ANDREW TATE

A Sentiment Analysis on One of The Most Controversial Figures

Team 19

LU TECK HII
TONY SMITH
AMAY VISWANATHAN IYER

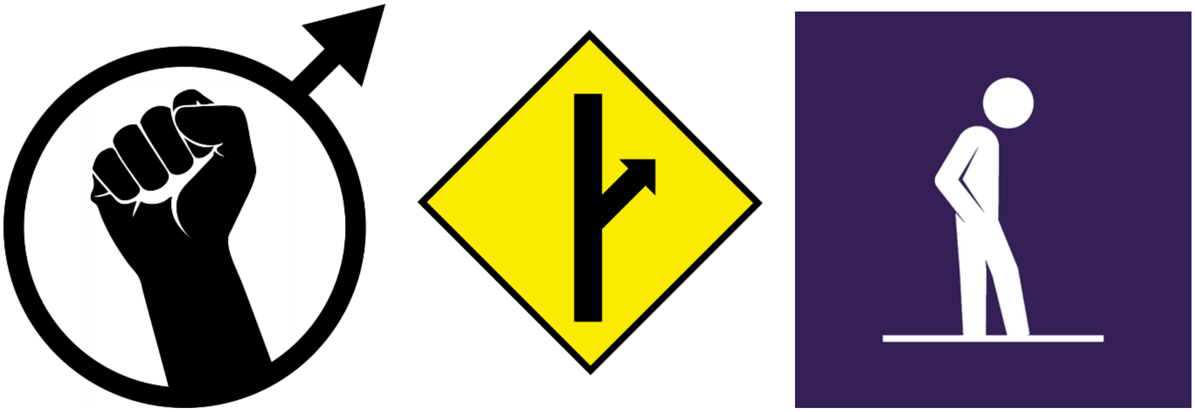
Course

SOCIAL MEDIA AND NETWORKS ANALYTICS

Table of Contents

Introduction

Andrew Tate, not just a former kickboxing world champion but a digital force to be reckoned with, stands tall and prominent within the labyrinthine world of the Manosphere. The Manosphere, a complex ecosystem of websites, blogs, forums, and social media groups, has carved out a significant space in the digital realm, promoting a strong pro-masculinity stance while challenging feminist ideologies. Encompassing diverse movements such as Men's Rights Activists (MRAs), Men Going Their Own Way (MGTOW), and Involuntary Celibates (Incels) this digital subculture serves as both a sanctuary for some and a subject of critique and controversy for others. (logos shown below in the movement's mentioned order) (7)



Figures 1-3: Men's Rights Activists, Men Going Their Own Way, Involuntary Celibates

As time has unraveled, Andrew Tate's presence within this milieu has only amplified. Whether it is due to his candid nature, his unabashed opinions, or his refusal to conform, he has cemented himself as a figure that is both revered and reviled. His YouTube content, an eclectic mix of personal views, debates, and life philosophies, has often been a catalyst for strong reactions. For some, he is a beacon of truth in an increasingly 'politically correct' world, while for others, he represents a problematic mindset. In this report, our lens will be focused keenly on the digital imprints of audiences — the myriad of YouTube comments scattered across Tate's videos. Our goal is to extract, analyze, and understand the public sentiment orbiting around him and the topics he delves into. The heart of our research lies in a singular, compelling question: Can the vast sea of YouTube and Reddit comments, with their nuances, emotions, and biases, serve as a reliable barometer to gauge public sentiment, especially when the subject is as contentious as Andrew Tate?

Digital realms like YouTube are not just platforms; for many, they are extensions of their identity, a digital playground where real-world beliefs, biases, and emotions play out. They offer a space where debates ignite, discussions flow, and occasionally, where digital skirmishes erupt. By navigating through these intricate digital networks, we strive to chart the journey of sentiments — from their origin, through their metamorphosis, to the ripple effects they create in both the online and offline worlds. To methodically dissect and understand this vast digital interaction, our research is structured into two pivotal segments:

Observation & Exploration

In this section, we embark on an expedition into the heart of Andrew Tate's digital domain. Who are the denizens of this space? From die-hard followers who hang on to every word Tate utters, the critics who analyze and challenge his every stance, to the neutral observers who are just passing by, we aim to understand the demographics, psychographics, and the dynamics at play. By leveraging sentiment analysis tools, we dissect the emotions that underpin every praise, critique, and casual comment about him.

Our exploration will involve a meticulous assessment of YouTube and Reddit comments, diving deep into the layers of conversations happening there. The objective? To trace patterns and anomalies. Are there recurring themes in the praises and criticisms? This observation will lay the groundwork for more in-depth analyses as we proceed.

Inferences

Building upon our observations, this segment will pivot towards drawing meaningful inferences from the amassed data. Comments are more than just text; they are a goldmine of insights, revealing the collective consciousness of the audience. With the vast dataset of YouTube comments at our disposal, our aim is to extrapolate larger sentiments and beliefs of Tate's followers.

Every comment, whether praising, critical, or neutral, contains within it a sentiment, a belief, or a perspective. By amalgamating and analyzing these sentiments, can we delineate the broader ethos of Tate's followers? Moreover, by juxtaposing the sentiments from different content, we aim to identify any shifts in public perception over time. Have certain events or videos drastically altered the sentiment landscape? And if so, what were they and why did they resonate (or repel) so strongly?

Background

The digital age has ushered in a myriad of online communities, providing spaces for individuals to share ideologies, offer support, and coalesce around particular worldviews. One such burgeoning online movement has been the 'Manosphere', a collective of online platforms and groups focused on perceived crises of masculinity and the ensuing responses to these perceptions. Grounded in the analytical work of Bujalka, Rich, and Bender (2022), this study seeks to delve deeper into specific influential nodes within the Manosphere, with a particular focus on the online influence network of Andrew Tate.

Bujalka et al. (2022) provided a foundational understanding of the Manosphere, highlighting the symbiotic cycles of ontological security and insecurity propagated by its 'thought leaders'. The notion of 'ontological security', as initially posited by Anthony Giddens, implies a stable mental state derived from a sense of continuity regarding the events in one's life. Ontological insecurity, conversely, arises when this continuity is disrupted, leading to feelings of anxiety and uncertainty about one's place in the world. Within the context of the Manosphere, thought leaders leverage these cycles, creating a sense of crisis, only to offer solutions to the very threats they highlight. This creates a protection racket-like structure wherein followers are both made aware of perceived threats and offered solutions, all while reinforcing the thought leader's influence and often translating into material or social gains for themselves.

This paper by Bujalka et al. brings forth a crucial observation about how thought leaders within the Manosphere, through platforms like YouTube, not only draw individuals into this space but also extract material and social resources, creating a retention mechanism. The language of catastrophe, combined with the promise of reclaiming lost power or status, presents a potent lure for many. Platforms such as the 21 Convention, framed as a bulwark against perceived threats to masculinity, epitomize this approach.

Andrew Tate emerges as a significant figure within this milieu, yet there remains a gap in understanding the nature of his influence network. Does it function primarily as an echo chamber, reinforcing and amplifying existing beliefs without challenge? Or does it serve more as a structure of ontological security online, offering its adherents a sense of stability and understanding in a rapidly changing world?

Objective

The primary aim of our study is to elucidate the phenomenon of "Ontological Racketeering within the Manosphere" and determine how thought leaders exploit ontological security theory to further their agendas. Drawing upon the presented flowchart and supporting literature, we intend to dissect the strategies employed by these thought leaders in fostering a specific environment of ontological insecurity within their target audience, subsequently capitalizing on this insecurity for material gain.

