Optimizing Social Media Engagement: A Comprehensive and Integrative Analysis of Reddit Communities

Jay Sarode
Computer Science
Virginia Tech
Virginia, USA
jaysarode@vt.edu

Rashmi Kulkarni
Computer Science
Virginia Tech
Virginia, USA
rashmi.kulkarni@vt.edu

Aditya Kavuluri
Computer Science
Virginia Tech
Virginia, USA
adityakavuluri@vt.edu

ABSTRACT

This pioneering study in social media analytics focuses on enhancing user engagement and interaction within Reddit's diverse communities. Amidst the evolving landscape of social media, understanding user behavior and content preferences is critical for fostering meaningful engagement. This research utilizes comprehensive datasets encompassing network dynamics, sentiment analysis, and content distribution across 100 subreddits. Through a blend of exploratory data analysis, sentiment assessment, and advanced modeling techniques, including recommendation systems and topic modeling, this study reveals insights into the intricate social interactions and content trends that define Reddit. By identifying the underlying patterns of user engagement and the effectiveness of content, the findings aim to guide community managers and content creators in developing targeted strategies that enhance user interaction and community growth on social media platforms. This investigation also serves as a cornerstone for future research in social media analytics, aiming to improve community engagement strategies and content relevance in an increasingly connected world.

Keywords:

networks; network dynamics; centrality; modularity; hierarchical clustering; sentiment analysis; hybrid recommendation systems; topic modeling;

1 INTRODUCTION

In the dynamic and continually evolving realm of social media analytics, understanding and enhancing user engagement within online communities has become crucial. This research paper delves into the multifaceted world of social media interactions, focusing on Reddit, a platform known for its vibrant and diverse user base. As social media's influence expands, the complexities of user interactions and the effectiveness of content in these digital communities become significant areas of inquiry. For marketers, community managers, and platform developers, the ability to understand and optimize user engagement is paramount. Engaged users not only contribute to vibrant online communities but also drive key metrics such as user retention, brand loyalty, and platform growth. Responding to the challenges of digital community management and content strategy, this study introduces an innovative analytical approach. It utilizes extensive datasets covering various aspects of Reddit, including user interactions, post dynamics, and community feedback across 100 subreddits. This integrative analysis employs advanced methodologies such as network dynamics assessment, sentiment analysis, and machine learning techniques encompassing recommendation systems and topic modeling.

The study is structured around several key functionalities, each pivotal to dissecting the social fabric of Reddit:

Analysis and Comparison of Different Network Dynamics: We explore the structural and interaction patterns within subreddits to understand how different network configurations affect user engagement and content dissemination.

Comparison of Sentiment Analysis on Comments: We analyze the sentiment expressed in user comments, this study assesses the emotional responses to content, which is instrumental in understanding user engagement levels and the overall atmosphere within various communities.

Hybrid Recommendation System: We introduce a sophisticated model combining collaborative and content-based filtering techniques to enhance content discovery and personalization on Reddit, aiming to significantly improve user satisfaction and interaction rates.

Comparison of the Effects of Topic Modeling Based on Posts and Comments: We examine the themes that emerge from posts and comments, this part of the study identifies key content trends and their impacts on user behavior and community dynamics.

2 RELATED WORK

2.1 Sentiment Analysis on Reddit Trading Data

Govinda K et. al. [1] studied and developed an advanced machine learning model to perform sentiment analysis on Reddit trading data, aiming to refine trading strategies by leveraging the rich, emotive content shared in Reddit's trading forums. The model combines VADER Analysis to assess the emotional tone of comments, Fourier transforms to smooth data and reduce noise, and Long-Short Term Memory (LSTM) networks to capture long-term dependencies in sentiment trends. These methodologies enable the model to process large volumes of textual data effectively, identifying and predicting market sentiments that could influence trading behaviors and market movements. The research highlights the potential of sentiment analysis in financial contexts, particularly for enhancing decision-making processes by interpreting the complex, dynamic discussions within online trading communities.

2.2 Sentiment Analysis Towards Russia - Ukrainian Conflict: Analysis of Comments on Reddit

Tanmay Nandurkar et. al. [2], delve into the sentiment analysis of Reddit comments concerning the Russia-Ukraine conflict, utilizing the VADER sentiment analysis tool alongside multinomial Naïve Bayes models to capture and evaluate the public's sentiment over time. The authors meticulously preprocess the comments to strip away irrelevant data, and then classify them into positive, negative, or neutral categories according to derived sentiment scores. This systematic approach allows for an in-depth understanding of how sentiments evolve in response to the conflict and underscores the significant role of social media platforms like Reddit in influencing and reflecting public opinion on major global political issues. The study not only maps the fluctuating public mood but also illustrates the power of sentiment analysis in parsing complex, large-scale data to glean insights into collective attitudes and emotions.

