

Trabalho Final de Mineração e Análise de Redes Sociais

Arthur Henrique M. Santos

Universidade Federal de São João Del Rei (UFSJ) – São João Del Rei, MG - Brasil

ms.arthurhenrique@gmail.com

Abstract. *The project falls within the scope of social network analysis and data mining, using the subreddit r/Helldivers as a case study to explore user interactions, identify trending topics, and understand how content clusters into communities and discussions. Examples include posts about in-game factions, updates, gear, and player feedback.*

Resumo. *O trabalho se encaixa na análise de redes sociais e mineração de dados, usando do subreddit r/Helldivers, para entender como os usuários interagem, quais temas são mais discutidos e como o conteúdo se organiza em comunidades e tópicos. Exemplos incluem posts sobre as raças do jogo, atualizações, itens do jogo e feedback dos jogadores.*

1. Objetivo

Este trabalho tem como objetivo estudar a estrutura e a dinâmica das discussões no subreddit r/Helldivers. Identificando temas recorrentes, comunidades de usuários, padrões de coocorrência de palavras e sentimentos dos comentários. Para assim, entender como o conteúdo se organiza e quais fatores influenciam a popularidade dos posts.

2. Metodologia

2.1 Descrição dos dados e modelagem

Os dados foram coletados do subreddit r/Helldivers por meio da API do Reddit (PRAW), abrangendo informações como títulos das postagens, pontuações (scores), número de comentários e datas de criação. Além disso, foram extraídos também os comentários, incluindo o conteúdo textual, nome do autor, data e a relação de resposta entre eles (parent_id).

Vale destacar que determinados elementos não foram considerados na coleta, como conteúdos multimídia (imagens e vídeos), postagens removidas ou excluídas, e quaisquer dados privados dos usuários.

As ferramentas utilizadas ao longo do processo incluíram a linguagem Python, com a biblioteca *pandas* para manipulação dos dados e o *NetworkX* para a modelagem das redes. Os dados foram organizados e armazenados em arquivos CSV, o que facilitou a leitura e análise posterior.

Não foi necessário aplicar nenhum tipo de parser especial para processar os dados. Ainda assim, houve um tratamento textual básico, que envolveu a remoção de *stopwords* e a normalização dos textos para garantir maior consistência nas análises.

Foram construídas duas redes principais a partir dos dados: a rede de coocorrência de palavras e a rede de usuários.

Na rede de coocorrência de palavras, os vértices representam palavras únicas presentes nos títulos dos posts. As arestas são estabelecidas entre duas palavras sempre que elas aparecem juntas em um mesmo título, indicando sua relação de coocorrência.

Já na rede de usuários, os vértices correspondem aos usuários que realizaram comentários. As arestas são direcionadas e conectam o autor de um comentário ao autor do comentário ao qual ele respondeu, representando assim a dinâmica de interação entre os participantes do subreddit.

2.2 Experimentos

Foi realizada uma análise descritiva básica da rede de coocorrência de palavras, composta por 1.978 nós — cada um representando uma palavra única presente nos

títulos dos posts — e 9.484 arestas, que conectam pares de palavras que aparecem juntas em um mesmo título. A densidade da rede, de aproximadamente 0,0048, indica que se trata de uma rede pouco conectada. A distribuição dos graus revela que poucas palavras são altamente conectadas, enquanto a maioria possui apenas poucos vínculos. A visualização dessa distribuição está representada por um histograma.

