

Análise de Comentários de Vídeos do YouTube utilizando Redes Textuais

Eduarda Moura, Fabíola Pereira
{eduarda.lopes, fabiola.pereira}@ufu.br

RESUMO

Este trabalho investiga como diferentes formas de representar textos em formato de rede impactam a qualidade das análises. O objetivo é avaliar qual estratégia de modelagem oferece melhores resultados para análise exploratória e mineração de opinião em dados de redes sociais. Como estudo de caso, utilizou-se 53.831 comentários do YouTube. A revisão bibliográfica possibilitou a construção inicial de uma rede textual com identificação dos nós de maior centralidade. Como produto final, pretende-se desenvolver diretrizes práticas que orientem pesquisadores na escolha da técnica mais adequada conforme o contexto da análise.

INTRODUÇÃO

Dados textuais podem ser analisados para além de nuvens de palavras quando modelados por redes, em que nós (palavras, termos ou sentenças) conectam-se por arestas através de relações ou coocorrências, permitindo identificar elementos centrais, descobrir comunidades temáticas e realizar mineração de opinião. A literatura carece de estudos comparativos sobre o impacto de diferentes modelagens nas análises resultantes, sendo dados de redes sociais especialmente relevantes devido à sua informalidade, diversidade de autores e dinâmica temporal. Diante disso, a pesquisa investiga: "Qual a melhor modelagem de rede textual para mineração de opinião em comentários de redes sociais? E para análise exploratória desses dados?".

MATERIAIS E MÉTODOS

O corpus utilizado consiste em 53.831 comentários extraídos do vídeo "Bolsonaro pergunta para Lula sobre corrupção na Petrobras | Band Eleições - Debate Presidencial 2022" (Band Jornalismo, 9,6 milhões de visualizações). A modelagem da rede textual fundamentou-se na coocorrência de palavras, em que nós representam palavras lematizadas e arestas indicam coocorrências dentro de um mesmo comentário, com pesos calculados pela frequência de coocorrência. A **Figura 1** apresenta a rede sem lematização, após filtragens iniciais.

