Analysis of bullying tweets in comparison to non bullying tweets with network approach

Background

Content filtering plays a crucial role in addressing cyberbullying on social media platforms such as Twitter. Despite various efforts to curb it, cyberbullying remains a persistent issue that significantly impacts users' mental health. Analyzing harmful tweets can provide insights into behavioral patterns, enabling the development of more effective prevention and intervention strategies. This project focuses on comparing networks of harmful tweets with those of non-harmful tweets, with the goal of identifying structural and linguistic properties unique to cyberbullying. Through this network-based analysis, the findings may help improve content filtering techniques, platform policies, and educational programs aimed at reducing online abuse.

The Data

The dataset used in this project is called **"Cyberbullying Classification"**, available on Kaggle: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.
It consists of more than 40,000 tweets, each classified by the type of bullying it represents—such as **age**, **religion**, **gender**, or **non-bullying**.

The Networks Built for the Analysis

Network 1: This network was constructed by representing tweets as nodes, with edges between them based on cosine similarity. Two separate networks were created—one for bullying tweets and one for non-bullying tweets.
Network 2: In this network, words serve as nodes, and edges between them are determined by cosine similarity. Each network represents to a single tweet. To address the sparsity caused by short tweet texts, all bullying tweets were concatenated and longer tweets were sampled. A separate network was then built for each sampled tweet by connecting the words within it. The same process was applied to the non-bullying tweets.

Technical background

Cosine similarity

Cosine similarity is used to calculate the distance between two vectors by measuring the cosine of the angle between them. Since the tweets in this dataset are not represented as vectors by default, I applied transformation methods to convert them into vector representations suitable for similarity calculations.

Transformation of the data

I applied two different data transformation approaches, each suits the specific requirements of the network type.

For network 1-tweets as nodes:

I transformed the tweets using the TfidfVectorizer class from the scikit-learn library .

This method converts a collection of text into a matrix of token counts weighted by Term Frequency–Inverse Document Frequency (TF-IDF).
A token refers to a basic unit of text, such as a word, number, or punctuation mark, that carries meaning.
TF-IDF assigns greater importance to words that are frequent in a particular tweet but rare across the entire dataset, helping to highlight words that are uniquely significant to each tweet.

This approach is commonly used for information retrieval and document similarity analysis. It captures the importance of terms within a tweet relative to the entire collection and is particularly useful for identifying key features in short texts. TF-IDF focuses on the **statistical importance** of words, not their semantic meaning.

For Network 2 – Words within a tweet as nodes

I transformed the data using the **fastText** library, developed by Facebook.

The fastText library represents words as vectors using **neural network-based word embeddings**.

Word embeddings are a way to represent words as numerical vectors. These vectors are learned from large amounts of text and are designed to capture both the **semantic meaning** and **contextual relationships** of words. In practice, words that appear in similar contexts will have similar embeddings.

The fastText model uses various techniques, such as predicting the context of a word and analyzing co-occurrence patterns, to learn these embeddings. It's commonly used in applications like text classification and language modeling.

Using fastText allowed me to convert each word in a tweet into a vector that reflects its meaning. These vectors were then used to calculate cosine similarity between words, forming the edges of the network.

Threshold values for the edges

To reduce noise in the networks, I removed edges with low similarity scores, keeping only those with a high degree of semantic similarity. Cosine similarity values range from 0 to 1, where higher values indicate stronger connections between nodes. I set the threshold at **0.6**, which reflects a relatively strong connection and helps preserve meaningful relationships between tweets or words..

Removing irrelevant nodes

In both networks, I removed isolated nodes and self-loops to eliminate elements that were not relevant to the analysis.

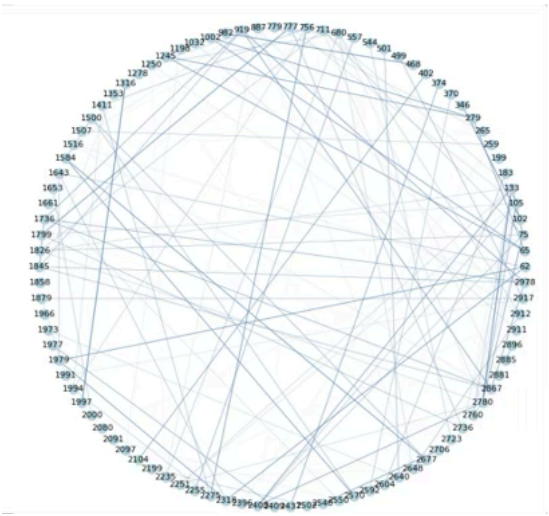
In **Network 2**, where words are represented as nodes, I performed additional preprocessing steps to refine the vocabulary. Specifically, I removed:

