Unraveling Social Network Complexity: A Multi-Dimensional Analysis of the Reddit Community

Mohammed A. Fulwala
York University
momo01@my.yorku.ca
Student Number: 217459744
Login: momo01

James Le
York University
jamesmqle@my.yorku.ca
Student Number: 217270943
Login: jamesmql

Stanley Ihesiulo
York University
ihesiulo@my.yorku.ca
Student Number: 216985236
Login: ihesiulo

Anika Prova
York University
anika98@my.yorku.ca
Student Number: 216474306
Login: anika98

ABSTRACT

This project explores the intricate dynamics of the Reddit network, employing advanced graph-based algorithms to unearth community structures, predict links, and analyze temporal evolution. Motivated by the need to comprehensively understand Reddit's complex information flow, our approach combines community detection, sentiment analysis, and temporal analysis. We pose key questions on subreddit connections, community identification, and temporal sentiment variations, driving a thorough exploration. Our research will hold significance as it provides a nuanced snapshot of diverse online communities, offering insights into global social dynamics. Methodologically, we acquire data, conduct sentiment analysis, employ community detection algorithms, and explore link prediction techniques. Experiments reveal correlations between community importance and negativity, the overlap of community structures and topic clusters, and the role of users in sentiment dynamics. The project lays a strong foundation for understanding Reddit's dynamic information flow. Notable results include revealing concentrated negativity contributions from a subset of subreddits, showcasing high cohesion within community-topic clusters, and highlighting the significant role of active users in predicting sentiment. These findings offer valuable insights for researchers and professionals navigating the complexities of online communities.

KEYWORDS

Statistical Significance, Cumulative Percentage, Page Rank, Negative Sentiment, Community Overlap, Topic Clustering, Accuracy, Precision, Link Prediction, Linear Regression, Active Users, Random Forest Classifier, Sentiment Prediction, Temporal Analysis, Community Dynamics, Growth Factor, Node Importance, Social Network Analysis, sentiment analysis, k-means clustering, Jaccard coefficient, Adamic-Adar coefficient, Clauset-Newman-Moore (CNM), bipartite graph

1 INTRODUCTION

1.1 Objective

This project aims to unravel the Reddit network through a multifaceted analysis. Firstly, the objective is to employ advanced graph-based algorithms for the identification of communities within the network, revealing the inherent structures and relationships among diverse subreddits. In addition to community detection, the project delves into link prediction, leveraging historical data and network characteristics to predict potential connections between subreddits. This predictive dimension not only enhances our understanding of emerging trends but also contributes to a proactive comprehension of network evolution. Finally, temporal analysis adds a dynamic perspective, displaying changes in community structures, link formations, and sentiment patterns over different years. By combining community detection, link prediction, and temporal analysis, this project aspires to provide a comprehensive and insightful exploration of the evolving dynamics within the Reddit network.

1.2 Problem Statement

The primary objective of this research is to comprehensively explore and understand the intricate dynamics within the Reddit network. The investigation unfolds on multiple fronts, aiming to unravel the structural intricacies of the network, track the evolution of communities over time, and delve into the nuanced patterns of sentiment in interactions between subreddits. By addressing these aspects, the research seeks to uncover the underlying mechanisms that shape the Reddit ecosystem, providing valuable insights into the interconnected nature of online communities, their temporal transformations, and the sentiments influencing user interactions. This multifaceted approach contributes to a holistic understanding of the Reddit network, shedding light on the intricate interplay between structure, evolution, and sentiment within this dynamic digital community.

1.3 Key Questions

To further explore the Reddit network, key questions must be asked. First and foremost, the project aims to illuminate the nature of connections between subreddits by investigating how they are interconnected. This encompasses understanding the patterns, strengths, and directions of interactions that shape the network's overall structure. Additionally, the question extends to identifying distinct communities within the Reddit network. The goal is to employ advanced graph-based algorithms to outline these communities, showcasing structures and relationships among diverse subreddits. Lastly, the project delves into the temporal dimension, seeking to unravel the variations in sentiment dynamics over time. By tracking changes in community structures, link formations, and sentiment patterns across different years, the project provides insights into the evolution of user sentiments within the Reddit network. In essence, these key questions drive a comprehensive exploration of how subreddits connect, form communities, and exhibit dynamic sentiment patterns over time.

1.4 Importance

This analysis is crucial as it provides a snapshot of diverse online communities within the ever-growing Reddit platform. By examining the structure and interactions, we gain pivotal insights into global online social dynamics. Delving into these communities allows us to understand how users engage and connect within this digital ecosystem. Deciphering emotional cues in interactions unveils subtle nuances in user engagement and factors contributing to community tension. The temporal dynamics of Reddit offer an opportunity to track changes over time, revealing long-term patterns and trends in community structures and sentiments. Addressing these challenges extends beyond understanding Reddit mechanics; it provides broader insights into the dynamic nature of online interactions, sentiments, and community evolution.

