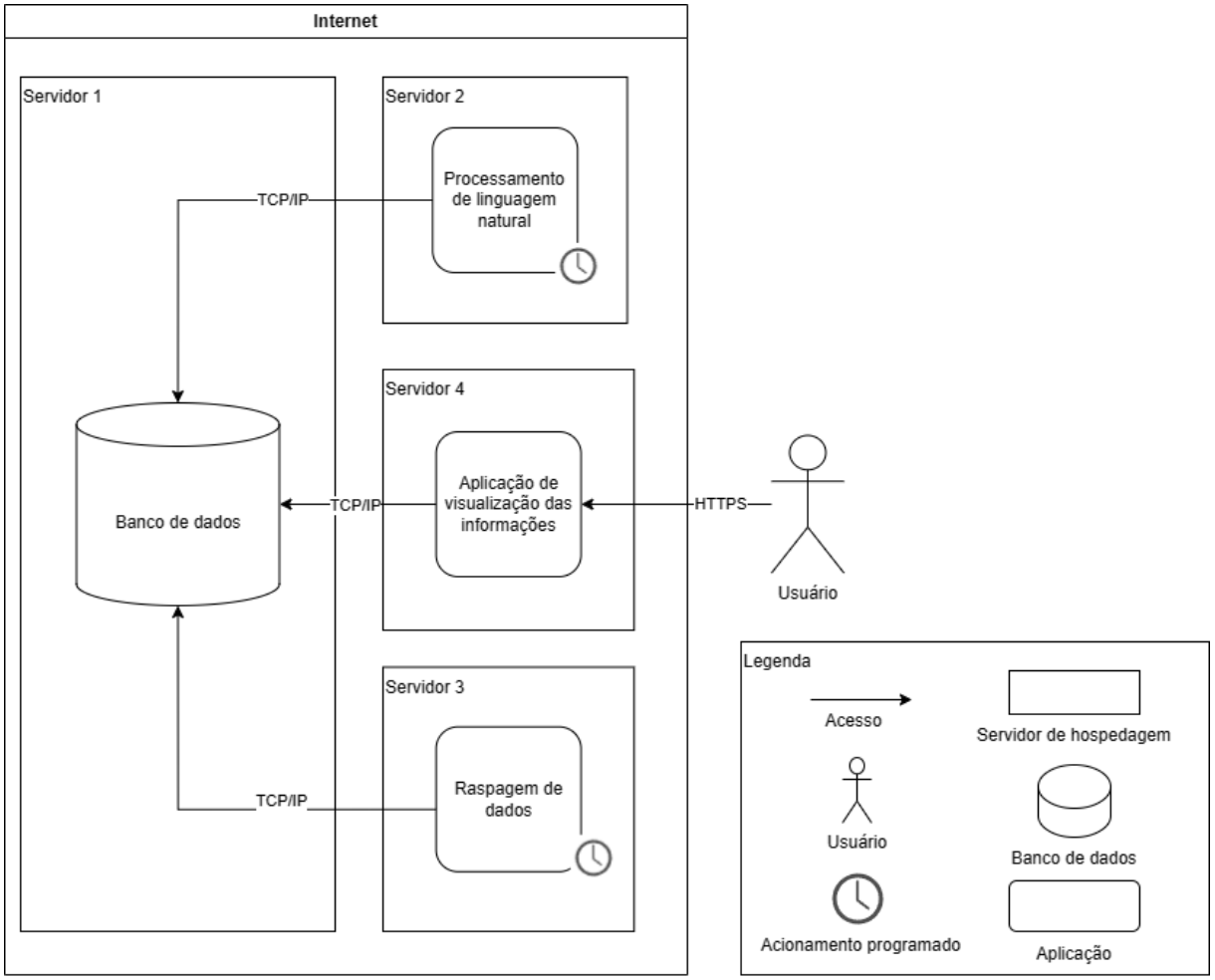
O segundo serviço Python, de processamento de dados, opera da seguinte maneira:

- O usuário cadastra os perfis de redes sociais que deseja monitorar na aplicação.
- O serviço Python lê os novos perfis cadastrados e realiza a raspagem de dados do Twitter.
- Os dados raspados são armazenados no banco de dados em uma tabela específica, sem que sejam inicialmente processados em termos de polaridade e lematização.

Posteriormente, o primeiro serviço Python lê os dados armazenados na tabela e realiza o processamento de polaridade e lematização. Esses dados processados são então armazenados na tabela final do banco de dados.

Um relatório é gerado a partir dos dados processados e é exibido na interface do front-end para o usuário final, completando o ciclo de informações dentro da ferramenta. Esta abordagem integra de forma eficiente as etapas de coleta, processamento e apresentação dos dados, garantindo que o usuário tenha acesso a análises de sentimentos precisas e atualizadas baseadas nos dados das redes sociais que ele próprio cadastrou na aplicação. Abaixo uma imagem para ilustrar a arquitetura da solução.

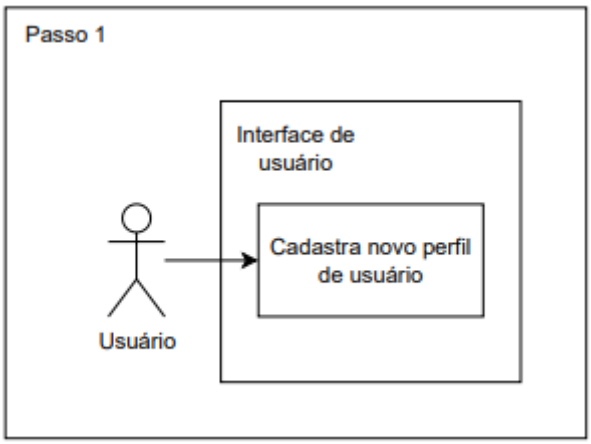
Figura 9 - Arquitetura do projeto



Fonte: Autoria própria

Abaixo uma sequência de figuras que representam o fluxo de dados da solução.

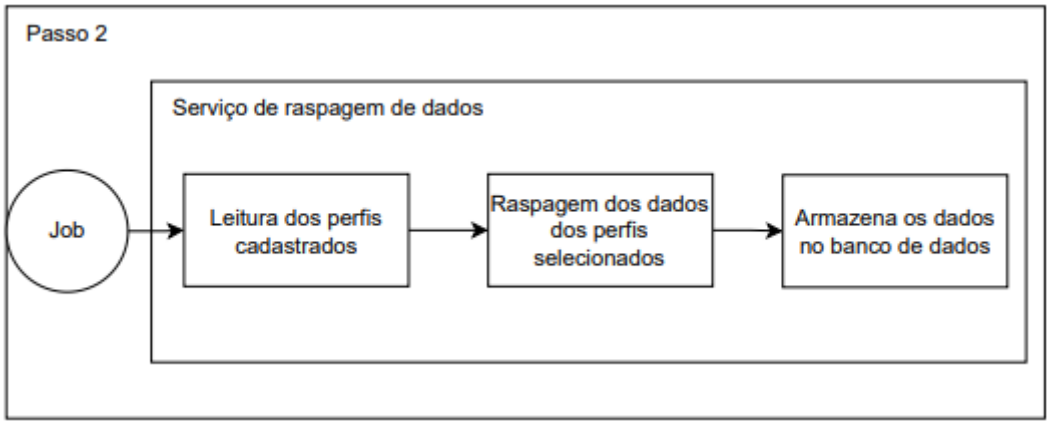
Figura 10 - Passo 1 do fluxo de dados



Fonte: Autoria própria

Usuário realiza o cadastramento de novos perfis de redes sociais.

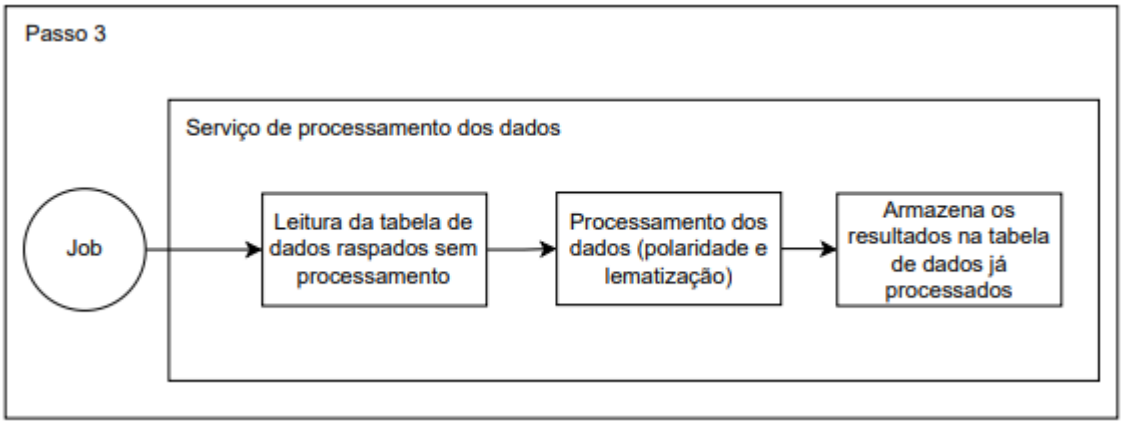
Figura 11 - Passo 2 do fluxo de dados



Fonte: Autoria própria

O serviço de acionamento programado realiza a chamada do programa Python responsável pela seleção dos novos dados, raspagem da rede social e armazenamento.

Figura 12 - Passo 3 do fluxo de dados



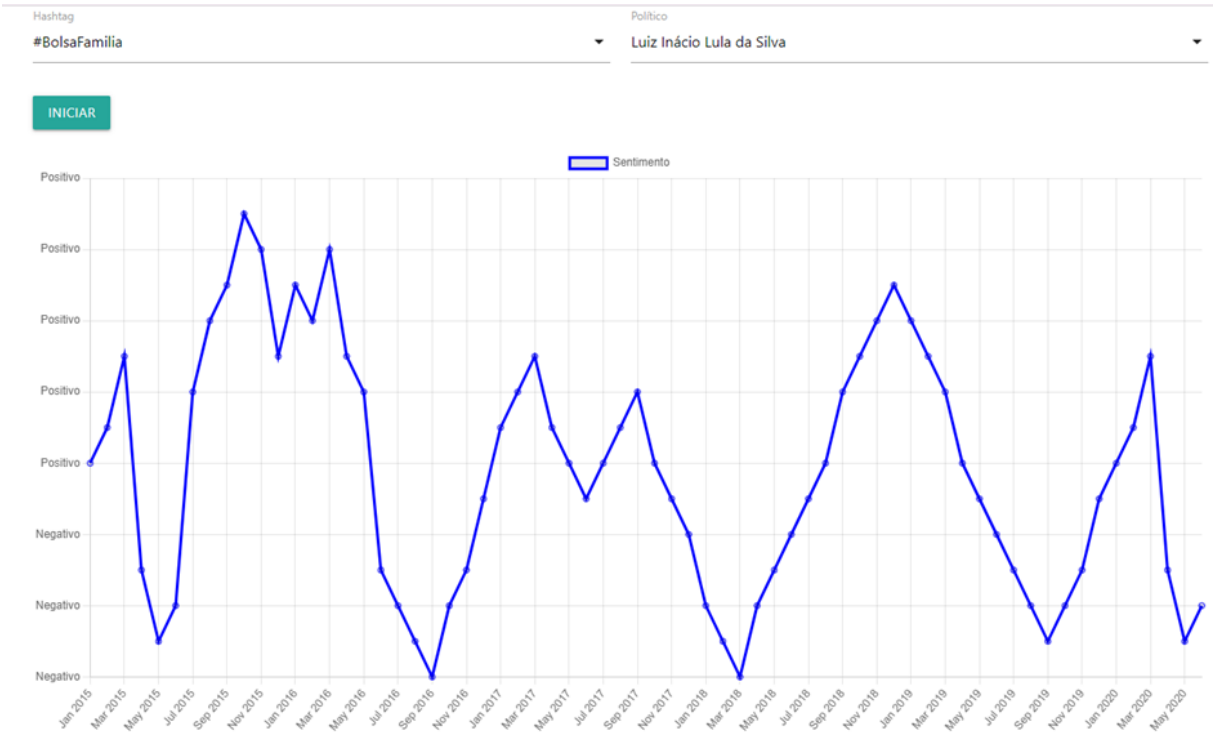
Fonte: Autoria própria

O serviço de acionamento programado realiza a chamada do segundo serviço Python que é responsável pelo processamento das informações armazenadas. Este processamento gera os dados de polaridade e lematização. Agora essas informações processadas são exibidas no gráfico com a análise de sentimentos conforme será demonstrado na próxima seção.

5.3. Prototipação das interfaces

Para propiciar um melhor entendimento do cenário proposto optou-se por iniciar o desenvolvimento pela prototipação das interfaces, bem como um esboço de base de dados que viria a ser implementada no projeto. A prototipação foi realizada diretamente em HTML com JavaScript. A tela abaixo não possui comportamentos e os dados são estáticos, a única finalidade é demonstração visual das funcionalidades principais da aplicação.

Figura 13 - Protótipo de interface



Fonte: Autoria própria

Através deste protótipo definiu-se alguns componentes essenciais que viriam a estar presentes nas demais versões da interface do sistema: Caixas combo “hashtag” e “político”, botão “iniciar” e o gráfico que exibirá os sentimentos presentes nos textos dos tweets. No gráfico temos as informações de: polaridade calculada no eixo Y e linha do tempo no eixo X, fazendo com que o leitor que esteja analisando as informações consiga perceber que os sentimentos relacionados ao tópico selecionado se alteraram ou não através do tempo.