The flowchart, labeled as "Ontological racketeering within the manosphere," presents a cyclic system that emphasizes four key stages:

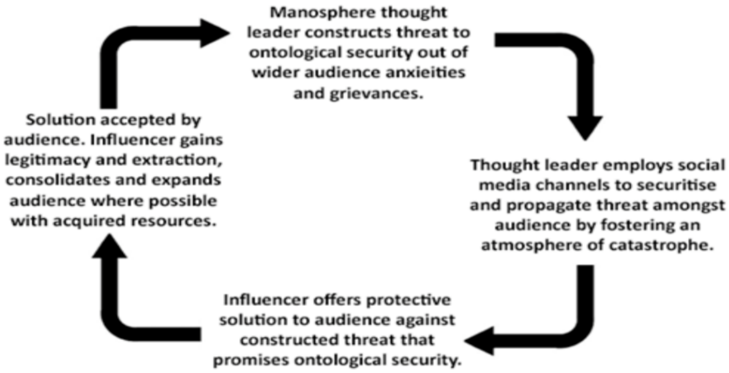

Manosphere thought leader constructs threat to ontological security out of wider audience anxieities and grievances.
Thought leader employs social media channels to securitise and propagate threat amongst audience by fostering an atmosphere of catastrophe.
Influencer offers protective solution to audience against constructed threat that promises ontological security.
Solution accepted by audience. Influencer gains legitimacy and extraction, consolidates and expands audience where possible with acquired resources.

Figure 4: Ontological racketeering within the manosphere

1. **Constructing a Threat:** The thought leaders of the Manosphere build an atmosphere of ontological insecurity by emphasizing existential threats to masculinity. By tapping into the wider audience's anxieties and grievances, they create an atmosphere where masculinity is portrayed as being under siege.

2. **Securitization Through Social Media:** Using their influential positions on platforms like YouTube, these leaders amplify and propagate the constructed threats, often exaggerating societal changes and portraying them as dire threats to the male identity. The use of provocative titles and content, as mentioned in the text, further serves to inflame these insecurities.

3. **Offering Protective Solutions:** Once the threat is established and propagated, these leaders present themselves as saviors by offering 'definitive' solutions to the 'problems' they have highlighted. These solutions, framed as essential actions, are marketed as the only way for their audience to reclaim and secure their threatened masculinity.

4. **Resource Extraction:** The final stage involves capitalizing on the established insecurity. The solutions, while presented as altruistic, often come with strings attached – be it in the form of purchasing online courses, books, or other merchandise. Thus, the thought leader benefits materially from the very insecurity they fostered.

Connecting these stages back to our research questions established in the Background, we intend to investigate:

a) Can we delineate the broader ethos of Tate's followers?
b) Do they mirror his beliefs, challenge them, or do they have a spectrum of perspectives of their own?

We aim to offer a nuanced understanding of the dynamics at play within influential nodes of the Manosphere and shed light on the mechanisms that drive their growth and impact in the digital age.

Data Gathering

Reasons for our Data Gathering Approach

In the ever-evolving world of the internet, understanding public sentiment is crucial. Social media platforms, with their vast repositories of user-generated content, have emerged as valuable tools for sentiment analysis. Of these platforms, YouTube and Reddit have consistently proven their worth as sources of diverse, global opinions, emotions, and insights. Here is a detailed exploration of our reasons for focusing on YouTube comments and subreddits for data gathering:

1. **Diverse User Demographics:** YouTube and Reddit encompass a global audience, catering to varied demographics — from teenagers to the elderly, from hobbyists to professionals. Such diversity offers a multifaceted viewpoint, enriching the data pool and facilitating comprehensive sentiment analysis.

2. **High Volume of Data:** The sheer user count — over 2 billion monthly users on YouTube and 430 million on Reddit — translates to an immense volume of comments, providing a vast dataset crucial for nuanced analysis.

3. **Topical Concentration:** Subreddits function as micro-communities, each revolving around niche topics. Gleaning data from these communities provides insights laser-focused on specific subjects. Meanwhile, YouTube comments, often sparked by the video content, offer direct, raw sentiments.

4. **Temporal Insights:** Comments on these platforms are not static; they morph over time, mirroring the evolution of public sentiment. This dynamism grants a view of the ebb and flow of public opinion over periods, pivotal for gauging reactions to events or trends.

5. **Rich Contextual Data:** Reddit's threaded discussions and YouTube's reply feature encourage dialogues, adding layers to individual comments. This depth is indispensable for sentiment analysis, preventing skewed interpretations based on decontextualized comments.

6. **Reputable Precedents:** Several studies have underscored the efficacy of these platforms for sentiment analysis. A report by Pew Research Center notes Reddit's popularity among young adults, making it a goldmine for this demographic's sentiment. A study in the Journal of Computer-Mediated Communication emphasized YouTube comments' potential in gauging public opinions, particularly on contentious subjects. Further, research in "PLOS ONE" highlighted the utility of YouTube comments in discerning public perceptions about climate change. Meanwhile, a paper in the "Journal of Medical Internet Research" discussed leveraging Reddit to grasp community dynamics and

opinions on health topics. (4)

7. **Ethical Considerations:** Ethical data gathering is paramount. Both YouTube and Reddit have strict community and privacy guidelines. By extracting publicly accessible comments and honoring user anonymity, we ensure our approach is ethically sound and conforms to digital research norms. (4)

8. **Flexibility of Analysis:** The textual data from YouTube and Reddit is intricate and often paired with metadata, like upvotes or timestamps. Such rich data provides leeway in adopting diverse analysis methods — from qualitative content breakdowns to quantitative sentiment metrics. (4)

9. **Direct Access to User Feedback:** Traditional data collection methods, like polls, can be marred by biases. Direct access to spontaneous user comments guarantees a candid reflection of public sentiment, devoid of such biases.

10. **Historical Reference and Future Predictions:** Both platforms house historical data, allowing researchers to track sentiment trends over time. This not only offers a retrospective view but can also aid in forecasting future sentiment trajectories. (4)

Our data-gathering methodology, anchored in YouTube comments and subreddits, is driven by our goal of a coherent and comprehensive social media sentiment analysis. It harnesses the inherent strengths of these platforms in terms of volume, diversity, and specificity. This strategy, backed by credible studies, ensures a thorough understanding of public sentiment, aligning seamlessly with our study's objectives.

Sources of Our Data Collection

In the vast landscape of the internet, myriad sources offer data ripe for sentiment analysis. However, our focus narrowed down to two primary platforms: YouTube and Reddit. (5) This decision was not arbitrary but rooted in a strategic understanding of the nature and quality of data these platforms offer. Below is an in-depth look at the kind of data we collected from these platforms and the reasons for their predominance in our research methodology.

YouTube Comments: A Mirror to Global Sentiments

YouTube, as the world's largest video-sharing platform, plays host to a medley of content – from educational tutorials to entertainment snippets. Each video, regardless of its nature, invites feedback in the form of comments, reactions, and shares. We specifically honed in on the comments section, a dynamic space where users not only respond to the content but also engage in layered discussions with other viewers. Our dataset from YouTube comprised textual comments, reactions (likes, dislikes), and any associated metadata like timestamps and usernames (though ensuring anonymity).

The rationale behind leveraging YouTube comments lies in their spontaneity. They are instant reactions, often mirroring visceral sentiments of viewers. Moreover, given the global reach of YouTube, these comments represent a diverse demographic. From an urban teenager in Tokyo to a senior citizen in Toronto, the spectrum of voices is vast and varied. This global representation enriches our dataset, ensuring that our sentiment analysis is not insular but broad-based and inclusive.

Reddit: A Treasure Trove of Niche Insights

Reddit, often dubbed the "front page of the internet," is a hub of community-driven discussions. It is structured around 'subreddits' – individual communities focused on specific topics. Our data collection on Reddit was two-pronged. Firstly, we sourced textual comments from various threads, capturing the essence of discussions. Secondly, we incorporated metadata like upvotes, downvotes, and user flair, which provides context to the sentiments expressed.

Reddit's strength lies in its topical concentration. Each subreddit, with its distinct user base, offers niche insights, making our data collection targeted and precise. For instance, if our sentiment analysis aimed to understand public opinions on a newly released gadget, mining data from a technology-focused subreddit would yield concentrated insights. Furthermore, the upvote and downvote system on Reddit offers a quantitative measure, allowing us to gauge the popularity or acceptability of a sentiment within the community.

Why YouTube and Reddit Predominantly?

