Mineração de Dados em Redes Sociais para Aplicações em Análise de Sentimentos

Rafael V. Curiel[*], Reginaldo C. de Souza[†], Marcus V. S. Monteiro[‡],

Lucas W. Molin[§], Camila Veggi[¶]

[1] Faculdade de Tecnologia – Universidade Estadual de Campinas (UNICAMP)
Limeira – SP – Brasil

Abstract. *This article describes the methodology used in a data mining study on social networks for sentiment analysis. The objective of the study was to understand the dynamics and feelings underlying technological discussions in the online community of the r/Technology subreddit. For this, specific techniques were used, such as Exploratory Data Analysis (EDA) and Sentiment Analysis. The results obtained provided valuable insights into users' opinions and feelings regarding certain topics in online communities.*

Resumo. *Este artigo descreve a metodologia utilizada em um estudo de mineração de dados em redes sociais para análise de sentimentos. O objetivo do estudo foi entender as dinâmicas e sentimentos subjacentes às discussões tecnológicas na comunidade online do subreddit r/Technology. Para isso, foram utilizadas técnicas específicas, como a Análise Exploratória de Dados (EDA) e a Análise de Sentimentos. Os resultados obtidos forneceram insights valiosos sobre as opiniões e sentimentos dos usuários em relação a determinados tópicos em comunidades online.*

1. Introdução

A análise de dados é uma ferramenta essencial na compreensão das dinâmicas e dos sentimentos subjacentes às discussões tecnológicas em comunidades online. Neste contexto, o subreddit r/Technology se destaca como um espaço popular para a troca de informações e opiniões sobre tecnologia. No entanto, entender a complexidade desse ambiente requer a aplicação de técnicas específicas, como a Análise Exploratória de Dados (EDA) e a Análise de Sentimentos.

A EDA é uma abordagem fundamental para desvendar os padrões e tendências que permeiam r/Technology. Através dessa análise, é possível identificar os tópicos mais recorrentes nas discussões, as palavras-chave que mais se destacam e os momentos de maior atividade na comunidade. Essas informações fornecem insights valiosos sobre os interesses e as preferências dos membros da comunidade em relação à tecnologia.

[*]ex165318
[†]ex165442
[‡]ex165833
[§]ex165314
[¶]ex165308

Além disso, a análise da frequência e dos padrões de postagem pode revelar eventos tecnológicos significativos e discussões em curso.

A Análise de Sentimentos é outra ferramenta poderosa para compreender r/Technology. Utilizando técnicas de Processamento de Linguagem Natural (NLP) e Aprendizado de Máquina, essa análise permite determinar o tom emocional das postagens, classificando-as como positivas, negativas ou neutras. Isso nos ajuda a compreender as respostas emocionais dos membros da comunidade em relação aos tópicos discutidos. Além disso, permite mapear as atitudes predominantes dos participantes em relação aos avanços tecnológicos.

Essas análises combinadas oferecem uma visão aprofundada do ambiente digital de r/Technology, ajudando-nos a entender como a comunidade percebe e reage à tecnologia. Essa compreensão é fundamental para apreciar o papel das comunidades online na disseminação do conhecimento e na formação de opiniões no cenário tecnológico contemporâneo.

2. Metodologia

A metodologia empregada neste estudo, conduzido no âmbito de uma pesquisa de doutorado, consistiu em duas etapas fundamentais: a extração de dados e a análise subsequente. A seguir, detalhamos minuciosamente cada uma dessas etapas.

Extração de dados:

1. A coleta de informações do Reddit foi a primeira etapa da metodologia. Inicialmente, foi necessário preparar o ambiente para a obtenção dos dados, o que implicou o desenvolvimento de um código no ambiente Google Colab. Esse código realizou as seguintes ações:
2. Recuperação de Credenciais: Para acessar o Reddit, foi necessário obter credenciais de autenticação. Esse processo envolveu o uso de segredos de usuário para garantir a autorização adequada.
3. Inicialização do Cliente da API do Reddit: Foi utilizado o Python Reddit API Wrapper (PRAW) para inicializar o cliente da API do Reddit. Essa biblioteca é fundamental para interagir com a plataforma Reddit de forma programática.
4. Funções de Coleta e Processamento: Foram definidas funções específicas para buscar informações sobre usuários do Reddit e processar postagens ou comentários. A função ”process_item” desempenhou um papel crucial, permitindo a extração de dados relevantes, como autor, data e hora, título ou texto do comentário, pontuação, número de comentários e detalhes adicionais, como o karma da postagem e a tag da postagem do autor.
5. Carregamento de Dados: Por fim, os dados coletados foram carregados a partir de um arquivo CSV em um DataFrame do pandas, uma estrutura de dados essencial para a manipulação e análise subsequentes.