5.4. Análise da estrutura do Tweet

Um tweet, conforme mencionado na revisão bibliográfica é composto de alguns componentes:

1) Foto de perfil do usuário;
2) Nome de usuário;
3) Conta do usuário;
4) Data de publicação do tweet;
5) Texto/Conteúdo do tweet;
6) Respostas ao tweets;
7) Republicações do tweet;
8) Curtidas no tweet;
9) Alcance do tweet.

Conforme a imagem abaixo:

Figura 14 - Exemplo de tweet

Fonte: Autoria própria

A partir destas informações definiu-se quais os campos seriam utilizados para realização do processamento dos dados, bem como quais o comportamento da ferramenta, como por exemplo um filtro por data da publicação do tweet.

5.5. Integração com o Twitter

Durante o projeto abordaram-se algumas formas de integração com o Twitter. No início, imaginou-se que a integração seria realizada através da API, contudo, com

as novas diretrizes da empresa “X Corp”, a API não estaria mais disponível para consumo gratuito. Sem a API gratuita, levantou-se as seguintes formas de integração:

- Bibliotecas de terceiros: Bibliotecas e ferramentas de terceiros que ajudam a simplificar o processo de coleta e processamento de dados do Twitter. Algumas dessas bibliotecas estão disponíveis em linguagens de programação populares, como Python e R[5], exemplo: Tweepy[6], Twitter4J[7], Twint[8].
- Web Scrapping: O web scrapping envolve a coleta de dados diretamente das páginas públicas do Twitter, por meio de scripts automatizados que acessam o site e extraem as informações desejadas. Tecnologias que implementam o Web Scrapping: Beautiful Soup[9], Scrapy[10], Selenium[11], Puppeteer[12].

As bibliotecas de terceiros desempenham o papel de simplificar o acesso aos dados do Twitter, tornando o processo mais eficiente. Entretanto, com a introdução de tarifas, muitas dessas bibliotecas foram descontinuadas e alguns serviços agora requerem autenticação para verificar se a conta do desenvolvedor está vinculada à versão premium de acesso à API, o que não se aplica ao projeto.

[5] R é uma linguagem de programação que, assim como o Python, é voltada para a ciência de dados e machine learning. Mais informações podem ser obtidas em: https://www.r-project.org/.
[6] https://www.tweepy.org/
[7] https://twitter4j.org/
[8] https://github.com/twintproject/twint
[9] https://beautiful-soup-4.readthedocs.io/
[10] https://scrapy.org/
[11] https://www.selenium.dev/
[12] https://pptr.dev/

Portanto optou-se pelo desenvolvimento de uma ferramenta que realize o web scrapping dos dados.

5.6. Modelagem de dados

No projeto optou-se pela modelagem de dados relacional que é uma abordagem para projetos de bancos de dados e representa os dados de forma estruturada e relacionada. Os dados são organizados em tabelas, onde cada tabela representa uma entidade do mundo real ou um conceito de negócio. Cada linha em uma tabela, também conhecida como tupla ou registro, representa uma ocorrência específica desta entidade, enquanto as colunas representam os atributos ou características dessa entidade. A relação entre tabelas é estabelecida por meio de chaves estrangeiras, que são atributos em uma tabela que fazem referência a chaves primárias em outras tabelas. Essa relação permite que os dados sejam vinculados e acessados de maneira eficiente por meio de consultas SQL (Structured Query Language).

Essa abordagem foi adotada no projeto devido a fácil estruturação do conteúdo das informações, que inicialmente serão de Tweets, e que tem em seu cerne fundamental um dado textual ligado a um autor, ou seja, há uma relação direta do conteúdo dos Tweet com a entidade de negócio autor. Ainda que novas redes sociais possam vir a ser integradas, a estratégia de banco de dados continua sendo válida.

Um ponto essencial da modelagem dos dados para o projeto é a estrutura da chamada de acionamento das funções responsáveis pela manutenção da base de dados. Essa rotina também é controlada por tabelas no banco e são executadas

através do serviço de agendamento do NodeJS. Cada função é responsável por uma parte do processamento dos dados tais quais: 1) raspagem de novos autores; 2) processamento de dados; 3) raspagem de autores já cadastrados. Demais questões ligadas a raspagem de dados serão detalhadas a posteriori.

5.7. Raspagem de dados

Através das interações com o Twitter notou-se que este utiliza-se de tecnologias de carregamento dinâmico. Esse método implica que os elementos da página são carregados e atualizados durante a navegação do usuário na página, através de chamadas AJAX ou manipulação direta do DOM, devido a este comportamento optou-se pela utilização do Selenium para realização da raspagem de dados.

A raspagem de dados ocorre através da ferramenta Selenium implementada em forma de serviço web utilizando a linguagem Python, este serviço pode ser acionado via chamadas HTTPS recorrentes (JOB), conforme utilizado na ferramenta ou através de chamadas HTTPS diretas convencionais.

Para a utilização do serviço é necessária a instalação das bibliotecas do Selenium selecionadas para o projeto, bem como a atualização das variáveis de ambientes para redirecionamento do serviço a base de dados da ferramenta, seja na nuvem ou local. Após a instalação das bibliotecas e configuração das variáveis de ambientes o serviço ao ser executado realizará os seguintes passos:

1) Autenticação: A autenticação está configurada para utilizar o usuário nominal previamente informado no arquivo “authentication.txt”;

2) Filtro com o nome do usuário: A API recebe o nome do usuário que será raspado, e filtra os usuários do Twitter com este nome;
3) Seleção do usuário: O primeiro resultado do filtro do Twitter é o usuário filtrado, este é selecionado pelo sistema;
4) Raspagem dos dados: Após a seleção do usuário o sistema realiza a raspagem das informações dos tweets até a data limite de três meses;
5) Registro dos dados: Após a leitura, os dados ainda não processados são armazenados no banco de dados PostgreSQL [13]disponível na nuvem ou localmente.

5.7.1. Métricas da raspagem de dados

Este subcapítulo apresenta as métricas utilizadas para avaliar o desempenho do processo de raspagem de dados no Twitter, implementado com Selenium. As métricas foram escolhidas com base nas etapas descritas no capítulo anterior, visando garantir a eficiência e a qualidade da coleta de dados.

- Tempo de resposta total: O tempo de resposta total refere-se ao tempo necessário para concluir todas as etapas da raspagem de dados, desde a autenticação até o registro dos dados no banco PostgreSQL. Está métrica varia de acordo com o parâmetro de data programa no algoritmo de raspagem, a métrica foi calculada com base em uma amostragem de dez execuções para

[13] PostgresSQL é um sistema de gerenciamento de banco de dados de código aberto. Mais informações disponíveis em: https://www.postgresql.org/

um intervalo de sessenta dias. Nestes termos a aplicação desempenhou uma média de tempo de dois minutos, sendo a média calculada entre dez execuções através de registro de log dentro da própria ferramenta.

- Taxa de sucesso de autenticação automática: A taxa de sucesso de autenticação indica a proporção de tentativas de autenticação bem-sucedidas em relação ao total de tentativas realizadas. Durante o desenvolvimento do projeto diversas interfaces de autenticação foram raspadas, durante as últimas semanas houveram duas principais variações, uma com sistema de identificação de humano/ robô e uma nova interface que solicitava o aumento de segurança no usuário autenticado. Portanto, preparou-se o sistema de raspagem para realizar um *sleep* de 40 segundos entre a abertura da interface do Twitter e a autenticação passou a ser feita de maneira manual com o intuito de burlar o sistema de identificação de robô. Isso influenciou a taxa de sucesso de autenticação automática que fôra de 100% para 0% de sucesso a partir do momento da inclusão deste sistema, contudo, com a intervenção humana o sistema continuou conseguir autenticar na plataforma sem problemas.
- Tempo Médio de Raspagem por Tweet: O tempo médio de raspagem por tweet é a média de tempo gasto para coletar os dados de cada tweet. A média abaixo foi calculada através da raspagem de tweets do Lula em um intervalo de dois meses que obteve trezentos tweets com um tempo de raspagem de dois minutos resultando em 2.5 tweets raspados por segundo.

5.8. Preparação dos dados

A qualidade dos dados desempenha um papel de extrema importância no campo do processamento de linguagem natural, especialmente em ambientes de redes sociais, onde a comunicação não segue uma padronização rígida e a expressão de sentimentos pode assumir diversas formas, como imagens, figuras, gírias, abreviações, hashtags, variações de caixa (maiúsculas e minúsculas), menções, entre outros.

Os dados não processados ou "sujos" coletados nestes ambientes podem apresentar uma série de imperfeições, tais como erros de digitação e inconsistências na codificação dos caracteres. Essas "impurezas" podem surgir de várias fontes, incluindo erros de entrada, formatação irregular e duplicação de informações. Portanto, a qualidade dos dados é o alicerce sobre o qual o projeto irá se basear para garantir a qualidade das informações advindas do sistema.

O processo de limpeza dos dados foi realizado em Python e utilizou-se a biblioteca NLTK, criada e utilizada para o processamento de linguagem natural. No escopo do projeto as seguintes formas de expressão encontradas no Twitter foram caracterizadas como descartáveis e removidas no processo de limpeza:

- URLs;
- Tweets repetidos do mesmo autor;
- Palavras de parada (stopwords).

Após a remoção das características definidas como descartáveis prosseguiu-se para formatação e padronização das informações, todos os tweets tratados foram

transformados em caixa alta e informações como tokenização e lematização foram realizadas para realização dos processamentos dos dados.

5.9. Processamento dos dados

Após a raspagem das informações realizada no primeiro serviço Python descrito no capítulo ocorre a etapa de processamento dos dados. Nesta etapa, realiza-se a lematização e análise de polaridade dos textos dos tweets. Para a lematização utilizou a biblioteca spaCy[14] que utiliza bases de dados com textos de notícias de jornais, o que melhor a sua capacidade de entender e processar a linguagem natural em diferentes contextos. As palavras “lematizadas” serão utilizadas pela camada de aplicação para a realização dos filtros do usuário. A biblioteca utiliza um dicionário interno para associação das palavras com suas formas lematizadas. Para desenvolvimento do projeto utilizou-se o dicionário “pt_core_news_sm”. Conforme exemplificado na seção 2.5.3.

Para o cálculo de polaridade utilizou-se a biblioteca NLTK, que realiza este cálculo através do Sentiment Intensity Analyzer.

O Sentiment Intensity Analyzer ou SIA é uma ferramenta de análise de sentimento utilizada no processamento de linguagem natural. O SIA opera com base no VADER (Valence Aware Dictionary and Sentiment Reasoner), um recurso que consiste em um léxico e um analisador de sentimentos com base em regras.

[14] https://spacy.io/

Especificamente desenvolvido para interpretar sentimentos expressos em mídias sociais e outros dados textuais, o VADER contém uma extensa lista de palavras, cada uma associada a um sentimento (positivo ou negativo). Durante a análise de uma sentença, o SIA verifica a presença de palavras com conotação positiva e negativa. Se houver uma predominância de palavras positivas, a sentença é classificada como mais positiva; caso contrário, se houver uma predominância de palavras negativas, é considerada mais negativa, sendo o sentimento negativo e positivo caracterizado por uma “pontuação” que varia no intervalo de números reais de -1 a 1, onde quanto menor o valor mais negativo é o sentimento.

5.10. Visualização dos dados

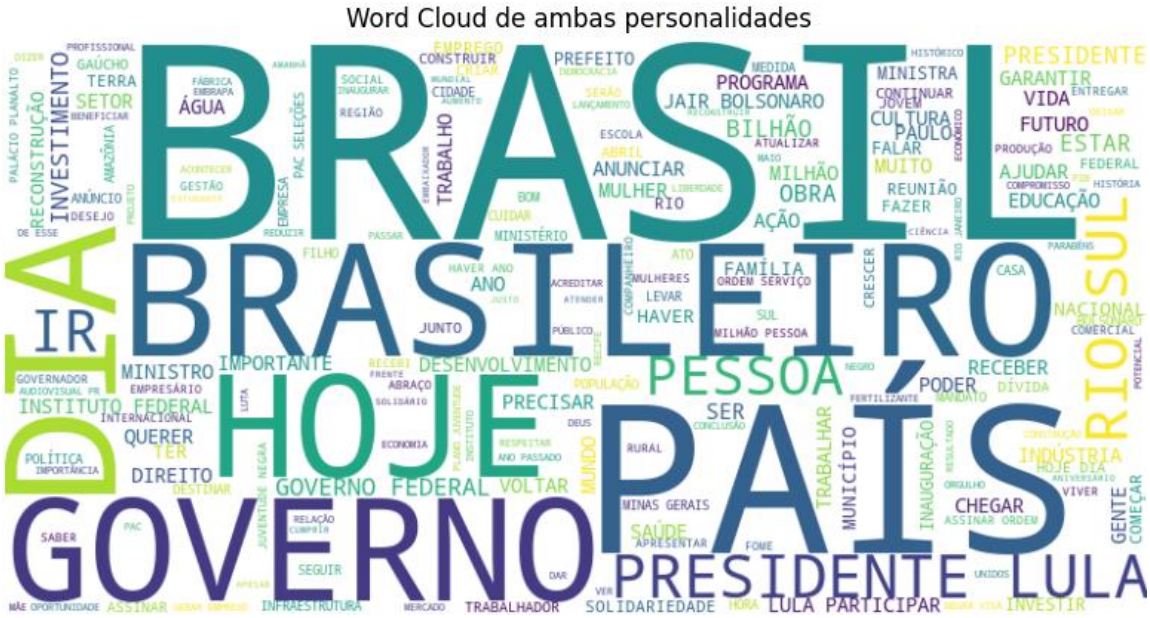
Nesta seção será analisada uma análise exploratória na base de dados e a demonstração do resultado obtido na camada da aplicação.