With numerous social media platforms available, one might wonder about our emphasis on YouTube and Reddit. The decision is rooted in the unique nature of data these platforms offer. While platforms like Instagram or Facebook are skewed towards visual content and personal updates, YouTube and Reddit prioritize discussions. Comments on YouTube are reactions to a shared visual stimulus, the video, making them a rich source of public opinion. On the other hand, Reddit's textual discussions, driven by community dynamics, offer deep dives into specific topics. (5)

Another significant factor is the sheer volume of data. Both platforms boast of massive active user bases, translating to an abundant and continuous stream of data. This volume ensures that our dataset is expansive, providing a holistic view of public sentiment. (4)

Quality Over Quantity

While volume is crucial, the quality of data is paramount. Both YouTube and Reddit, owing to their nature, ensure that the data is organic and genuine. The anonymity that Reddit offers encourages candid discussions, often leading to raw, unfiltered sentiments. YouTube comments, being public, might have a degree of posturing, but the sheer volume and diversity counterbalance that, providing a genuine reflection of public sentiment. (4)

Ethical Considerations in Source Selection

Our emphasis on YouTube and Reddit also stems from ethical considerations. Both platforms have robust community guidelines and user privacy norms. Our data gathering techniques respect these guidelines, ensuring that the data collection is transparent, ethical, and non-intrusive. By focusing solely on publicly available comments and discussions, we maintain a balance between data richness and user privacy. Anonymizing any identifiable information safeguards individual users while allowing us to tap into the collective sentiment of communities.

Reddit Data Retrieval, Pre-Processing and Bag-of-Words:

1. **RedditProcessing.py:** We used this Python script from our lab which is designed to preprocess Reddit posts. It involves the following steps:
    a. **Initialization:**

    We initialized the class RedditProcessing with three parameters: tokeniser, stemmer, and lStopwords. These will be used throughout the processing.

    - **tokeniser:** This was an object responsible for breaking down the text into individual tokens (usually words).
    - **stemmer:** This was an object to convert words into their base or root form (e.g., "running" becomes "run").
    - **lStopwords:** This was a list of words that will be excluded during processing as they do not add much semantic meaning to the text.

    b. **Text-Processing:**

    - The process function processed a given piece of text (text).
    - The text was first converted to lowercase using lower().
    - The text was then tokenized (split into individual words or tokens).
    - Whitespace was then stripped from each token.
    - Tokens that matched certain patterns were removed:
    - Strings of digits or fractions.
    - Any token starting with "http".

- Stopwords were also removed.

2. **BoW.ipynb:** This Jupyter notebook focused on the preprocessing of Reddit data to produce a Bag of Words (BoW) representation.
   a. **Setup and Dependencies:**
      i. Dependencies such as pandas, nltk, and other utility libraries were imported.
      ii. RedditProcessing was imported to process the Reddit posts.
      iii. Tokenization was set up using the TweetTokenizer from the nltk library.
      iv. A list of stopwords was defined, which includes standard English stopwords and some additional ones.
      v. The lemmatizer was initialized. Unlike stemmers, lemmatizers convert words to their base form based on the actual meaning in the dictionary (word + part of speech & tense).

   b. **Data Import:** Data was read from a CSV file named, "Tate_all.csv" into a dataframe.

   c. **Data Preprocessing:**
      i. The function preprocessing was defined to process the content of the Reddit data:
         1. Each Content from the dataframe was processed using the RedditProcessing class, converting them into a list of tokens.
         2. The processed tokens were then joined back together as a string and appended to the lPosts list.
         3. The lPosts list (which contains the processed content) was saved as a pickle file for later use.
      ii. The function tokenZtext was defined to tokenize, normalize (convert to lowercase and stem), and remove digits from the text. This function is a simplified version of the preprocessing process, primarily aimed at tokenizing text and producing stemmed tokens.

Summary

We designed a systematic approach to preprocess Reddit data, which involves tokenization, stemming, and stopword removal. For this, the data was imported from a CSV file. The data was then processed using the RedditProcessing class from the RedditProcessing.py script. This class tokenized the data, removed stopwords, and removed tokens that matched certain patterns (like digits or URLs). In the Jupyter notebook, we'd also set up dependencies and additional preprocessing utilities, including the WordNetLemmatizer for lemmatization. The main preprocessing happened in the preprocessing function, where each piece of content from the dataframe was processed to produce a list of tokens. This processed content was saved for later use. Additionally, we set up a function (tokenZtext) to tokenize and stem the Reddit data while removing any numeric tokens.

Youtube Data Retrieval from API

1. **Imports & Configuration:**
    a. Libraries such as requests, time, pandas, networkx were imported for data retrieval, processing, and storage.
    b. We also had configurations such as SEARCH_URL, COMMENTS_URL and API_KEY which pertain to the YouTube Data API V3.

2. **Utility Functions:**
    a. store_data and load_data: Used to save and load data respectively in JSON format.
    b. make_request: A robust function to handle YouTube API requests. It includes error handling for quota limits, where it will sleep for a day before retrying if it encounters a quota error.

3. **Video Retrieval:**
    a. First, we tried to load previously stored video data. If not available, it fetched data from the YouTube API.
    b. The search query being used is 'Andrew Tate' with a type filter set to 'video'.
    c. The search results are paginated, so the loop continued to fetch results as long as there's a nextPageToken.
    d. The data was continuously updated in a json file.
    e. Once all the results were obtained, they were saved in videos.json.

4. **Comment and Reply Retrieval:**
    a. For each video obtained, its comments and replies were fetched.
    b. Within a loop, we extracted key data from the comments and replies such as IDs, content, like counts, etc., and appended them to df_data.
    c. This data was then periodically stored in df_data.json.

5. **Dataframe Creation & Graph Building:**
    a. The collected data was converted into a Pandas DataFrame.
    b. Using NetworkX, we built a directed graph (G) where nodes were comments or replies, and the edges were determined by the relationship between the parent comment and their replies. Each edge's weight was set to the upvotes of the comment or reply.

   c. This dataframe was then saved as TATE_youtube.csv, and the graph was saved as Tate_youtube.graphml.

6. **Additional Data Analysis:**
   a. Comments were then grouped by their ParentID (i.e., the video they belong to).
   b. Comments without replies and those with replies were segregated into video_no_reply and video_reply respectively.
   c. For a given video_id, we then fetched the top-level comments and any related replies.

In summary, the Jupyter notebook implementation of our Youtube API scraper had a structured approach to extracting video, comment, and reply data related to a search term ('Andrew Tate'). The data was then processed and stored in a structured format, and relationships between comments and replies were visualized as a directed graph. This data was then utilized further for sentiment analysis, or any other analysis as needed.

Sentiment Analysis

Text Preprocessing

The function tokenZtext is used for the preprocessing of the Reddit submissions and comments to create a bag of words. It performs:

- **Text normalization**: By converting all characters to lowercase.
- **Tokenization**: The process of breaking down the text into individual words or tokens.
- **Stemming**: Reducing words to their base or root form to account for variations in word endings.
- **Filtering**: Removing digits and numbers.

Topic Modeling using LDA (Latent Dirichlet Allocation)

There are two approaches to Topic Modeling used in the code:

1. **Gensim LDA:**
   a. The function GensimPreprocess further preprocesses the text to prepare it for Gensim's LDA modeling. This includes lemmatization (converting words to their base or dictionary form) and removing custom stop words.

   b. GensimLDA then uses Gensim's LdaMulticore to perform LDA topic modeling and prints out the topics discovered.

c. The functions get_dominant_topic and describe_topic are utility functions to extract the dominant topic from a given document and describe a specific topic, respectively.

