Università Ca'Foscari Venezia

Sami Trabelsi

Sentiment Analysis on Interview between Andrew Tate and Andrew Piers

Abstract

In today's world of abundant information online, when there's a controversial topic, people often end up seeing only one side of the story because of algorithms that show them what they want to see. This makes it tough to get a true sense of public opinion on a subject, as various factors work to influence people's perspectives.

This paper aims to analyze the public opinion expressed in the interview between Andrew Tate and Piers Morgan, which took place on YouTube. The structure of the comments will also be analyzed using network analysis. The data we're using consists of thousands of comments from the YouTube video of the interview.

The findings reveal that the comments generally lean towards a positive view of Andrew Tate. However, it's important to note that overall, the sentiment expressed in the comments is quite balanced, indicating a mix of opinions among the viewers.

1 Introduction

In contemporary digital culture, Andrew Tate has garnered unprecedented attention as the most frequently searched individual globally on google, amassing billions of views for his online video content. His influence has engendered a spectrum of reactions, ranging from vehement criticism of both his persona and ideologies to fervent adoption of his worldviews by a diverse audience, incorporating them into their daily lives.

Despite the prolific discourse surrounding this figure, the prevailing sentiment within the public domain remains elusive. The ambiguity persists as to whether a majority disapproves of him or, conversely, aligns with his perspectives. Subsequent to attaining celebrity status, Mr. Tate faced the cessation of his social media presence due to allegations of misogynistic behavior. This pivotal development prompted an exclusive interview with British journalist Piers Morgan, generating millions of views and hundreds of thousands of comments.

This study endeavors to discern the prevailing public sentiment toward Andrew Tate by conducting a sentiment analysis on the comments associated with his interview video. Additionally, a network graph will be employed to elucidate the intricate relationships between individual comments and the underlying sentiments expressed. Combining these analytical methods is meant to provide insight into whether the public sentiment leans favorably or unfavorably towards Andrew Tate.

2 Methods

2.1 Gathering the Data

The dataset under consideration comprises textual data in the form of comments obtained from the publicly available interview between Piers Morgan and Andrew Tate, accessible on YouTube at the following URL: https://www.youtube.com/watch?v=VGWGcESPltM. Retrieval of the comments necessitated the utilization of an Application Programming Interface (API) to interface with the YouTube platform.

The resultant dataset encompasses a total of 105,000 individual comments. From this comprehensive dataset, a subset of columns was meticulously curated for relevance to the subsequent analyses. The selected columns deemed pivotal for the investigative purposes of this study are enumerated as follows:
"Comment": The content of the comment in the form of text
"AuthorDisplayName": The Author's Name
"ReplyCount": The number of Replies of the comment
"LikeCount": The number of Likes of the comment
"PublishedAt": The Date at which the comment was published
"CommentID": The unique ID of the Comment
"ParentID": The unique ID of the parent comment if any

2.2 Text Processing

Text processing is crucial before conducting sentiment analysis on any text data. This is because raw text data, especially from sources like social media, often contains noise, irrelevant information, and various elements that can affect the accuracy of sentiment analysis. Here are the various steps taken in text processing for the analysis, considering the presence of emojis and other factors:

- ***Removing NA Values:*** Elimination of instances with missing values for data integrity.
- ***Text Cleaning: Handle Special Characters:*** Special characters are not essential for sentiment analysis since they give no information.
- ***Lower-casing:*** Standardize the text by converting all letters to lowercase.
- ***Removing Stop Words:***
  - *Filter Out Common Words:* Systematic removal of common linguistic stopwords to focus on substantive content.
  - *Filter Out Custom Stop Words:* Expunging of internet slang-related custom stopwords for nuanced refinement. (s, t, m, re)
- ***Lemmatization or Stemming:*** Used to reduce words to their base or root form. This helps in reducing the dimensionality of the data and capturing the essence of the words.
- ***Tokenization:*** Break down the text into individual words or tokens. This step is essential for further analysis and feature extraction.

2.3 Sentiment Analysis

2.3.1 Sentiment Analysis per Comment:

The VADER (Valence Aware Dictionary and Sentiment Reasoner) will be used for this analysis.
VADER is designed for analyzing sentiment in social media text, making it potentially well-suited for YouTube comments. It is particularly effective in handling text with informal language, slang, and emojis which is why emojis were kept in the text data.

