Wildlife Conservation Network Analysis on Reddit Dataset

Thinakone Louangdy

1. Introduction

Understanding online discussions on wildlife conservation and environmental issues is crucial in an era where digital platforms significantly influence public opinion. These virtual spaces are instrumental in shaping perceptions and driving collaborative efforts in conservation communities. With increasing threats to biodiversity, analyzing these online dialogues can provide essential insights for developing effective conservation strategies.

Initially, our study intended to use Twitter data, a rich source for gauging public opinion. However, following the platform's rebranding to "X" and removing free developer packages for educational and institutional use, we shifted our focus to Reddit. This platform offers a diverse range of topic-specific forums, presenting a unique opportunity to explore in-depth conversations about wildlife conservation.

By examining Reddit's discussions, this report aims to uncover trends, identify influential voices, and understand public sentiments in the conservation domain. This shift in data sources allows us to continue our vital research into digital discourse's impact on environmental awareness and wildlife conservation efforts.

2. Data Collection

To comprehensively analyze online discussions on wildlife conservation, we utilized the Async PRAW library, an efficient API wrapper for Reddit's APIs. This choice was driven by the need to gather data from specific, highly engaged communities within the Reddit platform. Our focus centered on six key Subreddits renowned for their active participation in wildlife conservation topics: 'conservation,' 'Wildlife,' 'EndangeredSpecies,' 'AnimalRights,' 'invasivespecies', and 'wildlifebiology'.

The data collection process was designed to capture these Subreddits' most recent and relevant discussions. To achieve this, we set parameters within Async PRAW to scrape the newest posts from each selected community, extending our collection to the limit of approximately 1000 posts per Subreddit (Nature of PRAW library). This approach ensured we gathered the most current insights and trends within these communities.

The scope of data extracted included various critical details from each post, essential for our analysis. These details comprised the **post ID, post title, author, score**, number of **upvotes** and **downvotes**, **upvote ratio**, the **name of the Subreddit**, the time of post creation ('created_at'), and the **URL** to the original post. This wide array of data points allowed for a multifaceted analysis of the online discourse surrounding wildlife conservation.

All the collected data were meticulously organized and saved in a single CSV file. This format was chosen for its simplicity and compatibility, facilitating ease of access and analysis in subsequent stages of our research.

The methodology adopted for data collection provides a robust foundation for our analysis, ensuring a comprehensive understanding of the current state of online discussions in wildlife conservation.

3. Data Preprocessing

Before the analysis, the dataset underwent a rigorous cleaning process. This preprocessing stage was essential to refine the data, ensuring its readiness for effective and meaningful analysis. Utilizing the Pandas library, we first identified and removed any NaN or Null values, ensuring the integrity of the dataset. Initially, in timestamp format, the 'created_at' field was converted to a more readable datetime format for better analysis and interpretation.

Given the focus on sentiment analysis, the text data, specifically post titles, required further preprocessing. The Natural Language Toolkit (NLTK) library was instrumental in this phase. We employed it to strip away stop words and convert all text to lowercase, standardizing the dataset for accurate sentiment analysis.

To quantify the sentiment of the titles, we calculated scores for positive, negative, neutral, and compound sentiments. This step was crucial for labeling the data, which would later facilitate our Exploratory Data Analysis (EDA). The preprocessed and labeled data were then saved as a CSV file, laying the groundwork for in-depth analysis in the subsequent stages of our study.

4. Exploration Data Analysis

The histogram of post scores from the wildlife conservation Subreddits shows a clear right-skew, with most posts receiving lower scores and a few receiving very high scores. This indicates that most posts tend to have a low level of engagement, with only a small number of posts becoming highly visible or popular within the community.

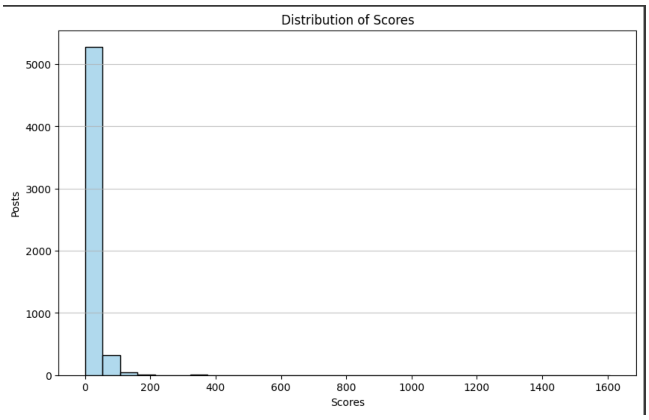


Figure 1. Distribution of Scores per Post

The bar chart comparing the number of posts across different Subreddits illustrates an equitable activity level among them. None of the Subreddits appears to dominate the discourse significantly, highlighting a distributed interest in wildlife conservation topics across these online communities.