5.10.1. Análise exploratória

No período de dois meses, de 05 de abril de 2024 até 06 de junho de 2024, foram publicados um total de 222 tweets pelas contas de @lulaoficial e @jairbolsonaro. O Lula foi o autor da maioria dessas publicações, com um total de 160 tweets, enquanto Bolsonaro publicou 62 tweets. Essa volumetria indica uma maior atividade de Lula no Twitter em comparação a Bolsonaro durante o período analisado.

A análise das Word Clouds para ambas as personalidades revela os termos mais frequentemente utilizados em seus tweets, proporcionando uma visão rápida dos principais tópicos abordados.

Figura 15 - World Cloud de ambas as personalidades



Fonte: Autoria própria

A Word Cloud conjunta de Lula e Bolsonaro destaca as palavras "Brasil", "brasileiro", "país", "governo", "hoje" e "dia" como as mais proeminentes. Esses termos sugerem uma ênfase em temas relacionados à nação e ao governo.

Figura 16 - World Cloud do Jair Bolsonaro

Fonte: Autoria própria

Na Word Cloud específica de Jair Bolsonaro, as palavras destacadas incluem "Gestão", "Bolsonaro", "Jair", "Brasil", "Atualizar", "Rio", "Governo", "Lula" e "Dia". A presença de seu próprio nome e variações, assim como referências a "gestão" e "atualizar", pode indicar uma narrativa focada em sua administração e atualizações sobre suas ações ou posições.

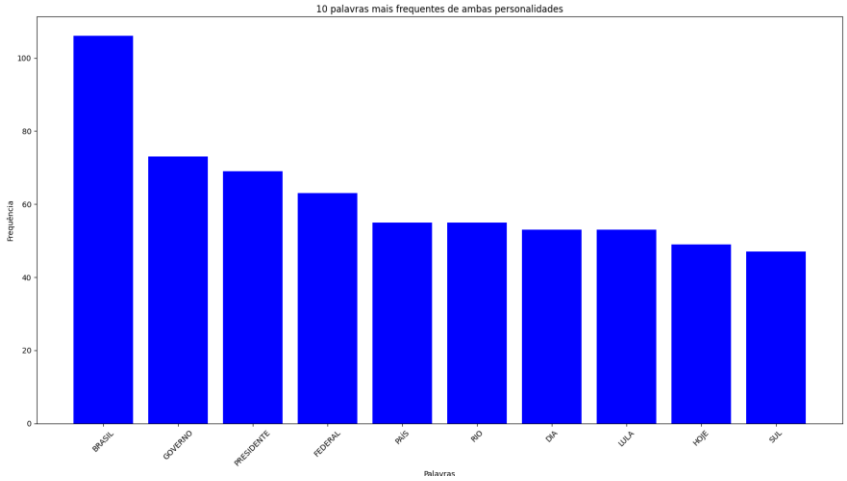
Figura 17 - World Cloud do Lula

Fonte: Autoria própria

A Word Cloud do Lula mostra palavras como "Brasil", "Dia", "Lula", "País", "Rio" e "Brasileiro". Isso sugere um foco em questões nacionais e em sua própria figura, reforçando a conexão com o público brasileiro e os eventos do dia a dia.

A análise de frequência de palavras nos tweets fornece uma visão mais quantitativa dos termos mais utilizados por cada personalidade.

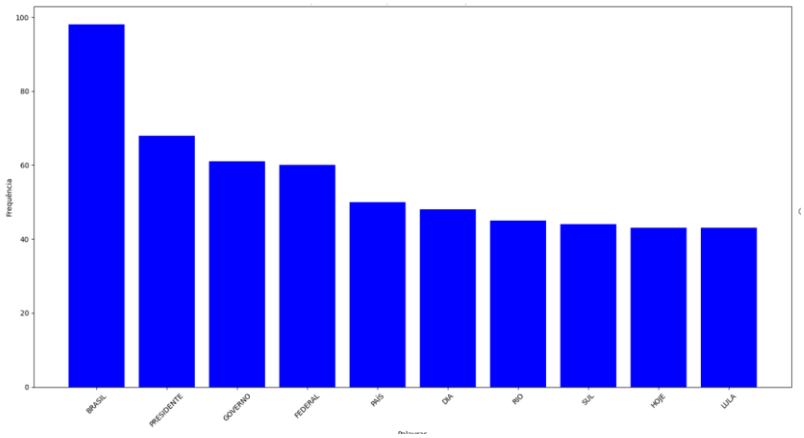
Figura 18 – Dez palavras mais frequentes de ambas personalidades



Fonte: Autoria própria

A análise conjunta das palavras mais frequentes em tweets de Lula e Bolsonaro destaca "Brasil", "Governo", "Presidente", "Federal", "País", "Rio", "dia", "lula", "Hoje" e "Sul". Estes termos refletem temas governamentais e nacionais comuns a ambos os perfis.

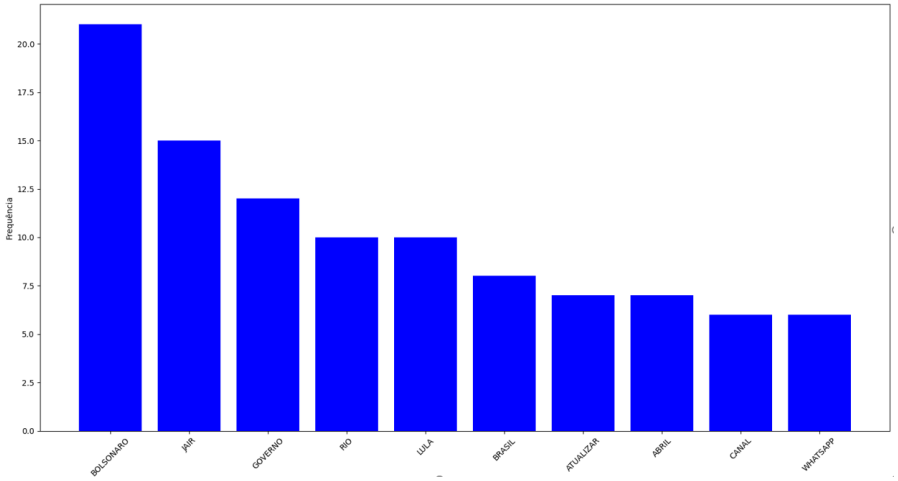
Figura 19 - Dez palavras frequentes do Lula



Fonte: Autoria própria

No caso de Lula, as palavras mais frequentes são "Brasil", "Presidente", "Governo", "Federal", "Dia", "Rio", "Sul", "Hoje" e "Lula". Isso confirma o foco de Lula em questões de governo, seu próprio papel e temas de interesse nacional.

Figura 20 - Dez palavras mais frequentes do Jair Bolsonaro



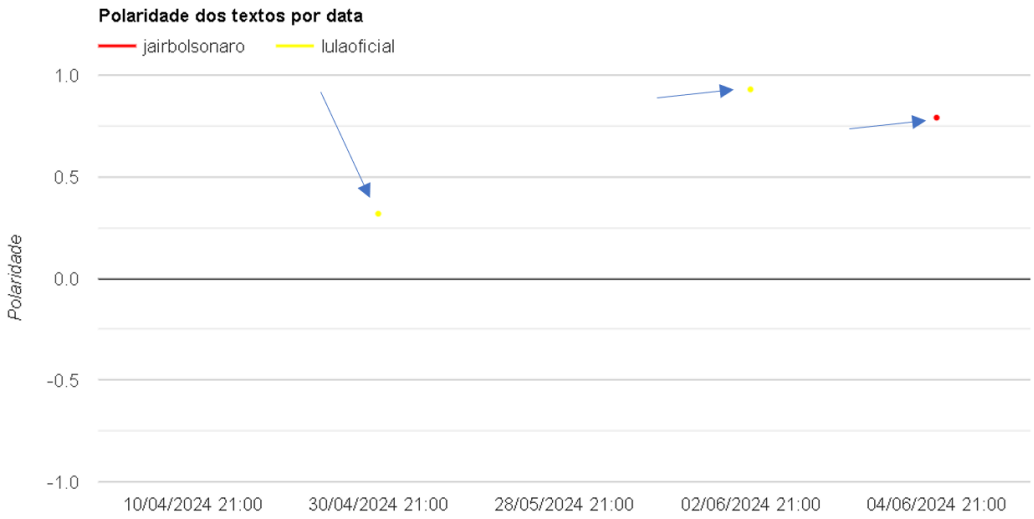
Fonte: Autoria própria

Para Jair Bolsonaro, as palavras mais frequentes incluem "Bolsonaro", "Jair", "Governo", "Rio", "Lula", "Brasil", "Atualizada", "abril", "Canal" e "WhatsApp". A ênfase em seu próprio nome, assim como em termos como "atualizada", "canal" e "WhatsApp", sugere uma comunicação direta e atualizações frequentes através de plataformas digitais.

5.10.2. Visualização dos dados na ferramenta

Através da ferramenta é possível a visualização dos gráficos que demonstram a evolução da polaridade em relação a linha do tempo, dentre os diversos lemas existentes na ferramenta, optou-se por incluir no relatório gráficos em que ambos os autores tivessem apresentado opinião. O seguinte gráfico foi gerado na ferramenta para os tweets onde há ocorrência do lema “educação”:

Figura 21 - Comparação polaridade em educação



Fonte: Autoria própria

A imagem acima foi retirada da ferramenta “análise de polaridade por autor” demonstra a polaridade por autor dos tweets processados pelo sistema através da biblioteca NLTK. Abaixo uma tabela com o conteúdo do tweet e a análise possível a partir dessas informações.

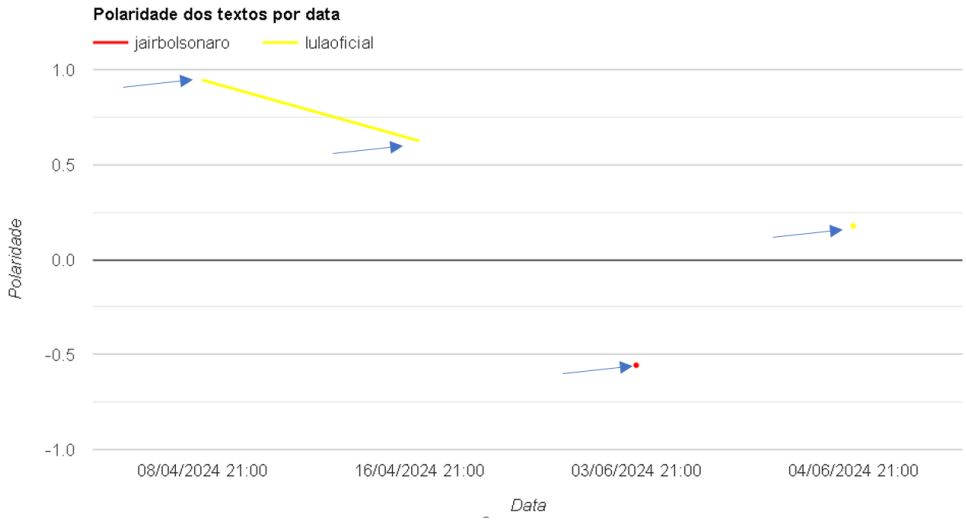
Autor	Data	Pol.	Conteúdo
lulaoficial	30/04/2024	0.3182	A Secom, chefiada pelo companheiro @Pimenta13Br, criou um site chamado ComunicaBR. Lá vocês podem ver tudo o que o governo federal tem feito, em cada canto por menor que seja. Investimentos em saúde,

			educação, cultura, segurança, agricultura, por exemplo. Acessem e mandem para as
lulaoficial	02/06/2024	0.9295	Hoje é aniversário do ministro da Educação @CamiloSantanaCE. Desejo muita força e saúde para continuar o trabalho pela educação brasileira, após tantos avanços do Ceará nessa área, e por um Brasil mais justo, desenvolvido e solidário. Forte abraço. @ricardostuckert
jairbolsonaro	04/06/2024	0.7906	- Criamos a histórica Secretaria Especial de Alfabetização, dissolvida pela gestão lula, concedemos aumento histórico no Piso Nacional dos Professores da Educação Básica de 33,24%, e nossa bancada, no Congresso Nacional segue o trabalho para valorização da educação

Puramente através do gráfico podemos notar que Lula e Bolsonaro demonstraram polaridades positivas quando o tema foi educação no período analisado.

Visualização do gráfico nos tweets onde há ocorrência do lema “Amazônia”.