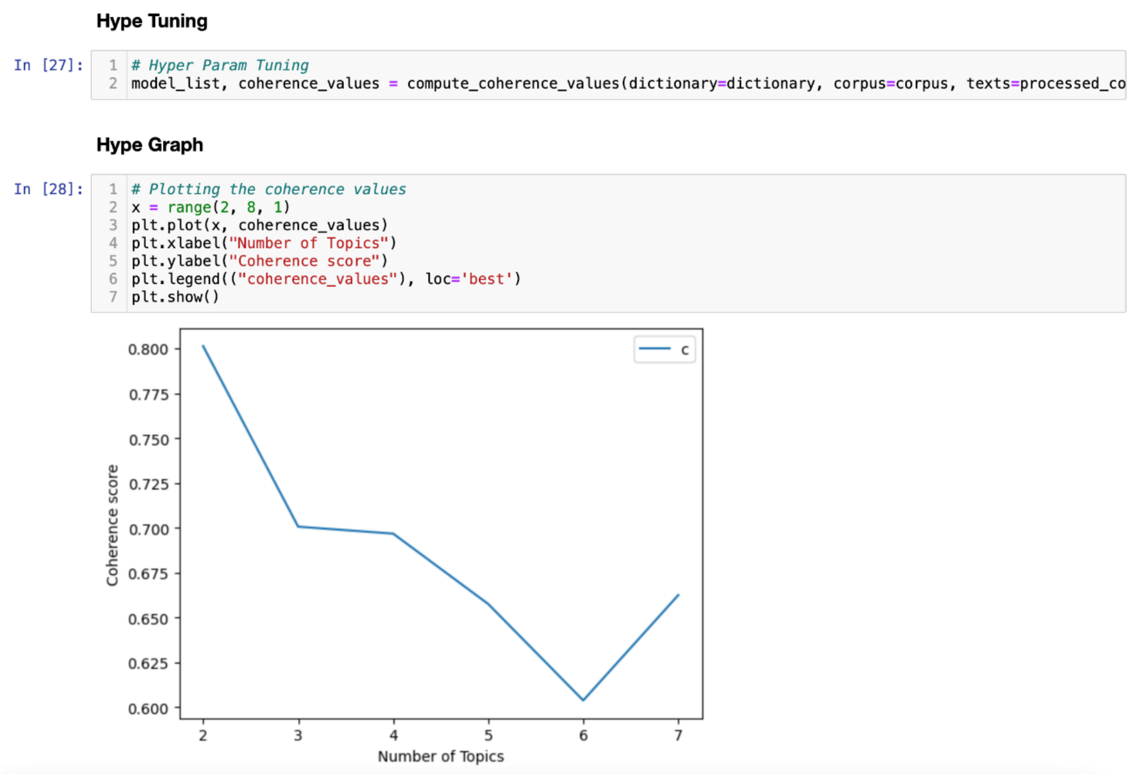
d. compute_coherence_values is used to determine the optimal number of topics by calculating the coherence values for different numbers of topics. A higher coherence value indicates a more interpretable model.

Overall, Topic modeling is a technique in text mining that seeks to discover abstract "topics" in a collection of documents. One of the most popular methods for topic modeling is Latent Dirichlet Allocation (LDA). This method assumes that there are K topics shared across a collection of documents. Each document is a mix of topics, and a topic is a mix of words. The task is to figure out what topics would make the documents look most like the observed documents. (6)

To tune hyperparameters for LDA, a range of topics is tested to determine the optimal number that gives the best coherence score. The coherence score is a measure of how well the words within a topic cohere together, with higher values indicating more coherent topics.

Hyperparameter Tuning Process

The given process begins with computing coherence values over a range of potential topic numbers, specifically between 2 and 7 inclusive. The resultant coherence values are then plotted against the number of topics, using the Matplotlib library:



Hype Tuning

In [27]:

```
# Hyper Param Tuning
model_list, coherence_values = compute_coherence_values(dictionary=dictionary, corpus=corpus, texts=processed_co
```