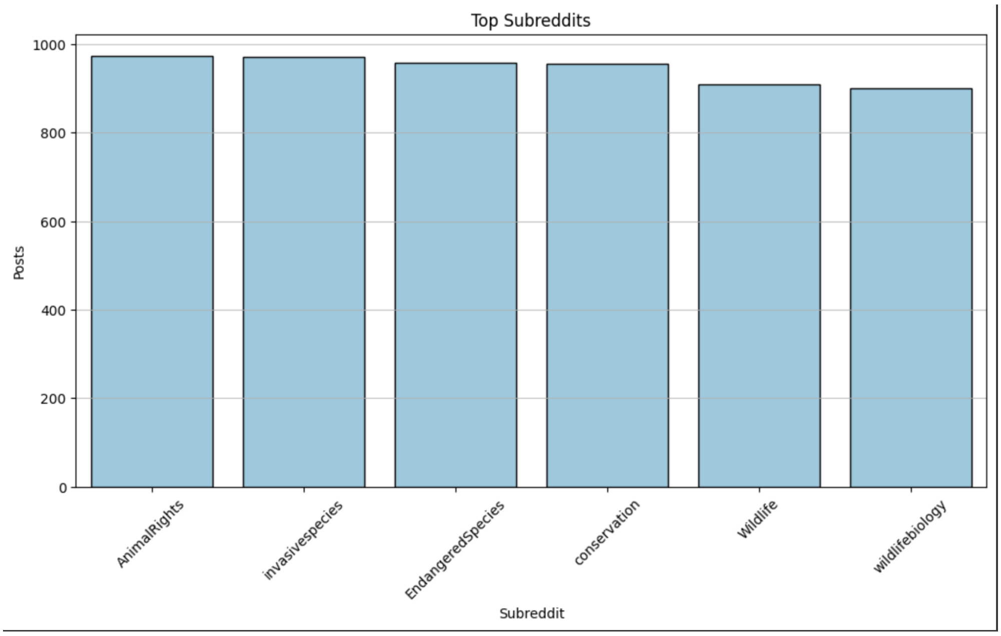


Figure 2. Top Subreddits Chart

A bar chart depicting the most active authors across all collected data shows that specific individuals are more prolific than others. The highest bars represent the top contributors, indicating that these users are potentially key opinion leaders in the wildlife conservation conversation on Reddit.

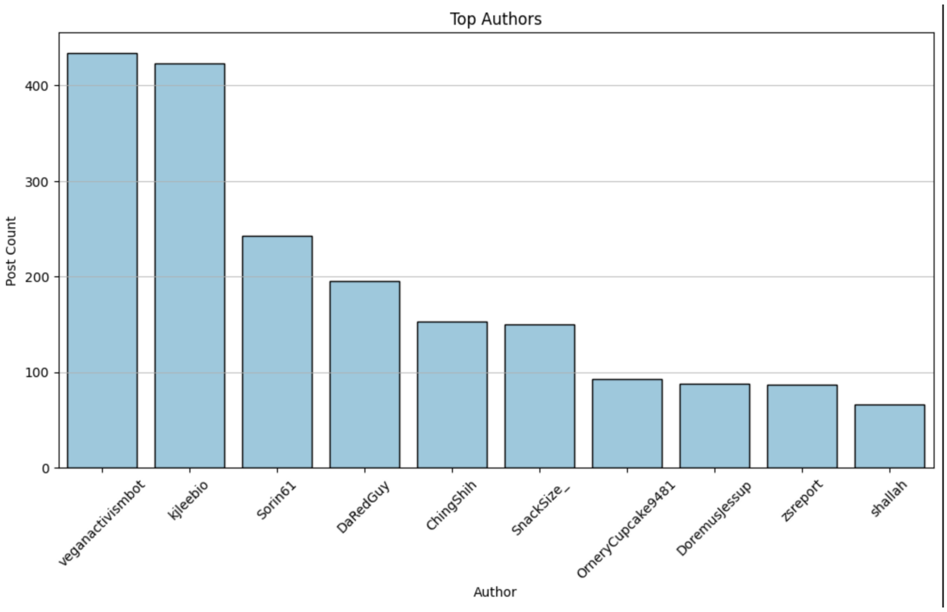


Figure 3. Top authors

A series of bar charts for each Subreddit displays the post count from top community users. These charts reveal that some users are mainly active within specific Subreddits, suggesting a specialized interest or influence within certain conservation topics.