Figura 22 - Comparação polaridade Amazônia



Fonte: Autoria própria

Abaixo a tabela com os tweets e seus conteúdos:

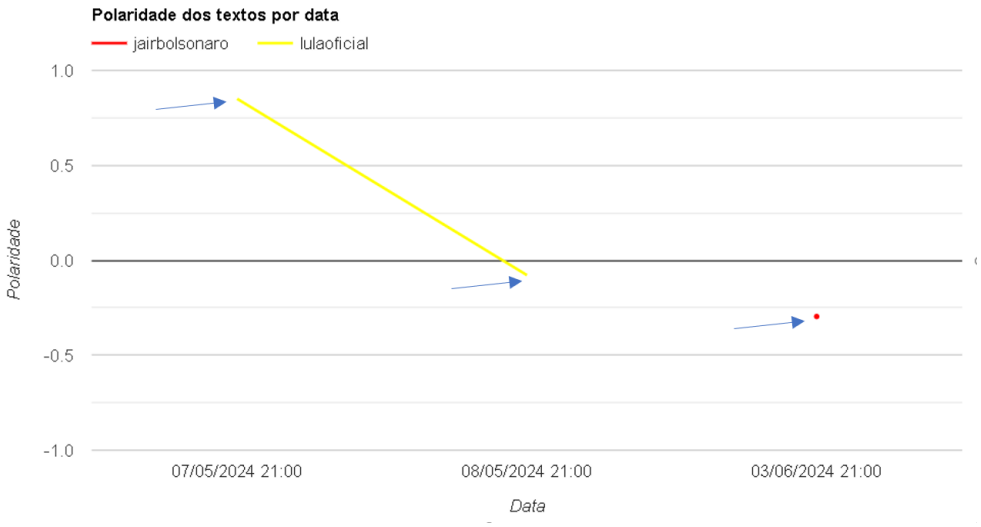
Autor	**Data**	**Pol.**	**Conteúdo**

jairbolsonaro	03/06/2024	-0.5562	Canção Pra Amazônia [Nando Reis] Maior floresta tropical da terra A toda hora sofre um duro golpe Contra trator, corrente, motosserra A bela flora clama em vão: Me poupe! [Diogo Nogueira] Porém, tem uma gente surda e cega Para a beleza e o valor da mata Embora o mundo grite que
lulaoficial	08/04/2024	0.9459	Aqui estamos cumprindo a nossa tarefa de cuidar da Amazônia, uma das coisas mais importantes do planeta. Cuidar da Amazônia significa cuidar da vida. Cuidar dos povos indígenas, dos pescadores, dos seringueiros. Das pessoas que vivem lá. O lançamento do Programa União com os
lulaoficial	16/04/2024	0.6249	A vocação para unir o Caribe, o Pacífico e a Amazônia tornam a Colômbia um sócio indispensável. Estamos bem posicionados para fazer frente ao imperativo da transição ecológica e da reindustrialização de nossas economias. A nova política industrial brasileira se propõe a aumentar a
lulaoficial	04/06/2024	0.1779	Dentre os 14 novos decretos e atos assinados neste Dia Mundial do Meio Ambiente pelo governo federal estão: - Assinatura do Pacto pela Prevenção e Controle de Incêndios com governadores do Pantanal e da Amazônia, para planejamento e implementação de ações colaborativas e
jairbolsonaro	25/04/2024	0.0772	- Lula regulamenta seu programa: *"Mais Impostos e mais ESG para Todos."* - O Projeto de Lei Complementar de regulamentação da Reforma Tributária, apresentado pelo PT, tem *Quase 500 artigos* e *360 páginas*. - E ainda virão pelo menos outros 2 PLs. E tem gente que ainda diz

Analisando o gráfico e a tabela vemos uma divergência na polaridade entre os autores, bem como uma alteração de muito positivo para quase neutro do autor Lula durante a linha do tempo.

Através da ferramenta é possível a visualização do seguinte gráfico nos tweets onde há ocorrência do lema “fake”.

Figura 23 - Comparação polaridade "Fake"



Fonte: Autoria própria

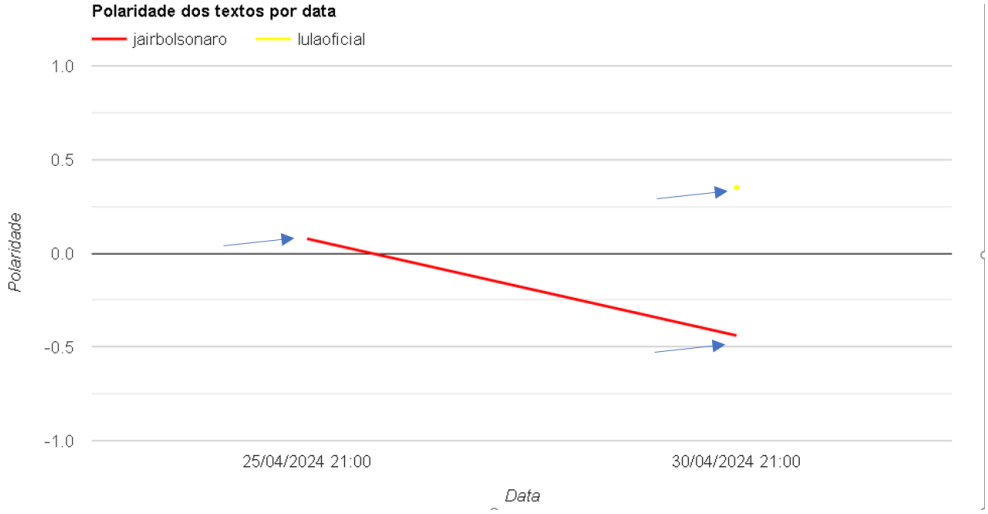
Abaixo a tabela com os tweets e seus conteúdos:

Autor	Data	Pol.	Conteúdo
lulaoficial	08/05/2024	-0.0781	O povo brasileiro é muito solidário e está se ajudando no Rio Grande do Sul. A solidariedade é uma coisa excepcional. Mas sabemos também que tem gente que aproveita tragédias como essas para espalhar fake news e atrapalhar as autoridades e voluntários que estão salvando vidas.
jairbolsonaro	03/06/2024	-0.296	- Frente fakenews diárias e permitidas por quem enche a boca alegando combatê-las, compartilhe a verdade:
lulaoficial	07/05/2024	0.8502	Brasil unido pelo Rio Grande do Sul. Sem fake news, com solidariedade e apoio ao povo gaúcho. #EquipeLula

Este gráfico mais uma vez demonstra uma queda abrupta na polaridade na linha do tempo do autor Lula, enquanto há apenas uma referência ao autor Bolsonaro que é negativa.

Visualização do seguinte gráfico nos tweets onde há ocorrência do lema “tributária”.

Figura 24 - Comparação polaridade tributária



Fonte: Autoria própria

Abaixo a tabela com os tweets e seus conteúdos:

Autor	Data	Pol.	Conteúdo
lulaoficial	30/04/2024	0.3506	Uma parte da imprensa insiste em dizer que existe uma relação ruim entre o Executivo e o Legislativo. Mas, até agora, todos os projetos que enviamos para o parlamento têm sido aprovados no Congresso Nacional, como a Reforma Tributária.
jairbolsonaro	30/04/2024	-0.4391	- Da Reforma Tributária em seu estágio atual temos apenas a certeza do seu imenso custo fiscal. - O PLP 68/24 enviado ao Congresso é longo, complexo, e sofrerá mudanças que podem torná-lo ainda mais confuso. Várias dessas mudanças serão questionadas na justiça gerando incerteza
jairbolsonaro	25/04/2024	0.0772	- Lula regulamenta seu programa: *"Mais Impostos e mais ESG para Todos."* - O Projeto de Lei Complementar de regulamentação da Reforma Tributária, apresentado pelo PT, tem *Quase 500 artigos*

			e *360 páginas*. - E ainda virão pelo menos outros 2 PLs. E tem gente que ainda diz

Desta vez temos uma alteração no gráfico onde o autor Bolsonaro saiu de neutro para negativo, enquanto a apenas um registro de polaridade de Lula.

5.11. Testes automatizados de software

O processo de testar um produto para verificar se ele atende às especificações e funciona corretamente no ambiente para o qual foi projetado é conhecido como teste de software. O objetivo do teste de software automatizado é mapear o maior número de falhas possíveis que uma solução pode apresentar e implementar estes cenários em um sistema “paralelo” que é responsável por testar o sistema principal. Este é caracterizado como uma das etapas mais custosas de um projeto de TI e por isso é comumente evitado dentre as fábricas de software.

Realizou-se neste projeto a implementação de testes de software na camada de serviços que atende a interface do usuário conforme abordado anteriormente, esta camada é responsável por implementar os casos de uso da ferramenta. Existem inúmeras tecnologias para testes de software para código JavaScript, por exemplo: Mocha[15], Chai[16], Cypress[17]. Para este projeto optou-se pela utilização do Jest[18].

[15] https://mochajs.org/
[16] https://www.chaijs.com/
[17] https://www.cypress.io/
[18] https://jestjs.io/pt-BR/

Jest é uma ferramenta de teste JavaScript desenvolvida pelo Facebook, especialmente para testar aplicações React[19]. Ela é amplamente utilizada devido à sua configuração simples e funcionalidade robusta. Dentre os recursos ele possui: suporte para mocks, snapshots, testes assíncronos e cobertura de código integrada.

A metodologia utilizada para definição do que deve ser testado e de quais dados devem ser criados foi a “técnica estrutural”. Como dispõe-se do código fonte é possível a realização de testes de casos de uso e de lógica, obtendo assim um maior número de cenários sobre controle. No projeto obteve-se uma cobertura de 100% para as rotas que realizam integração com o banco de dados. Desta forma, garante-se que novas funcionalidades possam ser movidas para o ambiente produtivo sem interferências negativas nas funcionalidades já em produção.

6. CONCLUSÃO

Em um ambiente controlado de comunicação oficial/formal, considerou-se que os jargões de internet e comunicações indiretas não estariam presentes na nossa base de avaliação. Assim, julgou-se adequado a utilização da biblioteca NLTK com os dados lematizados pelo spaCy. Além disso, é importante notar que os tweets são frequentemente gerenciados por agências e equipes de marketing, treinadas para influenciar a percepção pública. As opiniões expressas nos tweets podem não refletir

[19] React é um framework JavaScript criado pela Meta. Usado para criar interfaces de usuário. Mais informações podem ser obtidas em: https://react.dev/.

necessariamente os pensamentos dos autores, como evidenciado pelos tweets do Lula que contêm a hashtag #EquipeLula.

A análise exploratória dos tweets de Lula e Bolsonaro durante o período de dois meses revelou diferenças na atividade e no foco de suas comunicações. Lula mostrou-se mais ativo no Twitter, com um maior número de publicações, e ambos os ex-presidentes destacaram temas nacionais e governamentais em seus tweets. Nem sempre a polaridade obtida indica que o usuário sente algo positivo em relação ao tema principal. Por exemplo, em um dos tweets de Lula relacionado à educação, observamos que ele não está celebrando diretamente a educação, mas sim o aniversário do Ministro da Educação. Portanto, mudanças na polaridade dos tweets não indicam necessariamente uma mudança de ideia por parte dos políticos, e mesmo que houvesse uma mudança, isso não deve ser considerado negativo.

Um dos desafios encontrados foi a obtenção de dados devido à falta de acesso à API do Twitter, que agora é paga.

A análise de polaridade e comunicação no Twitter fornece uma visão valiosa das estratégias de comunicação e das percepções públicas dos atores políticos. A capacidade de interpretar corretamente a polaridade e o conteúdo dos tweets é crucial para uma compreensão mais profunda do impacto das redes sociais na política e na opinião pública. A implementação e adaptação de tecnologias mais avançadas podem melhorar ainda mais a precisão e a relevância dessas análises.

REFERÊNCIAS

1. PACETE, Luiz Gustavo. Brasil é o terceiro maior consumidor de redes sociais em todo o mundo. 2023. Disponível em: https://forbes.com.br/forbes-tech/2023/03/brasil-e-o-terceiro-pais-que-mais-consome-redes-sociais-em-todo-o-mundo/. Acesso em: 27 jun. 2023.

2. CONOVER, Michael; RATKIEWICZ, Jacob; FRANCISCO, Matthew; GONÇALVES, Bruno; FLAMMINI, Alessandro; MENCZER, Filippo. Political Polarization on Twitter. Fifth International Aaai Conference On Weblogs And Social Media. Bloomington, jul. 2011. p. 89-96.

3. CHOWDHARY, K. R. **Fundamentals of Artificial Intelligence**. [S. L.]: Springer New Delhi, 2020. 716 p. Disponível em: https://doi.org/10.1007/978-81-322-3972-7. Acesso em: 10 jun. 2024.

4. LASSEN, David S.; BROWN, Adam R. Twitter: The Electoral Connection? Social Science Computer Review: Social Science Computer Review. [S.L.], p. 420-436. nov. 2011.

5. RECUERO, Raquel; GRUZD, Anatoliy. Cascatas de Fake News Políticas: um estudo de caso no Twitter. Galaxia. São Paulo, maio 2019. p. 31-47. Disponível em: http://dx.doi.org/10.1590/1982-25542019239035. Acesso em: 28 jun. 2023.

6. Editorial G1 ELON Musk e Twitter: a cronologia da primeira negociação até a compra da rede social. [*S.l.*], 28 out. 2022. Disponível em: https://g1.globo.com/tecnologia/noticia/2022/10/28/elon-musk-e-twitter-a-cronologia-da-primeira-negociacao-ate-a-compra-da-rede-social.ghtml. Acesso em: 16 out. 2023.

7. Editorial G1. "Twitter demitiu 80% dos funcionários desde que Musk comprou a rede, diz CNBC". G1, 21 de janeiro de 2023. Disponível em:

https://g1.globo.com/tecnologia/noticia/2023/01/21/twitter-demitiu-80percent-dos-funcionarios-desde-que-musk-comprou-a-rede-diz-cnbc.ghtml. Acesso em: 08 out 2023.

8. FIGUEIREDO, Ana Luiza. "Twitter deixará de oferecer acesso gratuito à sua API: entenda". Olhar Digital, 02 de fevereiro de 2023. Disponível em: https://olhardigital.com.br/2023/02/02/internet-e-redes-sociais/twitter-deixara-de-oferecer-acesso-gratuito-a-sua-api-entenda/. Acesso em: 08 de outubro de 2023.

9. ZHAO, Bo. Web Scrapping. In: Encyclopedia of Big Data. Springer International Publishing, 2017, pp. 2-3.

10. LIDDY, Elizabeth Duross. Encyclopedia of Library and Information Science. 2. ed. Nova York: Marcel Decker, Inc, 2001. 13 p. Disponível em: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub. Acesso em: 02 out. 2023.

11. MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. **Sentiment analysis algorithms and applications**: a survey. 5. ed. Cairo: Elsevier B.V., 2014. 21 p. Disponível em: https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0005. Acesso em: 10 jun. 2024.