Analsie de dados:

1. Uma vez que os dados foram devidamente coletados e armazenados, a segunda etapa da metodologia foi a análise propriamente dita. Nesta fase, o código desenvolvido no Google Colab desempenhou as seguintes funções:

2. Importação de Bibliotecas: Foram importadas bibliotecas essenciais para a manipulação de dados, como expressões regulares, exibição de barras de progresso, geração de valores aleatórios, processamento de linguagem natural (NLP) para análise de sentimentos, criação de nuvens de palavras e visualização de dados por meio de diversas bibliotecas.
3. Configuração Inicial: O código realizou tarefas fundamentais, como a definição do estilo de plotagem, a inicialização de um analisador de sentimentos (SIA), a recuperação de palavras irrelevantes e a definição de um mapa de cores personalizado.
4. Coleta de Dados: O código coletou informações das postagens mais recentes do subreddit. Essas informações incluíram o ID da postagem, autor, data e hora, título, URL, pontuação, número de comentários, texto da postagem, texto original da postagem, karma de comentário do autor e uma tag associada à postagem. Todos esses dados foram organizados em um DataFrame chamado 'data', proporcionando uma base sólida para análises posteriores.
5. Preparação de Dados: Esta etapa envolveu uma série de operações nos dados para torná-los adequados à análise. Isso incluiu o preenchimento de tags ausentes, a contagem de ocorrências de cada tag e a aplicação de transformações de texto. As transformações de texto abrangeram a conversão para letras minúsculas, a remoção de nomes de usuário, hashtags e URLs, a extração de palavras relevantes, a eliminação de palavras irrelevantes e a adição de novas colunas para contagem de palavras, comprimento do texto e hora. Essas transformações visaram limpar e organizar os dados, preparando-os para análises e explorações posteriores.
6. Análise de Sentimentos: A análise de sentimentos foi realizada nos dados de texto utilizando um Sistema de Análise de Sentimento (SIA). Essa análise atribuiu pontuações de sentimento e classificações correspondentes (negativo, positivo ou neutro), que foram armazenadas nas colunas "sentiment_eval" e "class_sentiment".
7. Análise Exploratória de Dados: Para explorar e visualizar os dados, foram empregadas as bibliotecas Plotly Express e WordCloud. Histogramas, gráficos de barras empilhadas e nuvens de palavras foram criados para revelar insights sobre os resultados da análise de sentimento com base em tags e horários. O objetivo era comparar os sentimentos expressos nas postagens com os dados dos comentários, identificando padrões e semelhanças.

3. Resultados

Os resultados obtidos a partir da coleta de dados mostraram uma maior frequência de posts com tendência neutra, em comparação com as postagens que expressavam sentimentos negativos ou positivos. Essa constatação indica que os participantes procuram transmitir informações de maneira objetiva e equilibrada, demonstrando uma relativa neutralidade nas discussões.