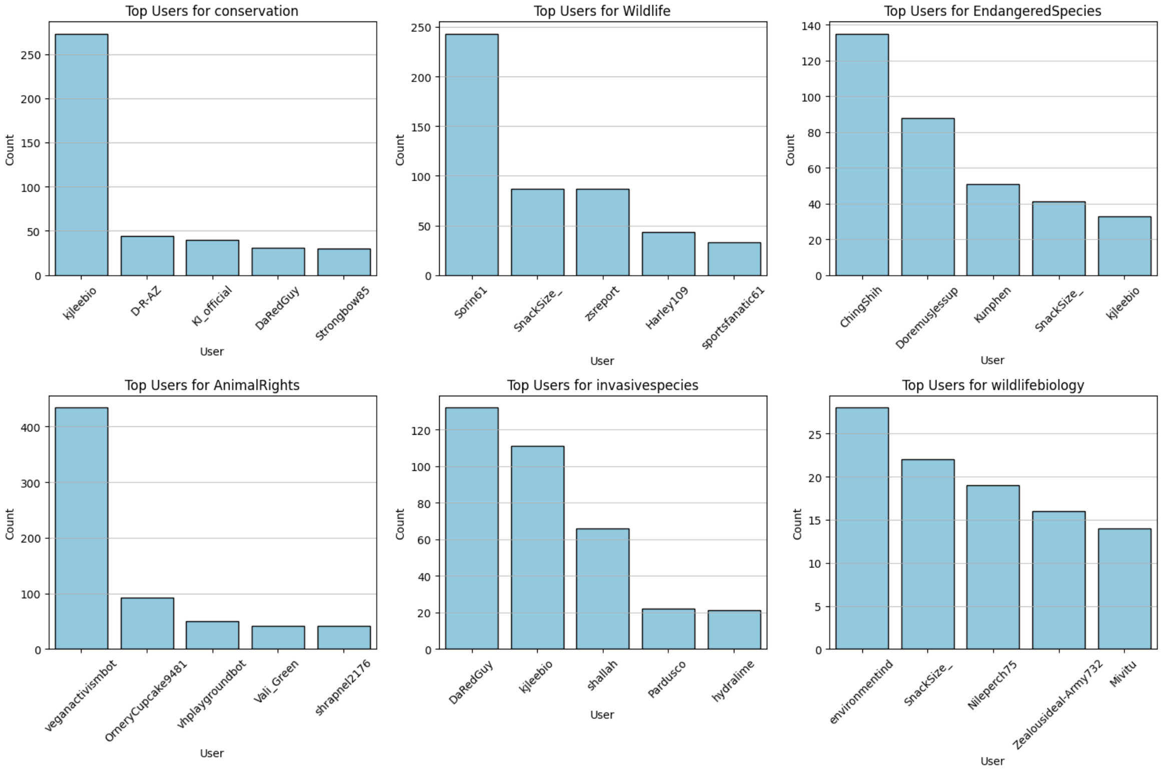


Figure 4. Top Users for Each Subreddit Community

These visual representations allow for a quick assessment of the community engagement patterns and highlight where the most active conversations occur within the wildlife conservation realm on Reddit. The distribution of engagement, both in terms of post scores and user activity, underscores the diversity and depth of participation within these communities.

The time series chart for daily post counts shows variability in the number of daily posts across the data collection period. There are noticeable spikes in activity on certain days, which could indicate specific events or discussions that triggered increased engagement. The overall trend does not follow a clear pattern, suggesting that posting is reactive to current events rather than consistent daily behavior.

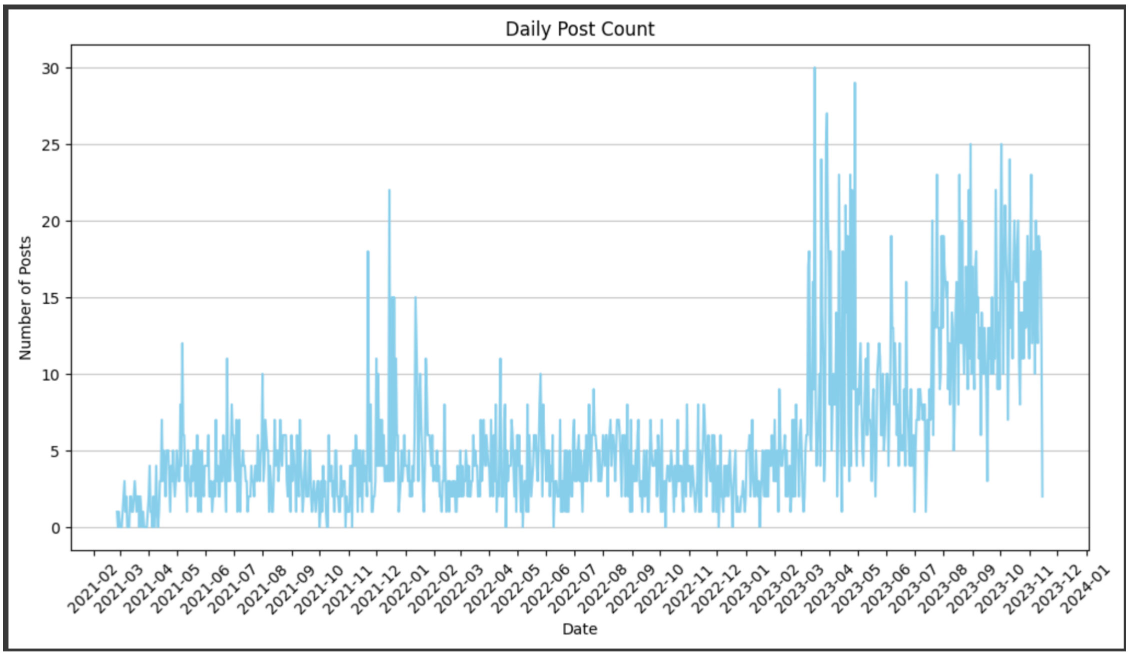


Figure 5. Daily Post Count

The second time series chart breaks down the daily posting activity of the top users. Each user's activity is represented by a different color, showing the individual contribution patterns over time. Some users have periods of intense activity, which could align with specific conservation events or personal posting schedules. This visualization highlights the variance in how different users contribute to the discourse and how their activity levels change over time.