12. BENEVENUTO, Fabrício; GONÇALVES, Pollyanna; ALMEIDA, Virgílio. O Que Tweets Contendo Emoticons Podem Revelar Sobre Sentimentos Coletivos? In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2., 2013, Maceió. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013. p. 128-139. ISSN 2595-6094.

13. WEBSTER, Jonathan J.; KIT, Chunyu. TOKENIZATION AS THE INITIAL PHASE IN NLP. Hong Kong: City Polytechnic Of Hong Kong, 1992. 5 p.

14. MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. An Introduction to Information Retrieval**.** Cambridge: Cambridge University Press, 2009. 581 p.

15. GOMES PESSANHA, G. R.; OLIVEIRA FIDELIS, T.; DOURADO FREIRE, C.; ALMEIDA SOARES, E. #FIQUEEMCASA: ANÁLISE DE SENTIMENTO DOS USUÁRIOS DO TWITTER EM RELAÇÃO AO COVID19. HOLOS, [S. l.], v. 5, p. 1–20, 2020. DOI: 10.15628/holos.2020.11147. Disponível em: https://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/11147. Acesso em: 28 jun. 2023.

16. Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An “Infodemic”: Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the COVID-19 Outbreak.Open Forum Infectious Diseases. Disponível em: https://www.medrxiv.org/content/10.1101/2020.04.03.20052936v1.doi: 10.1093/ofid/ofaa258. Acesso em: 23 nov. 2023.

17. MARQUES; ANDRADE; ARTHUR; FREIRE. Quem me representa? Disponível em: https://qmrepresenta.com.br/. Acesso em: 18 jul. 2023.

18. WASSERMAN, S. e FAUST, K. Social Network Analysis. Methods and Applications. Cambridge, UK: Cambridge University Press, 1994.

19. DEGENNE, A. & FORSÉ, M. Introducing Social Networks. Sage: London, 1999

20. MUSK, Elon. Elon Musk. [S. L.], 1 jul. 2023. Elon Musk. Twitter: @elonmusk. Disponível em: https://twitter.com/elonmusk. Acesso em: 10 jun. 2024.

21. Ranking dos Políticos. 2024. Disponível em: https://www.politicos.org.br/. Acesso em: 01 dez. 2023.

22. NLTK. 2024. Disponível em: https://www.nltk.org/. Acesso em: 10 jun. 2024.

23. Aaker, J., & Chang, V. (2009). Obama and the Power of Social Media and Technology. Stanford Graduate School of Business Case No. M327.

APÊNDICES

APÊNDICE A – ARTIGO

Ferramenta de Comparação de Sentimentos Entre Usuários no Twitter e Aplicação a Figuras Públicas

Denys P. Gaspar

Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)

denys.gaspar@grad.ufsc.br

Abstract. *Twitter has been instrumental in contemporary political communication, directly connecting voters and politicians through tweets and retweets (Conover et al., 2011). In this context, we propose a platform that graphically displays the evolution of sentiment on diverse topics on Twitter, offering insights into public perceptions and opinion dynamics. This platform uses Natural Language Processing (NLP) techniques to analyze sentiments expressed in tweets related to specific topics.*

Resumo. *O Twitter tem sido fundamental na comunicação política contemporânea, conectando diretamente eleitores e políticos através de tweets e retweets (Conover et al., 2011). Neste contexto. propõe-se uma plataforma que exibe graficamente a evolução dos sentimentos sobre temas diversificado no Twitter, oferecendo insights sobre percepções públicas e dinâmicas de opinião. Esta plataforma utiliza técnicas de Processamento de Linguagem Natural (PLN) para analisar sentimentos expressos em tweets relacionados a tópicos específicos.*

1. Introdução

As mídias sociais têm desempenhado um papel cada vez mais significativo na formação do discurso político em todo o mundo. No Brasil, em particular, as mídias sociais têm uma ampla adesão, com aproximadamente 60% da população utilizando essas plataformas (Pacete, 2023). Dentre as várias redes sociais disponíveis, o Twitter tem se destacado como um canal de comunicação entre eleitores e políticos. O Twitter permite que os usuários compartilhem mensagens curtas de até 280 caracteres, chamadas de tweets, e interajam por meio de retweets e menções (Conover et al., 2011). A importância das mídias sociais, especialmente do Twitter, como um espaço para o debate político é evidente há alguns anos. Desde 2011, pesquisadores têm analisado a utilização do Twitter para fins políticos, levando em consideração sua capacidade de conectar pessoas e promover interações entre usuários (Lassen & Brown, 2011). Essas interações no Twitter possuem um alto potencial de viralização, o que, combinado com a formação de bolhas políticas, torna essa rede social um ambiente relevante para a atuação dos homens públicos (Recuero & Gruzd, 2019).

O processamento de linguagem natural (PLN) é um subcampo da inteligência artificial que se dedica à análise e representação computacional de idiomas humanos (Chowdhary, 2020). Essa abordagem envolve a utilização de algoritmos e métodos para processar, compreender e extrair informações de textos escritos em linguagem natural.

No contexto deste trabalho, o PLN será aplicado para realizar a análise de sentimento dos tweets relacionados à hashtag selecionada pelos usuários. A ferramenta proposta neste trabalho é uma plataforma que permite aos usuários acompanhar e analisar os sentimentos expressos pelos usuários de redes sociais, especificamente no Twitter, em relação a um tópico específico ao longo do tempo.

2. Fundamentação Teórica

Este capítulo apresenta uma análise da literatura e referências relevantes ao trabalho, fornecendo conceitos essenciais para uma compreensão do escopo do trabalho e permitir a modelagem da proposta.

2.1 Análise de Sentimentos

A análise de sentimentos (SA, Sentiment Analysis) ou mineração de opinião (OM, Opinion Mining) é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade. A entidade pode variar entre indivíduos, eventos ou tópicos, sendo mais comum encontrá-la em resenhas. A mineração de opinião extrai e analisa as opiniões das pessoas sobre uma entidade, enquanto a análise de sentimento identifica o sentimento expresso em um texto e, em seguida, o analisa. Assim, o objetivo da SA é encontrar opiniões, identificar os sentimentos expressos e classificar sua polaridade (MEDHAT et al. 2014).

Polaridade que segundo Benevenuto et al. (2013) "representa o grau de positividade e negatividade de um texto", pode ser expressa de forma binária ou ternária, variando do método de medição e a tecnologia aplicada. A polaridade pode ser calculada de diversas formas, seja de forma humanizada através de classificação manual até RNNs, redes neurais recorrentes, neste trabalho serão utilizadas para o cálculo da polaridade dos tweets as bibliotecas NLTK e spaCy.

O Natural Language Toolkit (NLTK) é uma biblioteca de código aberto em Python amplamente reconhecida e utilizada para o processamento de linguagem natural (PLN). Com um conjunto de ferramentas e recursos. Entre as funcionalidades oferecidas pelo NLTK estão a tokenização, conforme será definida abaixo; a marcação POS (part-of-speech), que atribui a cada palavra uma classe gramatical, como substantivo, verbo ou adjetivo; a análise de sentimentos, que avalia a polaridade emocional do texto; a lematização, que reduz as palavras à sua forma base ou lema; e o stemming, que remove os sufixos das palavras para simplificá-las NLTK (2029).

VADER (Valence Aware Dictionary and sEntiment Reasoner) é um modelo baseado em regras desenvolvido para análise de sentimentos em textos de mídias sociais. Criado por C.J. Hutto e Eric Gilbert em 2014, o VADER combina uma lista padrão de características lexicais com cinco regras gerais que consideram convenções gramaticais e sintáticas que influenciam a intensidade do sentimento. Este modelo se destaca por sua capacidade de lidar eficientemente com textos curtos e informais, comuns em plataformas como o Twitter, e por ser capaz de interpretar gírias, pontuação e uso de maiúsculas para transmitir intensidade emocional (Hutto & Gilbert, 2014).

Utilizou-se o VADER para realizar a análise de sentimentos dos tweets relacionados aos usuários selecionados. Esta escolha se deu devido à alta precisão do modelo em contextos de mídias sociais, onde ele supera avaliadores humanos individuais em termos de precisão de classificação, com uma acurácia F1 de 0.96 em

comparação com 0.84 dos humanos. A implementação do VADER permitiu uma análise robusta dos sentimentos expressos nos tweets, fornecendo insights sobre as percepções e estratégias de comunicação dos atores políticos no Twitter (Hutto & Gilbert, 2014).

2.2 Processamento de Linguagem Natural

Segundo Chowdhary (2020), o processamento de linguagem natural (PLN) é um subcampo da inteligência artificial, localizado nos estudos que engloba um conjunto de técnicas computacionais para análise automática e representação de idiomas humanos. A história da PLN data da década de 40, no período pós-Segunda Guerra Mundial, Weaver e Booth foram pioneiros em um dos primeiros projetos de Tradução Automática (TA) em 1946, tendo suas contribuições amplamente reconhecidas como inspiradoras para o campo (Liddy, 2001). O PLN envolve o uso de algoritmos e métodos para processar, compreender e extrair informações de textos escritos em linguagem natural. Para o autor, acesso e aquisição de características lexicais, semânticas e episódicas, identificação de construções básicas de linguagem (por exemplo, objetos e ações), representação de conceitos abstratos são alguns dos recursos obrigatórios para uma PLN de alto nível. Quanto à sua utilização, esta pode variar em diversos sentidos:

- Indexação e pesquisa de textos grandes;
- Recuperação de informações (IR);
- Classificação de texto em categorias;
- Extração de informações (IE);
- Tradução automática de idiomas;
- Resumo automático de textos;
- Resposta a perguntas (QA);
- Aquisição de conhecimento;
- Geração de textos/diálogos.

No contexto deste trabalho o PLN será utilizado para extração de informações através de técnicas de análises sintática e semânticas, conforme descrito a seguir.

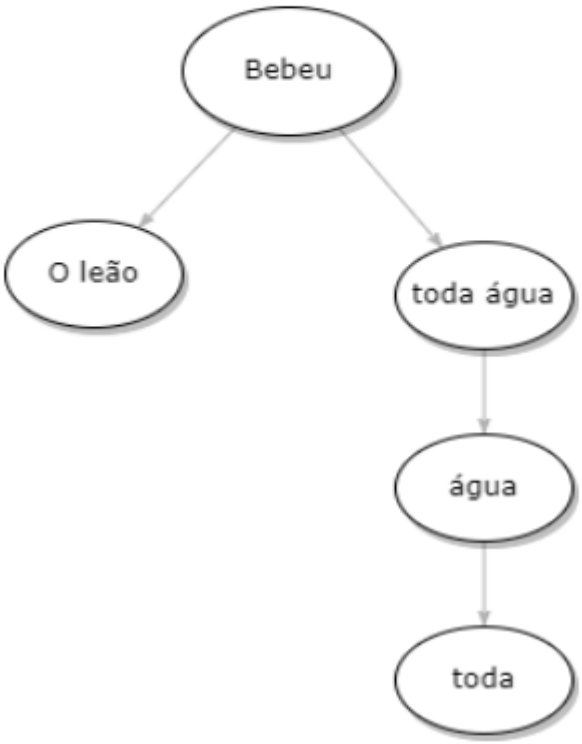
2.3. Análises Sintáticas e Semânticas

Se tratando de processamento de linguagem natural a análise sintática é responsável, tal qual no estudo linguístico de idiomas, das aplicações de regras gramaticais para análise de estrutura das frases. E esta, segundo Chowdhary (2020), pode ser dividida em duas funções: 1) Determinação da estrutura; 2) Regularização da estrutura de sintaxe;

A primeira função é a determinação da estrutura, na qual a análise de sintaxe identifica o sujeito, o objeto, as palavras modificadoras e a frase que cada palavra modifica, atribuindo uma estrutura de árvore durante o processo de análise.

Exemplo de análise sintática para a frase “O leão bebeu toda água.”:

- Sujeito da frase: "O leão", onde, "O" é artigo definido e "leão" exerce a função de substantivo.
- Verbo: "bebeu";
- Objeto direto: "toda água".



Fonte: Autoria própria

A segunda função é a regularização da estrutura de sintaxe, que simplifica o rocessamento subsequente, especialmente a análise semântica, ao mapear possíveis ntenças de entrada em um número reduzido de estruturas. Chowdhary exemplifica que rtos elementos nas frases podem ser omitidos, mantendo o significado das frases tacto.

Para o autor o objetivo da análise semântica é determinar o significado de uma ase, buscando as condições em que ela é verdadeira. Além disso, a análise semântica racteriza as regras de inferência para as sentenças da linguagem. No entanto, a racterização da semântica das sentenças interrogativas e imperativas apresenta maior mplexidade. O autor destaca a importância do significado de uma frase, mesmo ando esta está correta ou sintaticamente correta. Ao criar um sistema de linguagem tural, o objetivo é que o sistema execute alguma ação em resposta à entrada, como cuperar dados ou realizar comandos em um robô. Nesse sentido, a tradução da nguagem natural para linguagens formais, como sistemas de recuperação de banco de dos ou sistemas de comando de robô, é necessária. Essas linguagens formais possuem

propriedades não ambíguas, regras simples de interpretação e inferência, além de uma estrutura lógica determinada pela forma da sentença.

2.4. Mídias Sociais

As mídias sociais desempenham um papel importante na formação do discurso político em todo o mundo (Conover et al, 2011), e segundo Pacete (2023) publicado na revista Forbes "o levantamento da Comscore mostra que o país (Brasil) é o primeiro da América Latina em acesso às plataformas, o equivalente a 131,5 milhões de pessoas". Este indicativo de ampla adesão das redes sociais no Brasil, aproximadamente 60% da população, reforça o impacto das mídias no discurso público. E atualmente dentre as redes sociais o Twitter possui o papel de conectar eleitores e políticos. Diferentemente de outras plataformas de comunicação no Twitter existe uma via de mão dupla entre os usuários, cada perfil é locutor e interlocutor e as mensagens possuem um alto potencial de viralização (Lassen; Brown. 2011) que agregado com a construção de bolhas políticas tornam a rede social o habitat dos homens públicos). Portanto, nota-se a relevância das mídias sociais e em especial o Twitter que vem sendo analisado desde 2011 frente sua ampla utilização para fins políticos.