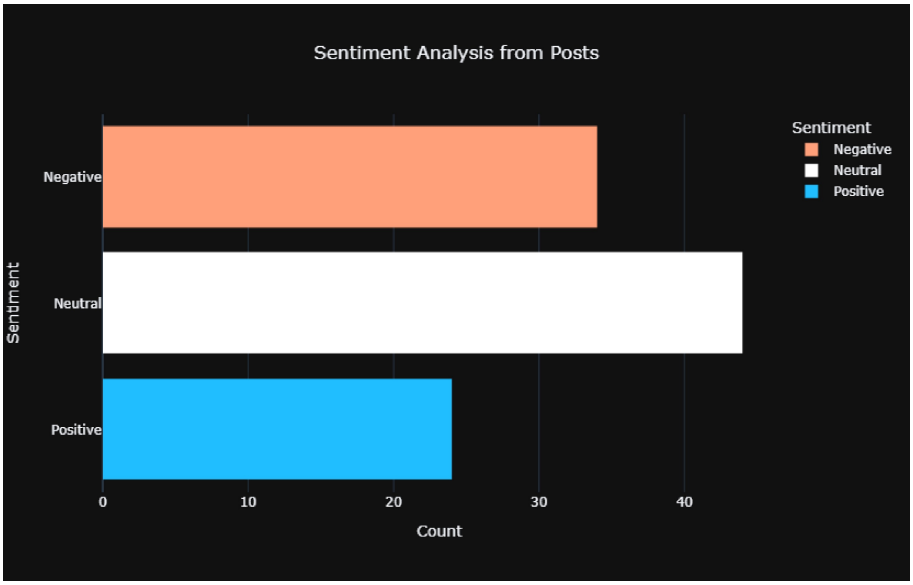

Figure 1. Contagem de posts por sentimento

Ao analisar os comentários, notamos uma tendência diferente em relação às postagens. Os comentários refletem predominantemente um tom positivo, indicando que a interação e as respostas dos membros da comunidade tendem a ser mais favoráveis e construtivas. Logo após os comentários positivos, há uma presença significativa de comentários neutros, demonstrando que muitos participantes optam por manter uma abordagem ponderada e informativa nas discussões. Por último, mas não menos importante, encontramos os comentários com sentimento negativo, que, embora menos comuns, ainda têm presença na comunidade e podem representar visões críticas ou discordantes. Esses resultados sugerem que a comunidade do subreddit é composta por membros que buscam manter um ambiente construtivo e equilibrado, mas que também estão dispostos a expressar opiniões críticas e discordantes quando necessário.

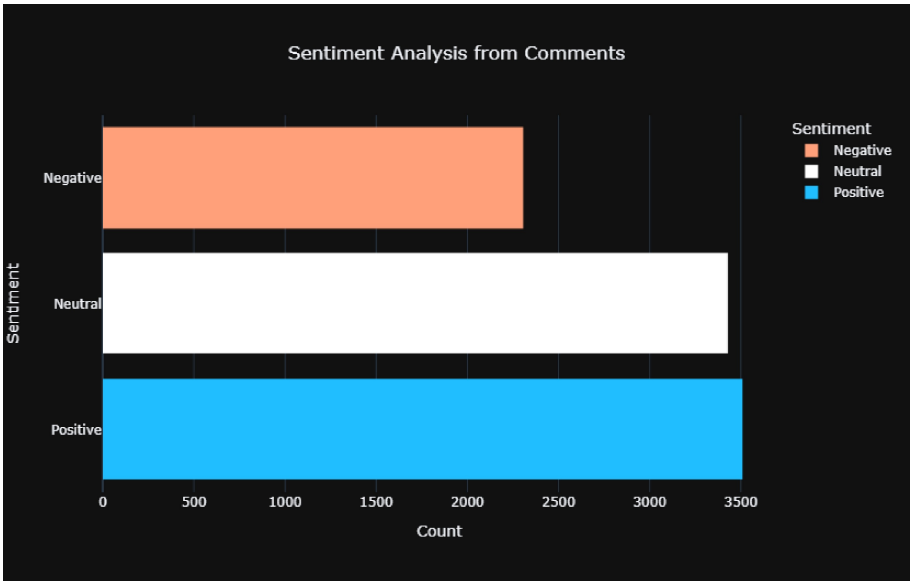

Figure 2. Contagem de comentários por sentimento

Outra forma de interpretar os dados dos posts é através da análise das frequências relativas dos sentimentos nas diferentes categorias de tópicos do subreddit. A categoria "Adblock Warning" apresenta uma expressiva negatividade, com 100% dos posts classificados como negativos. Outras categorias, como "Artificial Intelligence", "Biotechnology", e "Business", possuem uma maior proporção de posts neutros, indicando um menor envolvimento emocional dos participantes nessas áreas. Em contraste, categorias como "Politics", "Privacy", e "Security" demonstram uma tendência mais positiva, sugerindo uma maior satisfação ou otimismo dos membros nessas questões. Além disso, algumas categorias, como "Net Neutrality" e "Robotics/Automation", mostram extremos de negatividade ou neutralidade, revelando discussões acaloradas ou indiferentes sobre esses tópicos específicos na comunidade do subreddit. Esses resultados oferecem uma visão geral das dinâmicas de sentimentos relacionadas a diversas áreas de interesse, ressaltando as diferenças nas percepções e opiniões dos membros.