2.3 A Personalized Travel Recommendation System Using Social Media Analysis

Joseph Coelho et. al. [3], presents a personalized travel recommendation system that leverages social media analysis, with a specific focus on Twitter data. The system utilizes advanced machine learning techniques to classify tweets related to travel and to personalize travel recommendations according to user preferences, which are gleaned not only from the users' tweets but also from those of their followers. This methodological framework incorporates sentiment analysis to interpret the emotional content of tweets and collaborative filtering to refine and target recommendations more effectively. By analyzing the interactions and expressions within users' social networks, the system is able to offer tailored suggestions for destinations, enhancing the accuracy and relevance of travel recommendations. This innovative approach demonstrates the potential of integrating social media insights into practical applications, making travel planning more aligned with individual interests and social dynamics.

3 DESIGN OVERVIEW

3.1 Terminology Definitions

Definition 1. (Analysis and Comparison of Different Network Dynamics): Let $Dnwi$ represent the network insights dataset, which comprises network metrics for each subreddit indexed by i. These metrics include the Number of Nodes (NNi), Number of Edges (NEi), Average Degree (ADi), Average Network Density ($ANDi$), Average Degree Centrality (DCi), Average Closeness Centrality (CCi), Average Betweeness Centraility (BCi), Clustering Coefficient ($CCoeffi$), and Modularity ($Modi$). The final dataset is created from the union of the three datasets $D = Dnwi \cup Dpc \cup Dnw$.

Definition 2. (Comparison of Sentiment Analysis on Comments): From the combined dataset D, we extract comment data c and apply natural language preprocessing through the transformation function T to refine text quality. Specifically, for each comment c, the sentiment model will calculate the respective sentiment scores and classify them into negative ($sneg$), neutral ($sneu$), positive ($spos$), and compound ($scomp$) scores. These scores enable the classification of comments into sentiment categories based on predefined thresholds, enhancing the understanding of community emotional responses and engagement dynamics.

Definition 3. (Hybrid Recommendation System): The combined dataset D, undergoes natural language processing to enhance the textual data quality, facilitated by the transformation function T, leading to preprocessed subreddit descriptions S. The TF-IDF vectorization is applied to S, generating a feature matrix M. Subsequent cosine similarity computation between subreddit vectors in M yield a similarity matrix Sm. The recommendation function R utilizes Sm to suggest subreddits based on topic relevance, while a collaborative filtering method, based on user interactions with posts further refines these recommendations by integrating user-specific preferences and interaction process. The system effectively combines content-based and collaborative filtering approaches, optimizing subreddit recommendations tailored to user interests.

Definition 4. (Comparison of the Effects of Topic Modeling Based on Posts and Comments): Given the combined dataset D, the posts and comments are extracted, where preprocessing using natural language techniques to standardize and tokenize text is done, followed by constructing a dictionary $Dict$ and corpus C that facilitate the extraction of prevalent themes using the Latent Dirichlet Allocation model. The process results in the identification of dominant topics within each subreddit post and comment, which are subsequently compared to understand how user discussion and interactions vary thematically between two modes of communication.

3.2 Problem Definition

The primary objective of this research paper is multi-faceted, focusing on enhancing social media user engagement and interaction within Reddit's diverse communities through a comprehensive and integrative approach that spans several key dimensions of social media analytics. The research aims are articulated as follows:

Network Dynamics Analysis: The first goal is to dissect and understand the intricate network dynamics across various Reddit subreddits. This involves an in-depth analysis of structural and relational metrics, aiming to reveal patterns of user interactions and the organizational properties of social ties within communities.

Sentiment Analysis Comparison: The second objective focuses on the sentiment analysis of user-generated content, particularly comparing sentiments expressed in comments. By applying advanced sentiment analysis techniques, namely the Vader sentiment analysis and PMI sentiment analysis, the study aims to decode the emotional context of discussions, contributing to an understanding of how content sentiment influences user engagement.

Hybrid Recommendation Systems: The third aim is to develop and implement a hybrid recommendation system that incorporates both collaborative and content-based filtering techniques. This system is designed to enhance content discoverability and personalize the user experience by recommending relevant subreddit discussions based on user preferences and interaction history.