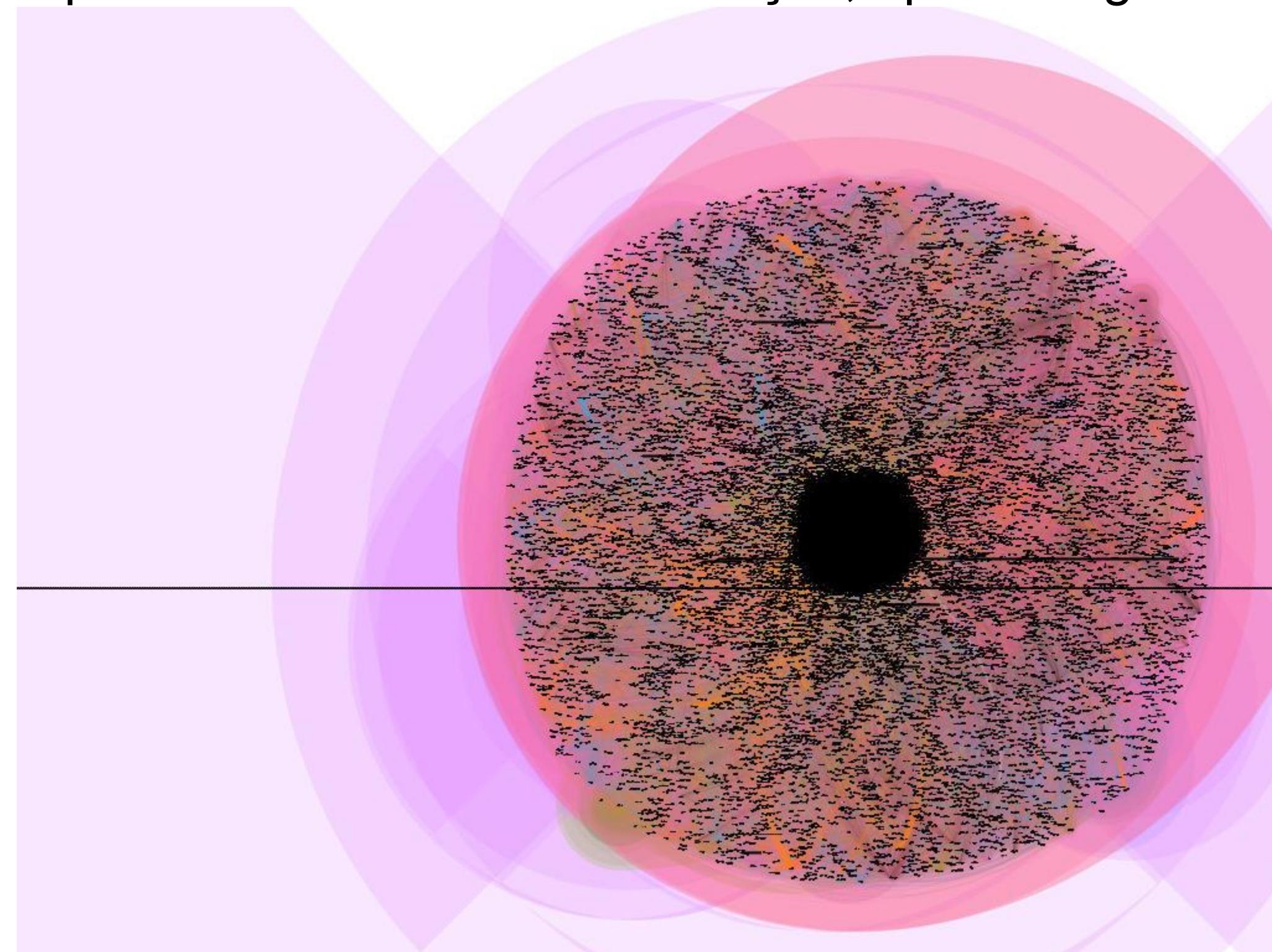


Figura 1. Rede textual completa. Visualização gerada no Gephi.

O pré-processamento dos dados compreendeu múltiplas etapas de filtragem e refinamento. Realizou-se a remoção de stopwords mediante a biblioteca NLTK e a lematização dos textos através da biblioteca Spacy. Na filtragem estrutural, removeram-se arestas com peso inferior a 15 coocorrências, eliminando conexões fracas e ruídos estatísticos. A seguir, selecionaram-se as 200 palavras com maior grau de centralidade, priorizando os termos mais relevantes e conectados da rede. Finalmente, extraiu-se o componente gigante, descartando nós isolados e subgrafos desconectados. A rede resultante desse processo está apresentada na **Figura 2**. Para a análise, aplicaram-se centralidade de grau, identificando os termos mais conectados, e modularidade (algoritmo de Louvain), detectando comunidades temáticas mediante agrupamento de palavras coocorrentes. Os pesos das arestas foram normalizados para análises.

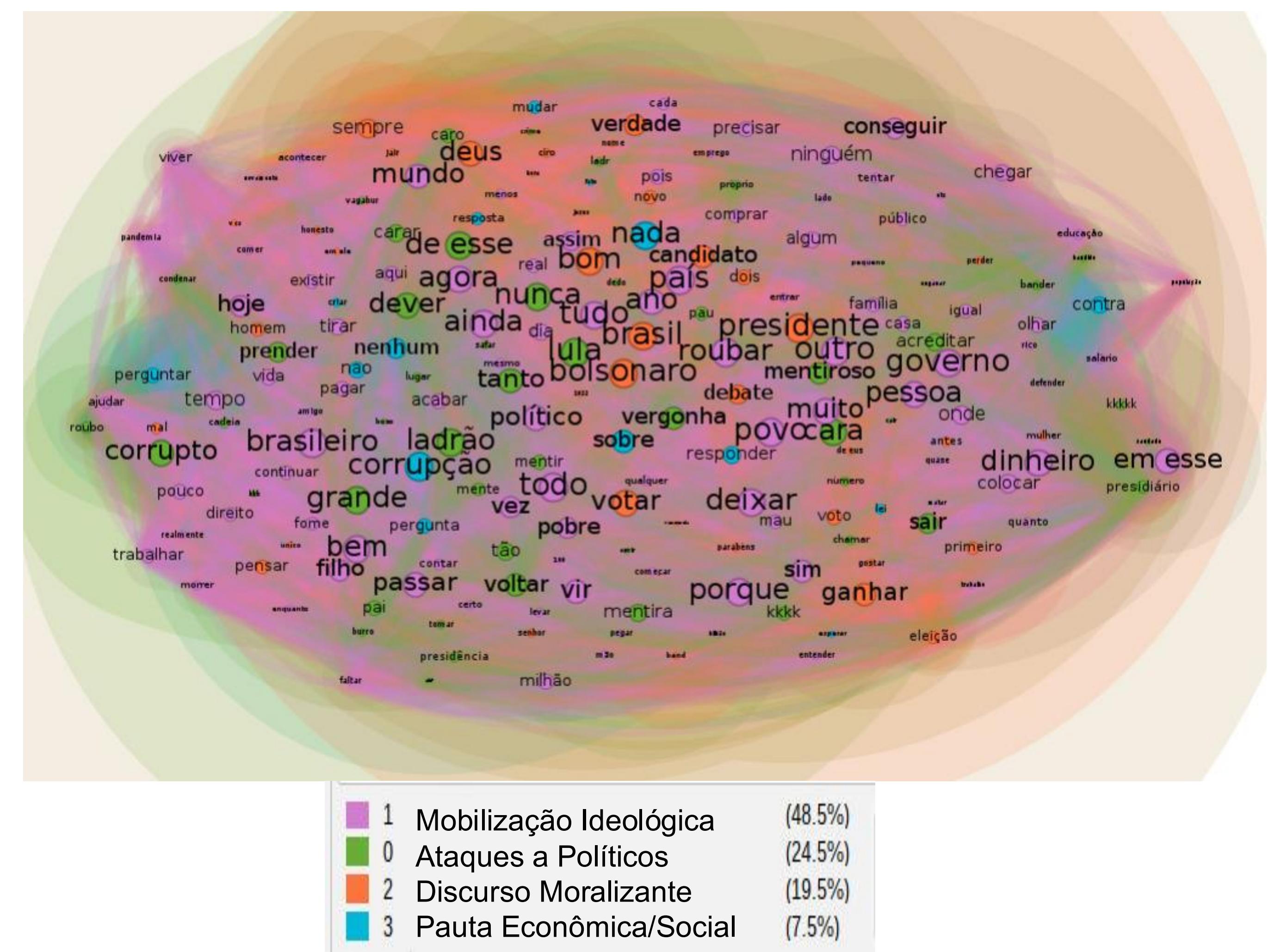


Figura 2. Rede textual processada. Visualização gerada no Gephi.

Os três maiores hubs semânticos de cada comunidade foram destacados na **Figura 3**. Cada termo com centralidade elevada possui importância estrutural para a rede.

Termo	Centralidade	Grupo
todo	1.0000	Mobilização Ideológica
povo	1.0000	Mobilização Ideológica
governo	0.9950	Mobilização Ideológica
lula	1.0000	Ataques a Políticos
de esse	0.9799	Ataques a Políticos
grande	0.9698	Ataques a Políticos
bolsonaro	1.0000	Discurso Moralizante
brasil	1.0000	Discurso Moralizante
presidente	1.0000	Discurso Moralizante
corrupção	0.9899	Pauta Econômica/Social
nada	0.9799	Pauta Econômica/Social
sobre	0.7940	Pauta Econômica/Social

Figura 3. Métricas de centralidade de grau e comunidades.

RESULTADOS ESPERADOS

A pesquisa encontra-se em andamento. Espera-se identificar quais estratégias de modelagem de redes textuais são mais adequadas para mineração de opinião e análise exploratória em dados de redes sociais, demonstrando como diferentes abordagens revelam padrões estruturais e temáticos distintos através de métricas como centralidade, modularidade e detecção de comunidades. Ao final, a sistematização comparativa das vantagens e limitações de cada modelagem deverá fornecer amparo metodológico que auxilie pesquisadores na seleção de técnicas mais apropriadas para seus contextos analíticos específicos.

BIBLIOGRAFIA

- [1] Bail, C. (2018). Text Networks. *Summer Institute in Computational Social Science*.
- [2] Vega, D. & Magnani, M. (2018). Foundations of Temporal Text Networks. *Applied Network Science*, 3(25).

AGRADECIMENTOS

Os autores agradecem ao PET/SESU/MEC e à PROPP/UFU pelo apoio concedido a esta pesquisa.