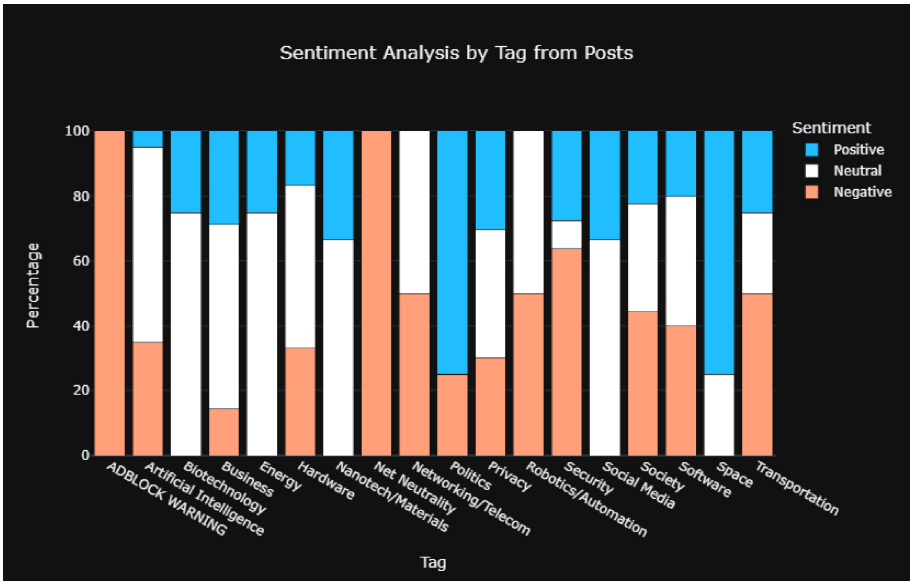

Figure 3. Composição percentual de sentimento por tag em posts

A composição percentual dos sentimentos por comentários mostra como os membros da comunidade se sentem sobre diferentes categorias de tópicos do subreddit. A categoria "Adblock Warning" tem mais sentimentos positivos, com 66,67% de positividade e 33,33% de neutralidade. Outras categorias, como "Artificial Intelligence" e "Business", têm uma mistura de sentimentos negativos, neutros e positivos. Algumas categorias, como "Energy", "Privacy", e "Social Media", recebem mais sentimentos positivos, indicando uma atitude positiva dos membros. Por outro lado, "Politics" tem mais sentimentos negativos, com 31,40% de negatividade, enquanto "Society" tem uma distribuição quase igual de sentimentos negativos, neutros e positivos. Esses dados revelam como os membros da comunidade expressam seus sentimentos sobre diferentes tópicos, mostrando a variedade de pontos de vista e opiniões dentro do subreddit.