2.3.2 Sentiment Analysis on the general public:

In this segment, the NRC (National Research Council) Emotion Lexicon was employed as a foundational resource for sentiment analysis.
The NRC lexicon, accessed using tidytext package in R, comprises a comprehensive set of words annotated with binary indicators for eight basic emotions and two sentiments. These emotions encompass anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, while sentiments include negative and positive valences. By utilizing this lexicon, the aim is to discern the emotional and sentiment nuances present in textual data. The resulting sentiment data frame, served as a valuable tool for investigating the emotional undertones of words.

2.4 Network Graph

2.4.1 Words Graph

The Words graph is constructed to analyze the connections among various adjectives employed in reference to both the interviewer, 'Piers Morgan,' and the interviewee, 'Andrew Tate,' facilitating the discernment of distinct opinions and sentiments surrounding both individuals

2.4.2 Comment Graph

The construction of the network graph for comment interactions involves a categorization and organization of comments within the dataset. Comments are stratified into two primary classifications: "Parent Comments" and "Comments." The former pertains to comments that elicit replies, whereas the latter encompasses both standalone comments and replies with no associated parent.

To facilitate a more insightful visualization of the network, the dataset undergoes ordering based on the ReplyCount metric. Specifically, comments boasting more than 40 replies, along with their respective reply chains, are retained in the dataset. This strategic selection ensures the inclusion of prolific comment threads, contributing to a more comprehensive representation of the network.

Distinguishing between Parent Comments and Reply Comments is achieved through the application of distinct colors, providing a visual demarcation between the initiating comments and their subsequent replies.

The subsequent graph construction is executed employing both the igraph and visNetwork library.
Igraph: A robust tool renowned for its efficacy in network analysis.
visNetwork: A versatile and interactive platform for creating and visualizing network graphs,

3 Results

In this section, a comprehensive presentation of the analytical outcomes will be undertaken. The elucidation will encompass an examination of the distribution of the initial dataset, an exploration of the results derived from the sentiment analysis, and a depiction of the network graph. Various plots will be strategically employed to elucidate the observed phenomena, providing a visually informative narrative of the analytical findings.

3.1 Descriptive Analysis

First plot (figure 1) will be the Distribution of Comments according to their ReplyCount and LikeCount. As Expected the vast majority of comments have 0 replies and 0 likes, therefore for the visualization to be somewhat meaningful, the log scale was implemented in these graphs.

The Maximum Reply Count is 490. The median of the Measure is 0 with a mean of 0.25

According to the second plot (figure 2) The Maximum Like Count is 119062. The median of the Measure is is 0 with a mean of 11.05