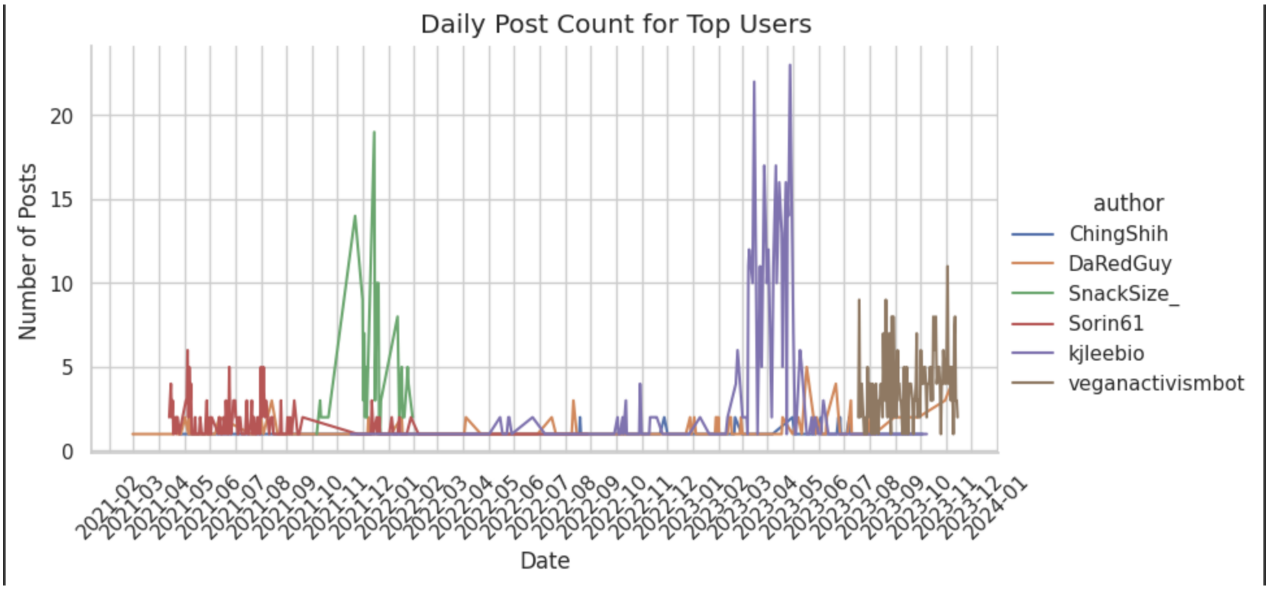
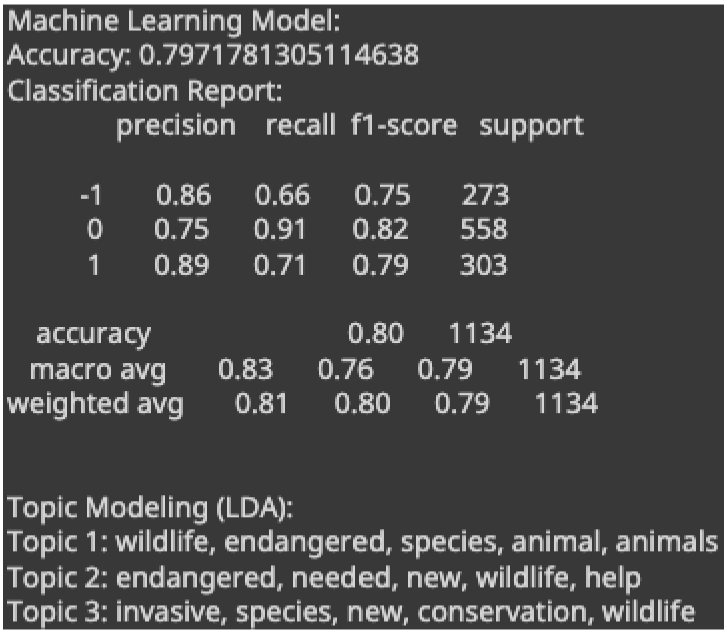


Figure 6. Daily Post Count for Top Users

These temporal analyses are essential for understanding the rhythm of the community's interactions and can guide the timing of targeted communications for maximum impact.

Our RandomForestClassifier achieved a 79.71% accuracy in classifying wildlife-related sentiment, showing a bias towards neutral sentiments and indicating room for improvement in detecting negative and positive sentiments. Latent Dirichlet Allocation revealed three key topics: conservation needs, endangered species, and invasive species effects, highlighted by prominent keywords. Additionally, Word2Vec associations suggest frequent discussions around 'wildlife' with related terms like 'new,' 'animal,' and 'conservation.'



```
Machine Learning Model:
Accuracy: 0.7971781305114638
Classification Report:
              precision    recall  f1-score   support

          -1       0.86      0.66      0.75       273
           0       0.75      0.91      0.82       558
           1       0.89      0.71      0.79       303

    accuracy                           0.80      1134
   macro avg       0.83      0.76      0.79      1134
weighted avg       0.81      0.80      0.79      1134


Topic Modeling (LDA):
Topic 1: wildlife, endangered, species, animal, animals
Topic 2: endangered, needed, new, wildlife, help
Topic 3: invasive, species, new, conservation, wildlife
```

Figure 7. Result of our model in sentimental labeling & topic modeling.

The histogram illustrates a significant skew toward neutral emotional expressions in the dataset, with fewer instances of strong negative or positive sentiments. This neutrality could reflect a balanced discourse or a tendency towards factual reporting in the content analyzed.