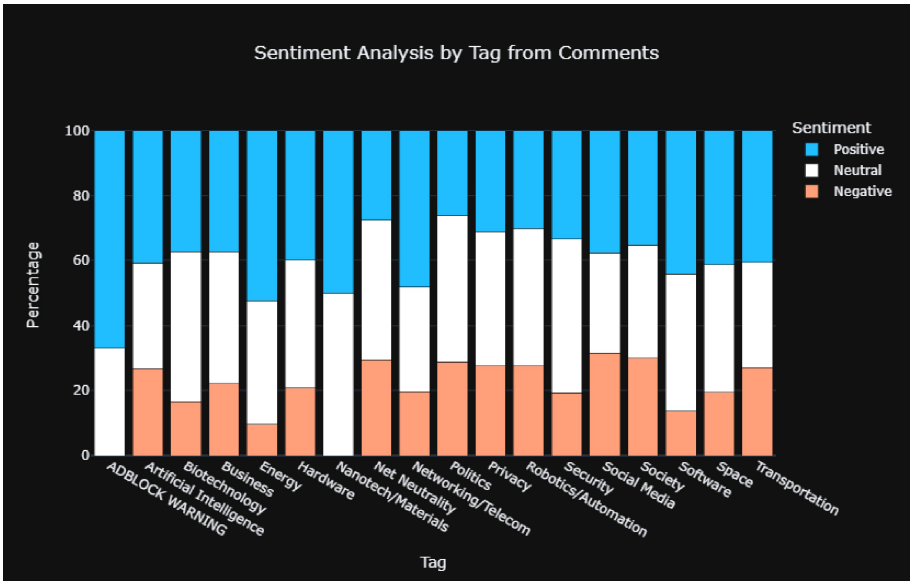

Figure 4. Composição percentual de sentimento por tag em comentários

A análise da contagem de sentimentos dos posts por hora revela padrões notáveis nas flutuações emocionais dos membros da comunidade no decorrer do dia. Os dados mostram que os sentimentos neutros e positivos predominam nas primeiras horas da manhã, com um pico de postagens positivas às 06:00. Entre 08:00 e 12:00, há uma mudança para sentimentos negativos, que atingem o máximo às 10:00, e depois se recuperam gradualmente para sentimentos neutros e positivos. À tarde, das 12:00 às 16:00, os sentimentos se estabilizam, enquanto à noite, das 16:00 às 23:00, os sentimentos positivos prevalecem, especialmente às 19:00. Essas mudanças mostram as variações emocionais na comunidade do subreddit, que podem estar relacionadas às diferentes atividades e estados de ânimo dos membros.

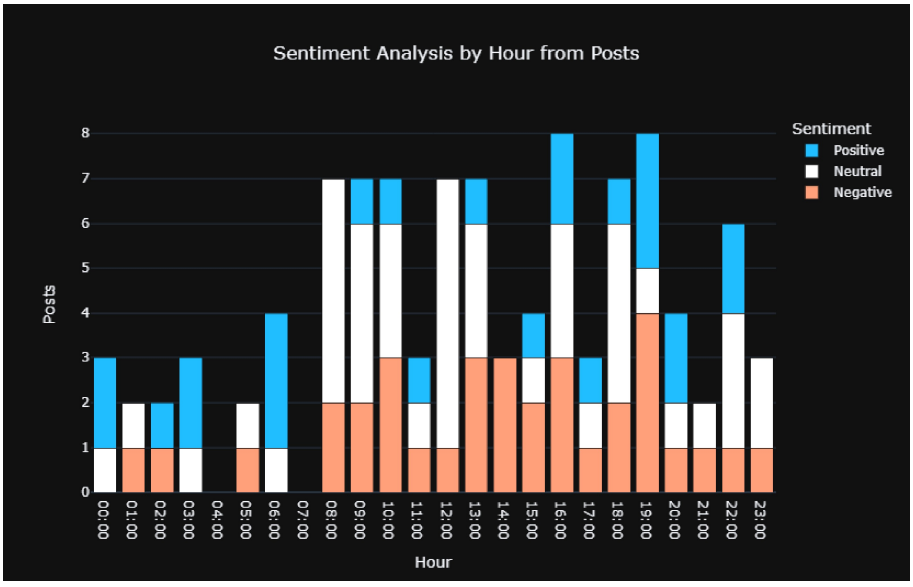

Figure 5. Contagem de posts por sentimento e hora

No que tange aos sentimentos dos comentários por hora, a imagem revela tendências interessantes nas variações emocionais dos membros da comunidade ao longo do tempo. Das 00:00 às 07:00, são marcadas por sentimentos positivos, com um crescimento das postagens positivas até às 03:00. Já as horas da manhã e da tarde, das 08:00 às 15:00, apresentam sentimentos neutros, com um aumento da neutralidade a partir das 08:00 e um equilíbrio entre sentimentos neutros e positivos nesse intervalo. Por outro lado, a tarde, das 16:00 às 18:00, é dominada por sentimentos positivos, com um pico de postagens positivas às 16:00. As horas da noite, das 19:00 às 23:00, também revelam uma predominância de sentimentos positivos, especialmente às 23:00.

