1. Data Collection

1.1. Topic Selection

Topic selected for social media data analysis is “AI Technology”. Given the latest development in the AI industry , social media generates more contracting views on usage of AI. To get contracting viewpoints from mastodon I used the mastodon server “mastodon.social”. To get appropriate posts and opinions related to AI, hashtags #aiethics", #aiforgood, #Artificial intelligence was used.
In short, Approach is:
1. Find users relevant to given hashtags. In the network this would be used to create nodes.
2. Then find relationships between users. In the network this would be used to construct edges.

1.2. User nodes relevant to given hashtags

To get an appropriate set of used selection of the right mastodon apis was necessary. By going through mastodon API documentation, I found three APIs which like to give a comprehensive set of users related to hashtags. User nodes extracted from following APIs.
1. Mastodon.search_v2(q, resolve=True, result_type=None, account_id=None, offset=None, min_id=None, max_id=None, exclude_unreviewed=True)
2. Mastodon.timeline_hashtag(hashtag, local=False, max_id=None, min_id=None, since_id=None, limit=None, only_media=False, remote=False)
3. Mastodon.account_search(q, limit=None, following=False, resolve=False)

Apart from account APIs, user nodes are also extracted from post authors, mentions in post etc. also retrieved only relevant attribute of users from these apis. So followings are user’s attribute:

Id - Same as <numerical id>
Username - The username (what you @ them with)
Url - Their URL; for example 'https://mastodon.social/users/<acct>'
Note - Their bio
Posts - User’s toots relevant to hashtags
Followers - ids of follower they have
Followings - ids of people they follow
Display_name - user’s display name
Following_count - How many people they follow
Followers_count - How many followers they have
Statuses_count - How many statuses they have
Bot - Boolean indicating whether this account is automated.
Group - A boolean indicating whether the account represents a group and not individual
Ai_sentiment - Category representing where the user sentiment is AI optimist, neutral, or AI pessimist.

During processing and scrapping

1. Parse the HTML text to normal readable text which can be understood by LLM.
2. Tried to remove stop words, blank spaces, unnecessary characters and url or links.
3. Converted text into english if any text is not in english
4. Removed hashtags and mentions in text and added it as attribute of posts.

In this manner we have collected more than 300 users. But for final dataset selected only 300 users having more activity in network like more follower, followings or more number of topic related posts.

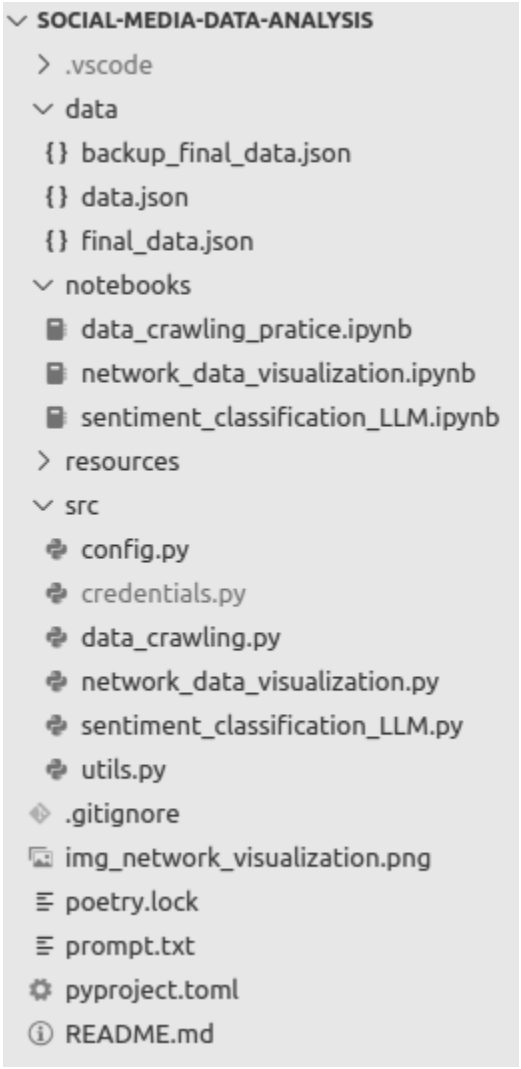
1.3. Find Relationships between users

For each user we retrieve user's followers and followings. It will acts as edges between users.