Topic Modeling Comparison: The final goal is to compare the effects of topic modeling applied to posts versus comments. This comparison seeks to identify prevailing themes and topics within each communication mode, providing insights into content strategy and community management practices that can boost user participation and content relevance.

Definition 5. (Comprehensive Social Media Optimization Problem): Given the combined dataset D, which includes subreddit interactions, post details, and user comments from sources Dnwi, Dpc, and Dnw, we aim to:

1. Analyze D to identify network dynamics to ascertain structural characteristics and interaction patterns across subreddits.

$$\text{Analyze } D = \bigcup_{i=1}^{n} (D_{\text{NWI}} \cup D_{\text{DAI}} \cup D_{\text{ND}})$$

2. Apply sentiment analysis in D to compare user emotions, focusing on extracting emotional contexts from user comments and aiding in the understanding of content reception.

$$\text{Sentiment Analysis}(D) = \{s \mid s = \text{Sentiment}(d), \forall d \in D\}$$

3. Develop a hybrid recommendation model M, where R represents recommended subreddits or topics tailored to user preferences derived from their interaction patterns within D.

$$M : D \to R$$

4. Conduct topic modeling to compare content themes in posts and comments, through which we derive topics from posts and comments, respectively.

$$T_{\text{posts}}, T_{\text{comments}} = \text{Topics}(D_{\text{posts}}), \text{Topics}(D_{\text{comments}})$$

4 APPROACH

4.1 Analysis and Comparison of Different Network Dynamics

The core of our analytical methodology revolves around the exhaustive examination of network dynamics across a diverse set of 100 subreddits. Utilizing an integrated dataset, this study employs both exploratory data analysis (EDA) and advanced network analytical techniques to extract and scrutinize the underlying patterns within these Reddit communities.

4.1.1 Network Structure and Composition Analysis

Initially, we delved into the dataset by examining the basic network structure and user interactions within the subreddit communities. Descriptive statistics provided a foundational understanding of the network's scale and scope, including metrics such as the number of nodes and edges. Histograms and density plots helped visualize the distribution of these metrics, providing a clear picture of community sizes and connectivity levels.

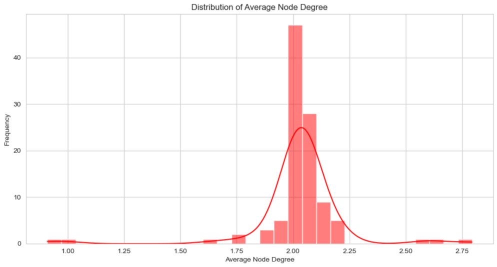


Fig. 1: Distribution of Node Degree across Subreddits

4.1.2 Centrality and Influence within Networks

Further, we explored the centrality measures to identify key influencers and pivotal elements within the network. Using calculations of degree, closeness, and betweenness centrality, we determined which subreddits wield significant influence over information flow and community engagement. These metrics were visualized using scatter plots and heat maps to highlight the variance in influence across the network.

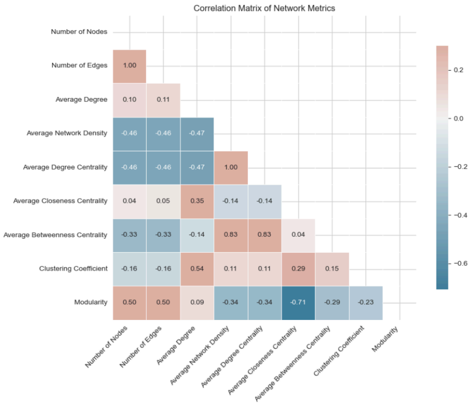


Fig. 2: Correlation of Network Metrics

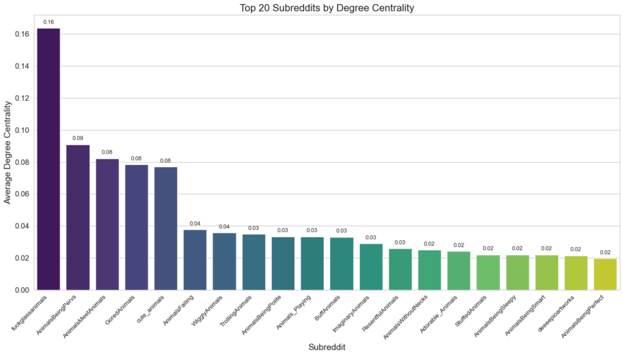


Fig. 3: Top 20 Subreddits by Degree Centrality

4.1.3 Hierarchical Clustering Analysis

We applied hierarchical clustering to systematically segment the network into groups of subreddits that demonstrate similar interaction patterns. This analytical method allowed us to observe the arrangement of subreddits into hierarchical structures, facilitating a deeper understanding of their interaction levels and community bonding. Dendrograms were utilized to visually represent these hierarchical relationships, showcasing the connections between various clusters within the network.