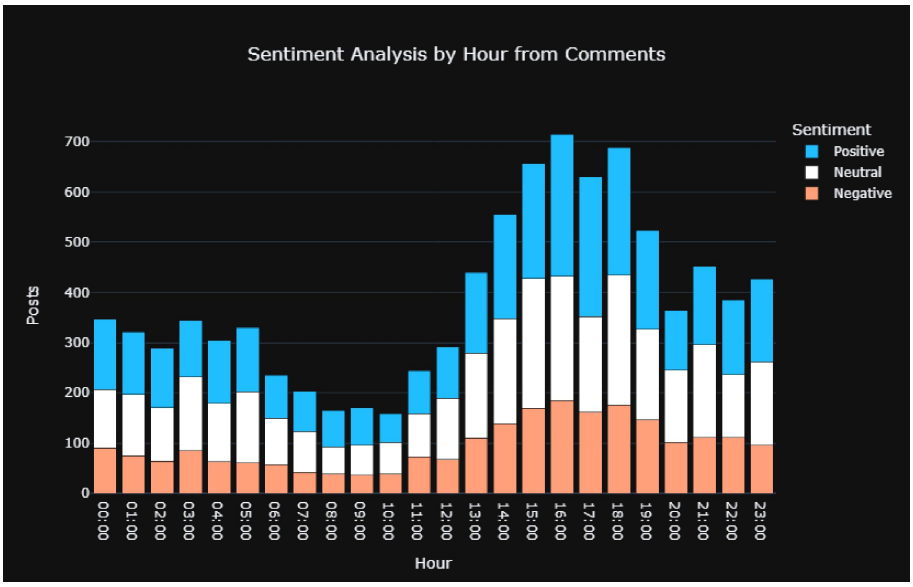

Figure 6. Composição percentual de sentimento por tag em comentários

A análise das palavras mais destacadas nos posts revela um cenário diversificado de sentimentos na comunidade. Termos como “google”, “new”, “india”, e “space” estão associados a sentimentos positivos, sugerindo entusiasmo em relação a avanços tecnológicos ou exploração espacial. No entanto, palavras como “nuclear”, “weapons”, e “privacy” refletem preocupações e sentimentos negativos relacionados à segurança e à privacidade de dados. Essas palavras oferecem insights sobre os tópicos e as emoções predominantes nas discussões dos posts.

Figure 7. Nuvem de palavras dos posts (top 100)

A análise das palavras mais destacadas nos comentários revela uma variedade de sentimentos presentes na comunidade. Termos como “one”, “people”, “make”, e “time” estão associados a sentimentos positivos, indicando discussões construtivas e sugestões. Por outro lado, a presença da palavra “law” sugere debates sobre questões legais, enquanto

"problem" e "getting" indicam desafios discutidos na comunidade. A palavra "dopamine" pode refletir conversas sobre neurociência e comportamento humano. Esses termos oferecem insights sobre as emoções e os temas predominantes nas interações dos comentários da comunidade.

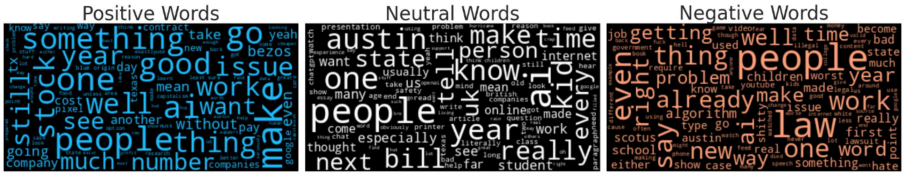

Figure 8. Nuvem de palavras dos comentários (top 100)

4. Discussão

A análise de sentimentos foi uma metodologia importante para este estudo, pois permitiu compreender melhor as emoções e opiniões dos participantes da comunidade r/Technology no Reddit. Um dos resultados mais surpreendentes foi a alta frequência de postagens neutras. Isso indica que a comunidade tende a ter uma postura equilibrada e racional ao debater temas tecnológicos. Esse comportamento de neutralidade pode estar relacionado ao caráter informativo e factual das discussões tecnológicas, nas quais os participantes trocam notícias, dados técnicos e avaliações.

