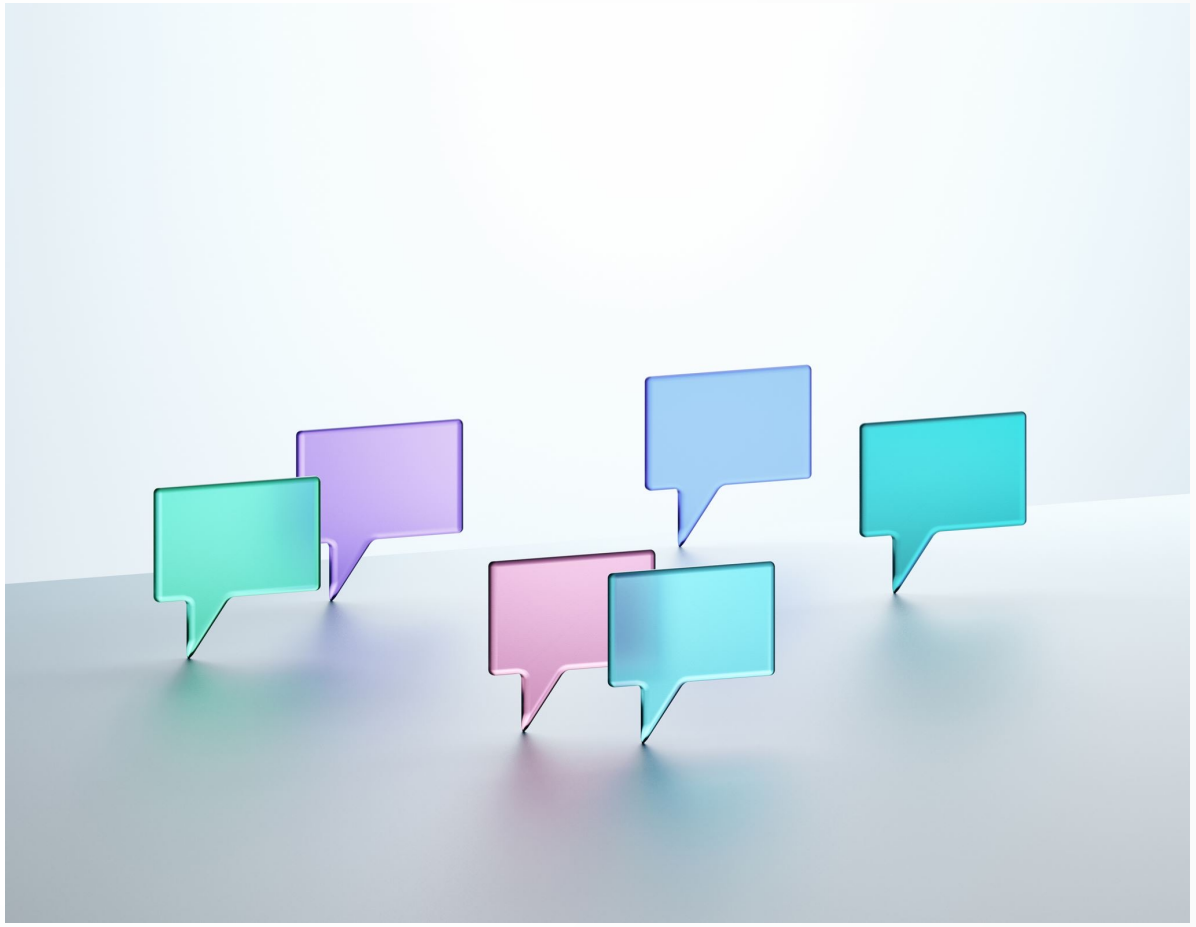
Unmasking Cyberbullying: Classifying comments and Revealing Communities



CIS 600: **Principles : Social Media and Data Mining**

Introduction

- Cyberbullying is bullying through digital technology, which can cause emotional distress, anxiety, and even suicide.
- A toxic comment classification model can distinguish patterns and attributes of various forms of toxicity, enabling identification and flagging of potentially damaging comments prior to posting.
- Community analysis can reveal groups with higher levels of cyberbullying and provide insights for targeted interventions and support.
- Advantages of the model include preventing harm, saving time and resources for moderators, and creating safer online environments.
- Together, toxic comment classification and community analysis can help prevent cyberbullying and promote respectful and constructive online interactions.