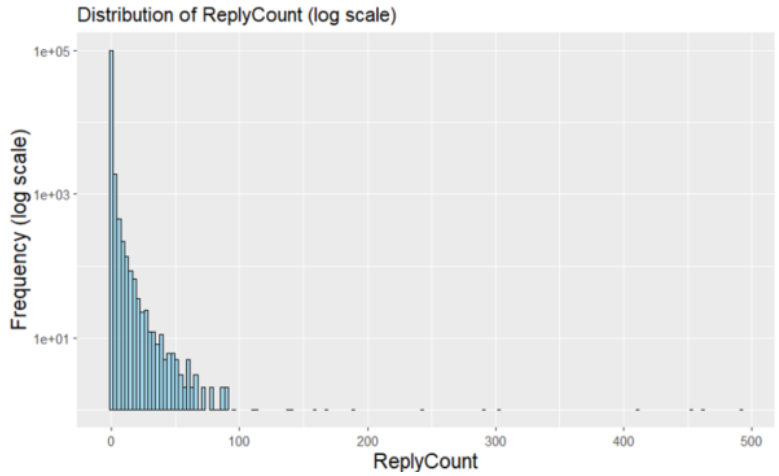

Figure 1: Frequency of ReplyCount Logscaled

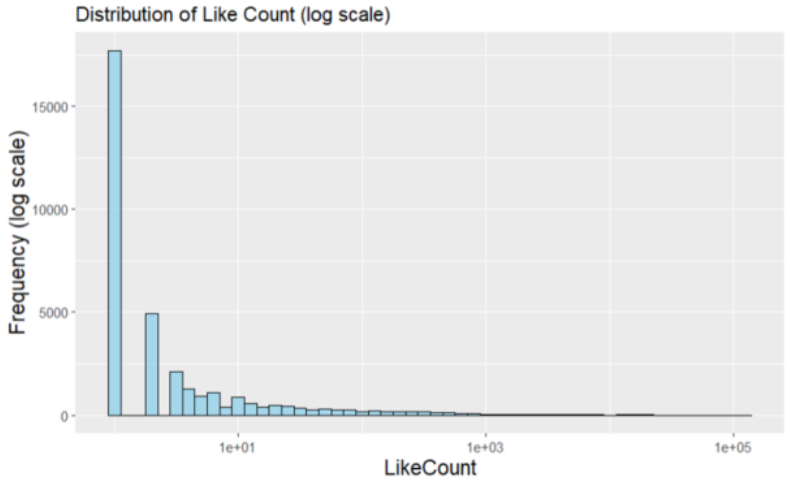

Figure 2: Frequency of LikeCount Logscaled

As depicted in Figure 3 below, there exists a discernible correlation between the "LikeCount" and "ReplyCount." This correlation is logically intuitive, as comments accruing a higher number of likes tend to be more prominently displayed, thereby inviting increased participation in the form of replies. Notably, comments that serve as catalysts for discussion often exhibit a higher count of likes, underscoring the intrinsic link between engagement and the appeal of the content.

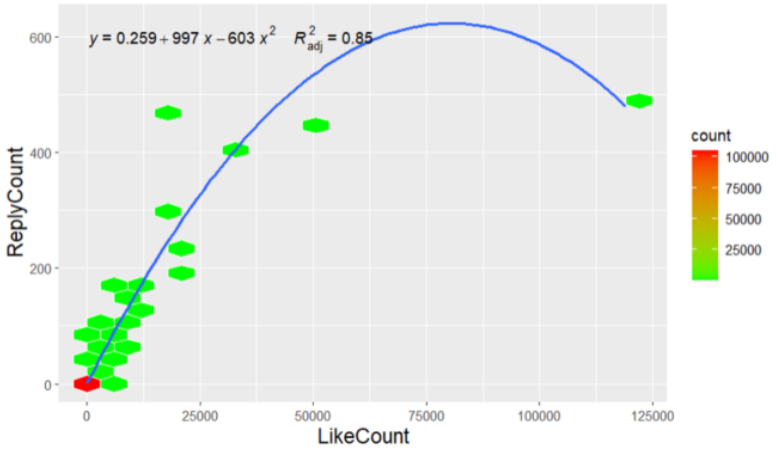

Figure 3: Density plot of Reply Count and Comment Likes

3.2 Most Common words

The initial dataset is inherently unrefined, particularly considering its origin from the comment section of a YouTube video. Given the informal nature of online interactions, the dataset is likely to contain a plethora of internet slang and an abundance of common stop words, presenting potential obstacles to our analysis. Nevertheless, prior to data cleaning, an examination of the most prevalent words—both pre and post-cleaning—offers an insightful perspective on the prevailing linguistic patterns within the dataset.

Figure 4: Most Frequent Words Before Data Cleaning

In the initial state of the raw dataset, prior to text cleaning procedures, the word cloud visually represents the most frequently occurring words within the entirety of the comments. Notably, the presented word cloud (Figure 4) lacks substantive information, as it predominantly features stopwords—words devoid of meaningful content in the context of sentiment or the intended meaning of the comments. Illustrative examples of such stopwords include "the," "and," "a," and "this."

Figure 5: Most Frequent Words After Data Cleaning

In the updated word cloud derived from the cleaned dataset (Figure 5), a discernible augmentation in informative content is immediately evident. Notably, prominent words emerge that depict descriptors associated with the interview, including terms such as "man," "like," "authority," "depression," and "think".

The presented word cloud encapsulates the most frequently used words without attributing them to specific individuals. Consequently, while it offers valuable insights into the general sentiment pervading the comments, it does not facilitate the differentiation of sentiments directed towards particular individuals.

To elucidate this aspect, a graph network will be employed, linking primary nodes representing 'Andrew Tate' and 'Piers Morgan' to additional nodes representing various adjectives used to characterize them. This network graph will provide a nuanced understanding of the sentiments associated with each individual within the interview context.

3.3 Word Graph

Within the commentary section, distinct sentiments have been expressed, some of which are directed negatively towards the interviewer, while others are aimed at the interviewee. To facilitate a nuanced differentiation between the two perspectives, a systematic analysis is essential. Specifically, the identification of the most frequently occurring words linked with the names of both individuals becomes instrumental in delineating the prevailing sentiments. This analytical outcome will be visually presented through the medium of a word graph, offering a comprehensive representation of the prevalent language associated with each party involved in the interview process.