Hype Graph

In [28]:

```
# Plotting the coherence values
x = range(2, 8, 1)
plt.plot(x, coherence_values)
plt.xlabel("Number of Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

Figure 5: Hyper Parameter Tuning for LDA Topics

From the attached coherence values graph, the coherence scores start high at two topics, decrease steadily, and then increase sharply after five topics. The number of topics corresponding to the highest coherence score is often selected as the optimal number. Although two topics have a high score, three topics were clearly identified and four started blurring the topics together.

The selected LDA model reveals the top 10 words for each of the three topics:

- **Topic 1:** traffic, charge, evidence, romania, arrest, sex, human, people, time, pizza
- **Topic 2**: tate, fuck, andrew, shit, guy, dude, watch, think, lmao, bro
- **Topic 3**: people, think, tate, women, men, want, good, guy, man, way

We will visualize and discuss the implications of these topics after the Gephi Topical modeling where we visualized a meta-network to evaluate the extent to which the above three most popular topics are aggregated.

Data-Preprocessing for Sentiment Frequency Visualization:

1. **Tokenization:**

Here, we used the TweetTokenizer from the Natural Language Toolkit (NLTK) library. This tokenizer is specialized for tokenizing tweets, making it adept at handling emoticons, hashtags, and mentions.

2. **Defining Punctuations and Stopwords:**
    a. lPunct is a list of standard punctuations.
    b. lStopwords is a combination of standard English stopwords, the above-mentioned punctuations, and additional terms like 'lol', 'hi', emoticons, and some common abbreviations or shorthand.

3. **Stemming:**

We used the Porter stemmer for stemming. Stemming trims words to their root form. For instance, "running" becomes "run."

4. **Reddit Processing Initialization:**

We have a custom Python script RedditProcessing (explained earlier) processes Reddit posts which we have used here. It's initialized with the tokenizer, stemmer, and list of stopwords.

5. **Loading Positive and Negative Words:**
    a. Two empty lists, lPosWords and lNegWords, are created.
    b. Positive and negative word files (paths not provided in the code) are then read. These files likely contain words that have positive or negative sentiments.
    c. The words from these files are added to the respective lists, which are then converted to sets for faster lookup.

6. **Processing Reddit Content:**
    a. The main loop processes each content entry in red_data['Content']:
        i. It checks if the content is a string; if not, it converts it to a string.
        ii. The content is then processed by the redditProcessor which likely tokenizes, removes stopwords, and stems the content.
        iii. The tokens are then updated in a termFreqCounter to keep track of word frequencies.
        iv. Processed tokens are joined together and stored in the lPosts list.

7. **Visualization of Word Frequency:** The provided visual plot shows the "Top 30 Term frequency distribution". It's a bar chart that represents the frequency of the top 30 terms in the processed Reddit dataset. The terms are on the x-axis, while their respective frequencies are on the y-axis. Terms like "scandal", "care", "case", and "someone" appear most frequently. The x-axis labels are rotated at a 45-degree angle for better readability.

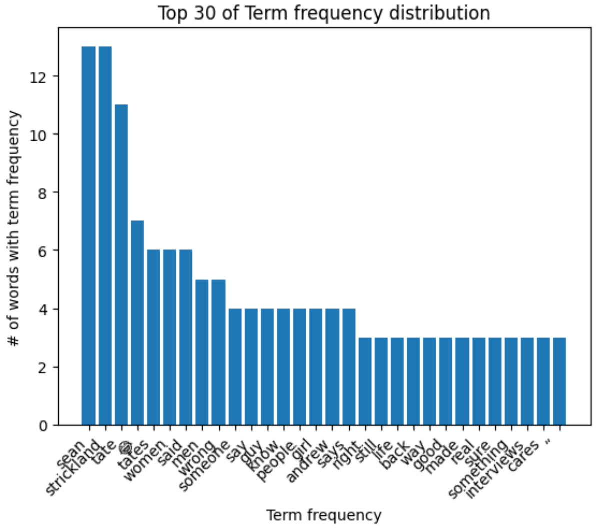


Figure 6: Top 30 Term Frequency Distribution

Discussing the Plot

The Frequency plot showcases the term frequency distribution for the top 30 terms from our processed Reddit data. Here's a deeper understanding:

a. **Most Frequent Terms**: The terms on the far left of the x-axis, like "scandal", "care", and "case", have the highest frequencies in the dataset. This implies that these terms are discussed or mentioned more often than others in the Reddit data we've processed.

b. **Decreasing Frequency**: As we move to the right on the x-axis, the term frequency gradually decreases. This means that terms like "some", "problem", and "politics" are mentioned less often than the ones on the far left.

c. **Interpretation**: The choice of the top terms can give an overview of what the dataset predominantly talks about. For example, if "scandal" and "politics" are among the top terms, it might hint that the dataset contains discussions related to political scandals or controversies.

d. **Visualization Details**:
   i. The x-axis labels (terms) are rotated 45 degrees to avoid overlap and ensure readability.
   ii. The y-axis represents the number of times each term appears in the dataset.
   iii. The title clearly states the nature of the data visualization, ensuring viewers quickly understand what they're observing.

In summary, the visual plot provides a quick way to understand which topics or terms are most prevalent in our Reddit dataset.

Time-Based Sentiment Analysis

1. **Counting Sentiments:**

   We first initialized an empty list named cSentiment. Later, we populated this list using the function countWordSentimentAnalysis that calculates sentiment scores for each data point (or text) in data_y[1]. The sentiment score is likely calculated based on the number of positive words (setPosWords) and negative words (setNegWords) present in each text.

2. **Printing Date:**

   Printing the 'Date' series from data_y[0]. The output shows datetime values starting from 2023-10-04 13:41:28 and ending at 2023-10-04 00:56:53 for a total of 4716 records.

3. **Converting the Sentiment to a Time-series Dataframe:**

   We created a pandas DataFrame series with two columns: 'date' and 'sentiment' using the data in cSentiment. Then, we set the 'date' column as the index and ensure that the 'sentiment' values are numeric.

4. **Resampling and Visualizing the Time Series Data:**

   We resampled the data to an hourly frequency, aggregating by sum. Then, we plotted the resampled data which results in the time-series graph below.

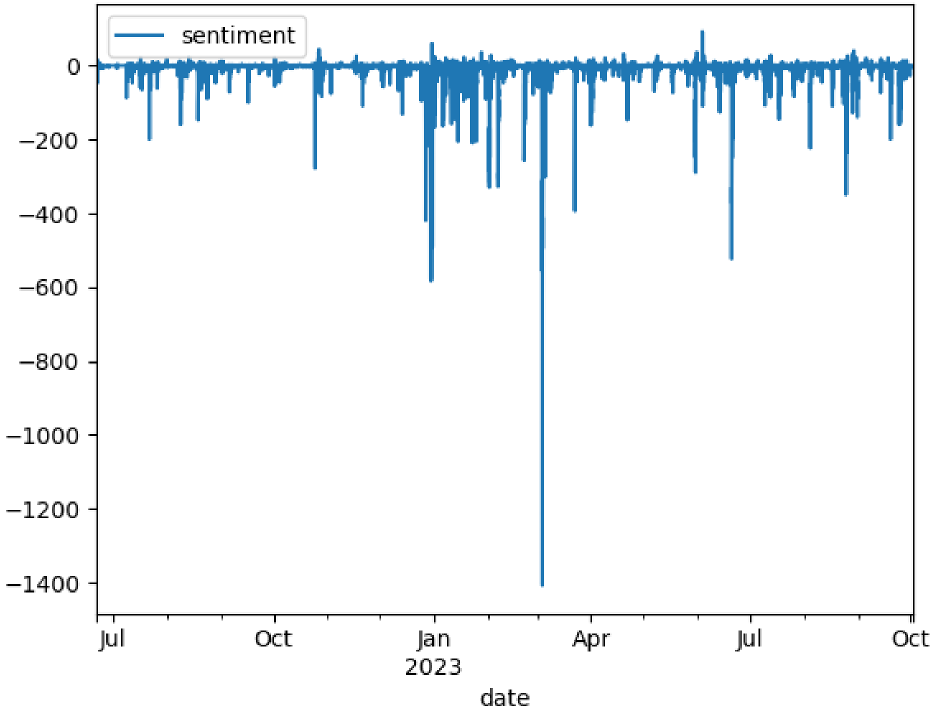


Figure 7: Sentiment of Andrew Tate from July to October 2023

The graph visually represents sentiment scores over time. Each point on the y-axis corresponds to the sum of sentiment scores in each hour. Positive values indicate positive sentiment, negative values indicate negative sentiment, and values around zero indicate neutral sentiment.