The network features were normalized:

$$X_{\text{normalized}} = \frac{X - \mu}{\sigma}$$

where X is the respective value in the dataset, μ is the mean, and σ is the standard deviation.

The linkage matrix Z is computed using Ward's method, which minimizes the total within-cluster variance. At each step, the pair of clusters with minimum between-cluster distance is merged:

$$Z = \text{linkage}(X_{\text{normalized}}, \text{'ward'})$$

After constructing the hierarchical clustering, flat clusters are formed based on a maximum number of clusters. This is achieved using the fcluster function:

$$C = \text{fcluster}(Z, k, \text{'maxclust'})$$

Where C is the cluster labels for the data points and k is the number of clusters desired.

4.2 Comparison of Sentiment Analysis on Comments

The section is dedicated to the comparison of sentiment analysis across comments and dives deep into assessing the emotional undertones within Reddit's diverse discussions. Our methodology starts by acquiring a cleaned and preprocessed dataset of comments from multiple subreddits, designated as Sci, from the combined union dataset D, where each comment c in the subreddit i undergoes sentiment evaluation.

For each comment c, the sentiment analysis process extracts four primary metrics: negative ($sneg$), neutral ($sneu$), positive ($spos$), and compound ($scomp$) scores.

The sentiment for each comment is computed using the Vader Sentiment Intensity Analyzer, formally represented as:

$$\text{Sentiment}(c) = \{s_{\text{neg}}(c), s_{\text{neu}}(c), s_{\text{pos}}(c), s_{\text{comp}}(c)\}$$

The aggregation of sentiment scores for a particular subreddit i involves calculating the average of these scores across all comments within that subreddit, denoted by:

$$\text{Overall Sentiment}(i) = \frac{1}{N_c} \sum_{c=1}^{N_c} \text{Sentiment}(c)$$

where Nc represents the total number of comments in the subreddit i.

To further refine the sentiment analysis, we classify the sentiment based on the compound score $scomp$ for each comment c as follows:

$$\text{Sentiment Label}(c) = \begin{cases} \text{'positive'} & \text{if } s_{\text{comp}}(c) > 0.5 \\ \text{'negative'} & \text{if } s_{\text{comp}}(c) < -0.5 \\ \text{'neutral'} & \text{otherwise} \end{cases}$$

Additionally, to understand the nuanced relationship between specific words and their sentiment implications, we employ the Pointwise Mutual Information (PMI) metric. The PMI quantifies the association strength between words within the comments and their sentiment labels, providing insights into the linguistic patterns that drive emotional expressions within Reddit's discussions. The PMI for a word w and a sentiment label S is mathematically calculated using:

$$\mathrm{PMI}(w, S) = \log_2 \frac{P(w \cap S)}{P(w)P(S)}$$

where $P(w)$ is the probability of occurrence of the word w, $P(S)$ is the probability of occurrence of the sentiment label S, and $P(w \cap S)$ represents the joint probability of both w and S. These probabilities are determined as follows:

$$P(w) = \frac{\text{frequency of word } w}{\text{total words in dataset}}$$

$$P(S) = \frac{\text{number of comments with sentiment } S}{\text{total comments in dataset}}$$

$$P(w \cap S) = \frac{\text{number of occurrences of word } w \text{ in comments with sent}}{\text{total comments in dataset}}$$

4.3 Hybrid Recommendation System

For this section, we incorporate an innovative hybrid recommendation system that combines content-based and collaborative filtering techniques to enhance recommendation accuracy and relevance within the context of Reddit's community discussions. This system leverages subreddit descriptions alongside user interactions to tailor content suggestions to individual preferences, promoting engaging and pertinent content discovery across the platform.

The subreddit descriptions Sdi, extracted from the combined union dataset D, undergo a text preprocessing phase using a series of natural language processing (NLP) techniques, which include tokenization, stop-word removal, stemming, and TF-IDF vectorization, represented mathematically as:

$$T(d) = \mathrm{TFIDF}(\mathrm{Stem}(\mathrm{Tokenize}(\mathrm{RemoveStopWords}(d))))$$

where $T(d)$ is the TF-IDF vector representation of the preprocessed text data from subreddit descriptions.