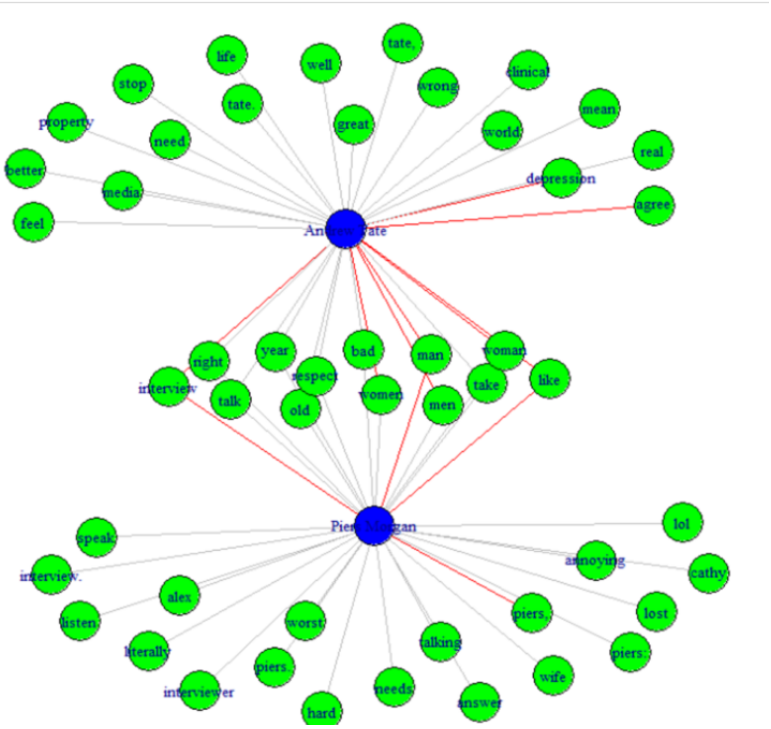

Figure 6: Word Graph of Most Common Words

In the depicted word graph, the primary nodes, distinguished by a blue node, represent the interviewer and interviewee, with associated words linked to each individual. Notably, words negatively and uniquely connected to Andrew Tate include "stop", "wrong", "depression," and "property" as well as positive descriptors such as "agree," "great," "well," "better," and "life."

Conversely, terms exclusively linked to Piers Morgan encompass "worst," "annoying," "lost," and "lol," along with words like "hard," "listen," and "talking."

Moreover, a set of common terms shared between both individuals consists of "old," "women," "men," "bad," and "right." This nuanced analysis serves to highlight distinct sentiments associated with each figure, as well as shared sentiments applicable to both.

The graph suggests a mixed sentiment towards both the interviewer and interviewee, encompassing both praise and criticism. While the visual representation indicates a relatively balanced sentiment, with a slight positive skew towards Andrew Tate, a more comprehensive understanding of the dynamics requires a genuine sentiment analysis of the comments. Such an analysis would provide a nuanced exploration of the sentiments expressed, enabling a clearer insight into the reception of both parties involved in the interview.

3.4 Sentiment Analysis on each comment

3.4.1 Distribution of the Sentiment Scores

Following the application of the VADER (Valence Aware Dictionary and Sentiment Reasoner), each comment has been assigned a numerical value that reflects its sentiment. A score of 0 corresponds to a neutral sentiment, while negative values indicate a negative sentiment and positive values signify a positive sentiment. This numerical representation allows for a quantitative assessment of the sentiments expressed in the comments, facilitating a more precise and systematic analysis of the overall sentiment dynamics associated with the interview and the individuals involved.