The above code and visualization provide insight into how sentiment varies over time. The graph displays sentiment fluctuation across different hours of the day on October 4, 2023. The spikes and troughs allow us to infer periods of positive and negative sentiment.

Sentiment for Andrew Tate has been overall negative across both social media platforms. This cause could be people who passionately dislike him are more likely to comment versus people who enjoy his content. The large amount of negativity in March was around the time he was arrested and charged with sex trafficking in Romania. Each spike after correlates with updates on his jail sentencing and information coming out from the victims.

Topic Graph Modeling using Gephi

Gephi is a powerful network visualization tool, which allows researchers to delve into these complexities, offering visual insights that can significantly augment traditional analytical methodologies. Constructing such a grand-scale representation necessitates substantial computational power, especially when handling big data with intricate linkages. After extracting data from Reddit and YouTube, it's processed using the Networkx library, which assists in creating the graph model. (6) Every comment, based on the LDA topic model, is classified, providing each node with an attribute that aids in the forthcoming visualization process.

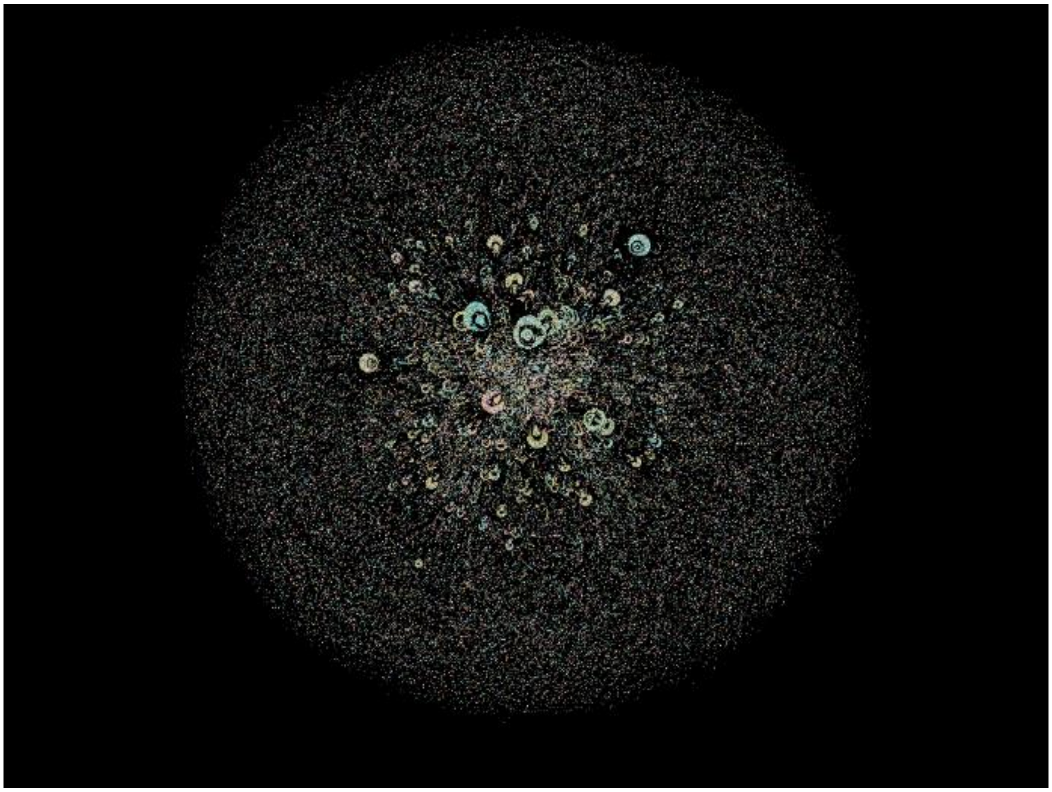


Figure 8: Graph ML Visualization with Force Atlas 2

The above image is a panoramic representation of the sentiment analysis outcomes from both Reddit comments. With a staggering 216,274 nodes and 158,124 edges, it showcases the interplay of the three identified topics:

- **Topic 1:** traffic, charge, evidence, romania, arrest, sex, human, people, time, pizza
- **Topic 2**: tate, fuck, andrew, shit, guy, dude, watch, think, lmao, bro
- **Topic 3**: people, think, tate, women, men, want, good, guy, man, way

This bird's-eye view serves as a macroscopic insight into the overarching sentimental trends and conversational threads revolving around Andrew Tate's online influence. It provides a preliminary understanding of which topics dominate the discourse and how they interconnect. Given that we have been able to generate the above model, we can zoom in on individual clusters based on the topic we want to explore, in this case, let's choose Topic 1 and find a cluster that is discussing Topic 1.

By applying selective filters and using Gephi's node partitioning functionalities, the initial mammoth dataset is slimmed down to 2,174 nodes and 1,493 edges for computational ease. This concentrated lens offers a clearer perspective on the dynamics surrounding a specific instance of Topic 1 being discussed among commenters. The interactions, relationships, and the overall influence trajectory related to Andrew Tate's arrest in Romania become discernible, shedding light on how sentiment oscillates in this context.

Zooming further into the discourse, the image on the left is a crystallized, singular instance of Topic 1 discussion amongst commenters, to be more specific, it is the dot visible in the previous slimmed down dataset. This directed visualization elucidates the gravitational pull of Topic 1.

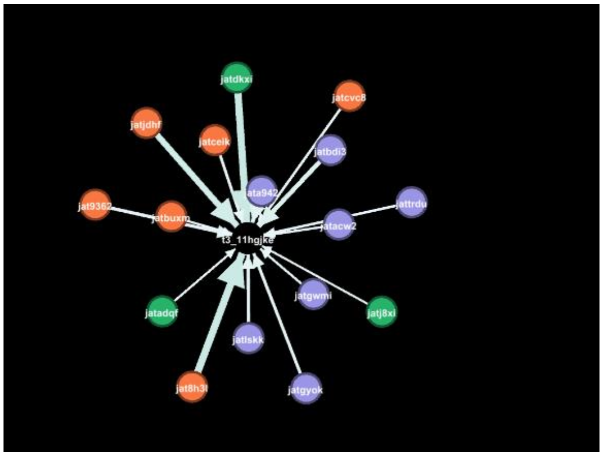
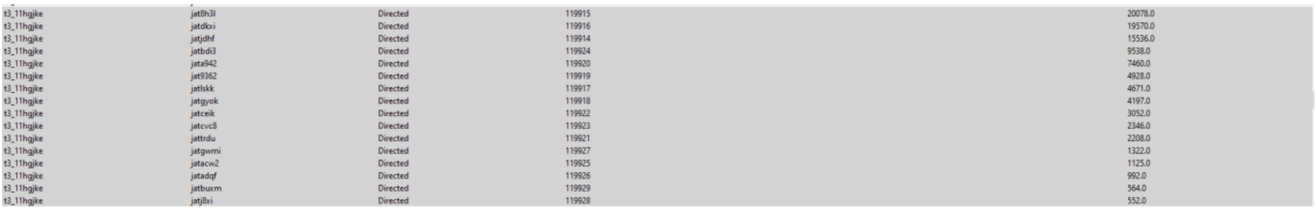


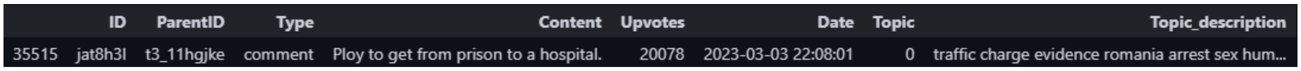
Figure 9: Closer look at Submission and Commenters' Topic

It's evident how conversations steer towards this specific issue, revealing the depth of sentiment and the underlying influencers propelling this topic's popularity. We can also see this pull exemplified when we see the tabulated representation of the graph on the left below.



t3_11hgjke	jat8h3l	Directed	119915	20078.0
t3_11hgjke	jatdkxi	Directed	119916	19570.0
t3_11hgjke	jatjdhf	Directed	119914	15536.0
t3_11hgjke	jatbdi3	Directed	119924	9538.0
t3_11hgjke	jata942	Directed	119920	7460.0
t3_11hgjke	jat9362	Directed	119919	4928.0
t3_11hgjke	jatlskk	Directed	119917	4671.0
t3_11hgjke	jatgyok	Directed	119918	4197.0
t3_11hgjke	jatceik	Directed	119922	3052.0
t3_11hgjke	jatcvc8	Directed	119923	2346.0
t3_11hgjke	jattrdu	Directed	119921	2208.0
t3_11hgjke	jatgwmi	Directed	119927	1322.0
t3_11hgjke	jatacw2	Directed	119925	1125.0
t3_11hgjke	jatadqf	Directed	119926	992.0
t3_11hgjke	jatbuxm	Directed	119929	564.0
t3_11hgjke	jatj8xi	Directed	119928	552.0

Figure 10: List of All Commenters from Submission



	ID	ParentID	Type	Content	Upvotes	Date	Topic	Topic_description
35515	jat8h3l	t3_11hgjke	comment	Ploy to get from prison to a hospital.	20078	2023-03-03 22:08:01	0	traffic charge evidence romania arrest sex hum...

Figure 11: Pinpointing Comment Being Replied To

As we can see, the centrality and popularity of Topic 1 is exemplified in the above comment referencing Andrew Tate's recently spread lie about him being diagnosed with Lung Cancer as a ploy to escape imprisonment. (3) The structured tabulation of the above users categorically presents each commenter's input, their corresponding upvotes, and the topic of discussion.

While the earlier images offer a visual treatise of sentiment dynamics, this table offers quantitative insights. By gauging upvotes, it provides a measure of consensus or popularity of a particular sentiment or opinion within the community. It aids in discerning which opinions resonate with the larger audience and which ones remain on the fringes.

Sentiment Model with Gephi

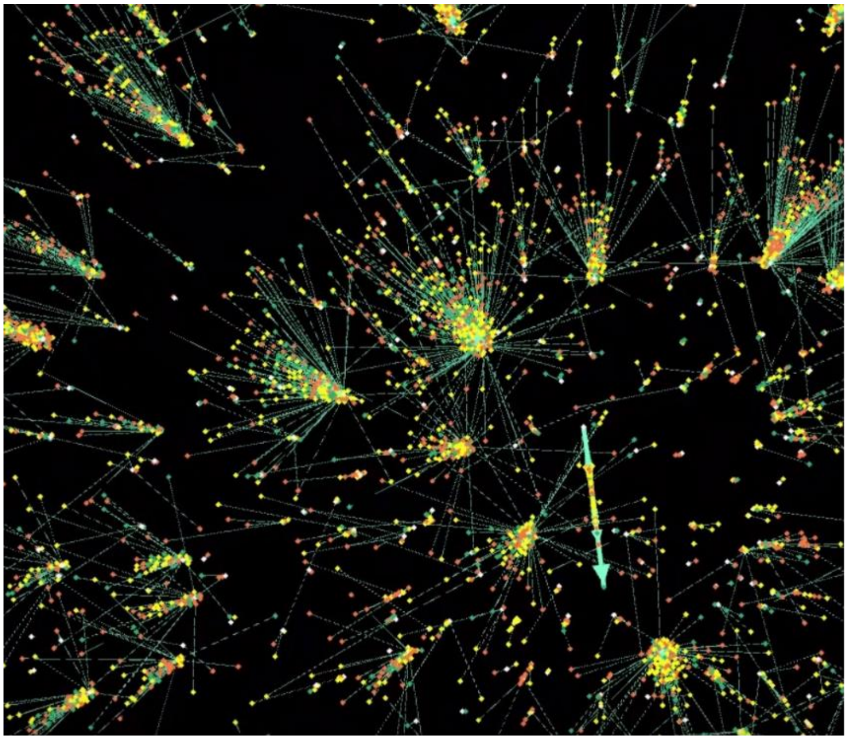


Figure 12: Graph Model Displaying Edges

Segmenting individuals by their feelings towards the subjects involving Andrew Tate reveals the diversity of opinions about him. While the predominant sentiment towards Andrew Tate is negative, a visual analysis paints a more nuanced picture. Notably, a significant number of positive remarks accompany Andrew Tate's Reddit posts, shedding light on the growth of his expansive online community. Graphics that depict users effectively simplify the complexity of these large numbers.

Community Detection

Having a modularity of 0.848 means the network is well-partitioned into distinct communities. The model has 26561 communities. Most of the communities are relatively small in size, with a significant number of them containing fewer than 40 nodes. There are a few outliers or larger communities, with some having over 200 nodes. The distribution appears to be somewhat random, suggesting there isn't a strong correlation between modularity class and community size in this specific dataset