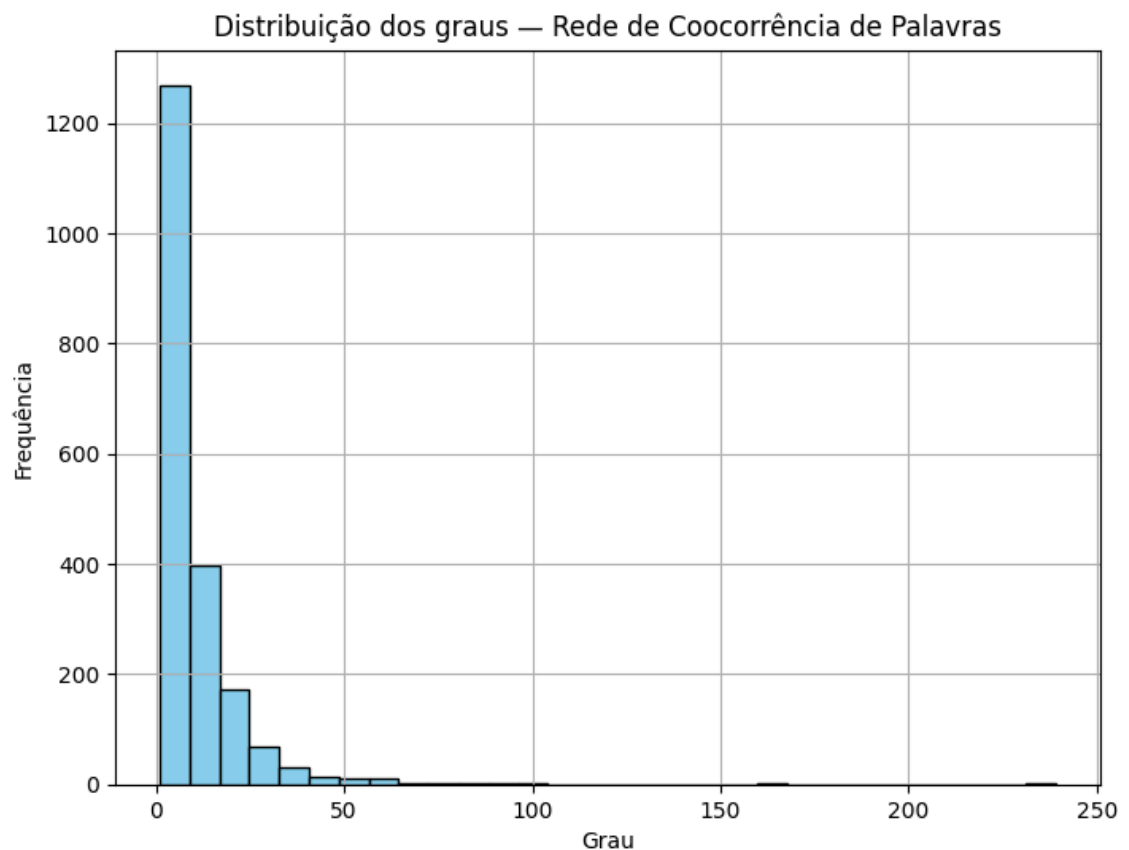


Figura 1. Frequência de palavras conforme o número de conexões na rede.

A visualização da rede de coocorrência mostra uma estrutura centralizada em torno de algumas palavras-chave. Uma segunda visualização, colorida com base nos valores de centralidade, destaca as palavras mais influentes dentro da rede.

Rede de Co-ocorrência de Palavras — r/Helldivers



Figura 2. Estrutura da rede de coocorrência entre palavras dos títulos.