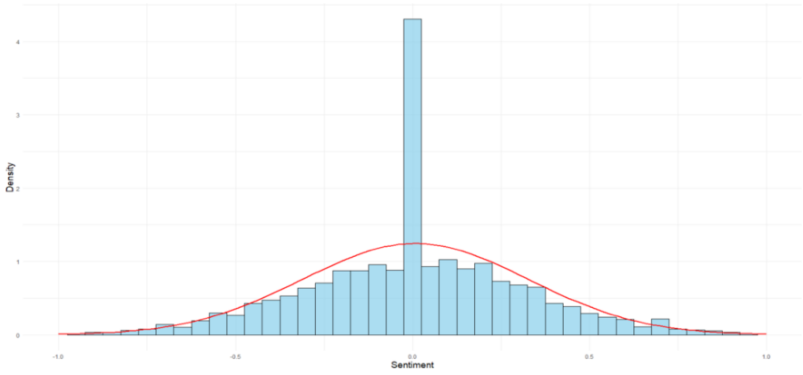

Figure 7: Distribution Of Sentiment Scores on each Comment

The mean sentiment score of 0.08 indicates a generally positive sentiment across the comments. The assertion that the sentiment scores follow a normal distribution is supported by a visual comparison between the actual data and simulated normal distribution curves, with the latter depicted in red.

Furthermore, the positive skewness of the distribution strengthens the observation that the sentiment distribution is right-skewed. This asymmetry indicates that the distribution is inclined towards the positive end, signifying a greater prevalence of positive sentiments in the comments compared to negative ones. The positive skewness contributes to the understanding that the overall sentiment expressed in the comments is more positively oriented.

Additionally, when considering that 51.6% of the sentiments are positive (excluding neutral comments), a robust conclusion can be drawn. The higher percentage of positive sentiments aligns with the right-skewed distribution, indicating that more comments express agreement or enjoyment regarding the interview. This convergence of statistical measures underscores a prevailing positive sentiment in the comments, suggesting a generally favorable reception of the interview content by the audience.

3.4.2 Sentiment Depending on the Like Count

Examining the correlation between the like count and the sentiment score provides valuable insights into the potential impact of individual comments on the overall sentiment analysis. If a single negative comment garners a substantial number of likes, it could significantly influence the overall sentiment despite being numerically negative.

Analyzing this correlation allows us to discern whether there is a consistent relationship between the sentiment expressed in comments and the level of engagement reflected by the like count. A positive correlation would imply that comments with higher sentiment scores tend to receive more likes, reinforcing the idea that the sentiment expressed carries weight in the perception of the audience. On the other hand, a weaker or negative correlation may suggest that like counts are less influenced by the sentiment expressed in the comments.

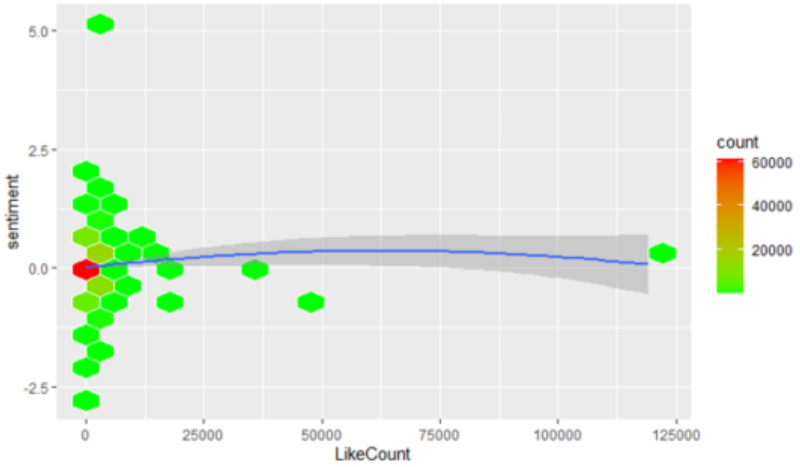

Figure 8: Density Plot of the Sentiment Scores by Comment Likes

The positive correlation of 0.075 between the like count and the sentiment score, despite being relatively low, indicates a discernible trend. This suggests that, to some extent, there is a tendency for comments with higher sentiment scores to attract more likes. While the correlation value may be modest, the positive direction of the correlation implies that, on average, as the sentiment score of a comment increases, there is an accompanying increase in the number of likes.

It's essential to recognize that other factors may also influence the like count, and the correlation value, though positive, may not be strong. Nevertheless, the observed trend supports the conclusion that there is a connection between the sentiment expressed in comments and the level of engagement reflected by the like count.