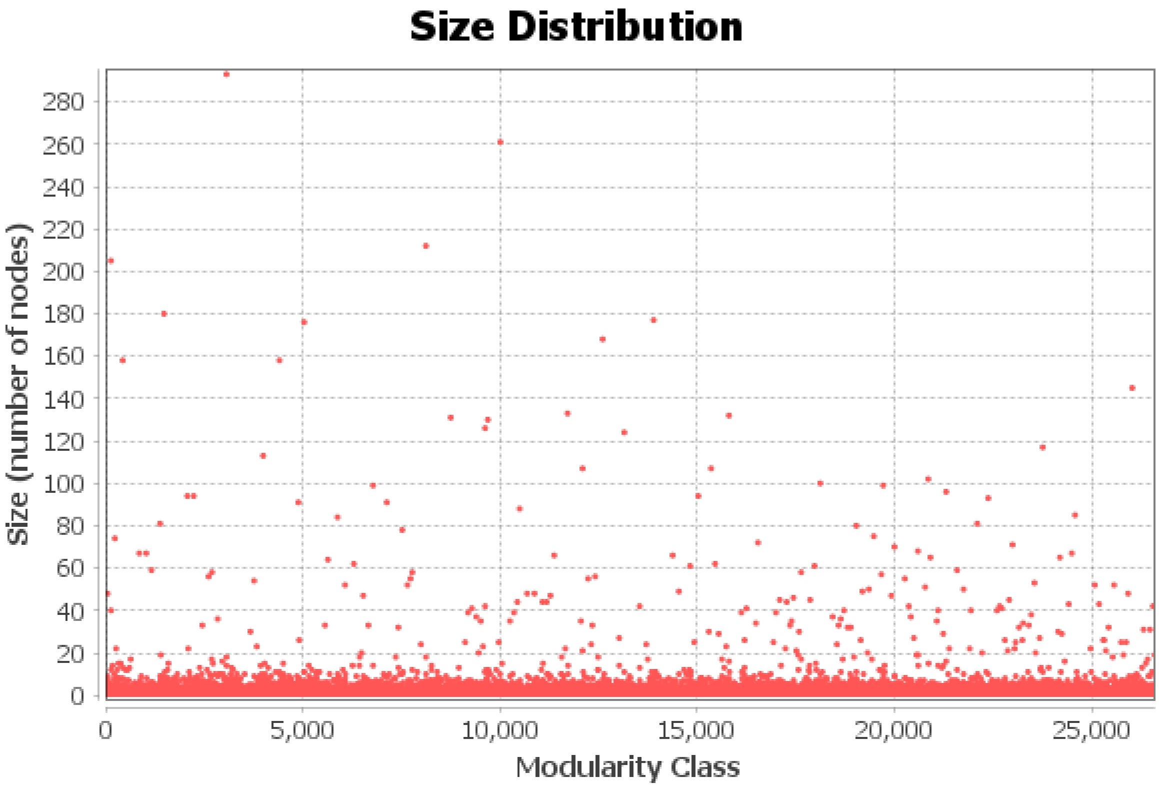


Figure 13: Graph Model Displaying Modularity

Computational Power and Requirements of Gephi

Handling vast datasets, especially graph data with intricate linkages, is computationally intensive. The initial dataset, comprising 216,274 nodes and 158,124 edges, demands significant memory and processing power. Gephi, while optimized for network visualization, can strain standard computer resources when rendering such extensive graphs. Let's delve into the computational aspects:

1. **Memory Management:**

   Large graphs demand robust memory handling. Rendering a graph of this magnitude would ideally require a computer with substantial RAM, preferably 32GB or higher, to ensure smooth visualization and interaction. (6)

2. **Processing Power:**

   The CPU's efficiency determines the speed at which nodes and edges are processed. A multi-core processor, such as an octa-core, aids in efficient multitasking, especially when partitioning nodes, filtering data, or applying layout algorithms.

3. **Graph Optimization:**

Before visualization, data preprocessing becomes imperative. This includes removing redundant nodes, converting negative weights to positive ones, and filtering out nodes with minimal interactions (like those with upvotes less than 500). These steps, while reducing computational strain, also enhance the clarity of the visual representation.

4. **Graphics Handling:**

For real-time rendering and interaction, a dedicated graphics card is beneficial. It ensures the visualizations are rendered seamlessly, especially when zooming in or out and interacting with specific nodes.

Filtered Graph vs Filtered	
Before	**After**

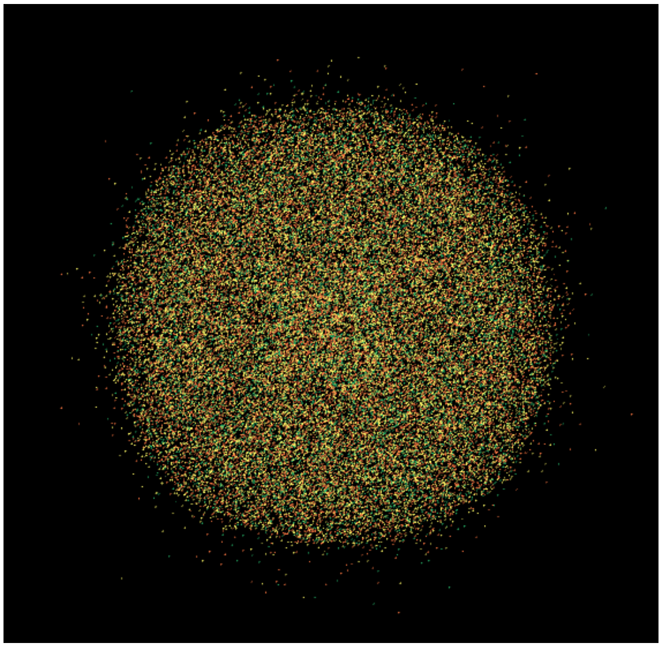
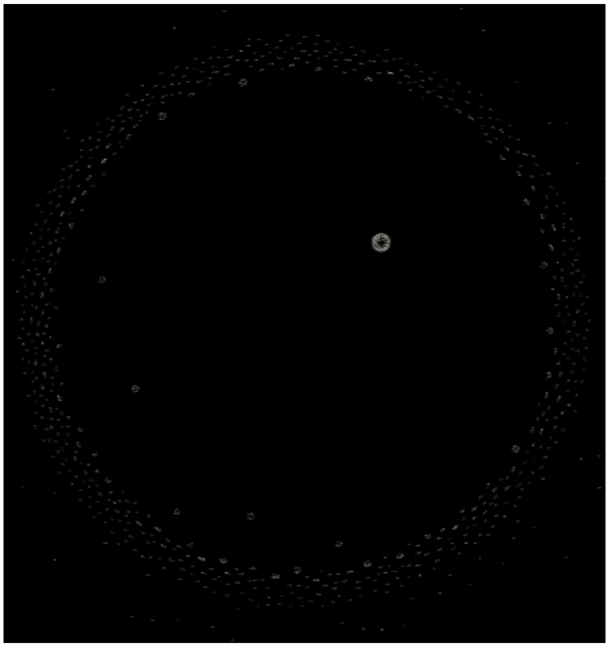

	

Figure 13: Un-filtered Graph	Figure 14: Filtered Graph

Purpose of Gephi Representation

1. **Holistic Understanding**

The vastness of digital conversations is distilled into comprehensible visuals. Researchers can grasp the magnitude of discourse and discern patterns that might remain obscured in tabulated data.

2. **Pinpointing Influencers**

   By visualizing node interconnectivity and gauging upvote metrics, influencers within the conversation become discernible. This aids in understanding how sentiments spread and which nodes (or participants) act as catalysts.

3. **Sentiment Oscillations**

   Observing how conversations steer between the three topics provides insights into sentiment dynamics. This oscillation can reveal triggers, influencers, or external events impacting the sentiment.

4. **Depth of Analysis**

   The multi-tiered visualization approach, transitioning from the macroscopic to the microscopic, ensures that while the broader patterns are understood, the granular details aren't lost.

In conclusion, the Gephi representation, backed by rigorous computational methodologies, is a potent tool in the arsenal of digital sentiment analysis. It not only visualizes the vastness of digital interactions but also offers tangible metrics, ensuring that the research is both comprehensive and precise. In the context of understanding Andrew Tate's online influence and the tactics of "Ontological Racketeering within the Manosphere," such a methodical approach becomes indispensable. It ensures that the strategies of thought leaders, their influence trajectory, and their impact on the digital community are comprehensively deciphered.

The three topics inferred from the dataset suggest:

- **Topic 1:** relates to legal issues, involving charges, evidence, and arrests, with specific mentions of trafficking and Romania. <- This was when Andrew Tate was arrested in Romania. People thought the police found out he was in the country because he doxed himself with a pizza box. The video was filmed on reaction, to attack climate activist Greta Thunberg. (2)
- **Topic 2:** Appears to be more casual and colloquial, relating to commentary.
- **Topic 3:** Seems to revolve around perceptions and opinions on Tate being a good guy for women.

To help understand these topics visually, word clouds were generated:

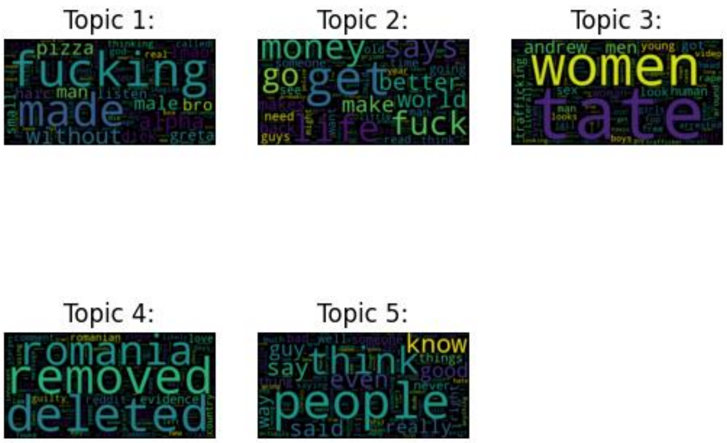


Figure 8: Word Clouds of Topics

Word Clouds offer a visual representation of text data. The prominence of each word in the cloud signifies its frequency or importance in the dataset. Larger and bolder words are mentioned more often or bear more weight in the topics.

Implications and Conclusions

1. **Topic Specificity:**

   Some topics are more specific than others. For instance, Topic 1 in the three-topic model and Topic 2 in the five-topic model both seem to revolve around similar themes, suggesting consistency in the model's ability to capture a specific set of discussions, centered around legal issues and controversies.

2. **Casual vs. Formal Discussions:**

   The dataset contains a mix of formal and informal discussions. While some topics highlight casual banter and colloquial language, others are more serious, focusing on societal issues or controversies.

3. **Relevance of Coherence Score:**

   The coherence score plot guided the selection of the number of topics. It is essential to note that while coherence scores provide a quantitative way to select the number of topics, qualitative judgment and domain knowledge can further refine this choice.

4. **Dataset's Broad Nature:**

   The range of topics identified underscores the dataset's diverse nature. From casual conversations to discussions on gender roles and legal issues, the dataset spans a wide array of themes.

5. **Role of Hyperparameters:**

The number of passes, workers, and the number of words chosen to represent topics play crucial roles in shaping the LDA model's output. More passes could increase accuracy, but there is also the risk of overfitting. The number of words chosen to represent topics can influence our interpretation. Too few words might not capture the essence of the topic, while too many could dilute its meaning.

In conclusion, topic modeling, particularly through LDA, offers a powerful way to sift through vast amounts of text data to extract underlying themes and topics. Through meticulous analysis, like the one undertaken, insights can be gleaned that provide a deeper understanding of the dataset's nature and content. Finely tuning and interpreting the LDA model is both an art and a science, requiring a combination of quantitative metrics and qualitative judgment.