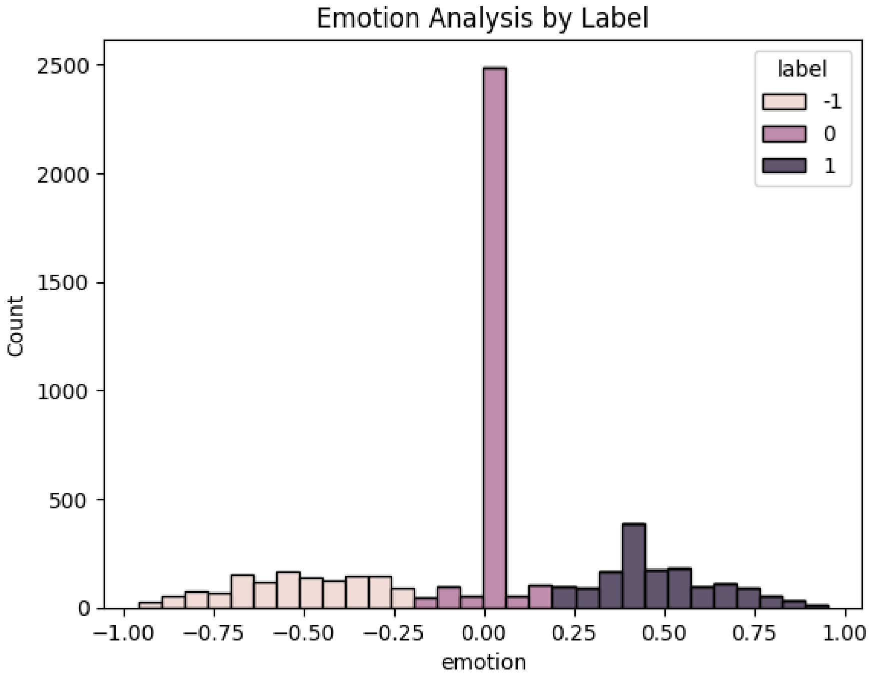


Figure 8. Emotion chart by post count.

The bar chart reveals the sentiment footprint of the top six authors in the dataset. It indicates that certain authors contribute predominantly to specific sentiment categories, which could display their personal biases or areas of interest—figure *9. Top six contributors in content sentimental*

Neutral sentiment is prevalent in the dataset, suggesting a discourse that is either evenly balanced or focused on factual information. Author contributions differ markedly across sentiment categories, with certain authors frequently posting content with a specific sentiment, indicating possible personal biases or specialized interests.

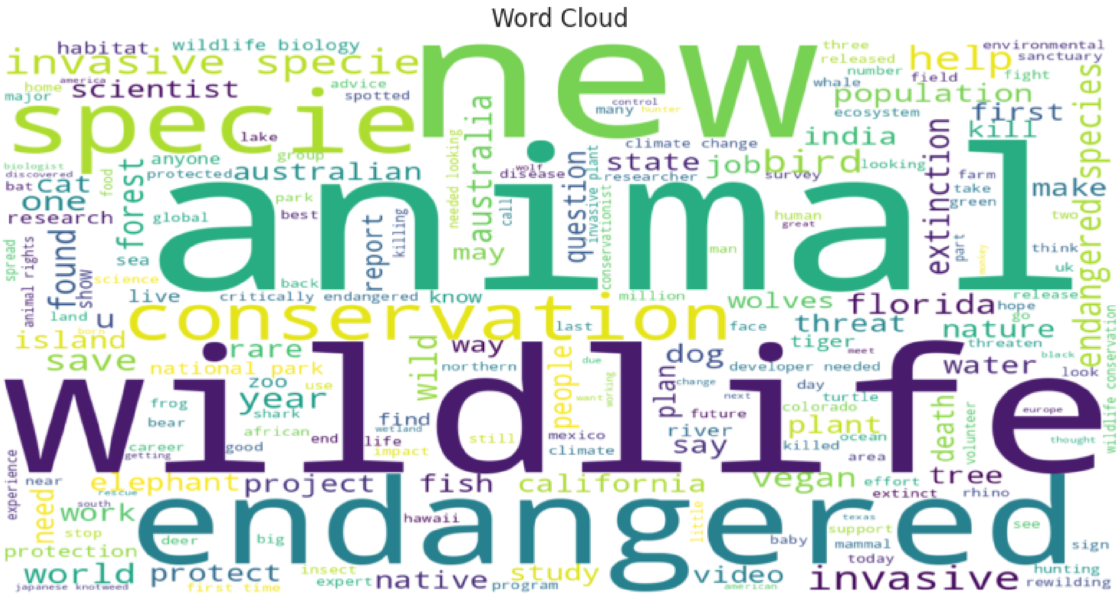


Figure 10. Word Cloud from the dataset.

The Word Cloud analysis from wildlife-related titles indicates key themes with 'animal' at a frequency of 1.0 and 'wildlife' and 'new' at 0.848 and 0.830, signaling concentrated discussions on wildlife and emerging topics. 'Endangered' follows at 0.732, underscoring the focus on at-risk species, with 'species' and 'conservation' also prominent, reflecting the network's dedication to biodiversity and preservation. 'Invasive species' at 0.565 and 'invasive' at 0.520 highlight concerns about ecosystem disruption. Terms like 'help,' 'save,' and 'found' reinforce the community's proactive stance. Geographical mentions such as 'Australia' and 'Florida,' with frequencies of 0.330 and 0.321, point to areas of specific interest or issue within the wildlife narrative. This data is pivotal for aligning the network's content and initiatives with its core interests and critical wildlife concerns. We will later use this to see more connections in our network analysis.