3.4.3 Variability of Sentiment over time

To ascertain the temporal evolution of sentiment regarding Andrew Tate, it would be beneficial to conduct a sentiment analysis over time. By observing changes in sentiment over time, one can gauge whether public opinion has shown a discernible trend of improvement or deterioration in response to unfolding events, such as the criminal investigation. the sentiment score graph over time indicates no significant difference in public

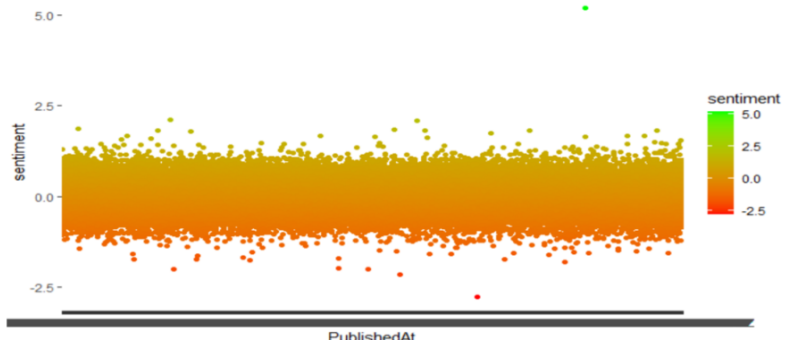

Figure 9: Scatterplot of Sentiment by Date of Publication of Comment

sentiment between the date of the video's publication and the present day, it suggests a relatively stable sentiment trend over the observed period.

3.5 Sentiment Analysis on all the comments

In this part of the study, we move from looking at comments as casual opinions to treating them as pieces of text that we can analyze more systematically. This shift allows us to explore the public's overall feelings on different emotional aspects. We make use of the National Research Council (NRC) Emotion Lexicon, which includes a set of words marked with indicators for eight basic emotions. This approach helps us understand the general sentiment in a more nuanced way compared to using numerical values.

Following the integration of the NRC library into the analysis of all comments, a newly incorporated column now showcases the assigned emotions for each word. This additional dataset will serve as the basis for our subsequent emotional analysis.

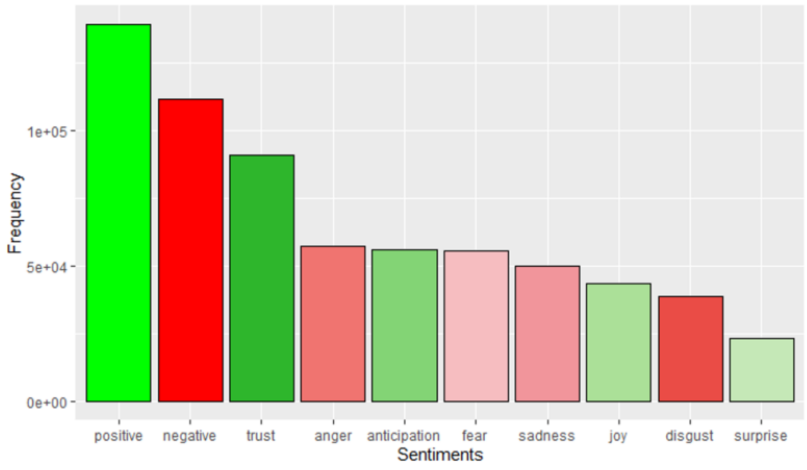

Figure 10: Distributionof Emotions

As evident from the distribution and our earlier exploration, there is a prevalent trend of positivity among individuals expressing their sentiments about the interview.

Notably, emotions such as trust, anger, anticipation, and fear emerge prominently, indicating their substantial presence in the overall emotional landscape of the comments. Conversely, the emotion of surprise appears less frequently, suggesting its comparative rarity in the expressed sentiments. This collective emotional disposition provides valuable insights into the predominant emotional tones and reactions conveyed by the audience toward the interview content.

To gain deeper insights into the dynamics of opinions, let's examine the words that are most frequently associated with both positive and negative emotions. This analysis will shed light on the specific language and expressions that contribute to the prevailing emotional tones in the comments.

Figure 11: Most Frequent Negative Words

The identification of the word "depression" as the most frequent negative term is noteworthy, its presence does not necessarily convey negative emotion since depression was a wide subject of the interview.

However, the presence of words such as "dumb," "idiot," "annoying," and "problem" suggests a distinct layer of negative sentiment in the comments. These terms are often associated with criticism, dissatisfaction, or disagreement, contributing to a more critical and unfavorable tone within the expressed opinions.

Figure 12: Most Frequent Negative Words

The prevalence of words such as "authority," "agree," and "respect" as the most frequent in positive emotions signals a positive and affirming sentiment within the comments.