Thus we have retrieved the user nodes and relations between them. Below is snapshot of our scrapped dataset.


```
{} final_data.json > ...
[
    {
        "id": 109332160649392039,
        "username": "molly0xfff",
        "url": "https://hachyderm.io/@molly0xfff",
        "note": "crypto researcher & critic, software engineer, wikipedian • @web3isgreat creator •  fellow @ Harvard Library Innovation Lab • writing  • :QueerCat_Bisexual:",
        "posts": [
            {
                "id": 110997528113760229,
                "url": "https://hachyderm.io/@molly0xfff/110997527177588715",
                "content": "going to start talking at great length online about how much i hate men to poison the dataset for anyone who tries to train one of these models on my social media",
                "attach_media_text": "",
                "tags": [
                    "ai",
                    "aiethics",
                    "techethics"
                ],
                "mentions": [],
                "language": "en",
                "in_reply_to_id": 110997457221466873,
                "in_reply_to_account_id": 109332160649392039,
                "replies_count": 14,
                "reblogs_count": 33,
                "favourites_count": 47
            },
            {
                "id": 110997457221466873,
                "url": "https://hachyderm.io/@molly0xfff/110997456129314827",
                "content": "excuse me what the fuck",
                "attach_media_text": "",
                "tags": [
                    "ai",
                    "aiethics",
                    "techethics"
                ],
                "mentions": [],
                "language": "en",
                "in_reply_to_id": null,
                "in_reply_to_account_id": null,
                "replies_count": 57,
                "reblogs_count": 167,
                "favourites_count": 92
            }
        ],
        "followers": [
            111058516394738086
        ],
        "followings": [],
        "display_name": "Molly White",
        "followers_count": 75326,
        "following_count": 573,
        "statuses_count": 2389,
        "created_at": "2022-11-12 00:00:00",
        "last_status_at": "2023-09-21 00:00:00",
        "acct": "molly0xfff@hachyderm.io",
        "bot": false,
        "group": false,
        "ai_sentiment": [
            0,
            1,
            0
        ]
    },
```

1.4. Folder structure of Source Code



Folder structure is self exploratory, but in brief:
Data: this folder contains all stage of dataset. `final_data.json` s main and final dataset
Notebook: this pratice notebook or notebook help understand code implementation
Src: This Core folder where all logic implementation have done
Resources: Useful documents and other file are here.