5. Network Analysis

In the first step of the network construction, the Python library NetworkX is used to create a graph representing the interactions between Reddit authors and subreddits. Nodes are added to the graph for each author and subreddit, with edges representing the authors' activities within specific subreddits. The graph is then visualized, distinguishing authors and subreddits by color to illustrate the network's complexity. Key metrics such as the total number of nodes, number of edges, and average degree provide an initial understanding of the network's density and connectivity. The visualization employs a spring layout to natural space out the nodes, offering insights into the community structure and the interaction patterns on Reddit.

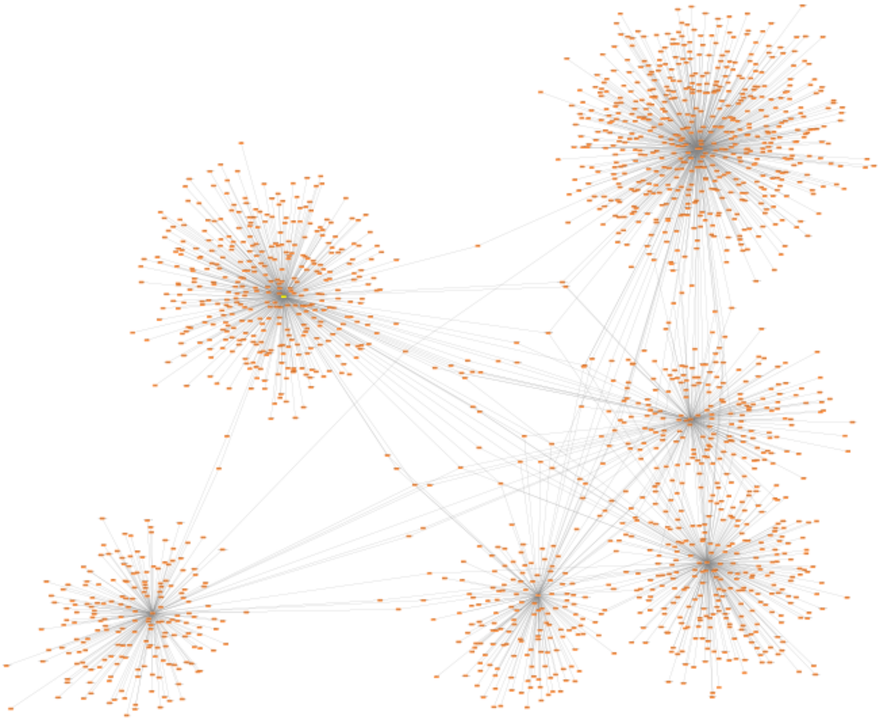


Figure 11. Reddit User in Subreddit Network.

Advancing the analysis, a subgraph is constructed to zoom in on the interactions of selected influential authors, including **'environmentind**,' identified as a top influencer from earlier exploratory data analysis. An ego graph for 'environmentind' delineates this user's immediate connections, spotlighting their engagement within specific subreddit communities. The visualization positions 'environmentind' centrally in green, with their subreddit interactions splayed around in light blue, offering a snapshot of the user's network influence on Reddit.

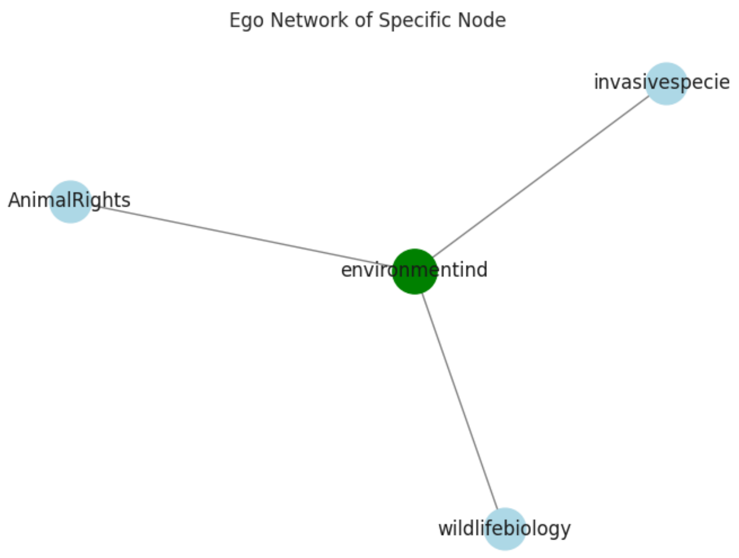


Figure 12. Network of Specific Node.

The network analysis advances by creating a combined ego graph that illustrates the connections between the top six contributors from our exploratory data analysis within Reddit communities. This composite graph highlights the authors in orange and their subreddit affiliations with gray edges, revealing the shared communities and collaborative interactions among these key influencers. The visualization, arranged using a spring layout for clarity, succinctly displays the networked relationships and potential sub-communities among these prominent Reddit users.