O impacto das mídias sociais na internet começou a ser notado nos Estados Unidos nas eleições de 2008, onde especialistas creditam a eleição do presidente Barak Hussein Obama ao trabalho da sua equipe de marketing. Naquele momento, a rede social mais utilizada pelos americanos era o Facebook, com 21 milhões de contas criadas. Além da divulgação dos objetivos da campanha de Obama, foi realizada uma campanha de vaquinha online para arrecadação de dinheiro para a campanha.

O texto "Obama and the Power of Social Media and Technology", escrito por Jennifer Aaker e Victoria Chang em 2009, analisa a estratégia de campanha de Barack Obama nas eleições presidenciais de 2008. Durante essa campanha, a equipe de Obama adotou uma abordagem inovadora, utilizando mídias sociais e tecnologia para arrecadar fundos e mobilizar voluntários.

De acordo com Aaker e Chang (2009), a campanha online de Obama foi fundamental para a eleição. Destacou-se que "Obama utilizou as mídias sociais de maneira eficaz para se conectar com os eleitores, mobilizar voluntários e arrecadar fundos". Essa estratégia permitiu que a campanha alcançasse um amplo público e criasse um movimento de apoio. Em termos de seguidores, Obama superou seu concorrente, John McCain, nas redes sociais. Ele tinha milhões de seguidores no Twitter, Facebook e Youtube, enquanto McCain tinha uma presença online menos expressiva. Segundo os autores essa diferença de alcance e engajamento foi crucial para a vitória de Obama.

3. Trabalhos Correlatos

A área de estudos e análises relacionados às mídias sociais e seu impacto na sociedade é ampla e conta com a contribuição de diversos autores e grupos. Nesta seção, serão apresentadas breves descrições de trabalhos que se dedicam a promover a transparência do poder público, explorar a integração com as mídias sociais e realizar análise de sentimentos. Esses estudos têm como objetivo aprofundar nosso entendimento sobre o papel das mídias sociais na sociedade e suas implicações, contribuindo para o avanço do conhecimento nessa área.

3.1. Quem Me Representa?

A página "Quem me representa?" é uma ferramenta online que permite aos usuários descobrir quais políticos e representantes estão atuando de acordo com suas opiniões (Marques et al., 2023). O objetivo do site é facilitar o acesso à informação sobre os representantes políticos eleitos em diferentes esferas governamentais, como vereadores, deputados estaduais, deputados federais e senadores.

Ao visitar o site, os usuários podem inserir seu endereço para obter uma lista dos políticos que os representam em cada nível de governo. A plataforma também fornece informações sobre o histórico político dos representantes, incluindo partido político, mandatos anteriores, projetos de lei de sua autoria e posicionamentos públicos.

A página visa promover a transparência e a participação cidadã, permitindo que os eleitores conheçam seus representantes e acompanhem suas atividades. Dessa forma, os cidadãos podem ter maior conhecimento sobre quem está tomando decisões políticas em seu nome e buscar uma participação mais informada na política. Pode ser acessado através do endereço: qmrepresenta.com.br.

3.2. Ranking dos Políticos

A página "Ranking dos Políticos" é uma plataforma online que tem como objetivo avaliar e classificar o desempenho dos políticos brasileiros (Ranking dos Políticos, 2024). Segundo os autores do site, a página busca fornecer informações aos cidadãos sobre o trabalho dos políticos eleitos, promovendo a transparência e a prestação de contas.

No site do Ranking dos Políticos, disponível em politicos.org.br, os usuários podem encontrar rankings atualizados dos políticos brasileiros, classificados de acordo com critérios como a presença nas sessões legislativas, a qualidade de seus projetos de lei e o uso consciente de recursos públicos. A página também disponibiliza informações sobre processos judiciais e condenações de políticos.

Para calcular a pontuação final dos políticos, são aplicados os seguintes pesos:

Votações: 3x

Gastos: 2÷

Processos Judiciais: quando há processos, subtrai pontos da nota final

Outros: soma ou subtrai pontos da nota final

Fórmula: [(V x 3) + (G / 3)] / 4 + P + OT = Nota Final

Onde:

V = Votações

G = Gastos

P = Processos Judiciais

OT = Outros

Exemplo prático para o deputado fictício João da Silva:

Votações: Fez 60 pontos em 100 possíveis nas votações = nota 6,0

Gastos: Teve 90 presenças em 100 sessões deliberativas possíveis = nota 9; economizou R$ 70 mil de R$ 100 mil disponíveis para Cota Parlamentar + Verba de Gabinete = nota 7

Nota: 8,0 (média das duas notas)

Processos Judiciais: Condenado em inquérito da Operação Lava-Jato = -1,0 ponto

Outros: Aprovou o projeto de sua autoria que cria o Marco das Ferrovias = +0,5 ponto

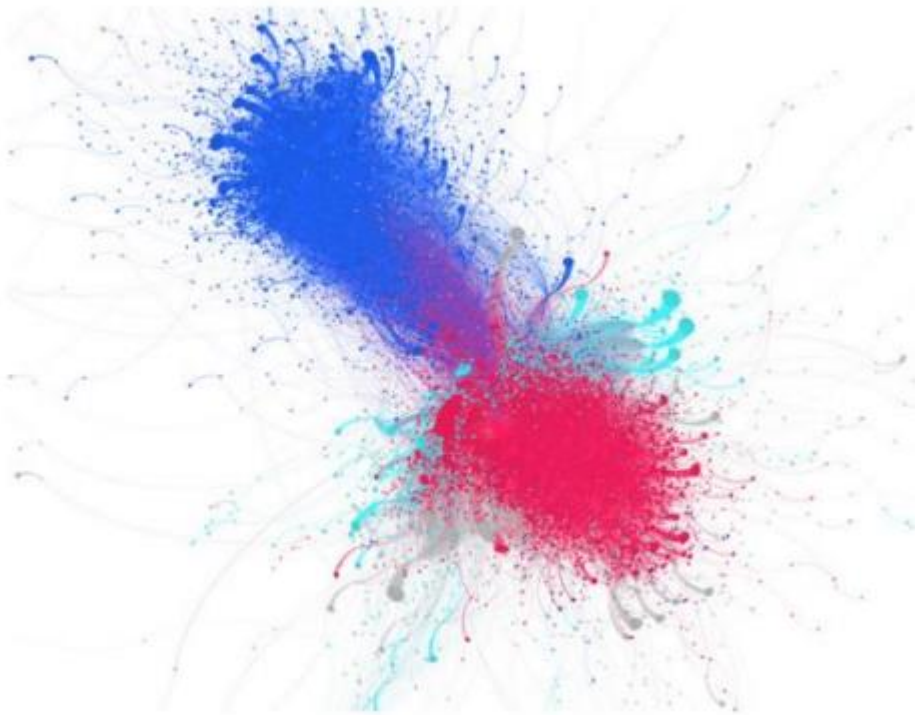
Cálculo da nota final: [(6,0 x 3) + (8,0 / 3)] / 4 - 1,0 + 0,5 = 4,66

3.3. Cascatas de Fake News Políticas: Um Estudo de Caso no Twitter

O estudo "Cascatas de Fake News Políticas: um estudo de caso no Twitter" (Recuero, 2019), teve como objetivo analisar a estrutura e difusão de "cascatas de fake news" relacionadas ao caso do julgamento e prisão do ex-presidente Lula. Para alcançar esse objetivo, a seguinte abordagem foi adotada:

Para a coleta de dados, utilizou-se um crawler com acesso à streaming API do Twitter, configurado para coletar tweets relacionados à palavra-chave "Lula". Ao final do processo de coleta, obteve-se um dataset inicial contendo 2.430.468 tweets, coletados entre 1 de fevereiro e 26 de abril de 2018.

A etapa de preparação dos dados incluiu a filtragem utilizando palavras-chave descritivas, permitindo identificar discussões pertinentes ao caso. Com o dataset filtrado, iniciou a análise da estrutura das redes de disseminação das fake news, utilizando métricas de análise de redes sociais, baseadas nos estudos de Degenne e Forsé (1999) e Wasserman e Faust (1994). Como ferramenta de análise, utilizou-se o software Gephi. Os nós foram associados às contas no Twitter e as conexões representadas pelos retweets e menções. Essa abordagem permitiu visualizar e estudar a estrutura das redes de disseminação das fake news.



Fonte: Recuero, 2019

Os resultados sugerem que as "cascatas de fake news" políticas tendem a ficar restritas aos seus próprios clusters ideológicos. No cluster vermelho, encontram-se os nós que demonstram maior afinidade com partidos de esquerda e apoiam o ex-presidente Lula, enquanto no cluster azul, estão os que demonstram posição contrária e são mais alinhados com partidos de direita. Os nós em azul claro representam grupos não vinculados a nenhum partido.

4. Metodologia e Desenvolvimento

No início do projeto, optou-se por definir objetivos mensais e realizar encontros semanais de acompanhamento para avaliar o avanço das atividades. No entanto, ao longo do desenvolvimento, foi constatado que certas fases da implementação exigiam mais tempo do que o inicialmente estimado. Por isso, ajustamos nossa abordagem, introduzindo sessões de planejamento nas reuniões semanais, que foram realizadas remotamente via Google Meet.

Nessas reuniões semanais, definíamos objetivos específicos e realizáveis, em colaboração efetiva entre a equipe de desenvolvimento e o orientador. As reuniões ocorriam às segundas-feiras, quando estabelecíamos metas para a semana seguinte, guiados pelo cronograma inicial. As entregas eram marcadas por demonstrações do funcionamento do sistema através do compartilhamento de tela. Falhas, dúvidas e próximos passos eram anotados informalmente e discutidos no próximo encontro semanal.

4.1. Arquitetura

A arquitetura do sistema é composta por várias camadas que interagem de forma coordenada para realizar a análise de sentimentos no Twitter: interface de usuário, camada de serviço NodeJS, serviço Python de raspagem de dados e serviço Python de processamento de dados.

A interação do usuário com o sistema inicia-se ao pressionar um botão no navegador, disparando uma chamada HTTPS que invoca uma função no back-end, implementada em NodeJS com o framework Express. As rotas definidas no back-end estabelecem a conexão com o banco de dados via TCP/IP. O banco de dados é alimentado por uma função Python responsável pelo processamento dos dados cadastrados. Esses dados têm origem em outro serviço Python que processa os novos perfis de redes sociais cadastrados pelos usuários na interface do frontend.

O serviço Python de processamento de dados opera da seguinte maneira:

- Usuário cadastra os perfis de redes sociais que deseja monitorar na aplicação;
- Serviço Python lê os novos perfis cadastrados e realiza a raspagem de dados do Twitter;
- Os dados raspados são armazenados no banco de dados em uma tabela específica, sem processamento inicial de polaridade e lematização.

Posteriormente, outro serviço Python lê os dados armazenados e realiza o processamento de polaridade e lematização, armazenando-os na tabela final do banco de dados.

Um relatório é gerado a partir dos dados processados e exibido na interface do frontend para o usuário final, completando o ciclo de informações dentro da ferramenta. Esta abordagem integra eficientemente as etapas de coleta, processamento e apresentação dos dados, garantindo acesso a análises de sentimentos precisas e atualizadas.

4.2. Métricas da Raspagem de Dados

As métricas utilizadas para avaliar o desempenho do processo de raspagem de dados no Twitter, implementado com Selenium, foram escolhidas com base nas etapas necessárias para a raspagem de dados: 1) autenticação; 2) filtro pelo nome de usuário; 3) seleção do usuário; 4) raspagem dos tweets; 5) registro dos dados.

- Tempo de resposta total: Refere-se ao tempo necessário para concluir todas as etapas da raspagem de dados, desde a autenticação até o registro dos dados no banco de dados PostgreSQL. Esta métrica varia conforme o parâmetro de data programado no algoritmo de raspagem. Em uma amostragem de dez execuções para um intervalo de sessenta dias, a aplicação apresentou um tempo médio de dois minutos, calculado a partir dos registros de log.
- Taxa de sucesso de autenticação automática: Indica a proporção de tentativas de autenticação bem-sucedidas em relação ao total de tentativas realizadas. Durante o desenvolvimento, enfrentamos desafios com diferentes interfaces de autenticação, incluindo sistemas de identificação de robôs. Isso reduziu a taxa de

sucesso de autenticação automática de 100% para 0%, necessitando de intervenção humana para continuar a autenticação.

- Tempo médio de raspagem por tweet: Refere-se à média de tempo gasto para coletar os dados de cada tweet. A média foi calculada através da raspagem de tweets de um intervalo de dois meses, resultando em 2,5 tweets raspados por segundo.

A arquitetura do sistema é composta por várias camadas que interagem de forma coordenada para realizar a análise de sentimentos no Twitter, interface de usuário, camada de serviço NodeJS , serviço Python de raspagem de dados e serviço Python de processamento de dados.