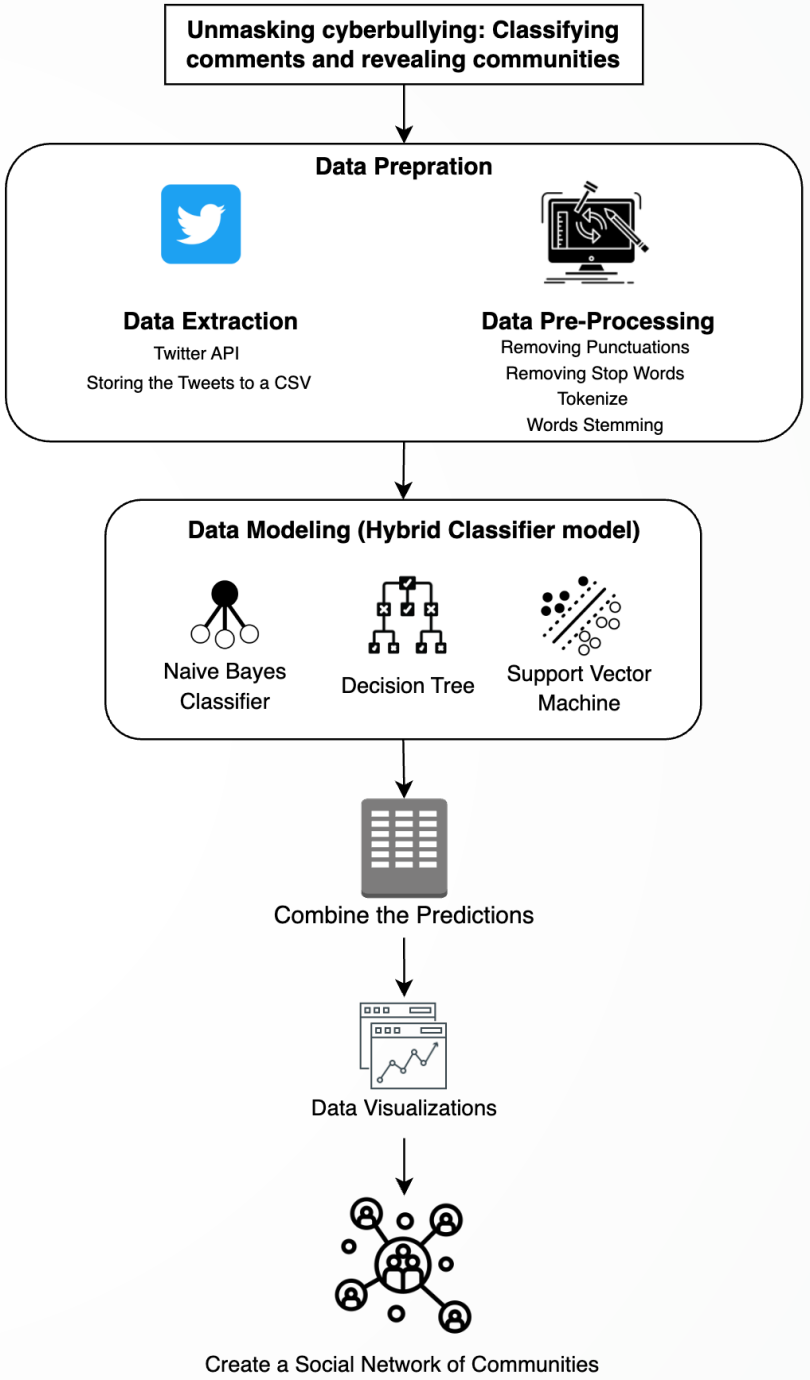
Objective: To develop a system that can classify cyberbullying comments and reveal the communities in which they belong.

Background: Cyberbullying is a serious problem that can have a devastating impact on victims. It can lead to depression, anxiety, and even suicide. Cyberbullying can take many forms, including name-calling, insults, threats, and even physical harm.

Approach: The system will use a combination of natural language processing and machine learning techniques to classify cyberbullying comments. It will also use social network analysis to reveal or identify the communities in which these users belongs to.

Benefits: The system will help to identify cyberbullying early on and prevent it from escalating. It will also help to identify the communities in which cyberbullying is taking place, so that interventions can be targeted at these communities.

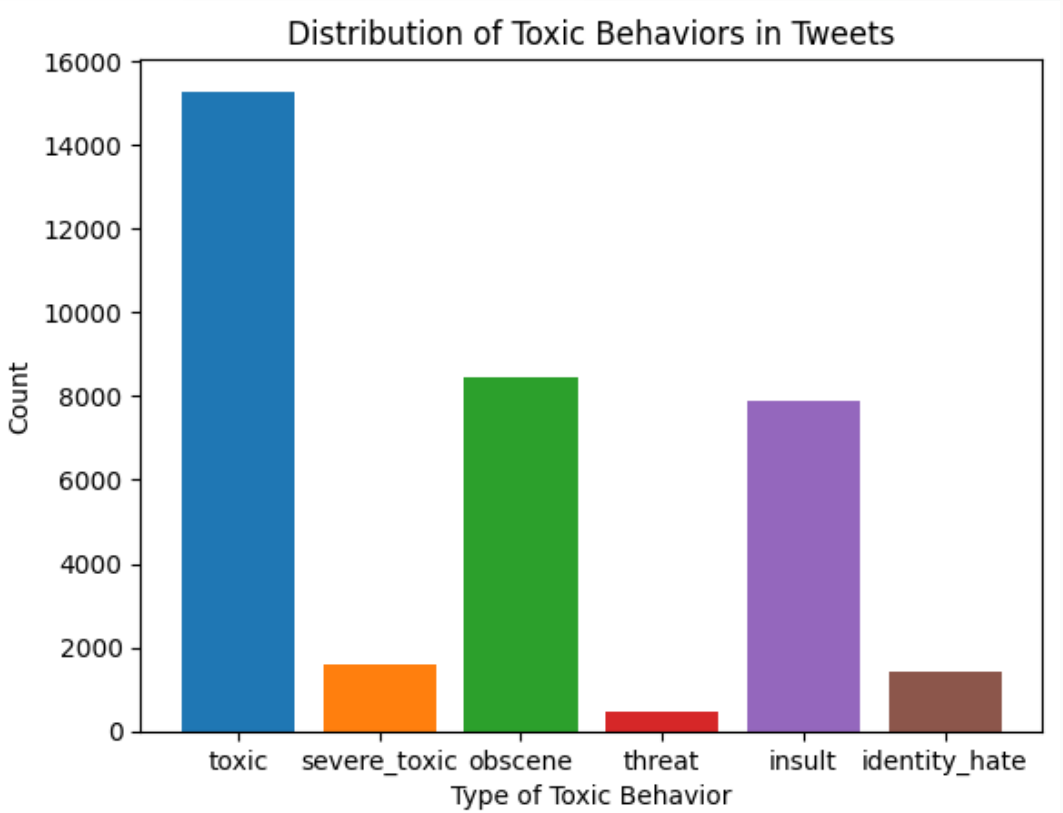
Architecture Design



Training Data Extraction

- The dataset for training and testing for our project is obtained from the Kaggle platform.
- The file train.csv is the training set, which contains comments with their binary labels.The file train.csv has a total of 151203 samples of comments and labeled data.
- The dataset consists of the following fields – id, comment_text, toxic, sever_toxic, obscene, threat, insult, identity_hate.
- The comment_text field will be preprocessed and fed into different classifiers to predict whether it falls under the label toxic or nontoxic.
- If any label has value 1, then the comment_text is classified as toxic comment.. Overall, 9790 samples are those that have at least one label, and 5957 samples have two or more labels.

Data Visualizations



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

Data Preprocessing

- Data preprocessing is required to transform or encode data for easy interpretation.
- Real-world datasets often contain missing data and noise, making data processing crucial for improving data quality and enabling pattern recognition.
- We performed the following preprocessing steps to our training data:

1. Removed punctuations
2. Removed the stop words
3. Stemming and lemmatization
4. Applied counter vectorizer

Data Extraction

- We began by utilizing the Twitter API in Python to retrieve tweets from the platform.
- We focused on collecting tweets with profanity by employing the track parameter, which allowed us to gather pertinent data for our analysis. The information gathered from Twitter was then stored in a CSV file.

Data Modelling: Multinomial Naives Bayes

- We have used the Multinomial Naïve Bayes classifier which is one of the popular choice for classification tasks involving discrete features. This classifier is particularly well-suited for text classification, where the features are typically represented as word frequencies or counts within documents.
- The algorithm operates by estimating the conditional probability of a specific word, given a particular class. It does this by calculating the relative frequency of the term t in documents that belong to class (c).
- This approach considers not only the presence of term t in training documents from class (c), but also the number of occurrences of term t within those documents. By considering multiple occurrences of a term, the classifier can more accurately capture the importance of a word in distinguishing between different classes.
- Accuracy 91.70

Data Modelling: Random Forest & SVM Classifier

Random Forest classifier

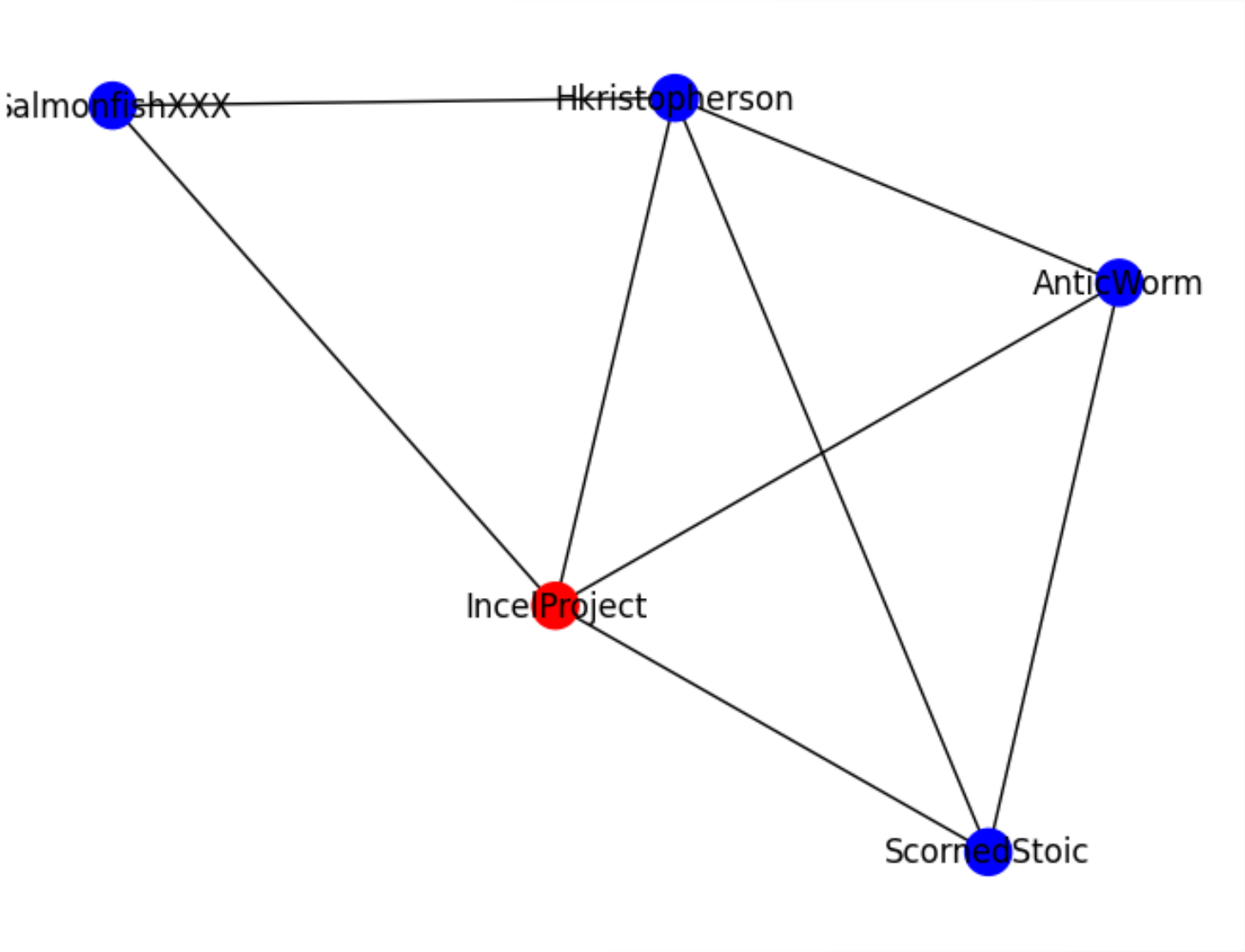
- Random Forest classifier using the labeled dataset. The classifier will create multiple decision trees, each trained on a random subset of the data and features.
- Evaluate the Random Forest classifier's performance on the testing set using metrics such as accuracy and precision.
- Apply the trained Random Forest model to comments in and reveal areas with high levels of cyberbullying.
- Random forest accuracy 95.20
- We Converted text data into a matrix of TF-IDF features using TfidfVectorizer.
- Reduced the dimensionality of the TF-IDF feature matrix using TruncatedSVD with 300 components.

SVM classifier (SVC)

- Trained an SVM classifier (SVC) with a linear kernel and fit the pipeline to the training data (X_train, Y_train).
- Predicted the class labels for the test data (X_test) using the trained pipeline.
- Calculate the accuracy by comparing the predicted labels (y_pred) with the true labels (Y_test).
- SVM Accuracy 94.12

Community graph (2-Clique)

Social network



ANALYSIS OF RESULTS

1 → 2 → 3 → 4 → 5

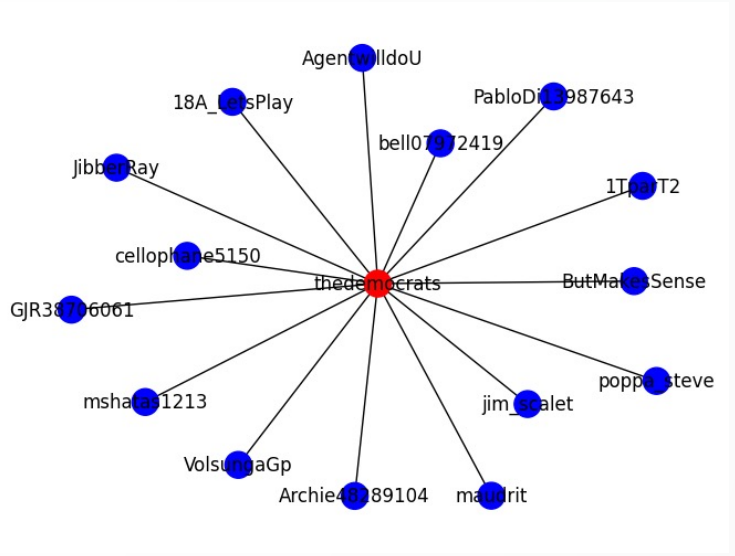
Visualization : To gain valuable insights into the relationships between the list of toxic users, we visualized the connections and interactions between them by creating a graph with the toxic users as nodes and the edges as their friends/followers.

Gathering insights : Some insights that can be drawn from the resultant graph are listed across.

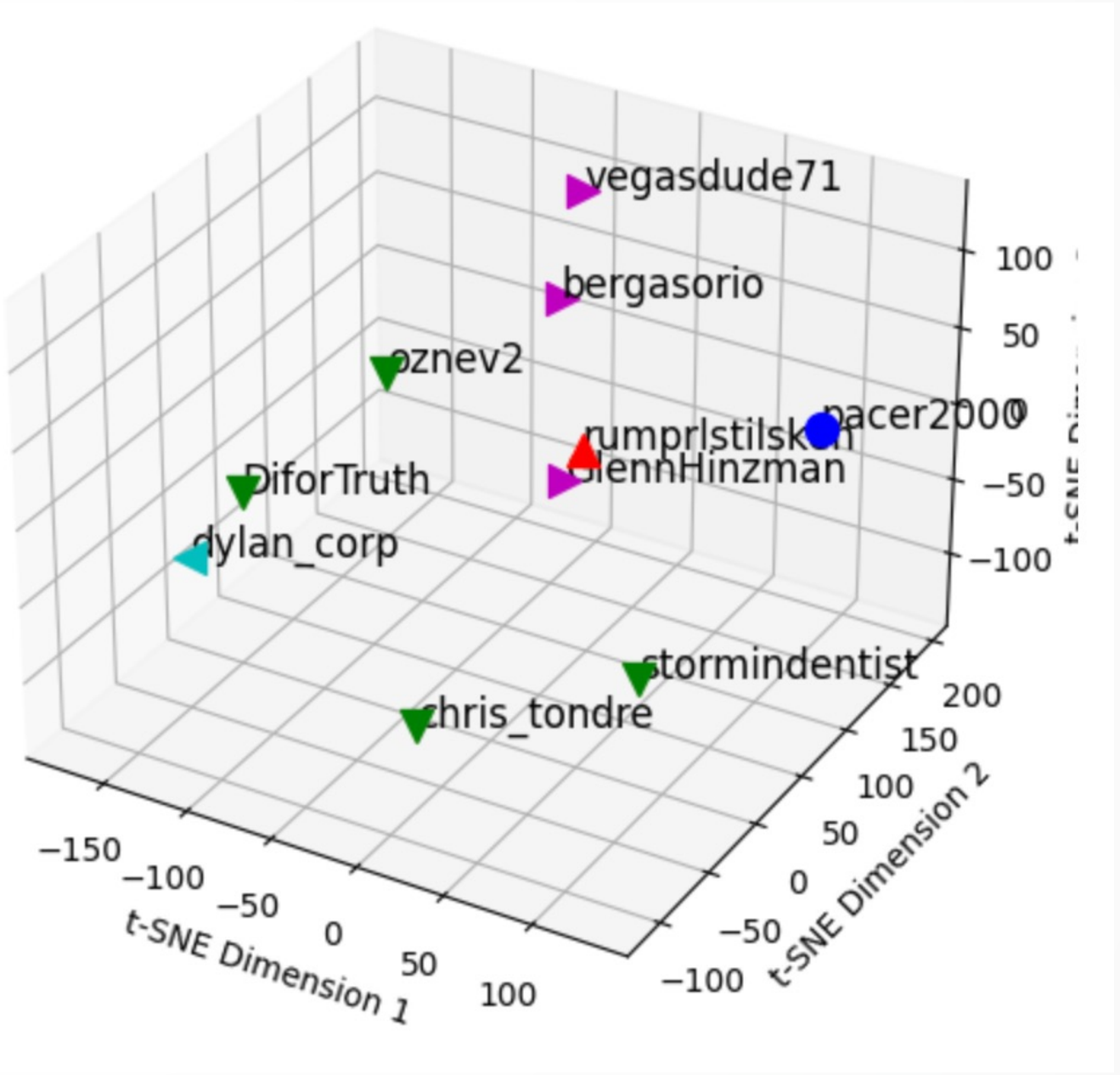
Identification of key players: By analyzing the graph, we can identify the most connected users, which could be the key players in the toxic community. These users could be driving the toxicity in the group. Eg IncelProject

Network structure: The graph reveals the structure of the toxic community, such as from our graph we can tell that this is a decentralized and clustered network, and all users follow each other.

Relationships: The graph depicts the nature of the relationships between the toxic users. Our graph indicates that the users follow each other and it cannot be a coincidence. It indicates that the group is more than an acquaintance.



Community Detection (clustering of users)



{'joebiden', 'go', 'racist', 'years', 'likes', 'big', 'male', 'politicians', 'chinese', 'world', '100', 'government', 'time', 'congress', 'idiot', 'problems', 'quit', 'border', 'money', 'gop', 'crack', 'biden', 'america', 'see', 'dead', 'place', 'better', 'https', 'woman', 'families', 'read', 'business', 'ban', 'china', 'always', 'piece', 'president', 'endwokeness', 'much', 'look', 'elonmusk', 'traitor', 'news', 'little', 'potus', 'country', 'back', 'day', 'must', 'take', 'people', 'krassenstein', 'know', 'senschumer', 'like', 'hey', 'seems', 'george', 'rest', 'smoking', 'getting', 'house', 'ukraine', 'first', 'start', 'would', 'black', 'history', 'class', 'got', 'us', 'term', 'get', 'co', 'thedemocrats', 'watch', 'ok', 'problem', 'dick', 'american', 'still', 'ass', 'inflation', 'poor', 'puppet', 'hunter', 'make', 'shit', 'bank', 'real', 'run', 'smart', 'criminal', 'want', 'bed', 'whitehouse', 'bill'}

'stop', 'hillaryclinton', 'living', 'girls', 'end', 'cause', 'use', 'anywhere', 'watch', 'drives', 'telling', 'barackobama', 'committee', 'inflation', 'screwed', 'governor', 'racism', 'goes', 'russian', 'ever', 'crazy', 'else', 'justice', 'exactly', 'democrat', 'irs', 'want', 'degree', 'easy', 'face', 'months', 'matter', 'police', 'instead', 'could', 'things', '2020', 'sense', 'course', 'dems', 'waiting', 'aoc', 'questions', 'control', 'worked', 'tax', 'broken', 'imagine', 'killing', 'given', 'politicians', 'research', 'thinking', 'chinese', 'went', 'giving', 'top', 'vote', 'call', 'threat', 'spend', 'might', 'thousands', 'may', 'blue', 'sorry', 'border', 'cbsnews', 'money', 'early', 'party', 'sell', 'davidhogg111', 'lives', 'see', 'place', 'deal', 'complete', 'better', 'wonder', 'resign', 'whatever', 'speak', 'corruption', 'saying', 'wing', 'weapons', 'read', 'happened', 'robreiner', 'soon', 'china', 'another', 'outside', 'report', 'disgusting', 'times', 'amp', 'care', 'fake', 'past', '10', 'mental', 'violence', 'jobs', 'traitor', 'needs', 'based', 'potus', 'dem', 'since', 'worst', 'liar', 'child', 'abortion', 'murder', 'free', 'voting', 'general', 'vp', 'asking', 'senschumer', 'men', 'try', 'left', 'foxnews', 'danrather', 'ridiculous', 'national', 'election', 'reporting', 'seems', 'trans', 'glad', 'amount', 'idea', 'action', 'repswalwell', 'owns', 'fool', 'tell', 'many', 'first', 'post', 'millions', 'joke', 'kids', 'history', 'maybe', 'love', 'died', 'others', 'jackposobiec', 'man', 'true', 'well', 'new', 'sit', 'involved', 'co', 'using', 'means', 'agree', 'claim', 'prove', 'american', 'enough', 'still', 'seriously', 'give', 'democracy', 'poor', 'hunter', 'mouth', 'daughter', 'sounds', 'hide', 'run', 'middle', 'fear', 'actually', 'failure', 'illegals', 'running', 'used', 'worse', 'bed', 'head', 'night', 'yes', 'even', 'win', 'right', 'bill', 'one', 'shot', 'girl', 'hillary', 'presidents', 'looks', 'red', 'looking', 'name', 'trying', 'word', 'oh', 'big', 'live', 'realjameswoods', 'statement', 'elected', 'find', 'current', 'repjeffries', 'human', 'mr', 'great', 'clear', 'young', 'democrats', 'full', 'came', 'fascist', 'also', 'absolutely', 'typical', 'touch', 'hold', 'person', 'reason', 'extremely', 'biden', 'gop', 'yet', 'borders', 'whole', 'able', 'hand', 'heard', 'thanks', 'families',

CHALLENGES

Limited access to data due to users privacy concern

Language barrier – If cyberbullying is done in other language, it cannot be identified by NLP.

Twitter's character limit on tweets can result in limited context for understanding the meaning and intent of a tweet

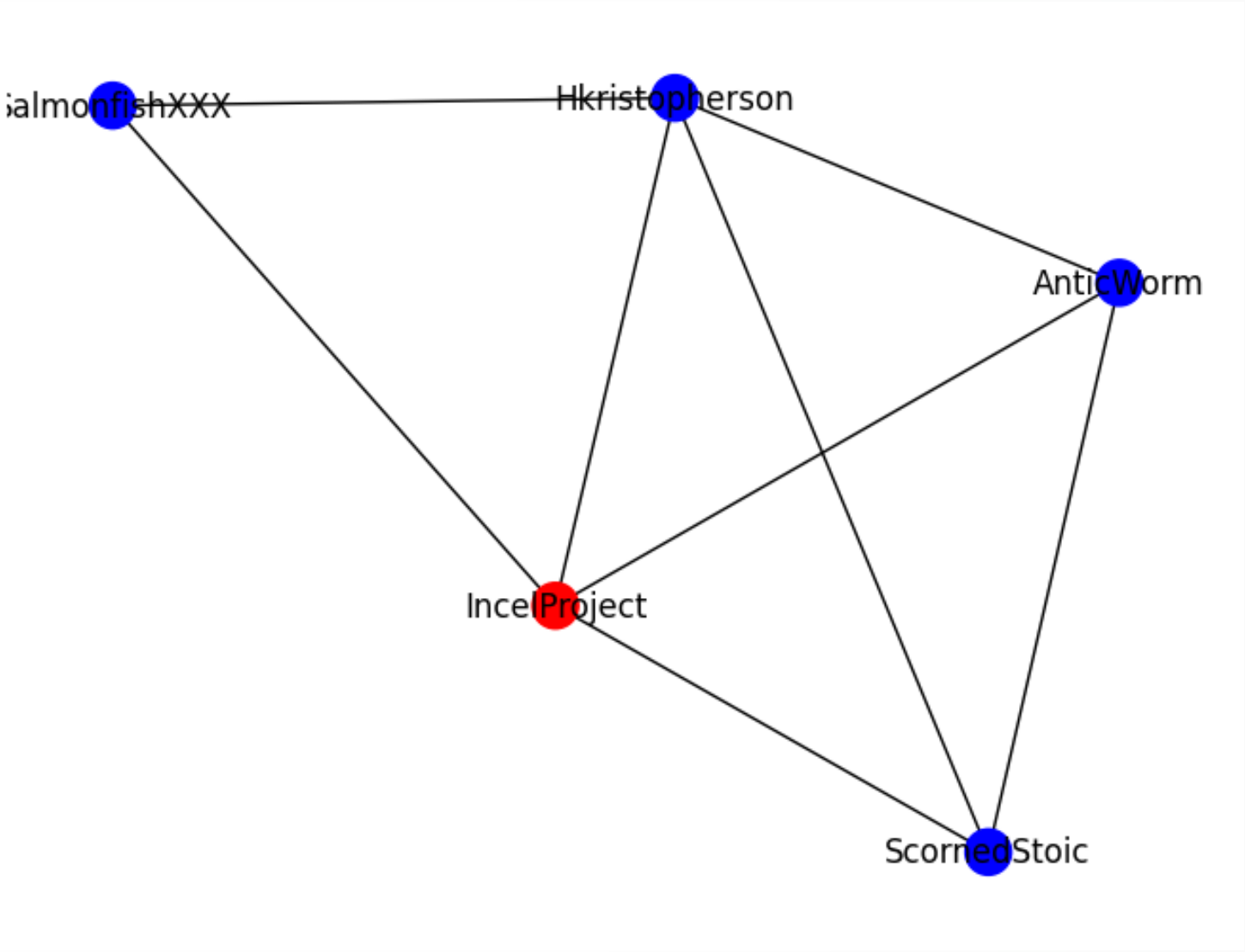
Due to restrictions on API, the data collection can be slow and tedious. As we can collect data only every 15 minutes.