- Words that were URLs
- Words containing non-alphabetic characters
- Common stopwords

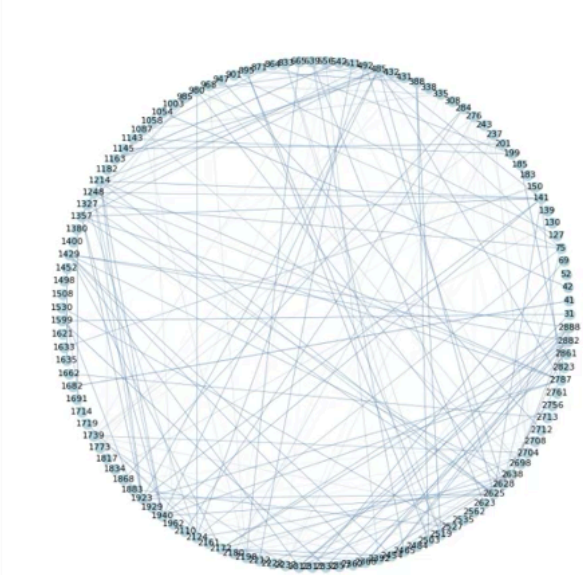
After this filtering process, the networks were constructed using the refined word set.

Visualizations:

Networks type 1: tweets as nodes

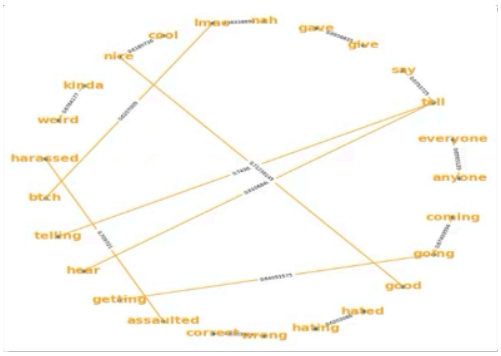


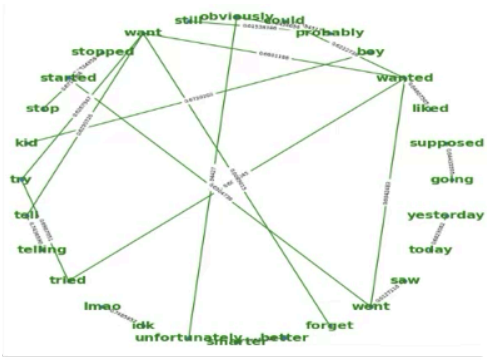
Bullying Tweets network:

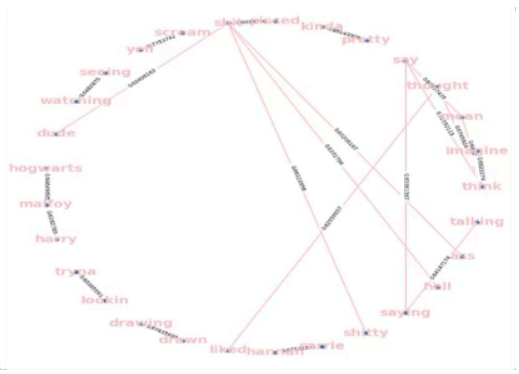


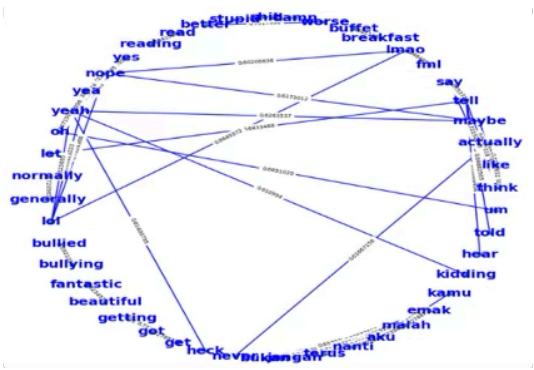
Non-Bullying Tweets network:

Networks of type 2- words as nodes:









Research questions:

1. What are the central and influential topics and key words of bullying and non-bullying tweets?
2. Is there a difference in the diversity of topics in bullying and non-bullying tweets?