Além disso, a análise mostrou a existência de sentimentos positivos e negativos em diferentes níveis. As palavras relacionadas a sentimentos positivos, como "google", "new", "india" e "space", revelam um interesse e uma admiração pelos progressos tecnológicos e pela exploração espacial. Em contrapartida, palavras como "nuclear", "weapons" e "privacy" expressam inquietações ligadas à segurança e à proteção de dados. Essa variedade de sentimentos evidencia que a comunidade r/Technology participa de discussões que não se limitam à informação, mas que também envolvem questões afetivas e preocupações relevantes sobre a tecnologia.

A mineração de dados foi um recurso fundamental para a coleta, organização e análise dos dados do Reddit. A fase de extração de dados consistiu na criação de um ambiente de coleta que envolveu a obtenção de credenciais de autenticação, inicialização do cliente da API do Reddit e definição de funções personalizadas para obter informações pertinentes. Essa abordagem programática possibilitou a obtenção de dados detalhados de postagens e comentários, incluindo autor, data, título, texto e informações adicionais. A preparação dos dados teve um papel essencial na análise, abrangendo a limpeza, transformação e organização dos dados. A utilização de técnicas de Processamento de Linguagem Natural (NLP) para análise de sentimentos foi crucial para atribuir pontuações de sentimento e classificações (positivo, negativo ou neutro) aos textos.

Essas descobertas têm implicações significativas para a compreensão da comunidade r/Technology e, por extensão, para a compreensão de como as comunidades online percebem e reagem à tecnologia. As análises realizadas neste estudo oferecem uma visão abrangente das dinâmicas emocionais e das tendências de sentimentos nessa comunidade específica. Essa compreensão pode ser valiosa para pesquisadores, profissionais de marketing e empresas que buscam compreender o envolvimento e as preferências dos usuários em discussões tecnológicas online.

Além disso, a metodologia de mineração de dados demonstrou ser uma abordagem eficaz para coletar e analisar dados em larga escala em comunidades online. A flexibilidade e a escalabilidade dessa abordagem podem ser aplicadas a diferentes contextos de pesquisa, permitindo a extração de insights em tempo real e a identificação de padrões de comportamento em comunidades online diversas.

5. Conclusão

Este estudo nos permitiu mergulhar no mundo da comunidade r/Technology no Reddit, examinando suas dinâmicas complexas através da mineração de dados e da análise de sentimentos. Nossa principal motivação era entender como os membros dessa comunidade veem e respondem à tecnologia, bem como detectar tendências emocionais que permeiam as discussões.

Com a análise de sentimentos, constatamos que a comunidade r/Technology se distingue por uma prevalência de postagens neutras, acompanhadas por aquelas com sentimentos negativos e positivos. Isso indica um ambiente onde se busca a imparcialidade e a objetividade, mas onde as expressões emocionais também têm espaço para se revelar.

As variações emocionais ao longo do dia nos mostraram padrões interessantes de como os membros da comunidade se relacionam com os assuntos tecnológicos. Desde o entusiasmo nas primeiras horas da manhã até os debates mais acirrados durante o dia, as emoções parecem seguir o fluxo das atividades diárias na comunidade.

Nossa análise por categorias de tópicos evidenciou a diversidade de percepções e opiniões dos membros sobre diferentes áreas de interesse. Cada categoria apresentou nuances nas expressões emocionais, refletindo a complexidade das discussões em torno de temas tecnológicos diversos.

Por fim, esta pesquisa nos ofereceu uma visão detalhada de como a mineração de dados e a análise de sentimentos podem revelar as sutilezas das interações humanas em comunidades online. Embora o foco da nossa investigação tenha sido a mineração de dados, é incontestável que a riqueza das descobertas não teria sido possível sem uma compreensão profunda das nuances emocionais subjacentes. Assim, esta análise não só nos aproximou do entendimento da comunidade r/Technology, mas também ressaltou a importância de levar em conta boas práticas na mineração de dados .

References

(2023). Python documentation. Online. Acesso em 16/09/2023: `https://docs.python.org/3/`.

(2023). Repositório do projeto de mineração de dados da unicamp. Online. Acesso em 16/09/2023: `https://github.com/r-curiel/unicamp_datamining`.