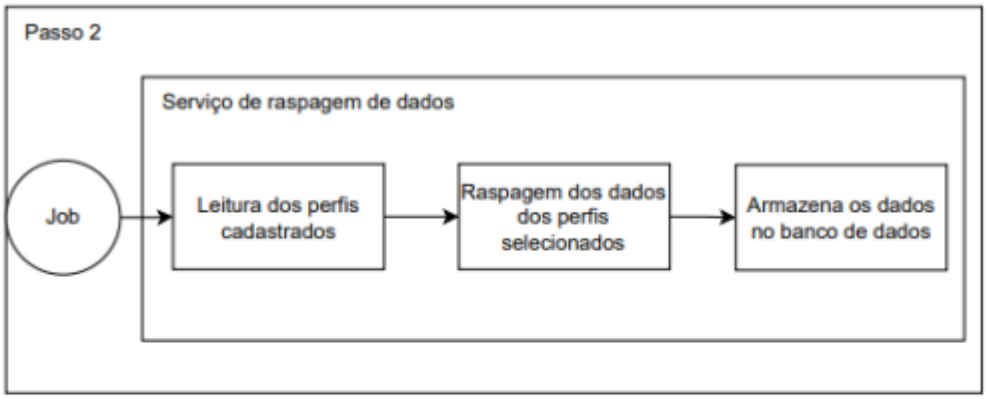
A interação do usuário com o sistema inicia-se ao pressionar um botão no navegador, o que dispara uma chamada HTTPS. Esta ação invoca uma função no back-end, implementada utilizando a tecnologia NodeJS em conjunto com o framework Express . As rotas definidas no back-end estabelecem a conexão com o banco de dados através de TCP/IP. O banco de dados é alimentado por uma função Python responsável pelo processamento dos dados cadastrados. Esses dados têm origem em outro serviço Python que processa os novos perfis de redes sociais cadastrados pelos usuários na interface do frontend.

O segundo serviço Python, de processamento de dados, opera da seguinte maneira:

- O usuário cadastra os perfis de redes sociais que deseja monitorar na aplicação.
- O serviço Python lê os novos perfis cadastrados e realiza a raspagem de dados do Twitter.
- Os dados raspados são armazenados no banco de dados em uma tabela específica, sem que sejam inicialmente processados em termos de polaridade e lematização.

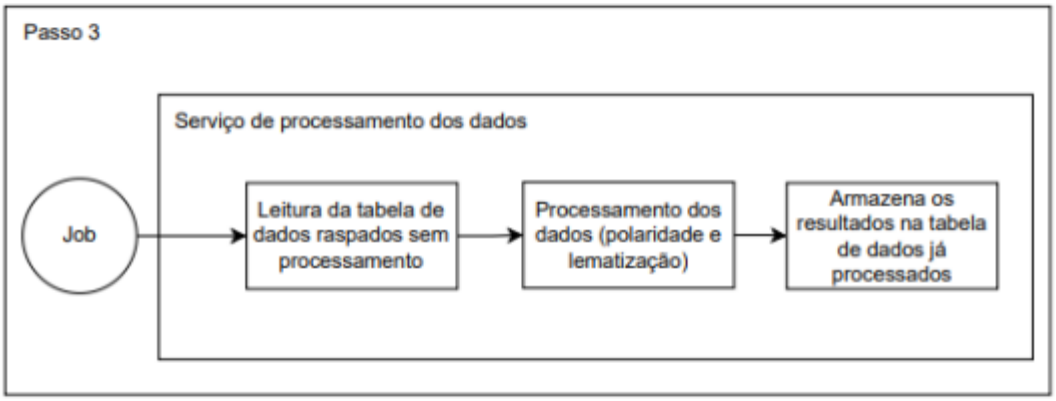
Posteriormente, o primeiro serviço Python lê os dados armazenados na tabela e realiza o processamento de polaridade e lematização. Esses dados processados são então armazenados na tabela final do banco de dados.

Um relatório é gerado a partir dos dados processados e é exibido na interface do front-end para o usuário final, completando o ciclo de informações dentro da ferramenta. Esta abordagem integra de forma eficiente as etapas de coleta, processamento e apresentação dos dados, garantindo que o usuário tenha acesso a análises de sentimentos precisas e atualizadas baseadas nos dados das redes sociais que ele próprio cadastrou na aplicação. Abaixo uma imagem para ilustrar a arquitetura da solução.



Fonte: Autoria própria

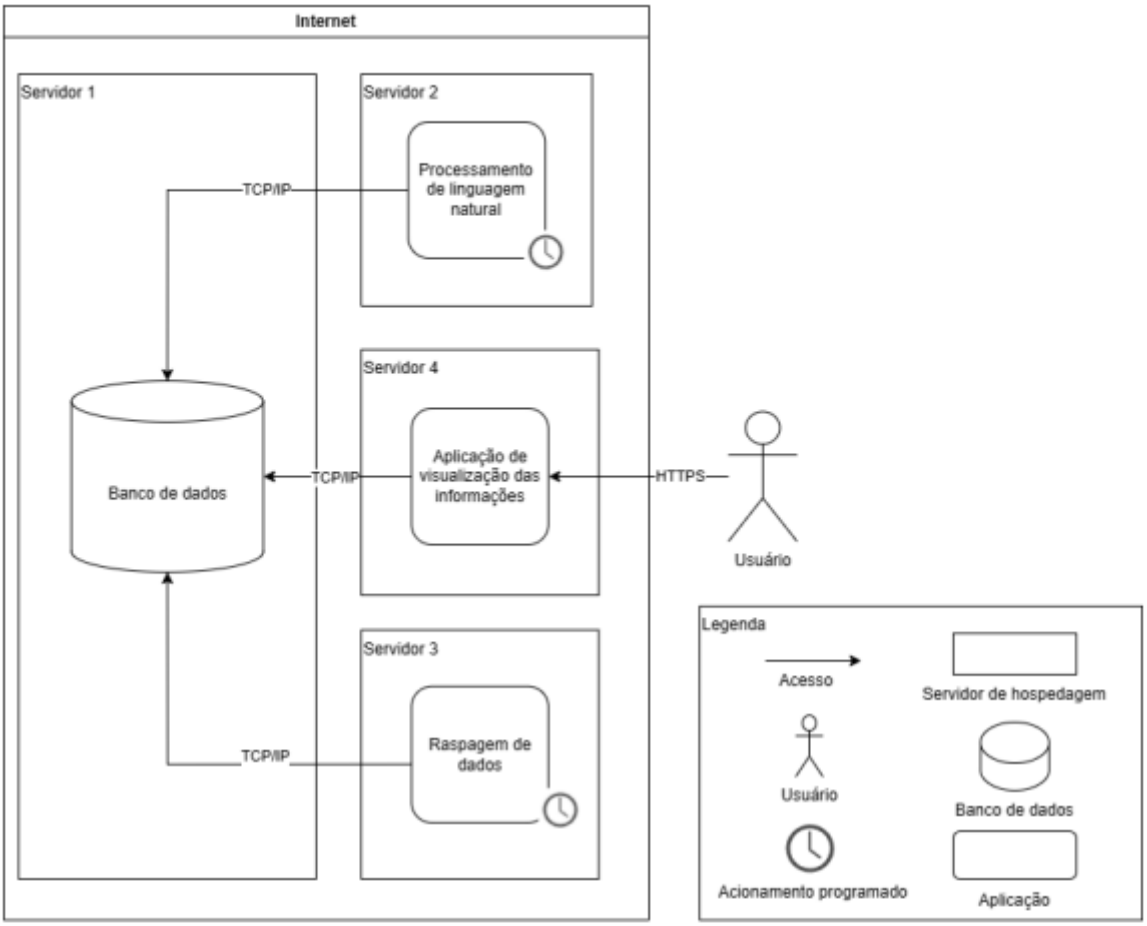
O serviço de acionamento programado realiza a chamada do programa Python responsável pela seleção dos novos dados, raspagem da rede social e armazenamento.



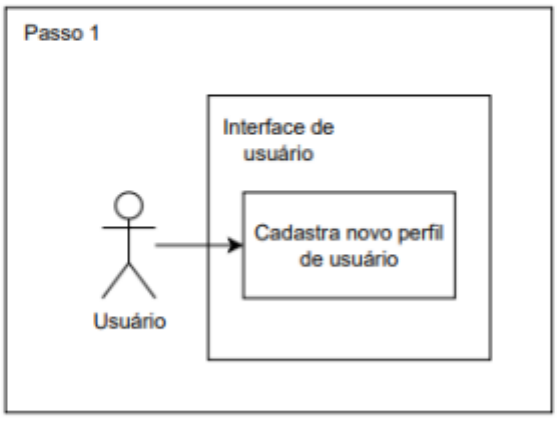
Fonte: Autoria própria

Para utilização dos serviços, é necessário instalar as bibliotecas do Selenium selecionadas para o projeto e atualizar as variáveis de ambiente para direcionar o serviço ao banco de dados da ferramenta, seja na nuvem ou localmente. Após a instalação e configuração, o serviço executa os seguintes passos:

- Autenticação: Utiliza o usuário nominal informado no arquivo "authentication.txt".
- Filtro com o nome do usuário: A API recebe o nome do usuário a ser raspado e filtra os usuários do Twitter.
- Seleção do usuário: O primeiro resultado do filtro é selecionado.
- Raspagem dos dados: O sistema raspa os tweets até a data limite de três meses.



Fonte: Autoria própria



Fonte: Autoria própria

Usuário realiza o cadastramento de novos perfis de redes sociais.

- Registro dos dados: Os dados não processados são armazenados no banco de dados PostgreSQL.
- Essas etapas garantem que a ferramenta possa coletar, processar e apresentar dados de maneira eficiente e precisa.

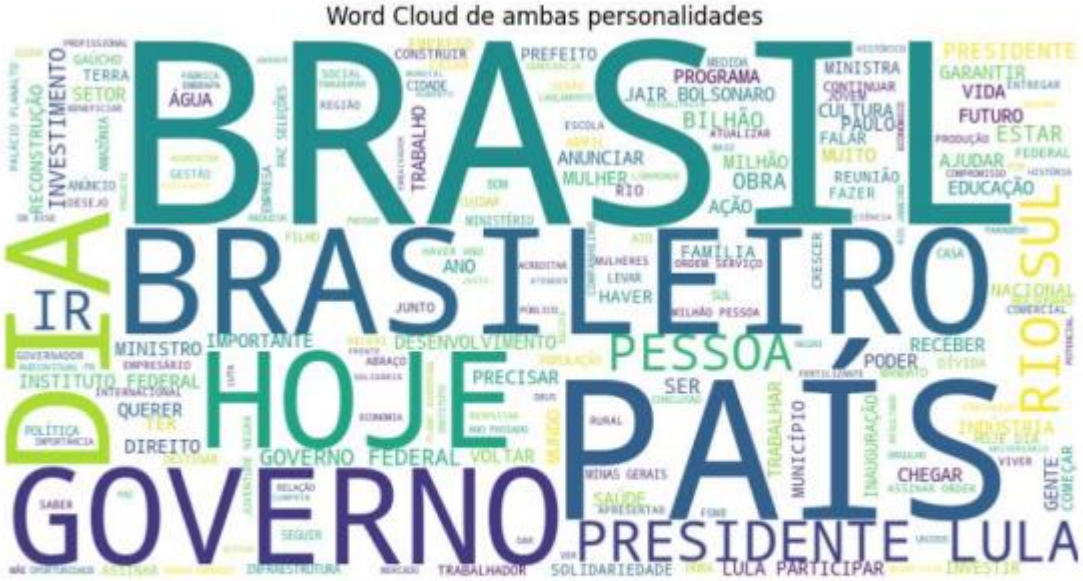
5. Resultados e Discussão

Nesta seção, será realizada uma análise exploratória na base de dados e uma demonstração dos resultados obtidos na camada da aplicação.

5.1. Análise Exploratória

Durante o período de dois meses, de 05 de abril de 2024 até 06 de junho de 2024, foram publicados um total de 222 tweets pelas contas de @lulaoficial e @jairbolsonaro. Lula foi o autor da maioria dessas publicações, com um total de 160 tweets, enquanto Bolsonaro publicou 62 tweets. Essa volumetria indica uma maior atividade de Lula no Twitter em comparação a Bolsonaro durante o período analisado.

A análise das Word Clouds para ambas as personalidades revela os termos mais frequentemente utilizados em seus tweets, proporcionando uma visão rápida dos principais tópicos abordados.



Fonte: Autoria própria

A Word Cloud conjunta de Lula e Bolsonaro destaca as palavras "Brasil", "brasileiro", "país", "governo", "hoje" e "dia" como as mais proeminentes. Esses termos sugerem uma ênfase em temas relacionados à nação e ao governo.

Word Cloud do Jair Bolsonaro

Fonte: Autoria própria

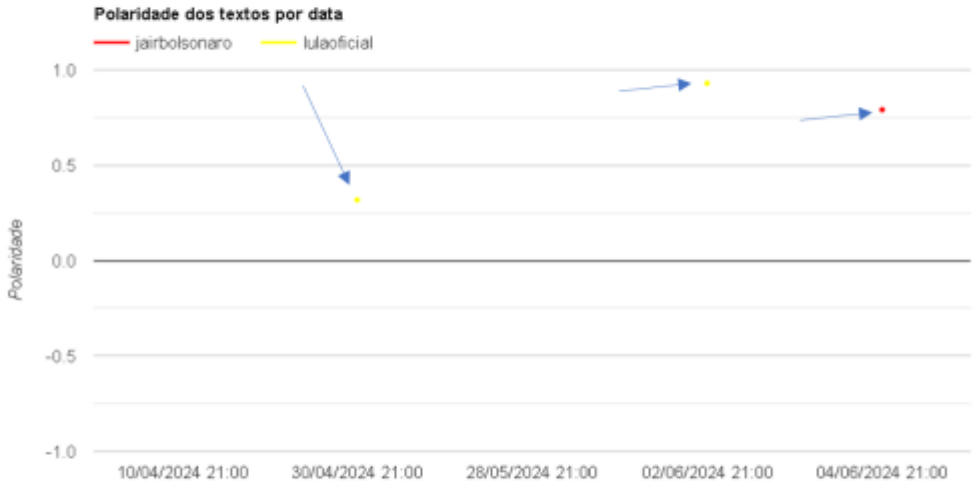
Na Word Cloud específica de Jair Bolsonaro, as palavras destacadas incluem "Gestão", "Bolsonaro", "Jair", "Brasil", "Atualizar", "Rio", "Governo", "Lula" e "Dia". A presença de seu próprio nome e variações, assim como referências a "gestão" e "atualizar", pode indicar uma narrativa focada em sua administração e atualizações sobre suas ações ou posições. A ênfase em seu próprio nome, assim como em termos como "atualizada", "canal" e "WhatsApp", sugere uma comunicação direta e atualizações frequentes através de plataformas digitais.