The planned analysis:

A. Centrality Measures for Network Type 1 – Tweets as Nodes.

1. Degree Centrality: highlights tweets that are directly connected to many others, potentially indicating central themes or widely discussed ideas within the network.
2. PageRank: PageRank is an algorithm that assigns a numerical importance score to each node in a directed or undirected graph. A node is considered important if it is connected to other important nodes. These central nodes, or tweets, are likely to have a higher impact, engagement, or influence within the network compared to other tweets. They can represent important topics, trends, or conversations that are widely discussed or shared among users.
3. Closeness Centrality: Closeness centrality measures the average distance of a node to all other nodes in the network. Nodes with high closeness centrality are more central and have shorter paths to other nodes in the network. This could suggest that the tweet covers a "central" or "popular" topic because it shares similarities with many other tweets.

   I computed the three measures and generated a list of the top 10–20 tweets (the number was adjusted according to the results) for each measure. Next, I identified the nodes that appeared in all the lists, which helped determine the most influential tweets.

   Once I identified the influential tweets, I planned to utilize **word frequency analysis** to uncover the topics discussed in those tweets. By examining words that frequently appeared in bullying tweets and less frequently in non-bullying tweets, I aimed to gain insights into the central topics and keywords associated with each category of influential nodes

   The three measures above were used to answer **Research Question 1**.

B. Phenomenon- community detection: Keywords extraction and modularity score (network type 1)

I will perform community detection on the network of tweets to identify their corresponding topics- will apply a community detection algorithm (Louvain method) to find communities inside the networks. Once communities are identified, I'll explore the content of the tweets within each community and check the modularity score.

- **Keywords extraction:** For each identified community, I will generate a word cloud to highlight the most frequent or important words. This should provide insights into the central and influential topics within each community. By comparing word clouds or extracted keywords from communities in the bullying and non-bullying networks, I can identify differences and similarities in the topics discussed. **This analysis will help answer Research Question 1.**
- **Modularity Score:** The modularity score measures the strength of division within a network. A high modularity score indicates well-defined communities, which may suggest a high diversity of topics in the network. Although the number of communities does not directly represent topic diversity, since multiple communities may form around similar themes, it can still provide a useful indication. **This will help answer research question number 2.**

C. Aggregated "number of connected components" measure (network type 2)-

I will compute the "number of connected components" measure for each network of type 2 to identify the diversity of topics within each kind of network (both bullying and non-bullying). A larger number of connected components in a network can indicate a greater diversity of topics, assuming that each connected component represents a topic or subtopic. I will calculate the average of this measure for each type of network and compare the results.

This will help answer research question number 2.

Analysis and results

First research Question:

Measures of network type 1 (method A)

I calculated all three centrality measures on the networks as planned. For each measure, I selected the top 10 tweets based on their scores.

The results:

Bullying Network :
Top 10 nodes with the highest Degree Centrality:
[75, 105, 133, 2780, 2408, 680, 887, 1250, 2912, 711]
Top 10 nodes with the highest PageRank Centrality:
[259, 75, 133, 2780, 2917, 105, 2640, 1353, 1278, 2408]
Top 10 nodes with the highest Closeness Centrality:
[75, 105, 133, 2780, 2408, 2251, 2760, 2736, 711, 1826]

Non Bullying Network :
Top 10 nodes with the highest Degree Centrality:
[1739, 2704, 2232, 141, 485, 556, 1214, 1248, 1357, 1923]
Top 10 nodes with the highest PageRank Centrality:
[1739, 2704, 2232, 1621, 968, 1087, 2180, 2392, 556, 335]
Top 10 nodes with the highest Closeness Centrality:
[1739, 2704, 2232, 556, 2180, 2392, 2562, 276, 237, 1929]

By analyzing these measures, I identified the top nodes with the highest centrality scores in each network. I then examined which nodes appeared across all three centrality rankings within each network, highlighting the tweets that were consistently influential according to multiple criteria.