1.5 Potential Applications

The project's outcomes offer practical applications for enhanced online platform management. Insights into community structures within the Reddit network provide valuable support for moderators and administrators in fostering healthy online communities. Additionally, understanding community dynamics improves content recommendation systems, refining algorithms for more personalized user experiences. The sentiment monitoring capabilities developed in this project find applications in businesses and organizations, offering a unique perspective on public opinion across different communities. This concise approach underscores the project's contributions to community management, content recommendations, and sentiment monitoring for a more effective and responsive online environment.

2 PROBLEM DEFINITION

2.1 Notation

Let $G = (V, E)$ be a directed graph representing the Reddit network, where V is the set of nodes (subreddits) and E is the set of directed edges representing interactions. The sentiment (i, j) represents the sentiment of the interaction from node i to node j. CmtyV is the set of communities detected in the network.

2.2 Formal Definition

Community Detection: Define CmtyV = $\{C_1, C_2, \ldots, C_k\}$ where C_i is a community in the network. A community is a subset of nodes, i.e., $C_i \subseteq V$. The goal is to find CmtyV that maximizes the modularity Q, where Q is a measure of the quality of the division of the network into communities.
Sentiment Analysis: Let sentiment(i, j) be the sentiment of the interaction from node i to node j. The sentiment can be binary, e.g., sentiment$(i, j) = -1$ or sentiment$(i, j) = 1$, indicating a negative or positive sentiment, respectively.
Nodes and Edges: Let each user-defined community, also known as a subreddit in Reddit, be a node in our network. Each edge in our directed graph depicts an interaction from the source community to the target community.

2.3 Constraints or Restrictions

The graph structure, denoted as G, is directed and signifies the connections between subreddits within the Reddit network. This directed graph captures the flow of interactions from one subreddit to another. Moreover, the sentiment analysis in this context is simplified to binary values. Each interaction is characterized as either positive or negative, offering a straightforward representation of the sentiment associated with the connections between subreddits. This binary sentiment approach streamlines the analysis, focusing on the polarity of interactions for ease of interpretation and computational efficiency.

2.4 Optimization Goal

Community Detection: Maximize the modularity Q of the partition CmtyV of the network.

2.5 Hardness of the Problem

Community Detection Complexity:

The community detection problem is known to be NP-hard. Finding the optimal partition that maximizes modularity involves exploring a combinatorially large solution space.

Sentiment Analysis Complexity:

Assigning sentiment to each edge involves a linear scan of the dataset, but the choice of sentiment values can impact the analysis. If sentiment assignment is part of an optimization process, it may introduce additional complexity.

2.6 Formal Problem Description

Given a directed graph $G = (V, E)$ representing the interactions between subreddits and sentiment values sentiment(i, j), the problem is to find a partition CmtyV of the graph that maximizes the modularity Q, considering the sentiment of interactions.

3 RELATED WORK

Li et al. discusses the problem of detecting social communities in large social networks, with a focus on communities with high outer influence. Outer influence refers to a community's ability to spread its internal information to external users, making it relevant for applications like Ads trending analytics, social opinion mining, and news propagation pattern discovery. Existing community detection techniques often overlook outer influence. To address this, they introduce the "Most Influential Community Search" problem, aimed at revealing communities with the highest outer influences. It proposes a new community model called "maximal kr-Clique community," which emphasizes cohesiveness, connectivity, and a minimum size of k nodes. This model overcomes the limitations of traditional community models like k-core and k-truss. The text also presents a tree-based index structure called "C-Tree" to efficiently maintain the offline computed r-cliques. It further introduces four advanced index-based algorithms to improve search performance. These algorithms enhance the efficiency and effectiveness of influential community searches. The problem discussed differs from traditional influence maximization problems, as it focuses on community-level influence rather than individual-level influence. They highlight that this problem is challenging, being NP-hard. [5]

Alexander Mantzaris discusses the analysis of large online social networks, focusing on community detection, boundary nodes, and the spread of information within these networks. He mentions the challenges posed by the size of these networks and the need for efficient analytical tools. He highlights the importance of community structures within these networks, which can be driven by factors like homophily and brand identity. The spread of information, such as news and viral content, is discussed as a critical aspect of social networks. He also outlines an algorithm for measuring boundary node proximity, which identifies nodes that play a key role in exchanging information between different communities within a network. The algorithm involves the use of random walks and community detection techniques to identify these influential nodes. The passage further discusses the results of applying the algorithm to synthetic datasets and real-world examples, demonstrating its effectiveness in identifying boundary nodes and the spread of information within networks. [7]