The similarity measure S is used to calculate the similarity between different subreddit descriptions based on their TF-IDF vectors. For a given subreddit r with a vector representation v_r, the similarity with another subreddit r' with a $v_{r'}$ is defined as:

$$S(r, r') = \frac{\vec{v}r \cdot \vec{v}r'}{\|\vec{v}r\| \|\vec{v}r'\|}$$

For collaborative filtering, an interaction matrix I is constructed from the user-post interactions across subreddits, where each entry Iu, s indicates the interaction strength of the user u with the subreddit s. This matrix is transformed into a sparse matrix representation M, which is used along with a nearest neighbors algorithm to find similar user profiles based on cosine distance:

$$M(u) = kNN(I(u), \text{'cosine'})$$

To introduce stochasticity and diversity in the recommendation process, a random sample of users is selected for generating recommendations. Let U be the set of all users, and u_r represent a random subset of users:

$$u_r = \mathrm{RandomSample}(U, n)$$

where n is the number of random users selected.

The hybrid recommendation for a user u is then generated by combining recommendations from both content-based and collaborative approaches, calculated as the union of the top N recommendations from both methods, ensuring the diversity and relevance of the recommended subreddits:

$$R(u) = \mathrm{TopN}(\mathrm{ContentBased}(u, S) \cup \mathrm{Collaborative}(u, M))$$

This systematic and integrative approach enables the recommendation system to leverage both the semantic content of subreddit descriptions and the behavioral patterns of users to provide tailored subreddit recommendations, thereby enhancing user experience and engagement on the platform.

4.4 Comparison of the Effects of Topic Modeling Based on Posts and Comments

This section is dedicated to analyzing the comparison of the effects of topic modeling on posts and comments. This methodology starts with extracting comments, from the combined dataset D. which are subjected to natural language preprocessing techniques, including tokenization, stemming, and removal of stopwords, collectively represented by the transformation T:

$$P = T(D_{\text{posts}}), \quad C = T(D_{\text{comments}})$$

where P and C represent the tokenized forms of posts and comments respectively.

Following preprocessing, a textual corpus is created for both posts and comments. Let $CorpusP$ and $CorpusC$ denote the corpora for posts and comments, respectively, formed by aggregating tokens:

$$\mathrm{Corpus}P = \bigcup p \in Pp, \quad \mathrm{Corpus}C = \bigcup c \in Cc$$

A dictionary D and bag-of-words model B are constructed for each corpus. The dictionary encapsulates the vocabulary of the corpus, and B transforms texts into vector representations based on D:

$$D_P, B_P = \text{Dictionary}(P), \text{BOW}(P)$$

$$D_C, B_C = \text{Dictionary}(C), \text{BOW}(C)$$

Topic modeling is then applied separately to B_p and B_c using Latent Dirichlet Allocation (LDA), aiming to identify thematic structures \tau{d} within the corpora:

$$\theta_P = \text{LDA}(B_P, D_P), \quad \theta_C = \text{LDA}(B_C, D_C)$$

Each document d within the corpora is assigned a dominant topic τ based on the highest probability topic distribution provided by LDA:

$$\tau(d) = \arg\max \Theta(d)$$

This methodological approach allows for the comparative analysis of thematic trends between posts and comments across subreddits, which are then visually represented through heatmaps to depict the frequency and distribution of topics, thereby illustrating the differences in content focus and user engagement between the two forms of communication:

$$\text{Heatmap}(\text{Frequency}(\theta_P)), \quad \text{Heatmap}(\text{Frequency}(\theta_C))$$

This approach not only highlights distinct patterns in user engagement but also aids in understanding how topics are differently addressed in posts versus comments within the Reddit community.

5 EXPERIMENTS

5.1 Datasets

In our analysis, we leveraged three dynamic pivotal datasets derived from social media interactions, specifically focusing on subreddit communities. The datasets are extracted from the Reddit API called "PRAW", through which we asked the user to select a topic and how many subreddits it should select from.