The common nodes for the bullying network:

```
Node 133: #Racism is when u select your White clothes to wash first before the Black colored ones...
Don't be a racist! Wash them all together

Node 2408: LMAO RT @2shorth: Racism is when u select ur white clothes to wash before d black colored
ones. Don't be a racist; wash them all together

Node 105: Racism is wen u select ur White clothes to wash first b4 the Black Colored ones... Don't be
a racist! Wash them all together #teamObama2012

Node 75: Racism is when u select yur white clothes to wash first before the black and colored one
s....Don't be a racist! Wash them all 2gether!!.

Node 2780: Racism is when u select your White clothes to wash first before the Black/ Colored ones...
Don't be a racist! Wash them all together
```

The common nodes for the non-bullying network:

```
Node 2704: @MisGrace @GBabeuf @RJennromao @DavidJo52951945 @Novorossiyan @gbazov @NewsCoverUp @rougek
68 Good night.

Node 1739: @MisGrace @GBabeuf @RJennromao @DavidJo52951945 @Novorossiyan @gbazov @NewsCoverUp @rougek
68 No it doesn't.

Node 556: @MisGrace @GBabeuf @RJennromao @DavidJo52951945 @Novorossiyan @gbazov @NewsCoverUp @rougek6
8 Russia, all media controlled by criminal Putin.

Node 2232: @MisGrace @GBabeuf @RJennromao @DavidJo52951945 @Novorossiyan @gbazov @NewsCoverUp @rougek
68 http://t.co/stp5NmjZRY
```

To gain insights into the topics discussed in these influential tweets, I initially planned to conduct a word frequency analysis. However, upon closer examination, I discovered that most of the highly ranked tweets were variations of a single tweet, differing only slightly in writing style. As a result, word frequency analysis would not yield meaningful results, and I decided to abandon that step.

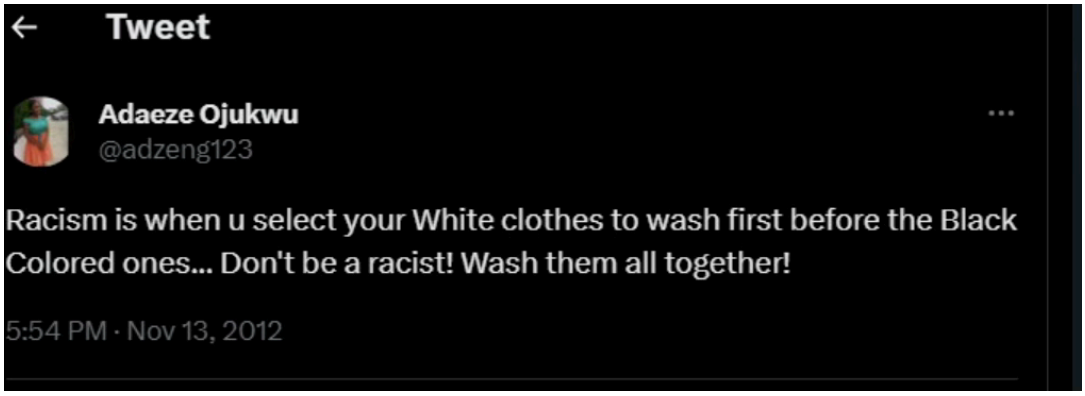
The influential topic in the non-bullying network: Finding the influential topics proved to be challenging, as the tweets mainly consisted of user tags. However, based on the content of node 556, I concluded that the discussion is likely related to Putin and Russia. Unfortunately, the rest of the tweets did not contain any significant or meaningful keywords that could provide further insight into the topic.

The influential topic in the bullying network: In the cyberbullying network, the main topic revolves around racism, particularly focusing on the distinction between White and Black skin tones.

Difficulties and limitations

- When analyzing the content of tweets, it is important to recognize that a tweet's intent may not always be offensive, even if it is classified as bullying. Interpretations can vary significantly between individuals.

  For example, during my analysis of the bullying network, I encountered a tweet whose meaning was unclear. It was difficult to determine whether it was intended ironically or seriously. Further investigation revealed that the tweet was posted by an African American girl with a dark skin tone. This challenged my initial assumption that it originated from a white individual trying to offend someone. It turned out the tweet was likely meant as a joke.