3.6 Network Graph With igraph

Now, as explained in the methods section, we're checking how comments are organized under the interview video. Comments are split into main comments and replies. Using the igraph library, we'll create a visual display to show how these comments are connected to each other. This will help us understand the order and relationships between the comments in a more straightforward way.

Once we've set up the igraph object, using the "DisplayAuthorName" as the names of the points and the like count as their importance, our next step is to create a network graph to see how everything connects.

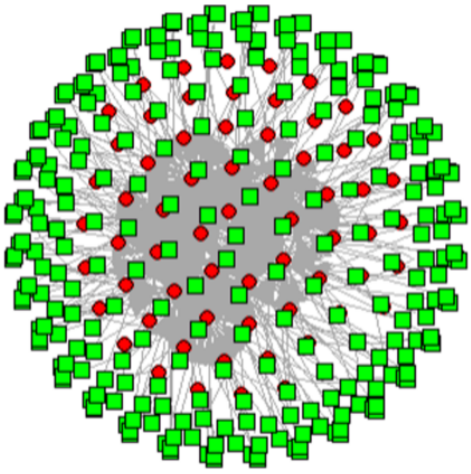

Figure 13: Igraph Object

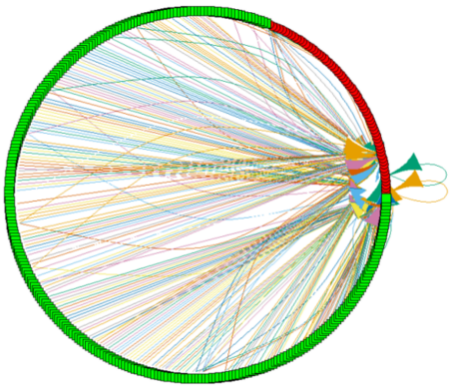

Figure 14: Igraph Object Circle Layout

Attributes of the igraph

- Mean Edge Weight:
  - Represents the average weight assigned to the edges connecting nodes in the graph. In this context, it reflects the average like count associated with connections between individuals.
  - The mean edge weight is 105, it is high compared to what was seen in the descriptive analysis but is it because in the graph only the popular comments were chosen.
- Diameter: The diameter of the graph is the longest shortest path between any two nodes. In this case, the diameter is 5242, indicating the maximum number of steps required to traverse from one node to another in the graph.
- Mean and Median Degree:
  - Mean Degree: The average number of edges connected to each node in the graph is 1.61
  - Median Degree: The middle value when all node degrees are arranged in ascending order. In this case, most nodes have a degree of 1.
- Mean Distance: Represents the average distance (number of edges) between pairs of nodes in the graph. In the graph the mean distance is 332. this value showcases that the graph presented here is not connected.

- Edge Density: Reflects the proportion of possible edges that exist in the graph. In the graph the edge density is 0.00238. this value showcases that the graph is not connected and has less connections.
- Reciprocity: Measures the extent to which connections in the graph are reciprocal (mutual). In the graph, the value is of 0 which indicates no reciprocity, meaning the author of the parent ID does not reply in other popular comments.
- Closeness Centrality: Measures how close a node is to all other nodes in the graph. The scores indicate the proximity of each node to others. The node closest to all other graph has a coefficient of 0.0076.

These attributes collectively offer a nuanced understanding of the structural characteristics, connectivity patterns, and centralization within the comment network.

Figure 15: Igraph Object Circle Layout

The graphical depiction derived from the comments, as observed in the network graph, manifests as a disconnected graph. This delineation signifies that the graph lacks complete connectivity, as evidenced by the presence of nine disjoint components

In the context of the observed degree distribution depicted in the histogram, it is evident that a substantial majority of nodes exhibit a degree of one. This prevalence strongly suggests that these nodes primarily function as replies to parent comments, consistent with the characterization provided earlier. Conversely, nodes with higher degrees are indicative of distinct patterns within the network. Such nodes may represent users who have engaged by commenting on multiple top-level comments or could potentially signify top-level comments that have garnered numerous replies, thus contributing to the network's overall structure and connectivity.