The first dataset, titled "Subreddit Data Info NW Insights", comprehensively captures subreddit information as well as network dynamics, which include may not be limited to degree distribution, centrality, modularity, etc. This dataset is instrumental in understanding the structure of each of the networks, which can be represented simply by:

$$D_{nwi} = \{(\text{subreddit information}_i, \text{network dynamics}_i) \mid i \in$$

The second data, titled "Subreddit Data Add Info", has comprehensive interaction data, which concludes the top 10 posts and top 20 comments for each of the subreddits. Also, these include karma values as well as crossposting subreddit names if possible. This dataset can be presented by:

$$D_{pc} = \{(\text{subreddit name}_i, \text{list}(\text{posts})_i, \text{list}(\text{comments})_i, \text{crossposting subreddits}_i) \mid i \in I\}$$

The third dataset, titled "Subreddit Network Data", has a list of subreddits along with the post authors and comment authors, for their representative posts and comments. This can be presented by:

$$D_{nw} = \{(\text{subreddit name}_i, \text{list}(\text{post authors})_i, \text{list}(\text{comment authors})_i) \mid i \in I\}$$

Finally, all the comprehensive datasets are combined together to form the final dataset, that we used for our functionalities. This can be respected as:

$$D = Dwi \cup Dpc \cup Dnw$$

5.2 Baseline and Metrics

We introduce some prediction techniques that comprise the baselines of the models:

Accuracy: It can be defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions Made}}$$

5.3 Results

5.3.1 Results of Analysis and Comparison of Different Network Dynamics

In this experiment, we cluster the users based on the network metrics: 'Number of Nodes', 'Number of Edges', 'Average Degree', 'Average Network Density', 'Average Degree Centrality', 'Average Closeness Centrality', 'Average Betweenness Centrality', 'Clustering Coefficient', 'Modularity', which were normalized and put through a linkage matrix. Through the linkage matrix, we cluster the subreddits into 4 different clusters:

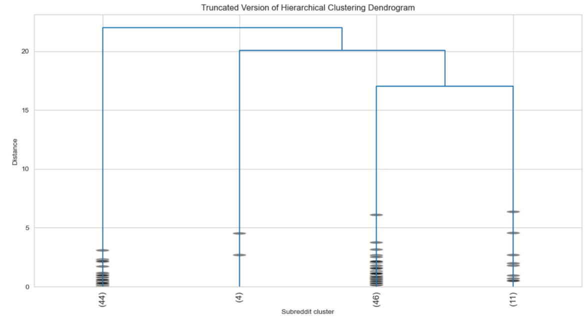


Fig. 4: Truncated Version of Hierarchical Clustering

Cluster 1 is made up of large and sparse networks. Subreddits like Animals, FunnyAnimals, and AskReddit feature in Cluster 1. These have large networks with many nodes and edges but low density, indicating sparse connections. High modularity points to distinct communities within these broadly themed subreddits.

Cluster 2 is made up of small and dense networks. These include niche groups such as AnimalsBeingPervs and cute_animals, characterized by small but highly dense networks. This setup reflects tightly-knit communities where nodes have strong influence, evident from high centrality measures and lower modularity.

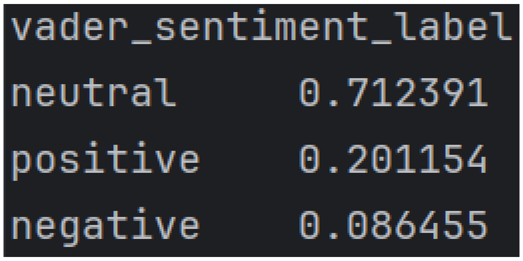
Cluster 3 is made up of moderately sized and moderately-connected networks. These include niche groups such as AnimalsBeingStrange and AnimalsOnReddit, which are characterized by medium-sized networks with moderate density and centrality, suggesting a balance between interconnection and sparsity. The fairly high modularity indicates the presence of distinct sub-communities.

Cluster 4 is made up of highly interactive, small networks. These include AnimalBeingAnimals and HybridAnimalsGame, which are small but interactively dense networks. High closeness centrality in these networks facilitates rapid information spread, with very low modularity reflecting a unified community theme or interaction style.

5.3.2 Results of Comparison of Sentiment Analysis of Comments

In this experiment, we performed sentiment analysis on subreddit comments using the Vader Sentiment Intensity Analyzer and the Pointwise Mutual Information (PMI) method.