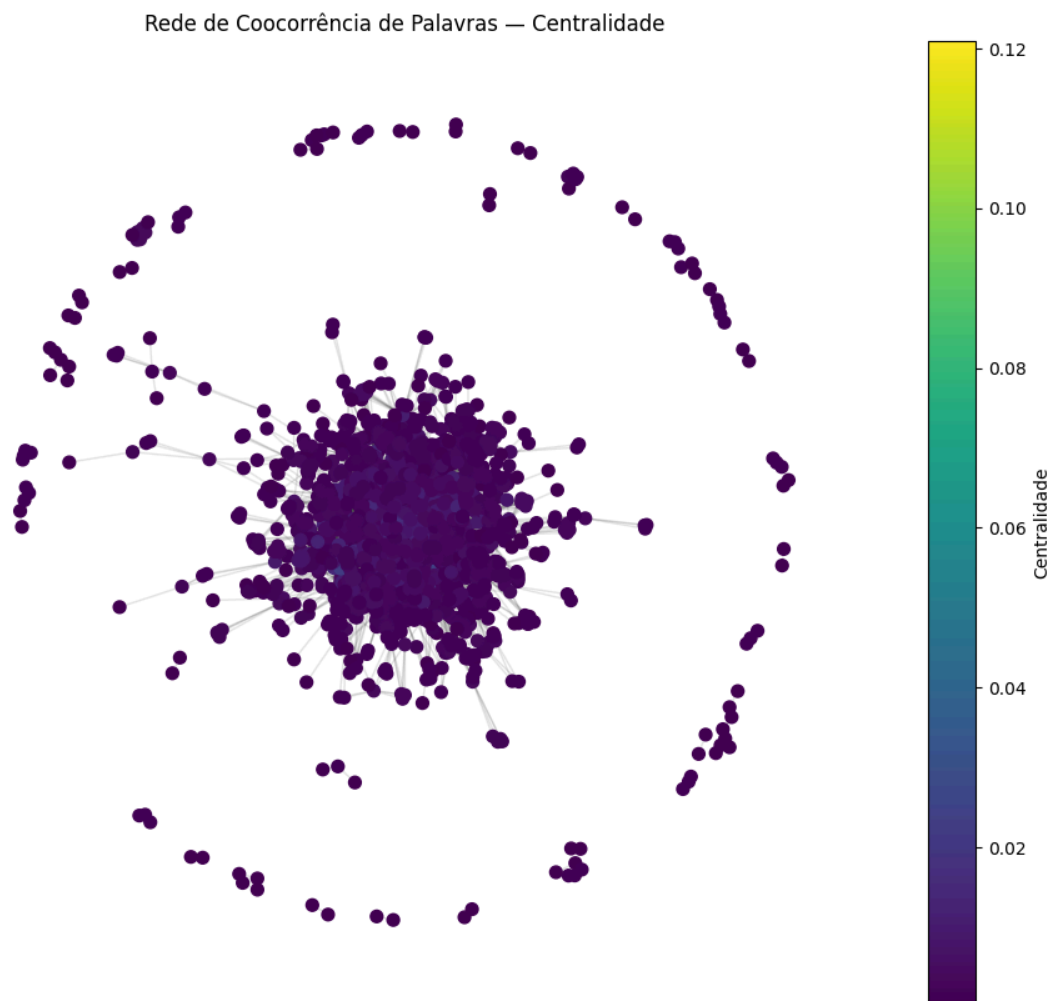


Figura 3. Rede de palavras destacando os termos mais influentes por cor.

Na análise topológica, foram detectadas 74 comunidades distintas de palavras, o que sugere a existência de agrupamentos temáticos recorrentes nos títulos. Entre as palavras com maior centralidade destacam-se: *helldivers*, *super*, *pack*, *warp* e *warbond*.

Também foi construída uma rede de usuários a partir das interações nos comentários. Nessa rede, os vértices representam os usuários, e as arestas direcionadas conectam o autor de um comentário ao autor da mensagem respondida. Por meio do método de Louvain, foi possível identificar comunidades de usuários com padrões de interação semelhantes.