Discussion of Sentiment Analysis Results

1. **Comparative Analysis**: Our research, through the lens of sentiment analysis, paints a distinct image of Andrew Tate's role within the Manosphere, further extending the foundation set by Bujalka et al. (2022). While Bujalka et al. outlined the broader landscape, our findings delve deeper into the textual sentiments and themes orbiting Tate's online discourse.

2. **Distinctive Contributions**: The sentiment analysis offers a more granular understanding, shedding light on the specific sentiments and discussions tied to Tate within the Manosphere, distinguishing between the broader themes identified by Bujalka et al. (2022) and the narrower, more specific narratives surrounding Tate.

3. **Key Findings**: Tate's influence has generated a varied landscape of sentiments. The data shows the presence of both support and opposition, evidence of an echo chamber, but also nuances of ontological security. The fluctuating sentiments underline the multifaceted perception of Tate, reinforcing our initial findings on the nature of online engagements around him.

4. **Implications for Future Research**: The sentiment analysis performed offers a blueprint for assessing individual digital personalities. The utilization of topic modeling, graph modeling, and sentiment computation provides a comprehensive method to evaluate other online figures. Our detailed analysis also suggests the importance of contextualizing topics and sentiments in the backdrop of real-life events, as seen with the discussions on Tate's legal issues in Romania.

Conclusion

1. **Hypothesis Verification**: The sentiment analysis corroborates our initial hypothesis and findings. The digital discourse around Tate oscillates between support and criticism, underscoring the effectiveness of ontological security theory in shaping and driving these sentiments, which further aligns with the model presented by Bujalka et al. (2022).

2. **Research Synopsis**: The sentiment analysis emphasizes Tate's prowess in manipulating ontological insecurities. Through the identified topics, it becomes evident that Tate manages to evoke a sense of threat to masculinity, offering remedies that potentially lead to tangible benefits for him. This dynamic foster a spectrum of sentiments, both positive and negative, around his persona.

3. **Recommendations for Future Exploration**: Given the efficacy of the sentiment analysis approach, future research might consider broadening its scope to other platforms where Tate is active. Additionally, integrating more qualitative methods, such as interviews or focused group discussions, would deepen the understanding of the sentiments and their origins.

4. **Learnings and Implications**: The sentiment analysis underscores the potent influence digital personalities wield in shaping online perceptions. Figures like Tate exploit the interplay of insecurities and solutions, crafting vast digital networks. Recognizing and understanding these patterns is vital as the digital realm becomes more deeply ingrained in societal dynamics. The research serves as a clarion call for stakeholders more robust digital literacy education and tools that can empower users to critically analyze and differentiate between genuine online interactions and those manipulated for personal or commercial gain.

Moreover, the proliferation of digital influences and personalities can have profound impacts on the mental well-being of users, especially younger audiences. The constant exposure to curated lives and narratives can foster unrealistic expectations, potentially leading to feelings of inadequacy, anxiety, and depression. Therefore, it becomes imperative for platforms hosting such content to introduce safeguards, guidelines, and support systems to mitigate potential harm. Furthermore, the nature of online interactions and their influence on real-world behavior underscores the need for ethical considerations. Digital personalities, while potentially offering a vast range of benefits like entertainment, education, and social connections, also carry the risk of propagating misinformation, biases, and harmful stereotypes. Platforms and influencers should take on the responsibility of ensuring the content they produce, and share aligns with ethical standards. Collaborative efforts between tech companies, policymakers, educators, and mental health professionals can pave the way for a more balanced and conscious online environment. This might include the creation of toolkits for parents and educators to navigate the digital landscape, guidelines for digital influencers to ensure they operate within ethical boundaries, and platform-based interventions that promote positive online behaviors.

In conclusion, the digital age presents both unprecedented opportunities and challenges. Embracing its benefits requires a proactive approach in addressing its potential pitfalls. The findings of this sentiment analysis reinforce the importance of continuous research, monitoring, and intervention to ensure a healthy and constructive digital experience for all users.

Bibliography

The paper that this is inspired by:

(1) The Manosphere as an Online Protection Racket: How the Red Pill Monetizes Male Need for Security in Modern Society – Fast Capitalism (2022)

(2) Murray, C. 2023. 'Andrew Tate's Rape And Human Trafficking Charges Explained: A Timeline Of The Social Media Star's Controversies', *Forbes*, 1 February. Available at: https://www.forbes.com/sites/conormurray/2023/02/01/andrew-tate-again-appeals-romanian-detention-his-human-trafficking-charges-explained-and-a-timeline-of-the-social-media-stars-controversies/?sh=20c60c6d4e6e [Accessed on (15, October 2023)].

(3) https://www.abc.net.au/news/2023-03-05/andrew-tate-lung-cancer-recruits-politicians-romania/102055620

(4) Anderson and Auxier (2021) Social Media Use 2021, Pew Research Center, https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [Accessed on 20th October, 2023]

(5) A Beginner's guide to LDS, Kulshreshtha 2019, https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

(6) Gephi explained - https://gephi.org/users/supported-graph-formats/graphml-format/

(7) Manosphere Explained, ISD, 2023 https://www.isdglobal.org/explainers/the-manosphere-explainer/