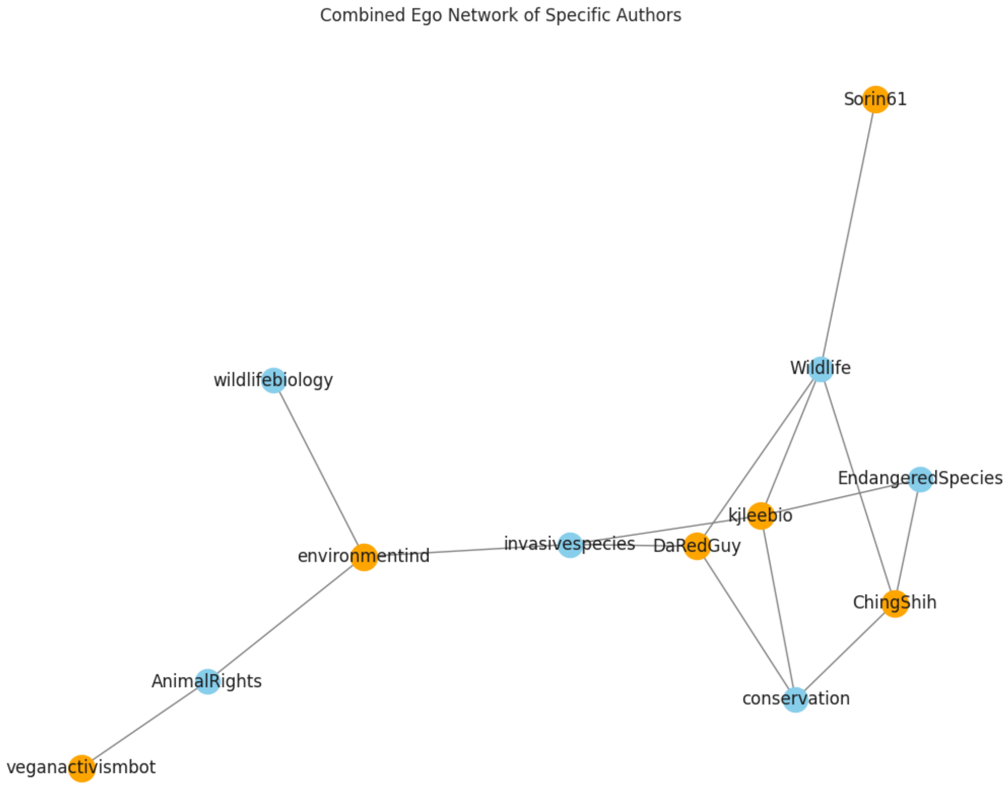


Figure 13. Connections Between The Top Six Contributors.

The network analysis continued with mapping connections between authors and keywords in their post titles. An extensive network comprised 12,123 nodes and 34,945 edges, representing authors and their associated keywords. A subgraph for the keyword 'wildlife' was drawn to distill this vast network, isolating authors who mentioned this term. The resulting visualization captures the nexus between 'wildlife' and relevant authors, offering a clear view of who is most actively discussing this topic within the network.

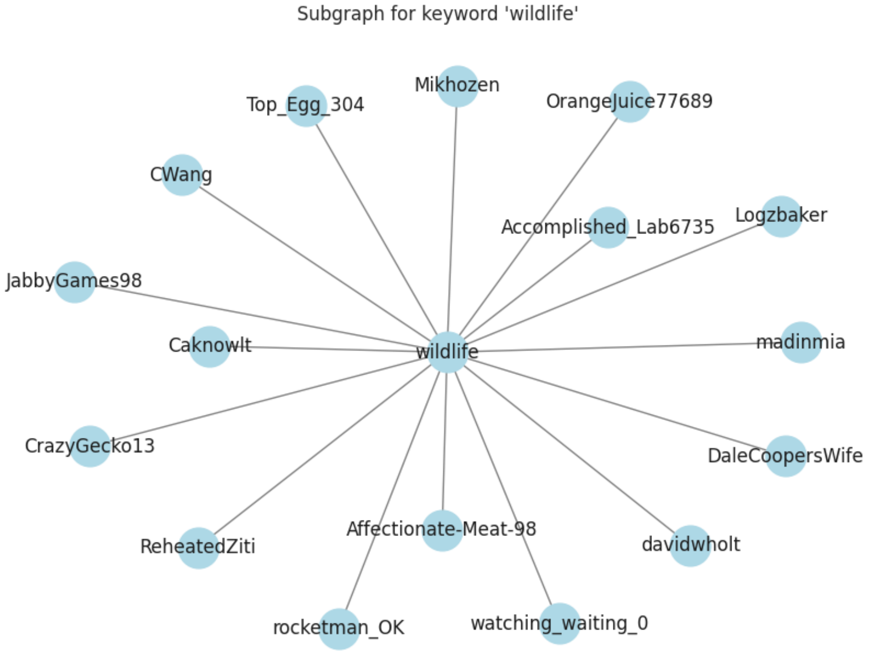


Figure 14. Subgraph for Specific Keyword.

Continuing with the network analysis, a targeted subgraph was created to explore authors' engagement with specific high-frequency keywords identified earlier. This focused subgraph helps to elucidate the connections between authors and key thematic terms. By employing color differentiation, the subgraph reveals the prevalent topics of discourse and the authors most engaged with these themes. The subgraph data was exported for more detailed visual analysis, enabling advanced graph visualization platforms like Cytoscape. This step enhances the depth and clarity of the report, providing a nuanced scientific understanding of the community's focal interests within the network.

