Project Proposal

Team Members:

1. Akshay Reddy Narra
2. Samhitha Pentaparthy
3. Sivakumarreddy Sangu

Dataset: Tweets of EPL Teams

The dataset contains tweets of all teams in the English Premier League based on their team hashtag. i.e. #Chelsea from July to October in the year 2020.

The dataset comprises 15 columns and nearly one million rows, with each row corresponding to an individual tweet. The columns represent the below information:

1. **Team**: This field represents the Premier League team associated with the tweet. It can be derived from the file_name field.
2. **Date**: The date at which the tweet was created. This information is available in the created_at field.
3. **Search Query**: The search query used to query the Twitter Search Engine. This can be found in the search_query field.
4. **Tweet Full Text**: The full text of the tweet, which can be accessed from the text field.
5. **Username**: The username of the account that posted the tweet. This can be accessed from the username field.
6. **Followers**: The number of followers of the account, which can be accessed from the followers field.
7. **Friends**: The number of friends (accounts followed by the user). This can be accessed from the friends field.
8. **Retweet Count**: The number of times the tweet has been retweeted. This information is available in the retweet_count field.
9. **Location**: The location of the account, which can be accessed from the location field.

The dataset consists of a diverse range of Twitter data with features indicating network dynamics, textual content, and endorsements.

Network-related metrics like followers and friends offer insights into connectivity and social influence within the Twitter system. The inclusion of tweet full text provides textual information, facilitating sentiment analysis and content interpretation. Additionally, retweet count provides engagement and potential endorsement information.

Problem Statement:

"Explore the community structure of the dataset to uncover distinct fan communities and clusters. Find any cohesive partitions in the network based on interactions and sentiments along with the underlying factors like geographic location that could influence community formation and dynamics."

1. **Network Analysis**:
   - We'll start by constructing a network using interactions such as retweets, mentions, or follows between users in the dataset. Each user represents a node in the network, and interactions between users represent edges.
   - With the network constructed, we can apply community detection algorithms to identify cohesive partitions or groups of users that frequently interact with each other. These communities represent distinct fan communities within the dataset.
2. **Sentiment Analysis**:
   - Alongside network analysis, we'll conduct sentiment analysis on the tweets in the dataset to determine the sentiment expressed by users towards different teams, matches, and events.
   - Analyzing sentiment within communities can provide insights into the emotional connections and attitudes shared among members, contributing to our understanding of community dynamics.
3. **Influence of Geographic Location**:
   - Additionally, we'll consider the influence of geographic location on community formation and dynamics. We can leverage location information available in the dataset to explore whether users within the same geographic region tend to form cohesive communities or exhibit similar sentiment patterns.

Integrating network analysis, sentiment analysis, and geographic information allows us to uncover the underlying factors driving community formation and dynamics within the English Premier League fan base on Twitter.