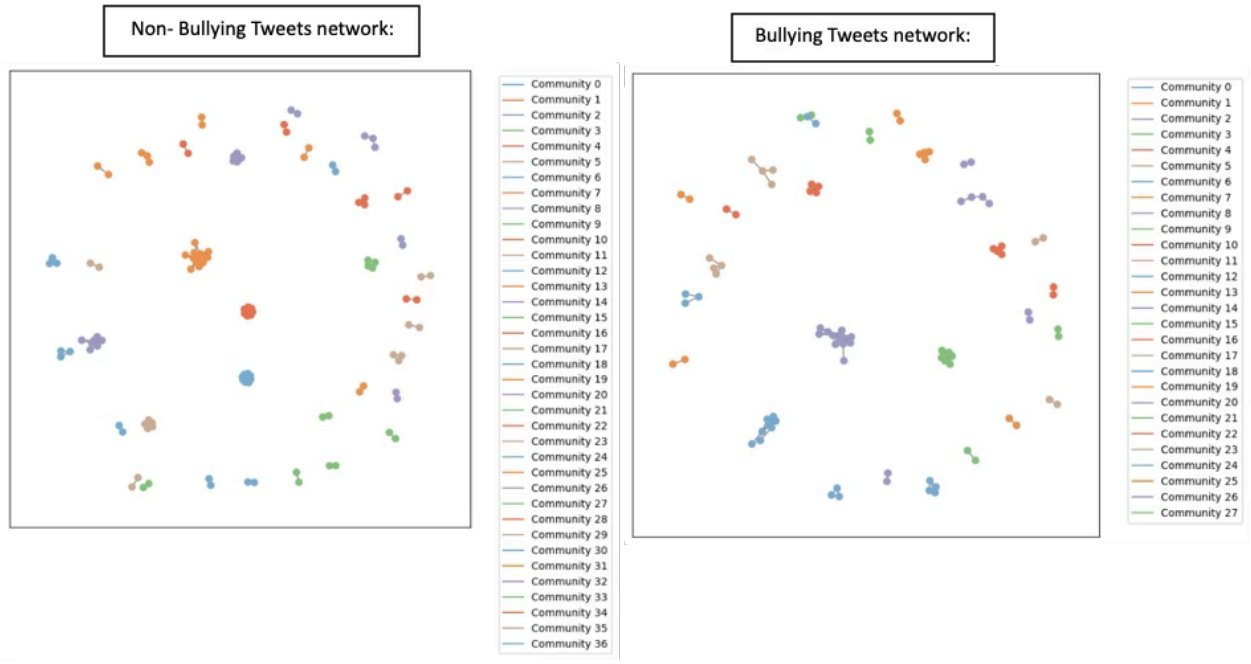
  This finding raises important questions about the difficulty of accurately distinguishing between genuinely offensive content and content that simply includes negative or sensitive language. It highlights the importance of context and nuance when evaluating online interactions.

- The networks used in this analysis were relatively small. After applying the filtering criteria, the dataset contained 119 non-bullying tweets and a similar-sized network of 94 bullying tweets. The limited size may have contributed to the repetition of certain tweets in slightly different forms, rather than revealing a broader range of distinct topics

In conclusion, the analysis of degree centrality, PageRank, and closeness centrality provided valuable insights into the influential nodes and the dissemination of content within both cyberbullying and non-cyberbullying networks.

Community detection of network type 1 (method B)

I applied the Louvain method to detect communities within each network and computed the modularity score.
To better understand how the tweets were divided into communities, I also visualized the resulting community structures.:



It can be observed that the non-bullying tweet network consists of more communities compared to the bullying tweet network. In both networks, there are a few larger communities alongside numerous smaller ones. This brief investigation, based on the visualizations, suggests that neither network contains a single dominant or significantly large community.