Word Cloud do Lula

Fonte: Autoria própria

A Word Cloud de Lula mostra palavras como "Brasil", "Dia", "Lula", "País", "Rio" e "Brasileiro". Isso sugere um foco em questões nacionais e em sua própria figura, reforçando a conexão com o público brasileiro e os eventos do dia a dia.

Através da ferramenta é possível a visualização dos gráficos que demonstram a evolução da polaridade em relação a linha do tempo, dentre os diversos lemas existentes na ferramenta, optou-se por incluir no relatório gráficos em que ambos os autores tivessem apresentado opinião. O seguinte gráfico foi gerado na ferramenta para os tweets onde há ocorrência do lema “educação”:

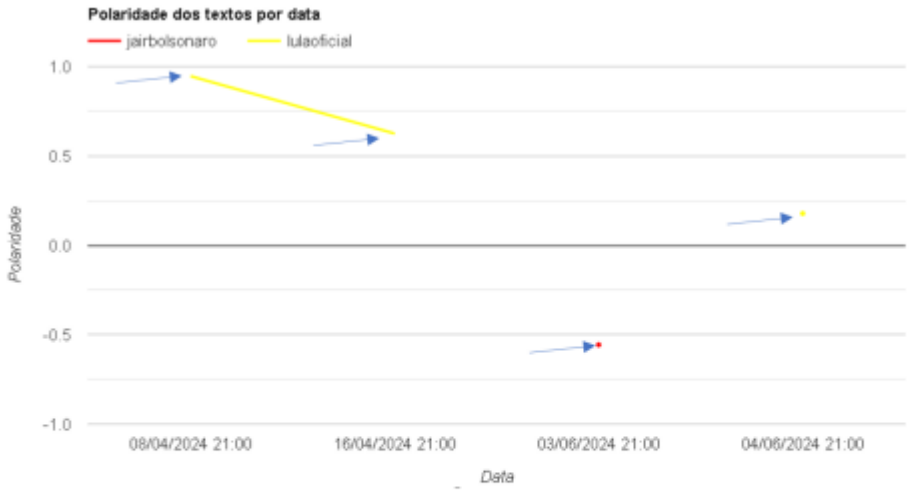


Fonte: Autoria própria

A imagem acima foi retirada da ferramenta “análise de polaridade por autor” demonstra a polaridade por autor dos tweets processados pelo sistema através da biblioteca NLTK. Abaixo uma tabela com o conteúdo do tweet e a análise possível a partir dessas informações.

Puramente através do gráfico podemos notar que Lula e Bolsonaro demonstraram polaridades positivas quando o tema foi educação no período analisado.

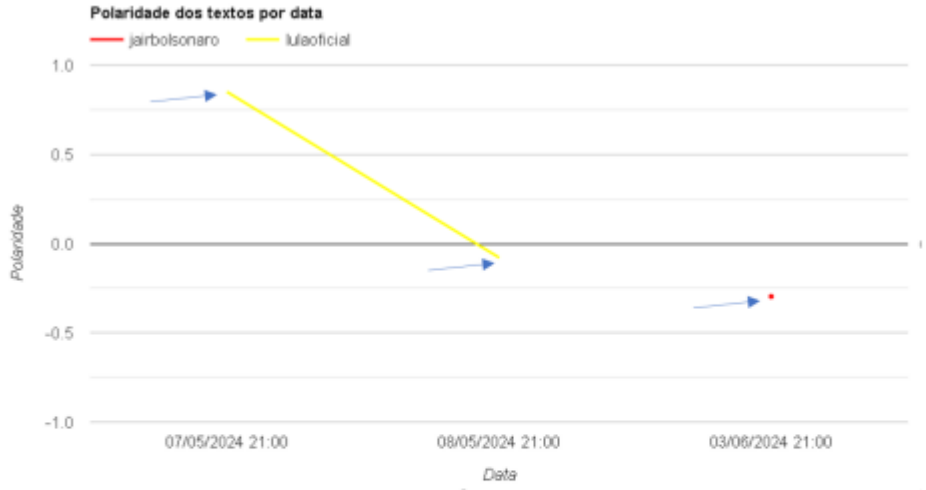
Visualização do gráfico nos tweets onde há ocorrência do lema “Amazônia”.



Fonte: Autoria própria

Analisando o gráfico e a tabela vemos uma divergência na polaridade entre os autores, bem como uma alteração de muito positivo para quase neutro do autor Lula durante a linha do tempo.

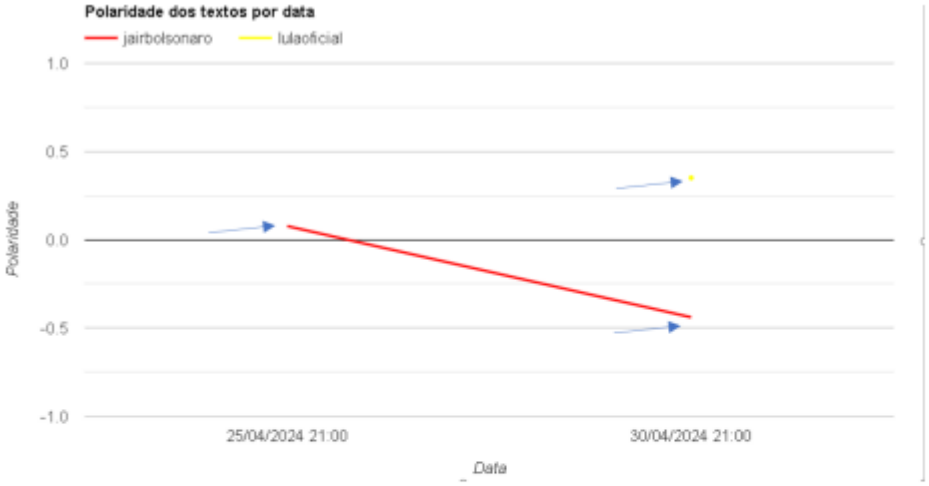
Através da ferramenta é possível a visualização do seguinte gráfico nos tweets onde há ocorrência do lema “fake”.



Fonte: Autoria própria

Este gráfico mais uma vez demonstra uma queda abrupta na polaridade na linha do tempo do autor Lula, enquanto há apenas uma referência ao autor Bolsonaro que é negativa.

Visualização do seguinte gráfico nos tweets onde há ocorrência do lema “tributária”.



Fonte: Autoria própria

Desta vez temos uma alteração no gráfico onde o autor Bolsonaro saiu de neutro para negativo, enquanto a apenas um registro de polaridade de Lula.

6. Conclusão e Trabalhos Futuros

Em um ambiente controlado de comunicação oficial/formal, considerou-se que os jargões de internet e comunicações indiretas não estariam presentes na nossa base de avaliação. Assim, julgou-se adequada a utilização da biblioteca NLTK com os dados lematizados pelo spaCy. Além disso, é importante notar que os tweets são frequentemente gerenciados por agências e equipes de marketing, treinadas para influenciar a percepção pública. As opiniões expressas nos tweets podem não refletir necessariamente os pensamentos dos autores, como evidenciado pelos tweets de Lula que contêm a hashtag #EquipeLula.

A análise exploratória dos tweets de Lula e Bolsonaro durante o período de dois meses revelou diferenças na atividade e no foco de suas comunicações. Lula mostrou-se mais ativo no Twitter, com um maior número de publicações, e ambos os ex-presidentes destacaram temas nacionais e governamentais em seus tweets. Nem sempre a polaridade obtida indica que o usuário sente algo positivo em relação ao tema principal. Por exemplo, em um dos tweets de Lula relacionado à educação, observamos que ele não está celebrando diretamente a educação, mas sim o aniversário do Ministro da Educação. Portanto, mudanças na polaridade dos tweets não indicam necessariamente uma mudança de ideia por parte dos políticos, e mesmo que houvesse uma mudança, isso não deve ser considerado negativo.

Um dos desafios encontrados foi a obtenção de dados devido à falta de acesso à API do Twitter, que agora é paga.

A análise de polaridade e comunicação no Twitter fornece uma visão valiosa das estratégias de comunicação e das percepções públicas dos atores políticos. A capacidade de interpretar corretamente a polaridade e o conteúdo dos tweets é crucial para uma compreensão mais profunda do impacto das redes sociais na política e na opinião

pública. A implementação e adaptação de tecnologias mais avançadas podem melhorar ainda mais a precisão e a relevância dessas análises.

Existem diversas possiblidades de melhorias e continuações para o trabalho apresentado, dentre essas possiblidades estão:

- Integração com LLM (Large Language Models): Incorporar modelos de linguagem como o GPT-3.5 para melhorar a compreensão e análise de sentimentos expressos em tweets. Esses modelos podem ajudar na identificação de nuances e contextos mais complexos, além de melhorem os tópicos alcançados pela biblioteca spaCy.
- Criação de cadastro de ferramenta de polaridade: Permitir aos usuários que selecionem quais as ferramentas de processamento de dados eles gostariam de visualizar, por exemplo, alterar o processamento que é realizado hoje pelo NLTK por um processamento realizado através do GPT.
- Alteração nos lemas: Implementar um pré-processamento avançado que inclua lematização aprimorada. Isso ajudaria na normalização de palavras e na melhoria da precisão da análise de sentimento, especialmente em contextos onde as variações linguísticas são comuns, abrangendo um maior cenário de figuras públicas, não apenas os políticos que trazem uma linguagem mais formal e direcionada.
- Integração em tempo real com o Twitter: Estabelecer um sistema de coleta e análise de dados em tempo real diretamente do Twitter. Isso envolve a utilização de APIs do Twitter para capturar tweets atualizados e aplicar análises de sentimento em tempo real, na implementação atual este ponto ficou prejudicado devido a inexistência de uma forma gratuita de conexão com o Twitter.
- Análise de sentimentos contextualizada para figuras públicas: Desenvolver algoritmos específicos para a análise de sentimentos direcionados a figuras públicas. Isso pode incluir a detecção de mudanças na percepção pública ao longo de eventos específicos ou políticas implementadas.

Referências

PACETE, Luiz Gustavo. Brasil é o terceiro maior consumidor de redes sociais em todo o mundo. 2023. Disponível em: https://forbes.com.br/forbes-tech/2023/03/brasil-e-o-terceiro-pais-que-mais-consome-redes-sociais-em-todo-o-mundo/. Acesso em: 27 jun. 2023.

CONOVER, Michael; RATKIEWICZ, Jacob; FRANCISCO, Matthew; GONÇALVES, Bruno; FLAMMINI, Alessandro; MENCZER, Filippo. Political Polarization on Twitter. Fifth International Aaai Conference On Weblogs And Social Media. Bloomington, jul. 2011. p. 89-96.

CHOWDHARY, K. R. Fundamentals of Artificial Intelligence. [S. L.]: Springer New Delhi, 2020. 716 p. Disponível em: https://doi.org/10.1007/978-81-322-3972-7. Acesso em: 10 jun. 2024.

LASSEN, David S.; BROWN, Adam R. Twitter: The Electoral Connection? Social Science Computer Review: Social Science Computer Review. [S.L.], p. 420-436. nov. 2011.

RECUERO, Raquel; GRUZD, Anatoliy. Cascatas de Fake News Políticas: um estudo de caso no Twitter. Galaxia. São Paulo, maio 2019. p. 31-47. Disponível em: http://dx.doi.org/10.1590/1982-25542019239035. Acesso em: 28 jun. 2023.

LIDDY, Elizabeth Duross. Encyclopedia of Library and Information Science. 2. ed. Nova York: Marcel Decker, Inc, 2001. 13 p. Disponível em: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub. Acesso em: 02 out. 2023.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. Sentiment analysis algorithms and applications: a survey. 5. ed. Cairo: Elsevier B.V., 2014. 21 p. Disponível em: https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0005. Acesso em: 10 jun. 2024.

BENEVENUTO, Fabrício; GONÇALVES, Pollyanna; ALMEIDA, Virgílio. O Que Tweets Contendo Emoticons Podem Revelar Sobre Sentimentos Coletivos? In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2., 2013, Maceió. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013. p. 128-139. ISSN 2595-6094.

WEBSTER, Jonathan J.; KIT, Chunyu. TOKENIZATION AS THE INITIAL PHASE IN NLP. Hong Kong: City Polytechnic Of Hong Kong, 1992. 5 p.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2009. 581 p.

MARQUES; ANDRADE; ARTHUR; FREIRE. Quem me representa? Disponível em: https://qmrepresenta.com.br/. Acesso em: 18 jul. 2023.

WASSERMAN, S. e FAUST, K. Social Network Analysis. Methods and Applications. Cambridge, UK: Cambridge University Press, 1994.

DEGENNE, A. & FORSÉ, M. Introducing Social Networks. Sage: London, 1999.

NLTK. 2024. Disponível em: https://www.nltk.org/. Acesso em: 10 jun. 2024.

AAKER, J., & Chang, V. (2009). Obama and the Power of Social Media and Technology. Stanford Graduate School of Business Case No. M327.