Lack of uniformity, as the definition of cyberbullying can vary across cultures and languages

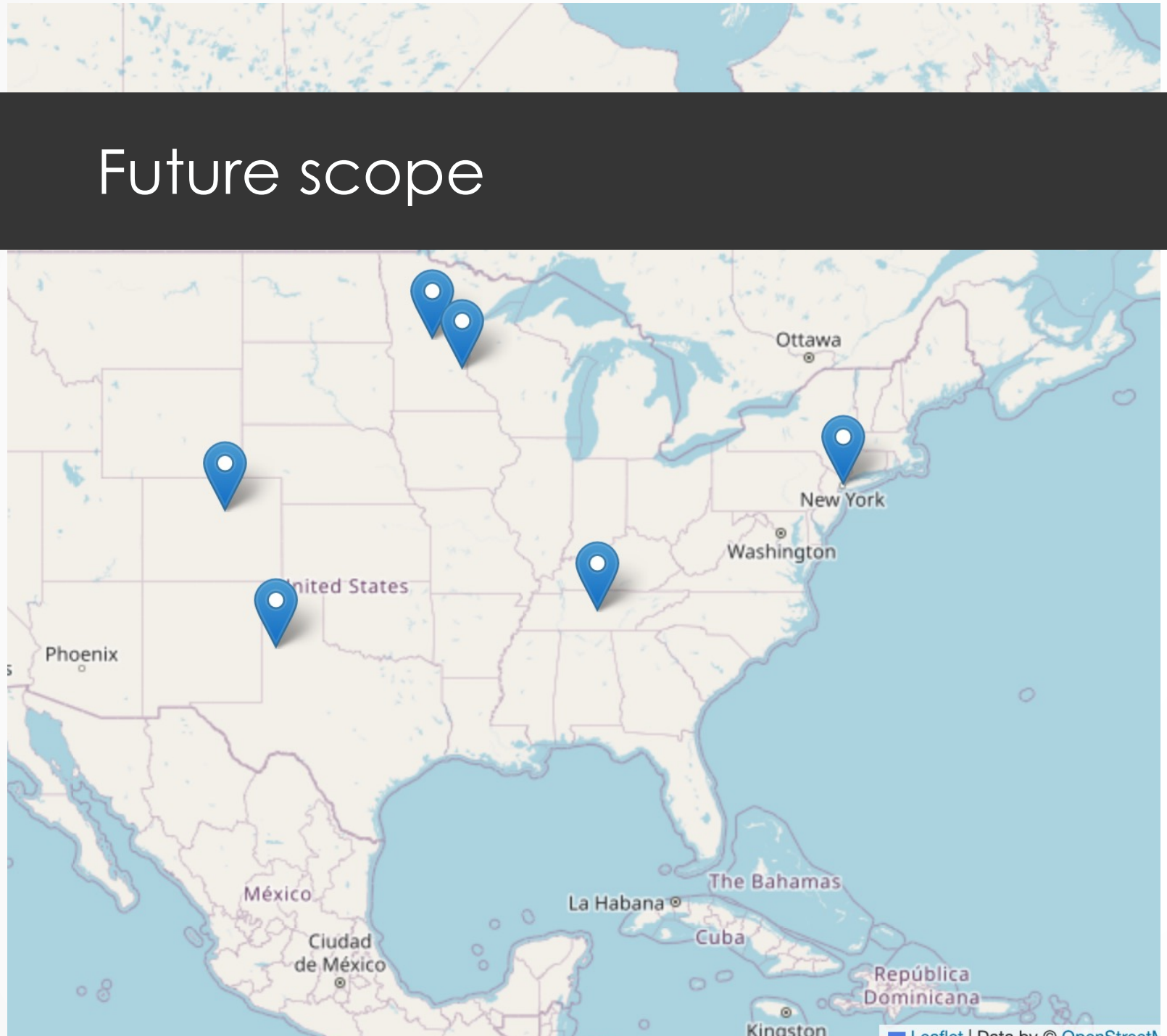
Rapid response required as Cyberbullying can spread quickly on Twitter is often necessary to address the issue.

Community graph (2-Clique)

Social network



Future scope



- We will be tagging people based on location.
- The data collected from the project can be used for predictive analytics, identifying the early signs of cyberbullying and preventing them before they escalate in real time.
- Apply this model to different social media platform

Conclusion

Our Project will help to identify cyberbullying early on and prevent it from escalating.

It will also help to identify the communities in which cyberbullying is taking place, so that interventions can be targeted at these communities.

Thank you