Comunidades detectadas na rede de usuários

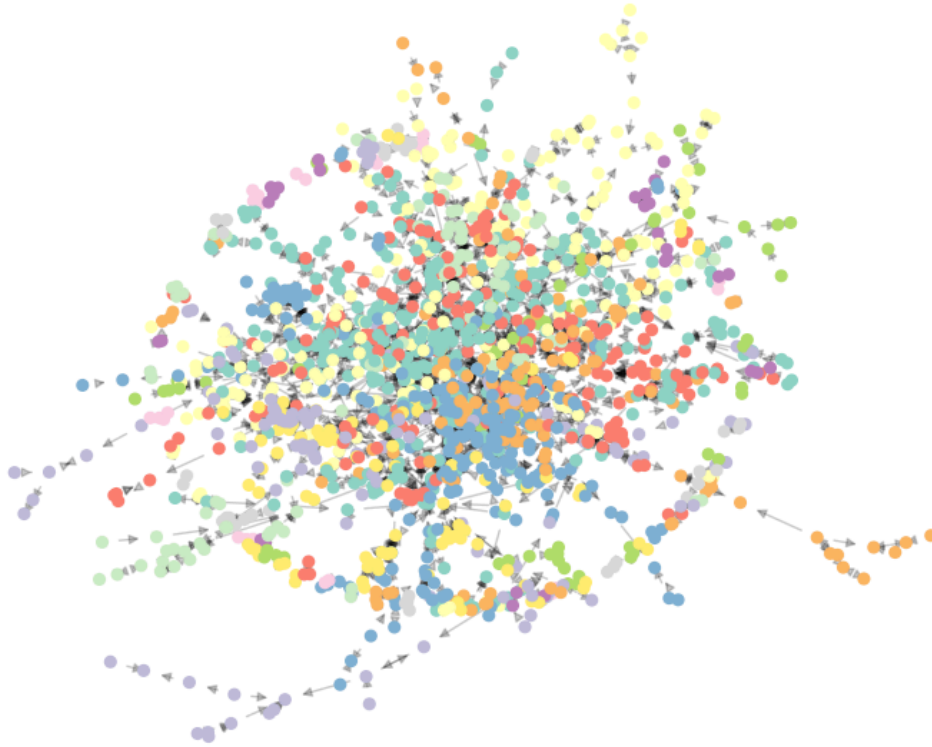


Figura 4. Rede de usuarios colorida por comunidades detectadas.

A distribuição dos graus nesta rede mostra que a maior parte dos usuários tem baixa atividade, enquanto poucos são responsáveis por grande parte das interações.