M. Li et al. highlights the importance of predictive models in comprehending information diffusion within social networks. They introduce three significant predictive models: the Independent Cascade Model (ICM), the Linear Threshold Model (LTM), and the Game Theory Model (GTM). The ICM is primarily concerned with forecasting and influencing research and incorporates scalability and practical applications. Conversely, the LTM places a strong emphasis on the cumulative influence and is frequently utilized for influence maximization studies, employing heuristic and greedy algorithms for identifying influential nodes. Game Theory models, the third category, delve into the strategic interactions among individuals or groups within the context of information dissemination, considering factors such as costs, benefits, and relationships. Researchers often combine elements from these models to construct more robust predictive frameworks, thus enhancing our understanding of information propagation within dynamic social networks. This introduction is part of a broader exploration of information diffusion within social networks, which also delves into various other aspects such as the emergence of social platforms, practical applications of diffusion models, and different categories of influence analysis, including individual, community, and influence maximization. The comprehensive understanding of these elements plays a vital role in the study of how information spreads through social networks. [8]

4 METHODOLOGY

The dataset was acquired through [6] and [21]. The 2 main datasets file that we used are: 'soc-redditHyperlinks-body.tsv' and 'post_crosslinks_info.tsv'. The 'soc-redditHyperlinks-body.tsv' dataset files are in tabulated format that contains the columns for SOURCE_SUBREDDIT, TARGET_SUBREDDIT, POST_ID, TIMESTAMP, LINK_SENTIMENT, PROPERTIES. The 'post_crosslinks_info.tsv' dataset file contains the raw data where each row contains data about a particular interaction - the source subreddit, target subreddit, the source and target posts, the timestamps of the source and target posts and the user who made the post. While we mostly use the 'soc-redditHyperlinks-body.tsv' for constructing subreddit to subreddit network, the 'post_crosslinks_info.tsv' is useful to construct the bipartite graph that we discuss later in this section.

4.1 Identification of Problematic Subreddits and Sentiment Analysis

For our first analysis, we explore the relationship between the number of outgoing negative sentiments from a subreddit and the pageRank of the subreddit (node). We also highlight the contribution of a subset of communities to the overall negative sentiment. For this analysis, we collected the data from the snap libary's post_crosslinks_info source. The data was cleaned, formatted, and stored in formatted_data_file.txt. We used the SNAP library to read the formatted data and create a directed graph (G), where each node represent a subreddit and the edges represent an interaction between the two communities. In this paper, an interaction is defined as a post in the source subreddit that refers to a post in the target subreddit. When constructing the graph, nodes, and edges were iteratively added based on the subreddit interactions described in the formatted_data_file.txt. This graph has:

Total Nodes	Total Edges
35776	286561

Alphanumeric post IDs and subreddit names were converted to integers using a base-36 conversion. The sentiment of each edge is defined as the sentiment of each interaction from the source subreddit to the target subreddit. The data for the sentiment was obtained from a data source named as 'label_info.tsv' where negative sentiments and non-negative sentiments are recorded as 'burst' and non-negative respectfully'. We later interpreted the sentiments as negative = 1 and non-negative = 0. Then, we used SNAP to calculate each node and performed a correlation analysis between the PageRank and the number of outgoing nodes per subreddit. Furthermore, we also evaluate the subset of subreddits to negative sentiments - what percentage of subreddits contribute to 75% of the negative sentiment? We do so by visualizing the cumulative percentage of negative sentiment and the number of subreddits that contributed to it. The results of this analysis are discussed in section 5.1

4.2 Community detection of subreddits and topic clustering

For our second analysis, we explore the overlap of community structure (created through community detection of the network graph) and the topic clusters (generated from word vectors of posts in each subreddit). The graph generation process is very similar to the process discussed in the first analysis and produces the same number of nodes and edges as before. We used the 'soc-redditHyperlinks-body.tsv' dataset for network graph generation. For topic clustering, we used the 'subreddit_embeddings.csv' dataset. This dataset is a Term Frequency-Inverse Document Frequency word vector - constructed for each subreddit from 300 posts made in that subreddit. To begin the topic clustering process, we first determined the cluster count by using silhouette score [22] to find the optimal number of clusters for k-means. Using the elbow method [23], we found that the cluster count is 7. Then, we used the 'sklearn' library's k-means clustering method to detect 7 clusters from the tf-idf vectors of the subreddits. The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), while the elbow method is a technique for finding the optimal number of clusters in a dataset by identifying the "elbow" point in a plot of the within-cluster sum of squares against the number of clusters. This gives us 7 clusters based on similar content or topic within each cluster. K-means clustering is an iterative algorithm that partitions data points into k distinct, non-overlapping subsets/ clusters based on feature similarity. For community detection within the network graph, we used the Clauset-Newman-Moore (CNM) algorithm that groups subreddits (nodes) that share strong connections [24]. Clauset-Newman-Moore (CNM) is a community detection algorithm that optimizes modularity to identify cohesive groups within complex networks.The CNM algorithm divided the graph into 2123 communities and achieved a modularity of 0.9987679294249004. Modularity is a measure of how well the communities within a network are separated. Finally, we analyze the overlap of network communities with the clusters. We compute the overlap matrix from the overlap coefficient of each community-cluster pair. The experimentation process and results of this analysis are described in section 5.2.