I also printed the tweet text from each community to review the results before analyzing them. Again, I found that most communities consisted of **the same posts with slight variations**.

```
Non-Bullying community tweets

Community 0:

Node 1058 tweet:
 #MKR star and chef Pete Evans says his critics will eat humble pie. http://t.co/LWC
u31ImVg http://t.co/JnYJYnM8Pz

Node 1962 tweet:
 'Bigger than Maccas' #MKR star &amp;amp; paleo chef Pete Evans says his critics will ea
t humble pie http://t.co/PFMPvItedY http://t.co/kMaSD5I3p2


Community 1:

Node 41 tweet:
 If your benchmark for success is Team Greek next door, then you're always going to
be feeling confident. #MKR

Node 1682 tweet:
 If your benchmark for success is Team Greek next door, then you're always going to
be feeling confident.


Community 2:

Node 42 tweet:
 RT @TVWEEKmag: #katandandre might have to eat her words after receiving some not so
great feedback on their ham and gruyere dish. #MKR #tvw…

Node 1054 tweet:
 #katandandre might have to eat her words after receiving some not so great feedback
on their ham and gruyere dish. #MKR #tvweekmag


Community 3:

Node 335 tweet:
 Quem mais sofre bullying: ( ) Loiras burras. ( ) Nerds. (X) O número “24” da chamad
a. -&amp;gt; @luuizaum DHUASHUDSAHUDAHSU

Node 665 tweet:
 Read my response to "Quem mais sofre bullying: ( ) Loiras burras. ( ) Nerds. (X) O
número “24” da chamada. ja…": http://4ms.me/pfcPAo

Node 968 tweet:
 Quem mais sofre bullying: - › ( ) Loiras burras. › › ( ) Nerds. › › (X) O número “2
4” da chamada. http://tumblr.com/xob3xk4pmc
```

```
Bullying community tweets

Community 0:

Node 62 tweet:
 #MileyCyrus Miley Cyrus causes controversy with date-rape 'joke' at GAY club gig in
London http://ift.tt/1jm2uyJ

Node 1643 tweet:
 #MileyCyrus Miley Cyrus causes controversy with date-rape 'joke' at GAY club gig in
London http://ift.tt/1l3jfBf'joke'-at-GAY-club-gig-in-London&amp;amp;c=2ANoySojn758d6ckA
cbPgEIBE3n0xG-Z3i4i-MK56ko&amp;amp;mkt=en-us

Node 1979 tweet:
 #MileyCyrus Miley Cyrus causes controversy with date-rape 'joke' at GAY club gig in
London http://ift.tt/1jUkLaR

Node 2318 tweet:
 #MileyCyrus Miley Cyrus causes controversy with date-rape 'joke' at GAY club gig in
London http://ift.tt/1lfdIGI


Community 1:

Node 65 tweet:
 #MileyCyrus Miley Cyrus Shocks With Insensitive 'Date Rape Joke' And 'Gay Comments'
[WATCH VIDEO] http://ift.tt/1laVtlV

Node 279 tweet:
 #MileyCyrus Miley Cyrus Shocks With Insensitive 'Date Rape Joke' And 'Gay Comments'
[WATCH VIDEO] http://ift.tt/1laJF2Z

Node 1002 tweet:
 #MileyCyrus Miley Cyrus Shocks With Insensitive 'Date Rape Joke' And 'Gay Comments'
[WATCH VIDEO] http://ift.tt/1jimNNo

Node 1245 tweet:
 #MileyCyrus Miley Cyrus Shocks With Insensitive 'Date Rape Joke' And 'Gay Comments'
[WATCH VIDEO] http://ift.tt/1nyO9qf


Community 2:

Node 2885 tweet:
 “@TheKidJR_: Suck some dick bitch RT“@tayyoung_: FUCK OBAMA, dumb ass nigger”” WHO
THE FUCK ARE YOU?

Node 2896 tweet:
 @TheKidJR_: Suck some dick bitch RT“@tayyoung_: FUCK OBAMA, dumb ass nigger” lmaooo
o she is U.P.S.E.T
```

Here is an example of a community that contained different posts, though such cases were relatively rare.

```
Community 7:

Node 370 tweet:
 @brittd1178: Everybody get out and vote so we can get this nigger out of office! wt
f dumb ass bitch fuck romney hoe

Node 2592 tweet:
 you dumb racist Bitch Fuck You &amp;amp;'nd Mitt Romney #TeamObama @brittd1178: Everybo
dy get out and vote so we can get this nigger out of office!"
```

I had initially planned to compute word clouds for each community to identify keywords and gain insights into the overall topics within each network. However, given the limitations and biases mentioned earlier, this approach was unlikely to yield meaningful results for topic identification, so I decided to abandon it**.**

Therefore, I manually reviewed the communities and identified the topics discussed in each one.

The main topics of the bullying network included Miley Cyrus and her joke relating to gay people, Barack Obama, and racism. These topics were highly repetitive across different communities.