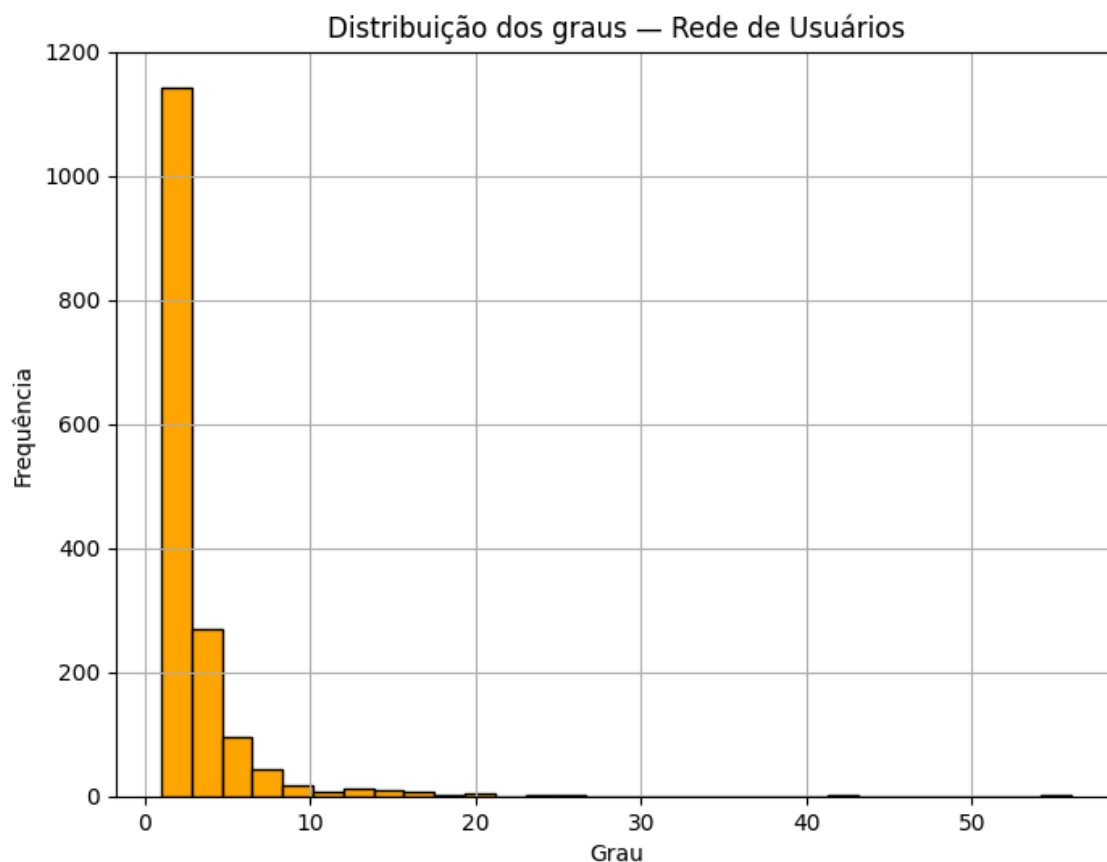


Figura 5. Frequência de usuários conforme o número de interações.

Na parte de modelagem textual, foram aplicadas técnicas de *TF-IDF* e agrupamento por *KMeans*, o que permitiu segmentar os títulos em cinco clusters principais. Essa segmentação é apresentada em um gráfico de dispersão. Também foi utilizada a técnica de *LDA* para identificação de cinco tópicos latentes presentes nos títulos dos posts.



Figura 6. Agrupamento dos títulos em cinco clusters principais.

Complementando a análise, foi realizada uma avaliação de sentimento nos comentários, revelando a presença de opiniões tanto positivas quanto negativas.

Por fim, dois experimentos adicionais foram conduzidos. O primeiro envolveu uma tarefa de classificação, utilizando uma árvore de decisão para prever se um post seria popular ou não, alcançando uma acurácia em torno de 53%. O segundo consistiu em um sistema de recomendação capaz de sugerir conteúdos similares ao post original. A análise de tópicos e a avaliação de sentimentos contribuíram para enriquecer a compreensão geral sobre os padrões e temas predominantes na comunidade analisada.

3. Discussão

A análise revelou que as discussões no subreddit r/Helldivers se concentram em poucos temas centrais, enquanto há grande diversidade de palavras periféricas. A estrutura das redes mostra que tanto os tópicos quanto as interações entre usuários tendem a formar comunidades distintas, refletindo interesses e padrões de engajamento específicos. Apesar disso, prever a popularidade dos posts apenas pelo título mostrou-se limitado, indicando que outros fatores influenciam o sucesso das publicações. A identificação de tópicos e sentimentos contribui para entender melhor o perfil e as demandas da

comunidade, podendo orientar futuras ações de moderação ou recomendação de conteúdo.

4. Conclusão

O trabalho possibilitou mapear e analisar as principais estruturas e padrões de interação no subreddit r/Helldivers, por meio da construção de redes de coocorrência de palavras e redes de usuários. A partir dessa abordagem, foram identificadas comunidades temáticas, tópicos recorrentes e dinâmicas de engajamento entre os participantes. A análise de sentimentos, aliada aos experimentos de classificação e recomendação, contribuíram para uma compreensão mais ampla do comportamento da comunidade.

Apesar dos avanços, não foi possível prever com alta precisão a popularidade dos posts considerando apenas os títulos, tampouco explorar conteúdos multimídia, que não foram incluídos na coleta. Como continuidade deste estudo, recomenda-se investigar aspectos temporais das interações, realizar análises de toxicidade nos comentários e ampliar a abordagem para outros tipos de dados e comunidades, a fim de enriquecer ainda mais as descobertas.

7. Referências

PRAW(Python Reddit API Wrapper), <https://praw.readthedocs.io/> .