2. Network Construction and Visualization

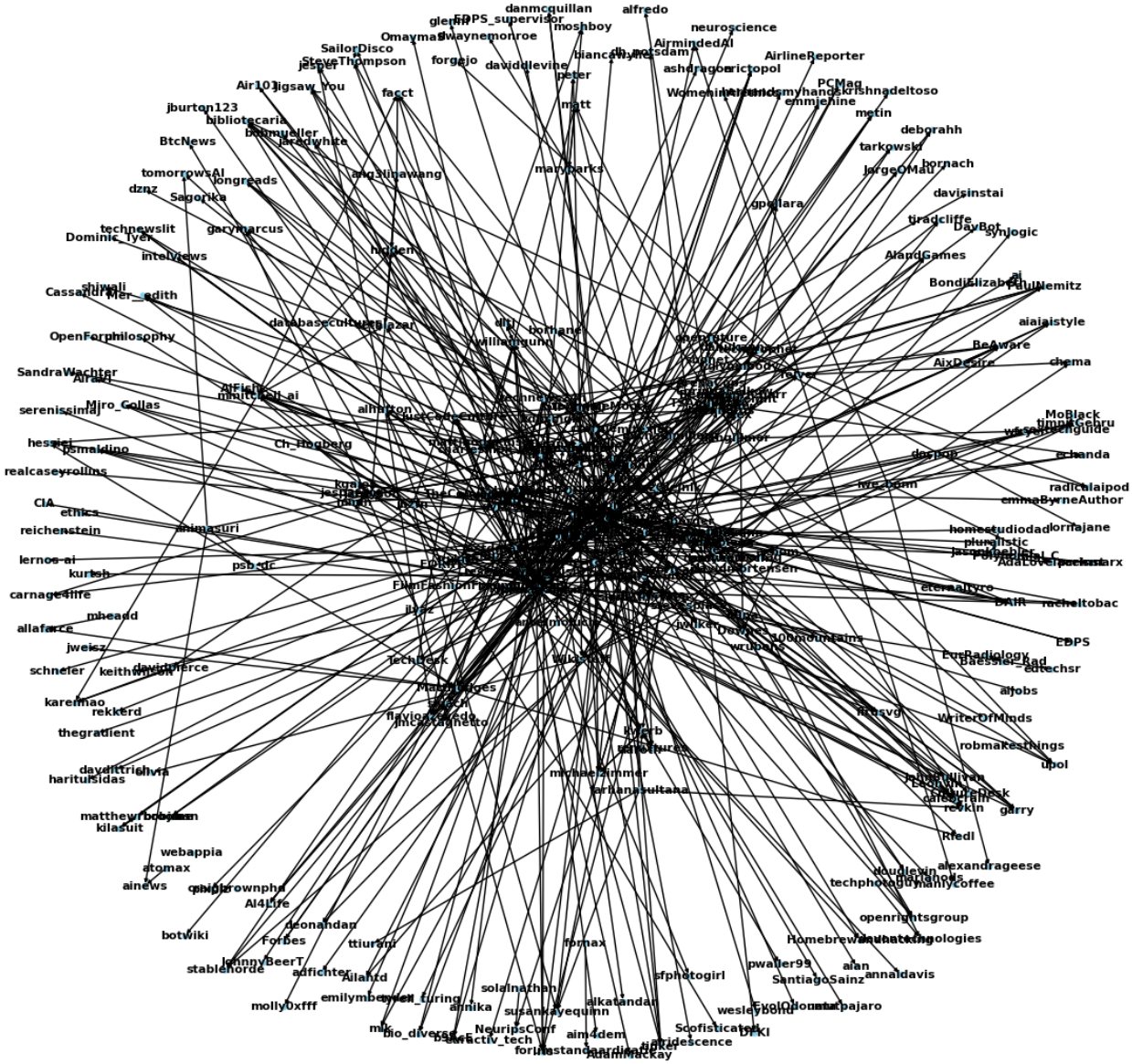
2.1. Friendship Network

A user’s friendship network can be represented as a graph in which the nodes represent the users, and the edges indicate the presence of a friendship relationship between them. The graph depicts the users and their follower and followee relationships in a directed manner.

2.2. Features used to construct network

Nodes are represented by user’s username instead of using user_id. This is done from a readability perspective. Then for each user, we extracted then following of that user. And connected the edge of current user’s to it’s following. Same process is followed for follower with check of if edge is already present of not.

2.3. Snapshot of generated network graph



3. Classification

To generate appropriate prompt for sentiment classification task we need good quality of prompt and user's data features which helps to make accurate decision. There are three categories for classification:

1. AI Optimist
2. Neutral
3. AI Pessimist

Llama2-7B model is used to do classification task.

3.1. User’s features provided to prompt

We have used following user’s data as feature for classification
1. User’s Bio/Notes
2. Posts - User’s toots relevant to hashtags
    2.1. Post content:
    2.2. Post Media Text
    2.3. Post’s Hashtags
    2.4. User who reboosted post
    2.5. Replies_count
    2.6. Reblogs_count
    2.7. favourites_count
3. Followers - ids of follower they have
4. Followings - ids of people they follow
5. Statuses_count - How many statuses they have
6. Bot - Boolean indicating whether this account is automated.
7. Group - A boolean indicating whether the account represents a group and not individual

3.2. Prompt used for classification

User Sentiment Classification Request
Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
User profile can be categories like below.
AI optimist, neutral, or AI pessimist

Below is some example classifications. the format is the user profile followed by classification tag.
1. **User Profile:**
- Bio: AI Ethusiat, working in AI for Good domain. Working toward making AI more accessable and safe

- Post 1 Content: It's not like AI is giving biased data. It human how given bias data. so blaming AI is not good thing. Ai just learning from human So calling Ai fdanger is wrong thing.
- Post 1 Hashtags: The post includes the following hashtags: #ai, #aiforgood.

<classification>AI optimist</classification>

2. **User Profile:**
- Bio: General Attorny, Work in risk evalution, regulation and audit department

- Post 1 Content: Resarch should stop working on AI technology recklessly. We never no small bug in system can endagours billions of human life. there should be proper evalution of risk at each step of AI development and if it causing any possble harm to living being. it should stop quickly.

- Post 1 Hashtags: The post includes the following hashtags: #airegulation, #aiethics.

<classification>AI pessimist</classification>

Classify below user profile.
User Profile:
- Bio: crypto researcher & critic, software engineer, wikipedian • @web3isgreat creator • fellow @ Harvard Library Innovation Lab • writing • :QueerCat_Bisexual:

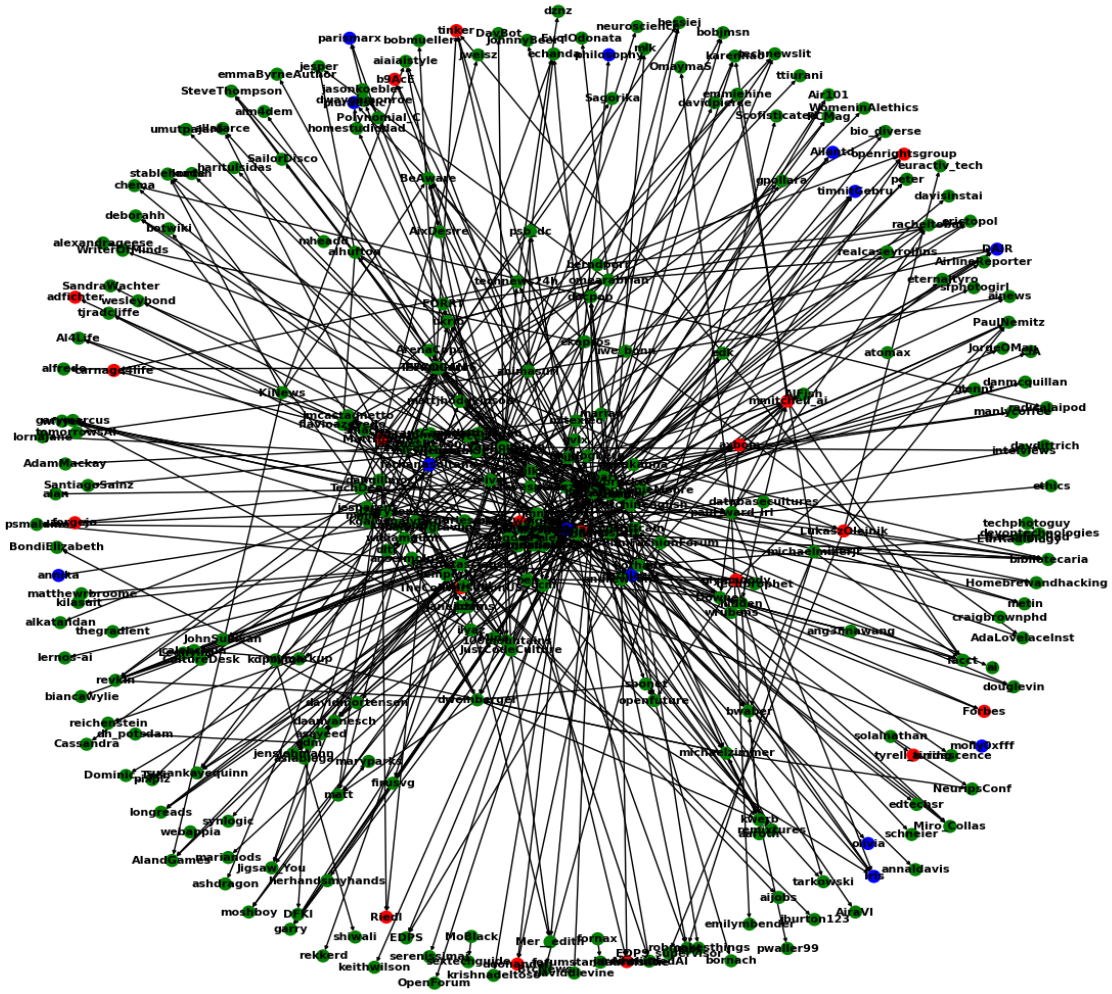
- Post 1 Content: going to start talking at great length online about how much i hate men to poison the dataset for anyone who tries to train one of these models on my social media
- Post 1 Hashtags: The post includes the following hashtags: #ai, #aiethics, #techethics.

- Post 2 Content: excuse me what the fuck
- Post 2 Hashtags: The post includes the following hashtags: #ai, #aiethics, #techethics.
Response:

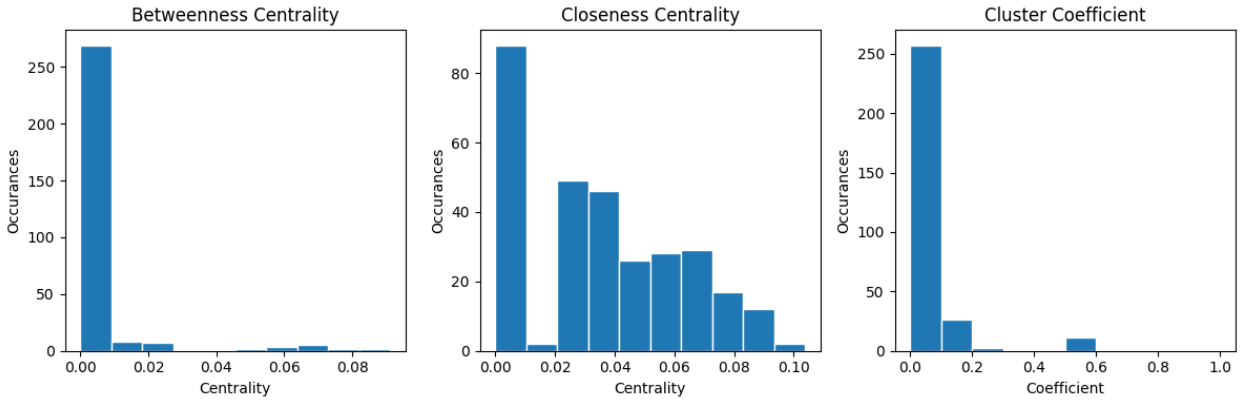
<classification>and here is the category response from you</classification>

give me catgory in first line and do rest of explaination below
---

3.3. Snapshot of classified network



3.4 Metrics



3.5 Distribution