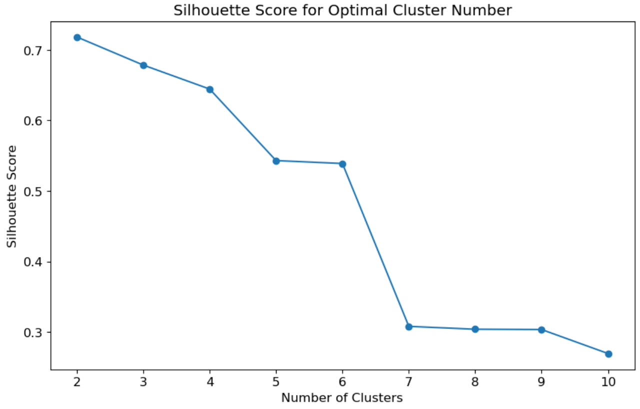

Figure 1: Elbow method to detect the number of clusters for topic modelling

4.3 Link Prediction within the unipartite graph

For this analysis, we explore link prediction within the network graph by using various link prediction techniques. Jaccard coefficient is a measure of similarity between two sets, calculated as the size of the intersection divided by the size of the union of the sets. Adamic-Adar coefficient is a measure of similarity between two nodes in a network, calculated by summing the reciprocals of the logarithm of the degree of common neighbors. We use the same unipartite network graph that we have been using in the previous analysis. However, in order to compute the jaccard coefficient, we convert the graph into a simple graph. At first we extract the positive and negative edges based on the sentiment. To begin the training process, we split the data into training (0.8) and test sets (0.2). We then compute jaccard coefficient of each edge in the training and test set. We use the coefficient as feature for training the classifier. We used the sklearn's Random Forest Classifier [26], and assigned a higher weight to the minority class (negative sentiments in our case). Random Forest Classifier is an ensemble learning method that uses multiple decision trees during training and outputs the class by using a combination of classification and regression to improve the predictive accuracy and control overfitting. We then repeat the training process using Adamic/ Adar coefficient of the edges. The accuracy and precision report of the both the models are briefly described in section 5.3.

4.4 Bipartite graph with user nodes and community nodes

In this analysis, we explore link prediction further by introducing user nodes in the network graph. We use the data in the 'formatted_data_file.txt' to create a bipartite graph [25]. A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets and, every edge connects a vertex in to one in. The nodes in each disjoint set represent a user and subreddit respectfully. The directed graph has a edge from the user node to the target subreddit node where they post. The sentiment of each edge is extracted from the 'label_info.tsv' file. At a higher level, this is what the bipartite graph looks like the figure shown below.

The graph has:

Total Nodes	Total Edges
180283	286561

Once the graph generation completed, we iterate through each edge of the bipartite graph. For each edge we get the source node id (user node id) and store it in a user_out_degree dictionary and increments it to keep track of the number of outgoing edges for each user. We also keep track of the sentiment of the edge and store it in a user_negative_sentiment dictionary that keeps track of the number of negative edges going out per user node. To eliminate noise, we only consider user nodes with more than 1 outgoing edge. In the next step, we train a linear regression model using the values of the dictionaries. The results of the visualisation show a positive correlation between number of out-going edges and number of negative sentiments outgoing per user node. To add more reliability to the experiments, we also trained a random forest classifier using number of outgoing edges for a user node, in order to predict the

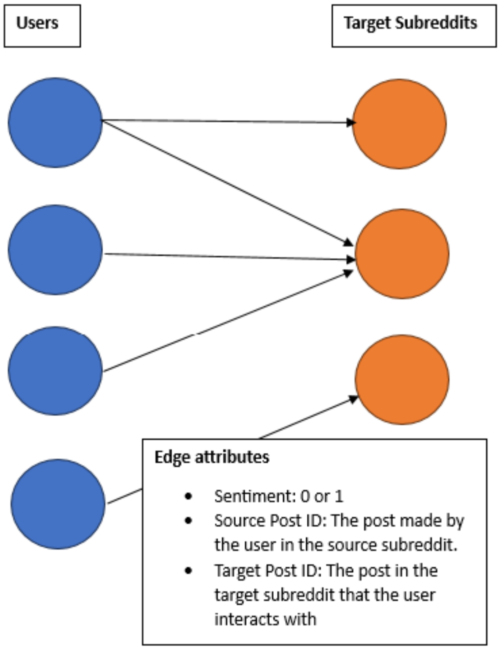
Users
Target Subreddits
Edge attributes
Sentiment: 0 or 1
Source Post ID: The post made by the user in the source subreddit.
Target Post ID: The post in the target subreddit that the user interacts with

Figure 2: High level overview of the bipartite graph

sentiment of a edge. The results of the models are discussed in section 5.4

4.5 Temporal Analysis

We undertook the work of analyzing how sentiments across communities changed over time and the effects of inter-community sentiment on a community’s importance over time. We first analyzed the negativity trends across communities. One such analysis involved looking at the top negative subreddits in 2014. We identified the negativity of subreddits by finding out how many negative posts had been made on that subreddit linking to another subreddit. We then tracked the top 10 negative subreddits in 2014, analyzing how their negativity changed over time, through the years up to and including 2017. As can be seen in the figure below, the subreddits under examination seem to exhibit a decreasing trend in negativity across time. It does seem that the “iama,” “dogecoin,” and “explainlikeimfive” subreddits had a spike in negativity in the year 2016, only to decrease to a nominal range the following year. The “iama” subreddit reached a sudden peak of 140 negative posts this year. It may be of importance that this year, 2016, was the year of the US presidential election between candidates Donald J. Trump and Hillary Clinton. Given the widespread media coverage and attention around this topic, it is plausible that it could have been a cause of the spike in inter-community conflict during this period.

We conducted a similar analysis of negativity trends, but this time tracking the most important subreddits, as decided on by their Page-Rank score. Page rank is a key algorithm used to assess the importance of web pages, originally developed by Larry Page and Sergey Brin at Google. It assigns each page a numerical weight based on its link structure, determining its relative significance within the network. In our analysis, we leveraged Page-Rank scores to identify and track the most influential subreddits, providing valuable insights into the trends and dynamics of negativity within these prominent online communities. As can be seen in the graph below, the top 10 important subreddits have their negativity levels remain fairly stable across the years. There does exist a spike in negativity in the year 2016, but the overall negativity of the most important subreddits seems to have remained relatively stable across the years.

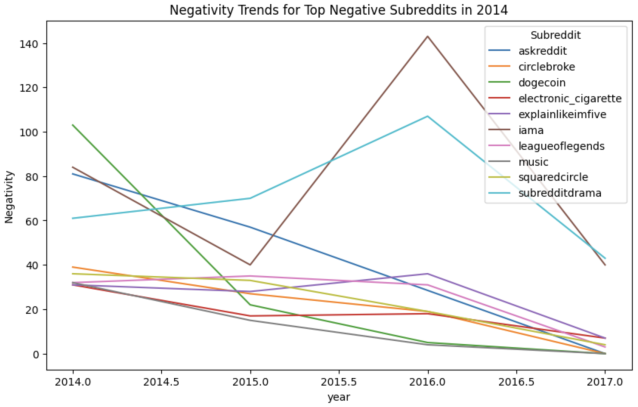

Figure 3: Negativity Trends for Top Negative Subreddits

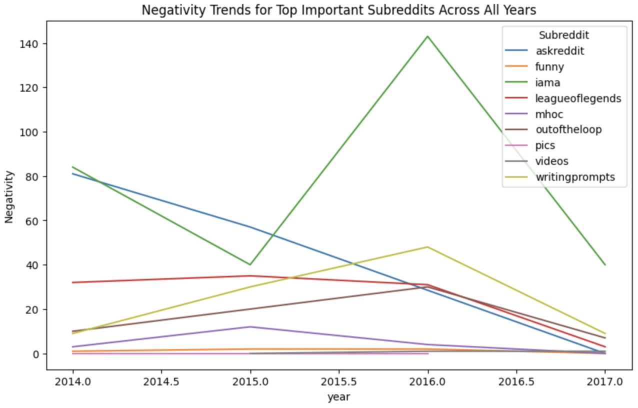

Figure 4: Negativity Trends for Top Important Subreddits

We also analyze how the importance of highly negative subreddits changes over time. We considered 10 top negative subreddits. As can be seen in the graph below, the importance of highly negative subreddits consistently went down over the years, reaching a new baseline.

5 EXPERIMENTS / EVALUATION

5.1 Unveiling community dynamics

The correlation coefficient of PageRank of the nodes (subreddits) and the number of negative edges is -0.0571646992463337. The

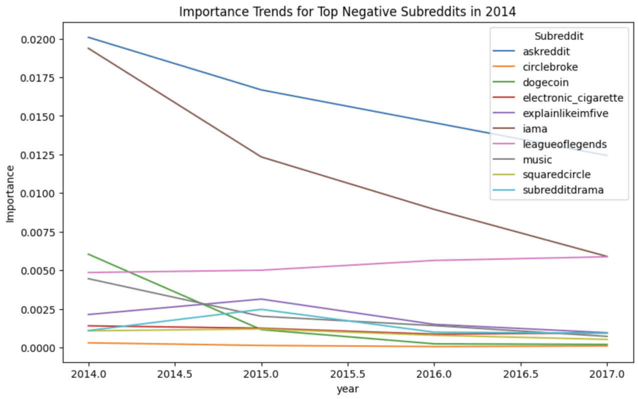

Figure 5: Important Trends for Top Negative Subreddits

negative correlation coefficient suggests a weak negative relationship and the number of outgoing negative sentiment edges per node. A very low p-value of 2.740326011169841e-27 implies that the observed correlation is statistically significant.

The visualization of the cumulative percentage of negative sentiment and the number number of subreddits highlights the concentrated contribution of a subset of subreddits to negative sentiments - 3% of subreddits are responsible for 75% of the conflict.

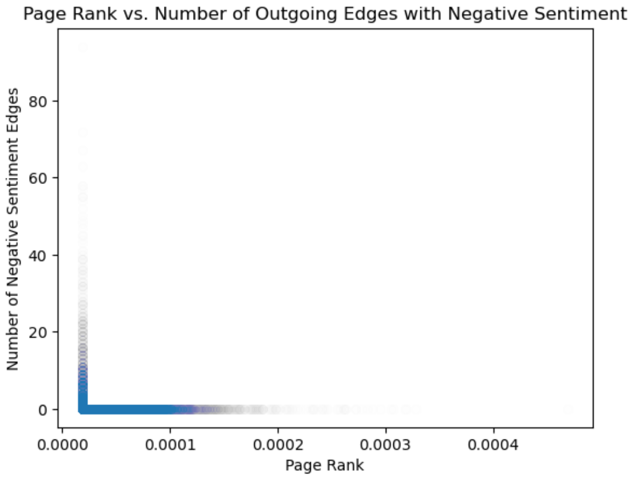

Figure 6: Page Rank vs. Number of Outgoing Edges with Negative Sentiment

5.2 Overlap of community and Topic Cluster

To generate the overlap matrix, our algorithm iterates through each community and cluster pair. For each pair, it calculates the overlap coefficient, which is the size of the intersection of community and cluster subreddits divided by the minimum size of either set. Intuitively, we tabulated the data where each row shows the community ID along with its overlap coefficients for each cluster. Then we computed the standard deviation of each row - 81.82% of the communities had a low standard deviation (below 0.2), indicating high cohesion - the subreddits within a community tend to come from the same topic cluster.

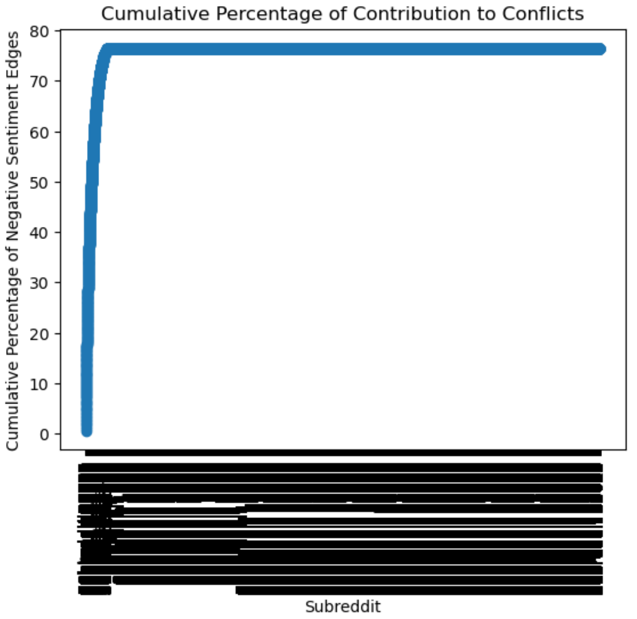

Figure 7: Cumalative Percentage of Contribution to Conflicts

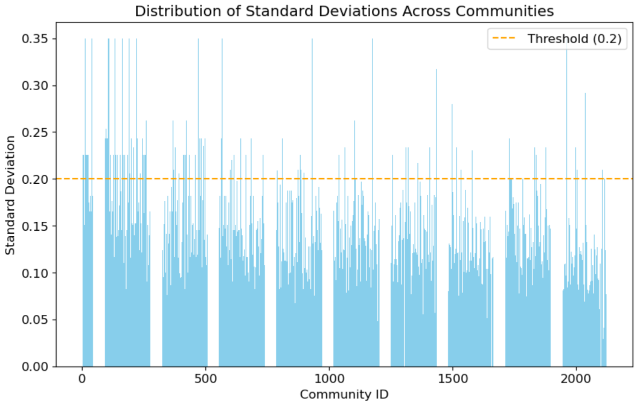

Figure 8: Subreddits of a community tends to belongs to the same topic cluster

To further our experiment on topic modeling of interacting subreddits (nodes), we iterate through the edges of the network graph and report the number of edges for which the source and target subreddit come from the same topic cluster and same community.

Category	Number of Edges
source and target subreddit from same cluster	91,486
source and target subreddit from community	9,668

From this analysis, we can safely conclude that interaction between subreddits tends to happen within subreddits of similar topics.

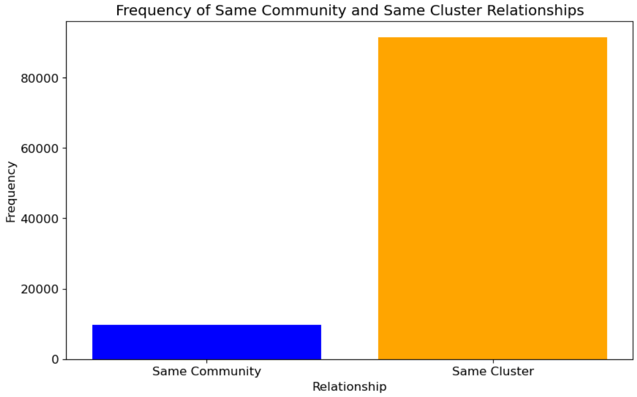

Figure 9: Frequency of Same Community and Same Cluster Relationships

5.3 Accuracy and precision report of link prediction in the unipartite (community node only) graph

Both the models using jaccard and Adamic/ Adar coefficients exhibit moderate overall accuracy, but shows low precision when detecting edges with negative sentiment. The accuracy report of the models are shown below:

	Accuracy	Precision (predicting negative sentiment edges)
Jaccard coefficient model	0.489	0.090
Adamic/Adar model	0.926	0.074

The class imbalance due the number of positive sentiments being much higher than negative sentiment edges could affect model learning, leading to biased predictions.The choice of features (Jaccard coefficients, Adamic/Adar coefficients) might not be sufficient for capturing sentiment dynamics effectively.

5.4 Link sentiment prediction and active users

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The visualisation of the line of best fit of the linear regression [27] implies that the more active a user is (higher number of outgoing edge), the more likely they are to initiate a conflict.

The accuracy report of the random forest classifier model are shown below:

Accuracy	Precision for negative sentiments
0.8277665995975855	0.22580645161290322

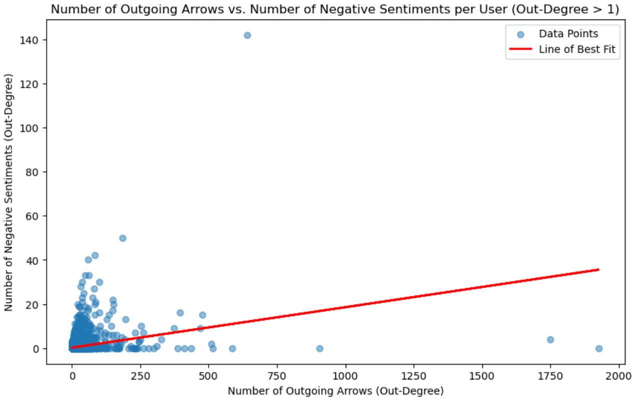

Figure 10: Active user are more likely to start a negative sentiment

This shows that users are important players in the network graph. Introducing users nodes can help to make sentiment prediction more accurately

6 CONCLUSION

This project has set out to gain a comprehensive understanding of the dynamics of information flow within the Reddit network, with a specific focus on the influence of individual users, subreddits and and the evolution of communities. The project was structured into several key phases, including data collection and preprocessing, community detection and analysis, the identification of influential users, and temporal analysis of community structures.

The utilization of the "soc-RedditHyperlinks" dataset provided valuable insights into the complex interactions and information dissemination within the Reddit network. The use of community detection algorithms allowed us to uncover and analyze the groups of users with shared interests, shedding light on the intricacies of community-level statistics and how these communities change over time. Identifying influential individuals was a critical aspect, and the choice of betweenness centrality as a measure of node importance aligned with the project's focus on information dissemination.

The network's subtle patterns and trends were exposed by the analysis. The study of the cumulative percentage of negative sentiment brought even more attention to this concentrated negativity by showing that just a small percentage of subreddits contribute to majority of conflicts.

Communities of subreddits tend to form around the same topic cluster. The overlap analysis between topic cluster and community structures indicates a high degree of cohesion within communities. This suggests that Reddit communities are built around shared content or themes, as well as patterns of interaction. Our experiments on link prediction in bipartite with user nodes and unipartite with nodes provided important insight into interaction dynamics. Models had issues due to feature selection and class imbalance, even though they were relatively accurate. Introducing users as nodes in bipartite graph showed that active users play an important role in stoking negative opinions.

Temporal analysis revealed patterns of negativity in prominent subreddits and communities. The increase in hostility that accompanied the 2016 US presidential election serves as a reminder of how world events affect the dynamics of online communities. Furthermore, monitoring the prominence of extremely negative subreddits over time revealed a steady decline, suggesting a possible change in the dynamics of the network. In summary, our project adds to our understanding of the intricacy of the Reddit network by providing information on sentiment patterns, community dynamics, and user roles. The analysis's multifaceted approach lays the groundwork for future studies in sentiment monitoring, community detection and online community management, all of which will enhance the effectiveness and responsiveness of the online environment.

7 GITHUB REPOSITORY

Visit https://github.com/mohful/Social-Media-Analysis/tree/anika to view our full code.