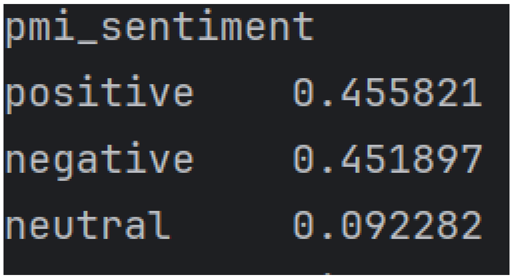
The distribution of the sentiment labels across the comments using the Vader Sentiment Analysis is:



```
vader_sentiment_label
neutral      0.712391
positive     0.201154
negative     0.086455
```

Fig. 5: Distribution of Sentiment Labels using Vader Sentiment

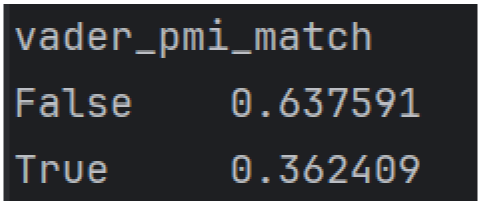
The distribution of the sentiment labels across the comments using the Pointwise Mutual Information is:



```
pmi_sentiment
positive     0.455821
negative     0.451897
neutral      0.092282
```

Fig. 6: Distribution of Sentiment Labels using Vader Sentiment

Now to see the comparison of the two methods, we compare the matching of the overall sentiment in a comparison table:



```
vader_pmi_match
False     0.637591
True      0.362409
```

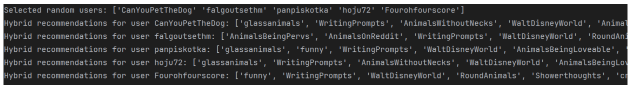
Fig. 7: Comparison of the matching of the two methods

This table shows only an accuracy of 36.24% agreement on sentiment classifications across the dataset. This suggests that each method interprets sentiments differently, potentially due to the unique mechanisms and contexts they consider. The 63.76% mismatch tells a notable discrepancy between how each method interprets the sentiment of the same text.

5.3.3 Results of Hybrid Recommendation System

In this segment of our study, we leverage the use of both content-based and collaborative filtering approaches to create unique recommendations for five random sets of users,

The output for 5 random users are:



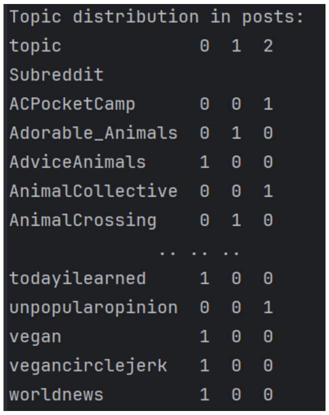
```
Selected random users: ['CanYouPetTheDog' 'falgoutsethm' 'panpiskotka' 'hoju72' 'Fourohfourscore']
Hybrid recommendations for user CanYouPetTheDog: ['glassanimals', 'WritingPrompts', 'AnimalsWithoutNecks', 'WaltDisneyWorld', 'Animal
Hybrid recommendations for user falgoutsethm: ['AnimalsBeingPervs', 'AnimalsOnReddit', 'WritingPrompts', 'WaltDisneyWorld', 'RoundAni
Hybrid recommendations for user panpiskotka: ['glassanimals', 'funny', 'WritingPrompts', 'WaltDisneyWorld', 'AnimalsBeingLoveable', '
Hybrid recommendations for user hoju72: ['glassanimals', 'WritingPrompts', 'AnimalsWithoutNecks', 'WaltDisneyWorld', 'AnimalsBeingLov
Hybrid recommendations for user Fourohfourscore: ['funny', 'WritingPrompts', 'WaltDisneyWorld', 'RoundAnimals', 'Showerthoughts', 'cr
```

Fig. 8: Hybrid recommendations for 5 random users

5.3.4 Results of Comparison of the Effects of Topic Modeling Based on Posts and Comments

In this experiments, we performed topic modeling on both of the topics and comments and saw their effects on how these topics are applied on the subreddits.

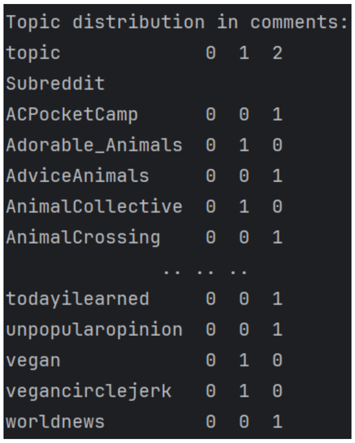
Here is the topic assigned to subreddits through posts:



```
Topic distribution in posts:
topic            0  1  2
Subreddit
ACPocketCamp     0  0  1
Adorable_Animals 0  1  0
AdviceAnimals    1  0  0
AnimalCollective 0  0  1
AnimalCrossing   0  1  0
                .. .. ..
todayilearned    1  0  0
unpopularopinion 0  0  1
vegan            1  0  0
vegancirclejerk  1  0  0
worldnews        1  0  0
```