The main topics of the non bullying included criticism of reality shows, wellness, anti-bullying advocacy on Tumblr, religion, and politics

* After analyzing each community, I found that almost every bullying-related community contained strong curse words, whereas the non-bullying tweets had very few or none at all.

* Identifying clear topics was especially challenging in the bullying communities, which were largely dominated by profanity.

Difficulties and limitations

- I had to manually examine and analyze the content of each community. This approach is susceptible to bias, as different interpreters might identify slightly different topics within the same community.
- The tweets are quite short and often lack contextual information, which makes it challenging to accurately determine the main topic of a tweet based solely on its content.
- The fact that most communities revolved around the same tweet suggests that the data representation is limited. It likely reflects a narrow subset of tweets and may not capture the full range of topics and discussions occurring on Twitter.

Second research Question:

Computing The modularity score to assess diversity (method B):

Upon examining the modularity scores, I found that they were relatively similar in both networks.

modularity score bullying tweets: 0.8680052674639873
modularity score non bullying tweets 0.8808472558737166

As a result, using this measure alone is not sufficient to determine a significant difference in topic diversity between the networks.

Limitations:

- If the topics of discussion overlap significantly across different communities, a network may still have high modularity but low topic diversity. In this analysis, I observed that many subjects appeared repeatedly across multiple communities, which made it difficult to assess diversity using the modularity score alone.
- Modularity doesn't account for diversity within communities. I found that most of the communities were centered around the same tweet, so I didn’t need to face this issue in this analysis.

Average “number of connected components” (method C)

I computed the number of connected components within each small network of type 2 (each representing a sampled tweet) and then calculated the average for the bullying networks and the non-bullying networks.

The results are:

average_connected_components bull: 10.984257357973991
average_connected_components non bull: 13.25462962962963

The average number of connected components in the non-bullying tweets is slightly higher, suggesting a modestly greater diversity of topics compared to the bullying tweets. However, the difference is not significant.

Limitations

- The “number of connected components” measure assumes that each connected component represents a distinct topic or subtopic. I support this assumption, as the edges calculated using cosine similarity capture both semantic and contextual relationships between words.
- Building networks from short and sparse text, such as tweets, can result in limited connectivity between words. These sparse networks are more likely to produce many small connected components, which may lead to an overestimation of topic diversity.

Conclusion of the results

Research question no.1: What are the central and influential topics and key words of bullying and non-bullying tweets?

In the non- bullying tweets:
Based on the centrality measures (**A**), the main topics discussed were related to Russia and Putin. Based on community analysis (**B**), the main topics included criticism of reality shows, wellness, stopping bullying on Tumblr, religion, and politics.

In the bullying tweets: Based on the centrality measures (**A**), the main topic was racism, particularly regarding the distinction between Black and White people. Based on community analysis (**B**), the main topics included Miley Cyrus and her joke relating to gay people, Barack Obama, and racism..

Research question no.2: Is there a difference in the diversity of topics in bullying and non-bullying tweets?

Based on the connected components and modularity analysis (**C**), the non-bullying tweets appeared to be slightly more diverse in terms of topics, but the difference was not significant.

Limitations- regarding the whole project

- The classification process of the data is unclear to me. I am unsure whether it was done by a computer or through manual annotation. Some classifications, such as the incident at Seabreeze High School involving a girl with special needs being attacked, were labeled as "age" cyberbullying. However, the connection to the category of "age" is not entirely clear in this case. It is possible that the classification was influenced by the cursing directed towards the girls who bullied the special needs student, but the direct association with "age" is questionable..
- The size of the networks is relatively small, consisting of approximately 100 nodes (tweets) after applying all constraints. As a result, the limited number of nodes may have contributed to the lack of diverse results in the analysis.

Final notes

While the findings alone may not be sufficient to improve content filtering without further investigation, I believe that the underlying investigative framework established here can be

utilized on larger datasets. With some adjustments to the data processing methods, it is possible to achieve improved results in content analysis and filtering.

Further research

- Choosing other platforms Data: Replace data from Twitter with data from other platforms like Instagram, YouTube, or Reddit to examine the same research questions in different environments.
- Filtering out similar posts- remove near-duplicate tweets to avoid the main problem encountered in this analysis.
- Analysis of hashtags instead of tweet content.