8 ACKNOWLEDGEMENT

ChatGPT was used to generate sections of the report, mainly summaries of research papers to assist us in writing the "Related Work" section.

9 REFERENCES

[1]"Betweenness centrality and other essential centrality measures in network analysis," Memgraph, (accessed Nov. 4, 2023).

[2]Jung Hyuk Seo et al., "Finding influential communities in networks with multiple influence types," Information Sciences, (accessed Nov. 4, 2023).

[3] P. Ganesh, "Influential communities in Social Networknbsp;: Simplified," Medium, https://towardsdatascience.com/influential-communities-in-social-network-simplified-fe5050dbe5a4 (accessed Nov. 4, 2023).

[4] Farzaneh Kazemzadeh et al., "Influence maximization in social networks using effective community detection," Physica A: Statistical Mechanics and its Applications, (accessed Nov. 4, 2023).

[5] Jianxin Li et al., "Most influential community search over large social networks" | IEEE ...,
https://ieeexplore.ieee.org/document/7930032/ (accessed Nov. 5, 2023).

[6] "Social Network: Reddit hyperlink network," SNAP,
https://snap.stanford.edu/data/soc-RedditHyperlinks.html
(accessed Nov. 4, 2023).

[7] A. Mantzaris, Uncovering nodes that spread information between communities in social networks,
https://link.springer.com/content/pdf/10.1140/epjds/s13688-014-0026-9.pdf (accessed Nov. 5, 2023).

[8] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," MDPI, https://www.mdpi.com/2078-2489/8/4/118 (accessed Nov. 4, 2023).

[9] B. Rozemberczki, R. Davies, R. Sarkar and C. Sutton. GEMSEC: Graph Embedding with Self Clustering. 2018.

[10] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

[11] B. Rozemberczki, C. Allen and R. Sarkar. Multi-scale Attributed Node Embedding. 2019.

[12]"Facebook network analysis," Facebook Network Analysis - NetworkX Notebooks (accessed Oct. 14, 2023).

[13] Srivastav, Manoj & Nath, Asoke. (2015). Study on Mathematical Modeling of Social Networks. International Journal of Emerging Technology and Advanced Engineering ISSN-2250-2459 I ISSN-2250-2459. Volume 5,. 611-618.

[14] A. Lua, "21 top social media sites to consider for your brand -," Buffer Library (accessed Oct. 14, 2023).

[15] S. J. Dixon, "Facebook mau worldwide 2023," Statista (accessed Oct. 14, 2023).

[16] Author links open overlay panelJooho Kim et al., "Social network analysis: Characteristics of online social networks after a disaster," International Journal of Information Management,
https://www.sciencedirect.com/science/article/abs/pii/S026840121730525X
(accessed Oct. 14, 2023).

[17] Author links open overlay panelQiuju Luo a b et al., "Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites," Tourism Management,
https://www.sciencedirect.com/science/article/abs/pii/
S0261517714001393 (accessed Oct. 14, 2023).

[18] Zhan, J., Guidibande, V. & Parsa, S.P.K. Identification of top-K influential communities in big networks. J Big Data 3, 16 (2016).
https://doi.org/10.1186/s40537-016-0050-7

[19] C. Egan, "Closeness and communities: Analyzing social networks with Python and NetworkX-part 3," Medium,
https://towardsdatascience.com/closeness-and-communities-analyzing-social-networks-with-python-and-networkx-part-3-c19feeb38223 (accessed Oct. 14, 2023).

[20] "Girvan-Newman algorithm," Memgraph's Guide for NetworkX library, https://networkx.guide/algorithms/community-detection/girvan-newman/ (accessed Oct. 14, 2023).

[21] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Community Interaction and Conflict on the Web. The Web Conference (WWW). 2018.

[22] Sklearn.metrics.silhouette$_{s}$*core.scikit.(n.d.).*
https : //scikit − learn.org/stable/modules/generated/
sklearn.metrics.silhouette_score.html

[23]*Elbowmethod.ElbowMethod−Yellowbrickv1.5documentation.(n.d.).*
https : //www.scikit − yb.org/en/latest/api/cluster/elbow.html

[24]*Greedy*$_{m}$*odularity*$_{c}$*ommunities.greedy*$_{m}$*odularity*$_{c}$*ommunities−NetworkX3.2.1documentation.(n.d.).*
https : //networkx.org/documentation/stable/reference/algorithms/
generated/networkx.algorithms.community.
modularity_max.greedy_modularity_communities.html

[25]"*Bipartitegraph.fromWolframMathWorld.(n.d.).*
https : //mathworld.wolfram.com/BipartiteGraph.html

[26]*Sklearn.ensemble.randomforestclassifier.scikit.(n.d. − a).*
https : //scikit − learn.org/stable/modules/generated/
sklearn.ensemble.RandomForestClassifier.html

[27]*Aboutlinearregression.IBM.(n.d.).*
https : //www.ibm.com/topics/linear − regression