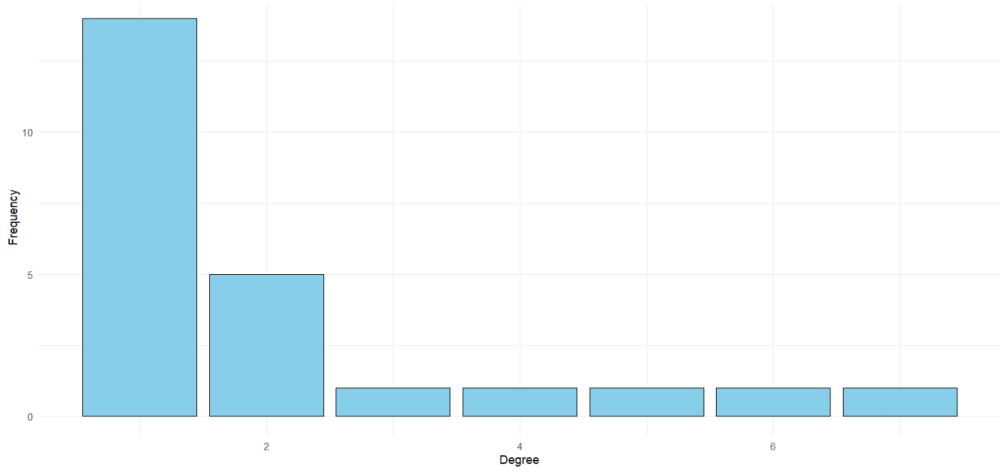

Figure 16: Degree Distribution

3.7 Network Graph using VisNetwork

Due to the high volume of data, we might need another graph that might be interactive and clearer to use. With this library, we can afford to have more comments and more replies.

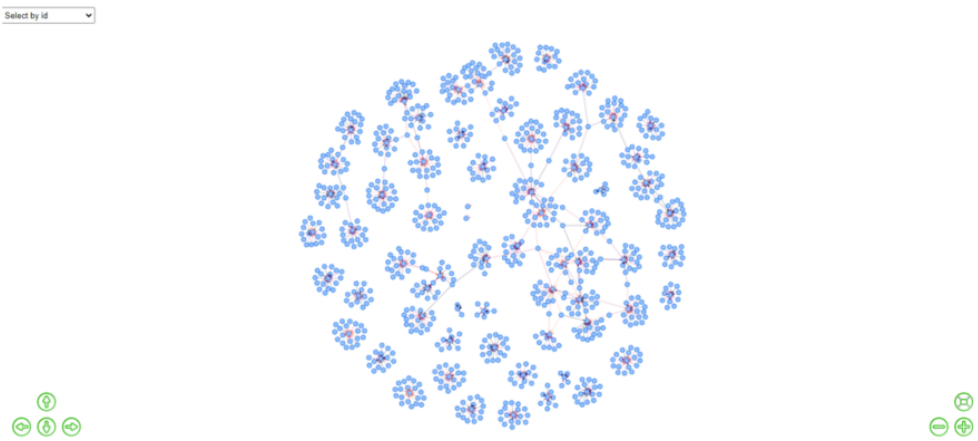

Figure 17: Igraph Object Circle Layout

We are provided with an interactive network visualization, offering the capability to navigate, drag, and zoom in and out. This functionality enhances our ability to clearly illustrate the nodes and their relationships within the network.

4 Discussion

Following the completion of sentiment analysis, the findings reveal a nuanced perspective within the public discourse regarding Andrew Tate. Although there is a discernible balance in divergent opinions, the sentiment analysis indicates a prevailing positivity towards his figure. This assertion is substantiated by both the NRC Emotion Lexicon and the VADER analysis.

Moreover, the network analysis serves a dual purpose by illuminating the directional focus of specific words in relation to the interview's participants. Additionally, the network visualization provides a structural representation of the comments on the YouTube video, distinguishing between parent comments and replies.

It is pertinent to acknowledge that the present analysis of public sentiment regarding Andrew Tate is confined solely to comments associated with the specific interview under consideration. For a more comprehensive and representative understanding of the broader public sentiment, it is necessary to conduct a sentiment analysis across various platforms and social media channels, encompassing a diverse range of posts.

5 Bibliography

James Cook, Associate Professor Of Sociology https://www.uma.edu/directory/staff/james-m-cook/

igraph Library https://cran.r-project.org/web/packages/igraph/index.html

VisNetwork Library https://datastorm-open.github.io/visNetwork/

Andrew Piers. (2022). Andrew Tate vs Piers Morgan — The Full Interview.
Retrieved from https://www.youtube.com/watch?v=VGWGcESPltM

Wordcloud2 Library https://cran.r-project.org/web/packages/wordcloud2/vignettes/wordcloud.html

Sentimentr Library https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf