Networks and Dynamics of Reddit Political Discussions

Yilin Zhou
The University of Texas at Austin
maeby_zhou@utexas.edu

Abstract

This project constructs networks from Reddit to identify user interaction patterns based on political engagement levels. By representing comments as nodes and measuring linguistic and semantic similarities to form edges, we capture comprehensive interaction patterns. Using Graph2Vec for embeddings, we captured high-level structural features and achieved robust and scalable representations. The results show a significant prediction accuracy of up to 0.8, highlighting differences between political and non-political subreddits. Future work will explore advanced graph embedding techniques and GNN architectures to enhance model performance and interoperability.

1. Introduction

In today's digital age, social media platforms like Reddit have become vital spaces for people to exchange ideas, share opinions, and engage in discussions on a wide range of topics. Among these topics, political discussions stand out due to their intensity and the diverse viewpoints they attract. However, understanding the underlying patterns and dynamics of these interactions can be challenging, especially when trying to distinguish between political and non-political engagement.

Our project addresses this challenge by creating networks from Reddit, where user comments are represented as nodes. We form connections based on linguistic and semantic similarities, allowing us to visualize and analyze user interactions according to their political engagement levels. We use Graph2Vec for embeddings, which capture high-level structural features for robust and scalable representations of these interaction patterns. Moving forward, we aim to enhance our model's performance and interoperability by exploring advanced graph embedding techniques and Graph Neural Network (GNN) architectures.

2. Previous Work

Unsupervised graph classification has been a significant research interest, leveraging structural graph properties for diverse applications. Traditional algorithms, such as K-Means [1] and spectral clustering [2], while effective in certain scenarios, often struggle with the high dimensionality and sparsity of large-scale graph data.

Recent advances include embedding-based methods like Node2vec [3] and DeepWalk [4], which represent graphs in a lower-dimensional vector space. Graph neural networks (GNNs) [5] are also explored for unsupervised graph classification, which has demonstrated an improved performance by capturing both local and global patterns.

Our approach builds on these methods by integrating TF-IDF [6] and SentenceTransformer [7] to measure both linguistic and semantic similarities, capturing comprehensive interaction patterns. Unlike traditional methods, our technique also considers semantic relationships, providing a nuanced understanding of interactions. We use Graph2Vec [8] for embeddings, combining the Weisfeiler-Lehman subtree kernel and Doc2Vec to create scalable and robust graph representations. This allows us to analyze how content, especially political topics, influences online discussion patterns in subreddits, offering insights into community dynamics.

3. Approach

3.1 Network Construction

The network in this project is built based on comment data from various subreddits, each representing a unique sub-community within the broader Reddit platform. These networks can encapsulate the diverse interactions and connections across different subreddits, providing a comprehensive view of the relationships and features within the Reddit ecosystem for deeper analysis and insights into community behavior.

Given that the *username_encoded* in the original dataset lacks duplicate entries, we considered each comment to represent a unique featureless node. Assuming that linguis-

tic similarity between comments is a valid way to represent interaction patterns between users in Reddit, connections between nodes were established by linking each node to nodes that meet the threshold within the current graph or with its most similar node. The similarity was determined by employing the *Term Frequency-Inverse Document Frequency(TF-IDF)* [6], which can measure how important a word is to a document relative to a collection of documents, and the pre-trained semantic model, *SentenceTransformer(allMiniLML6v2)* [7] to capture the semantic similarity between comments. We can create a robust representation of user interactions by integrating both methods using a weighted average controller by a designated parameter.

We employ two methods to construct the networks. The first method involves connecting each node to its most similar counterpart based on the similarity matrix. The second method involves setting a default threshold and connecting each node to other nodes whose similarities surpass this threshold.

For example, consider the following texts:

- I really love this it makes me happy.
- I just checked out her work awesome.
- I think her personality is great.

The corresponding similarity matrix after using TF-IDF and semantic model is given:

$$\begin{bmatrix} 0 & 0.33704219 & 0.33346806 \\ 0.33704219 & 0 & 0.20380523 \\ 0.33346806 & 0.20380523 & 0 \end{bmatrix}$$

Figure 1 shows the graph constructed by the first method. Figure 2 shows the graph constructed using the second method with a threshold of 0.2.

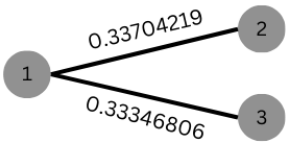


Figure 1: Non-Threshold Graph Based on Similarity.

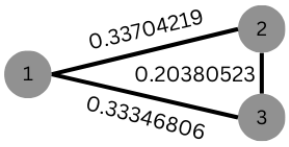


Figure 2: 0.2 Threshold Graph Based on Similarity.

3.2 Graph2Vec Model

We chose Graph2Vec [8] for generating graph embeddings because it captures the structural features of graphs, including the relationships and connection patterns between nodes. This is important for analyzing subreddits, as different subreddits may exhibit varying interaction patterns. For example, some subreddits might have dense discussion clusters, while others may feature more dispersely.

Graph2Vec is highly scalable and versatile. For large subreddits with thousands of nodes, it can efficiently generate embedding vectors without being hindered by the volume. For small subreddits with fewer nodes, it can still produce high-quality embeddings. By leveraging the embeddings generated by Graph2Vec, differences between subreddits can be better understood, improving the clustering and classification performance, regardless of the input size.

3.2.1 Weisfeiler Lehman

By using the Weisfeiler-Lehman (WL) subtree kernel method, Graph2Vec model generates labels for each node, reflecting the structural features of the nodes and their neighbors. These labels are then mapped to high-dimensional feature vectors, resulting in global embedding vectors for the graphs.

For instance, consider a fully connected three-node graph, shown in Figure 3, where each node has an initial feature. By hashing these initial features and iteratively updating them based on the features of neighboring nodes, the WL method generates refined labels that reflect higher-order structural information.

Figure 3: Weisfeiler Lehman labeling Process.

3.2.2 Doc2Vec

Distributed Memory is a variant of the Doc2Vec model. It learns fixed-length vector representations for text data by considering the context in which words appear. In the DM architecture, the neural network takes context words and a unique document ID as inputs. The context words are used to predict a target word, the document ID captures the document's overall meaning. The projection layer generates vectors for both, which are optimized through training to minimize the difference between predicted and actual words. The simplified model structure is shown in Figure 10.

In our model, the WL-generated labels are treated as words (such as Node 2: (2_1_3)), and with each graph as a unique document. We use the vectors of Node 2: (2_1_3), Node 3: (3_1_2), and the document vector to predict Node 1's label while minimizing the difference.

3.3 Main Tasks and Assumptions

Our primary objective is to ascertain whether online discussion patterns are independent of their content or if

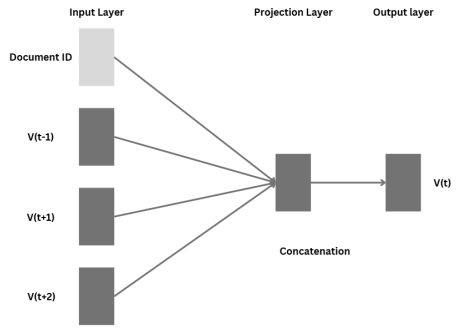
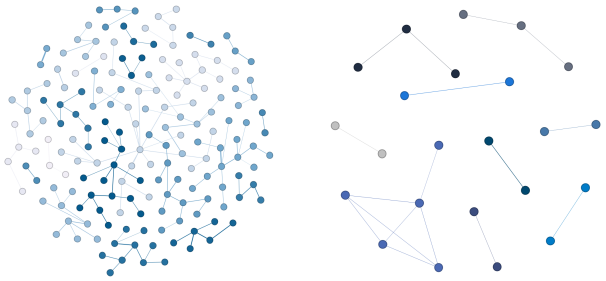


Figure 4: Distributed Memory Architecture.



(a) Actives Subreddit Graph of Non-Threshold

(b) Actives Subreddit Graph of Threshold 0.4

Figure 5: Actives Subreddit Graphs

political-related topics have a significant influence on these patterns. If political topics do impact discussion patterns, this implies that different social clusters may exhibit unique interaction behaviors.

In scenarios where we assume that political engagement within a topic affects interaction patterns, our goal is to identify and analyze these influences. We also assume that the findings from this analysis can be generalized to broader social networking behaviors. By understanding the relationship between political engagement and social interaction patterns, we aim to uncover deeper insights into how content context shapes online community dynamics.

4. Experiment and Result

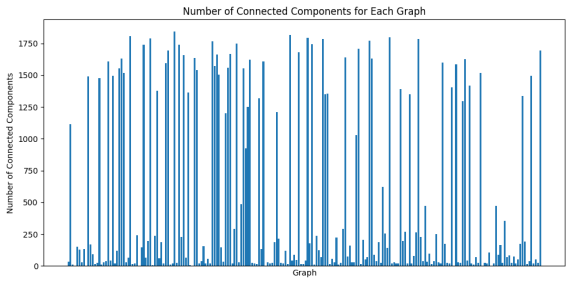
4.1 Dataset

We utilized a Reddit dataset from HuggingFace [9], comprising 214 cleaned subreddits. Each subreddit was processed to eliminate noise and ensure data integrity. The dataset was then employed to generate undirected networks for each module. The raw data was filtered to include only comments within a specific date range (from January 1, 2022, to December 31, 2024). Comment text was cleaned to remove unnecessary characters and standardize the format. We can get average path length, modularity, and other network characteristics to quantify and characterize the network's features and communication patterns. Figures 5 below are illustrative graphs of subreddits generated by Gephi, with different modularities indicated by color coding and weights represented by link thickness. Networks of the same subreddit are shown in Figure 5.

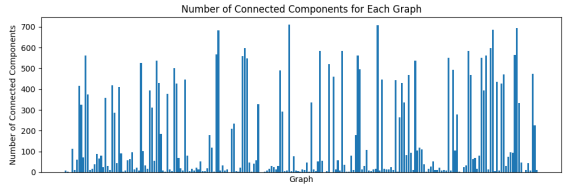
It can be observed that, when we set the threshold to 0.4, we tend to ignore nodes that lack the necessary similarity to form links. This method allows us to capture multiple relationships between nodes compared to the first method. This helps in identifying secondary relationships and substructures within the network, providing a more nuanced view of node interactions.

On the other hand, the first method, connecting each node to its most similar counterpart, focuses on the strongest similarities between nodes. This method emphasizes core relationships but potentially overlooks less prominent, yet still relevant interactions.

The numbers of connected components for each graph exhibit noticeable variation, as illustrated in Figure 6, showing different feature distributions.



(a) Number of Connected Components for Each Graph with Non-Threshold



(b) Number of Connected Components for Each Graph with a Threshold 0.4

Figure 6: Number of Connected Components for Each Graph

4.2 Experiment and Evaluation

4.2.1 Unsupervised Learning Classification Models

We performed clustering analysis using Gaussian Mixture Model (GMM) [10] and K-Means. Results are shown in Table 1 and Table 2. The results highlight distinct pat-

terns in clustering behavior under these two scenarios. Subsequently, we utilized t-SNE for dimensionality reduction and visualization, as shown in Figure 7.

The results demonstrate that applying the threshold significantly influences the clustering distribution. This impact is particularly evident with the increase in non-political items in Cluster 1. Additionally, the threshold application enhances the separation between two clusters, as shown by the clearer distinction in the t-SNE plot. The t-SNE plots reveal two separated clusters, however, it is apparent that clustering algorithms encounter difficulties in adequately distinguishing between political and non-political subreddits, especially within Cluster 0.

Table 1: Clustering Results of GMM

Method	Clusters	Politics	Non-Politics
Threshold 0.4	Cluster 0	67	82
	Cluster 1	5	60
Non-Threshold	Cluster 0	59	78
	Cluster 1	13	64

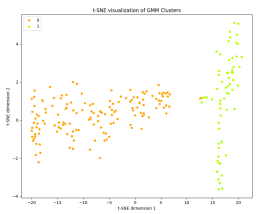
Table 2: Clustering Results of K-Means

Method	Clusters	Politics	Non-Politics
Threshold 0.4	Cluster 0	67	82
	Cluster 1	5	60
Non-Threshold	Cluster 0	60	80
	Cluster 1	12	62

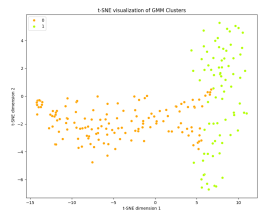
4.2.2 Supervised Learning Classification Model

To further evaluate the performance of our model, we manually categorized these subreddits into 2 labels, Politics and Non-Politics. We then implemented supervised learning models for classification [11]. Logistic regression [12] and Random Forest [13] are widely used statistical methods for binary classification tasks. The dataset was split into training and testing sets with an 80/20 ratio. The performance of this model was evaluated using metrics such as accuracy, precision, recall, and F1-score, shown in Table 3 and Table 4.

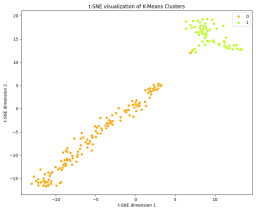
Networks with a threshold of 0.4 exhibit superior performance across all metrics, indicating that this approach more effectively captures the relationships between nodes and distinguishes between different classes. This method enhances the network's ability to identify and leverage multiple node connections, thereby providing a more nuanced understanding of the network's structure.



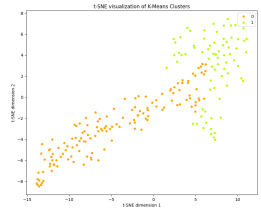
(a) t-SNE visualization of GMM with a Threshold 0.4



(b) t-SNE visualization of GMM with Non-Threshold



(c) t-SNE visualization of K-Means with a Threshold 0.4



(d) t-SNE visualization of K-Means with Non-Threshold

Figure 7: t-SNE visualization of Clusters

Table 3: Logistic Regression Evaluation Results

	0.4	N/A
Accuracy	0.83720930232558	0.6279069767442
Precision	0.86642814549791	0.66998892580288
Recall	0.83720930232558	0.6279069767442
F1-Score	0.80724969974888	0.64412641621944

Table 4: Random Forest Evaluation Results

	0.4	N/A
Accuracy	0.77209302325581	0.6279069767442
Precision	0.75339084438605	0.66998892580288
Recall	0.77209302325581	0.6279069767442
F1-Score	0.75719401866432	0.64412641621944

Conversely, the method of connecting each node to only its most similar counterpart shows lower performance. This approach may miss some crucial relationships between nodes, which can result in a decrease in overall model performance. By focusing solely on the strongest connections, this method potentially overlooks significant yet subtler interactions that are essential for a comprehensive understanding of the network dynamics.

4.3 Result

The data and charts reveal that different subreddits exhibit distinct interaction patterns. Supervised learning clas-

sification models for graph label prediction demonstrate high accuracy, particularly with Logistic Regression, where all metrics exceed 0.8 in the dataset with a threshold of 0.4. While clustering algorithms can identify certain patterns, their overall performance lags behind that of supervised classification methods.

This disparity arises from the inherent limitations of unsupervised learning methods. Without the benefit of labeled data for optimization, unsupervised approaches struggle to capture the structural features as effectively as supervised methods. Additionally, clustering algorithms often find it challenging to discern subtle differences in complex, high-dimensional datasets. Consequently, supervised classification models provide a more reliable and accurate means of predicting graph labels in this context.

Table 5: Avg. Features of Political and Non-Political Subreddits

	Non-Politics	Politics
Degree Centrality	0.0158	0.0323
Betweenness Centrality	0.0024	0.0025
Network Density	0.0158	0.0323
Modularity	0.9604	0.9244

From the metrics of networks with a threshold of 0.4 shown in Table 5, it is evident that political-related subreddits exhibit more frequent and dense interactions compared to non-political ones. These subreddits demonstrate a more integrated community structure, whereas non-political topics lead to more dispersed and independent interactions. This suggests that political discussions tend to generate higher levels of engagement and interconnectedness among users, while non-political discussions are more isolated and less cohesive.

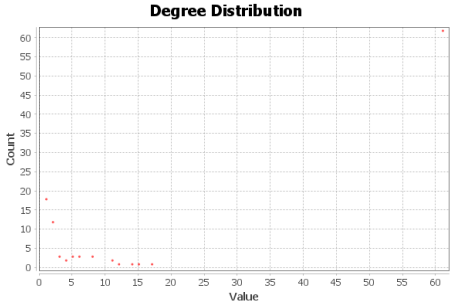
To further illustrate these differences, we selected two typical subreddits: the ask_politics subreddit representing the political type and the biology subreddit representing the non-political type. The average node degree of the ask_politics subreddit is 35.518, with a total node number of 112. In contrast, the non-political subreddit has an average node degree of 6.497, with a total node number of 459. This significant difference in average node degree highlights the varying interaction patterns between political and non-political subreddits.

The visualization plots are shown in Figure 8 and the degree distribution plots are shown in Figure 9. The integrated community structure of political subreddits is visually distinct, indicating stronger and more frequent connections among users. On the other hand, non-political subreddits appear more fragmented, reflecting the less cohesive nature of their interactions.

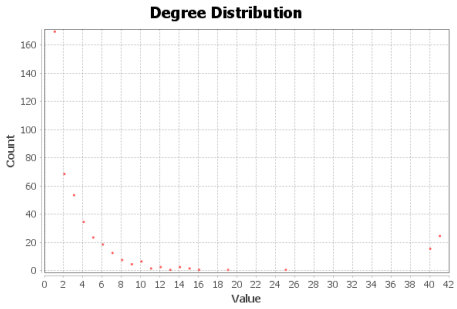
(a) Ask Politics Subreddit Graph with the Threshold of 0.4

(b) Biology Subreddit Graph with the Threshold of 0.4

Figure 8: Visualization of Node Degree in Political and Non-Political Graphs



(a) Average Node Degree of Ask Politics Subreddit



(b) Average Node Degree of Biology Subreddit

Figure 9: Average Node Degree of Two Subreddits

5. Conclusion

Our experiments demonstrated that political-related subreddits exhibit more frequent and dense interactions compared to non-political ones. This suggests that political discussions tend to generate higher levels of engagement and interconnectedness among users, while non-political discussions are more isolated and less cohesive.

The potential impact of this work lies in its ability to provide deeper insights into how content context shapes online community dynamics. By understanding the relationship between political engagement and social interaction patterns, we can better comprehend the behavior of online

communities. This knowledge can be applied to improve content moderation, enhance user experience, and foster healthier online discussions.

From this project, we've gained a comprehensive understanding of how online communities, particularly on Reddit, interact based on varying levels of political engagement. By employing advanced network construction techniques and graph embeddings, we've identified key differences in interaction patterns that highlight the unique dynamics of political discussions. This knowledge can influence how we approach content moderation, enhance user experience, and foster healthier, more constructive online discussions.

References

[1] Education Ecosystem (LEDU). Understanding k-means clustering in machine learning, 2018. Published in Towards Data Science, Accessed: 11.10.2024.

[2] William Fleshman. Spectral clustering: Foundation and application, 2019. Published in Towards Data Science, Accessed: 11.10.2024.

[3] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 855–864, New York, NY, USA, 2016.

[4] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 701–710, 2014.

[5] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1025–1035, 2017.

[6] Fatih Karabiber. Tf-idf — term frequency-inverse document frequency, 2021. Published in LearnDataSci, Accessed: 04.10.2024.

[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019.

[8] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

[9] Arrmlet. Reddit dataset, 2024. Accessed: 24.09.2024.

[10] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[11] Renuka Saravanan and Pothula Sujatha. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)*, pages 945–949. IEEE, 2018.

[12] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Klein. *Logistic regression*. Springer, 2002.

[13] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

Figure 10: Course Evaluation.