Fig. 9: Topic Modeling on Posts

Here is the topic assigned to subreddits through comments:



```
Topic distribution in comments:
topic             0  1  2
Subreddit
ACPocketCamp      0  0  1
Adorable_Animals  0  1  0
AdviceAnimals     0  0  1
AnimalCollective  0  1  0
AnimalCrossing    0  0  1
                 .. .. ..
todayilearned     0  0  1
unpopularopinion  0  0  1
vegan             0  1  0
vegancirclejerk   0  1  0
worldnews         0  0  1
```

Fig.10: Topic Modeling on Comments

To observe the comparison as a visualization, we see that:

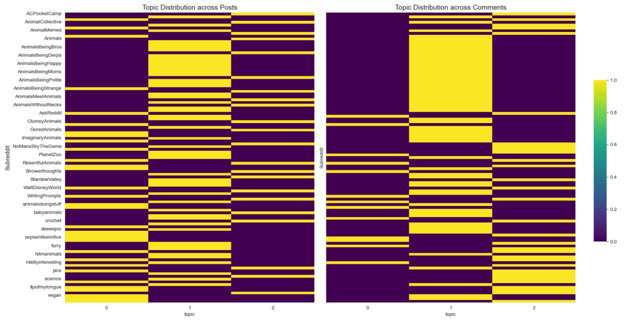


Fig. 11: Topic Distribution of Posts and Comments

In the posts (left side), subreddits like "AnimalsBeingDerps," "WaltDisneyWorld," and "WritingPrompts" show a varied distribution of topics, indicated by multiple colors across the bars. This diversity suggests a wide range of subjects being discussed within individual posts. Conversely, the comments (on the right side) display a more uniform color distribution, suggesting a narrower focus in discussions or possibly a dominant theme that pervades the comments within those subreddits. Notably, some subreddits have a significant presence of one or two topics, both in posts and comments, hinting at highly focused community interests or specific content that drives engagement in those areas. This differential pattern between posts and comments could reflect the dynamic nature of how topics are initiated versus how they are discussed within the community.

6 CONCLUSION

In this study, we delved into optimizing social media engagement by analyzing Reddit's most influential communities, called Subreddits. Our research revealed distinct network patterns between the subreddits. We successfully compared the sentiment of posts and labeled them by using the two methods and seeing how different these methods work. Moreover, we developed a hybrid recommendation system, that utilizes both content-based and collaborative filtering techniques to create recommendations. Furthermore, we also studied the effect of topic modeling on posts and comments and how different topics are assigned to the subreddits. These insights not only enhance our understanding of social media behaviors but also improve strategies for community management and engagement. This research underscores the potential of data-driven approaches to advance the customization and relevance of social media platforms, leading to more meaningful user experiences and increased user engagement. Through this project, we have not only gained insights into the intricate dynamics of Reddit communities but also learned valuable lessons about data analysis and interpretation. We have learned the importance of methodological rigor and interdisciplinary collaboration in conducting meaningful research in the realm of social media analytics. Looking to the future, there are numerous exciting possibilities for further exploration of Reddit data. One avenue for future research could involve examining the evolution of community norms and behaviors over time within specific Subreddits, shedding light on how online communities adapt and change in response to various factors.

7 ACKNOWLEDGEMENTS

We would like to express our deep gratitude to professors Naren Ramakrishnan and Patrick Butler for their unwavering support, guidance, and insightful feedback throughout this study. Their expertise and encouragement have been instrumental in shaping the direction of our research, enriching its scope and depth. Additionally, we acknowledge Reddit for granting us access to their data through the Reddit API, which significantly contributed to the breadth and depth of our research.

8 REFERENCES

[1] Govinda K, Akhil Chintalapati, Aparajita Senapatii, and Khashbhat Enkhbat
Sentiment Analysis on Reddit Trading Data
[2] Tanmay Nandurkar, Spandan Nagare, Shreyash Hake, and Kotadi Chinnaiah
Sentiment Analysis Towards Russia - Ukrainian Conflict: Analysis of Comments on Reddit
[3] Joseph Coelho, Paromita Nitu, and Praveen Madiraju
A Personalized Travel Recommendation System Using Social Media Analysis