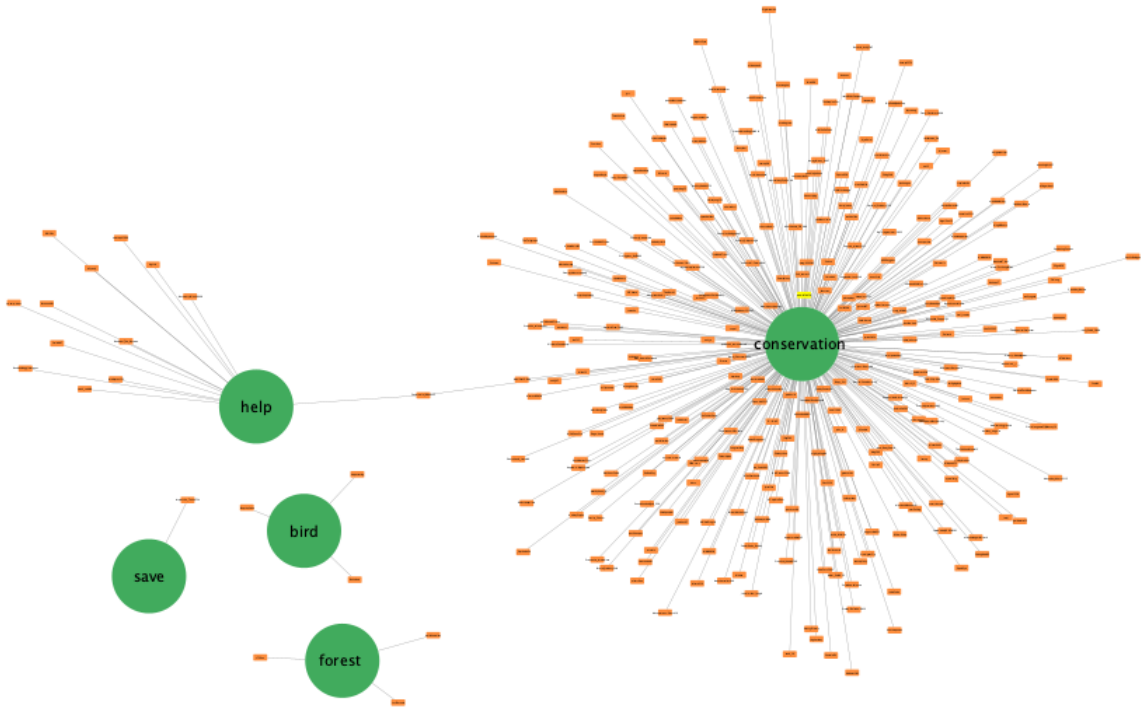


Figure 15. Subgraph on Author connection to each keyword.

The analysis reveals that certain users within the network have not only numerous connections but also hold strategic positions that control the flow of information. **'kjleebio**,' topping the list with a betweenness centrality score of 0.147, acts as a primary gatekeeper through which much of the network's information passes. **'veganactivismbot'** is a close second, also serving as a central hub within the network's communication channels.

Betweenness centrality offers insight into the influence of users beyond just their number of direct connections by highlighting their role in connecting different groups or communities within the network. For example, **'wildlifebiology'** and **'invasivespecies'** may not be the most connected regarding sheer numbers, but they are crucial for linking disparate network parts.

The analysis identifies who is famous or active (as shown by the number of connections) and who is pivotal for the network's cohesion. These key users could be targeted in efforts to spread important information quickly or to ensure that various parts of the network remain integrated. Understanding these roles is crucial for any strategy to effectively utilize the network's structure, whether for educational campaigns, information dissemination, or community engagement efforts.

6. Conclusion

The network analysis of this project revealed an intricate landscape of interactions where specific individuals hold significant sway over the flow of information. Key figures like **'kjleebio'** and **'veganactivismbot,'** with their high betweenness centrality scores, have emerged as central hubs in this network. Their roles go beyond having numerous connections; they are pivotal in connecting different groups or communities within the network. This finding is crucial as it highlights the most active or popular users and those who play a vital role in the network's cohesion. Communities like **'wildlifebiology'** and **'invasivespecies**,' although not the most connected in terms of sheer numbers, are essential for bridging different parts of the network.

A major challenge during this analysis was the network's size and complexity, which created a bottleneck in processing and visualizing the data. Managing a network with thousands of nodes and edges requires careful consideration, especially regarding computing resources and the effectiveness of visualization techniques. This issue was particularly evident when attempting to draw comprehensive insights from the vast amount of data. To address this, focused approaches like creating subgraphs for specific keywords and influential users were employed. These methods allowed for a more manageable and detailed examination of the network's key aspects, though they also highlighted the limitations of handling such extensive data sets.

In conclusion, this network analysis has provided valuable insights into the community dynamics. Identifying key influencers and understanding their roles makes it possible to strategize more effectively for information dissemination, community engagement, or educational campaigns. The challenges faced in managing and interpreting the extensive data emphasize the need for robust analytical approaches in network studies. This analysis not only sheds light on the structural intricacies of the network but also underscores the importance of strategic planning in leveraging such complex social structures for community development and targeted interventions.