AITA Subreddit Exploration
Web & Social Media Search and Analysis

Antonello Di Rita 514510
Alessandro Longato 512876
Emirhan Kayar 513222

June 9, 2025

Contents

1 Objective

The goal of this analysis is to explore and examine the subreddit `r/AmITheAsshole`. In this subreddit, users submit posts describing conflicts or misunderstandings with family members, friends, strangers, and so on. Each post ends with the question, "Am I the Asshole?". In the comments, other users can share their opinions and judge the original poster by explicitly including one of several tags:

YTA	You're the Asshole
YWBTA	You Would Be the Asshole
NTA	Not the Asshole (and the other person is)
YWNBTA	You Would Not Be the Asshole (and the other person would)
ESH	Everyone Sucks Here
NAH	No Assholes Here
INFO	Not Enough Info

In this project we focus on the two main, diametrically opposed verdicts: **YTA** and **NTA**.

We build interaction graphs and extract comment data from multiple submissions in order to address the following research questions:

- How, and to what extent, do reply graphs differ depending on the final verdict (YTA vs. NTA)?
- Do OPs (Original posters) judged as YTA reply to comments more than those judged as NTA? If so, how and by how much?
- Can the topology of user–interaction graphs reveal distinctive patterns among users who tend to vote YTA rather than NTA? Is there assortativity? Do clusters emerge around which discussions develop? What do centrality measures tell us?
- Is the final verdict (assigned by a bot that counts YTA and NTA tags) correlated with the average sentiment of the responses to a post?
- Do voting trends evolve over time within the subreddit?

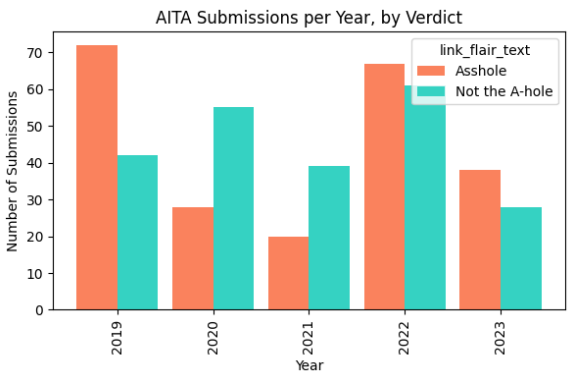
By combining graph-theoretic methods, the Reddit API, `NetworkX`, sentiment analysis, and visualization techniques, we investigate these questions.

2 Data

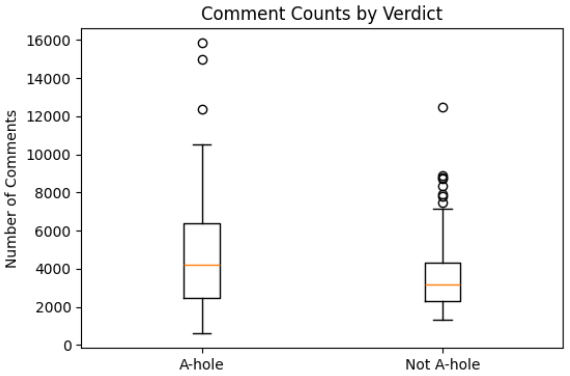
2.1 Main Dataset

To build the reply graphs and evaluate comment sentiment, we used the Reddit API to collect 500 posts from the subreddit, evenly split between the **YTA** and **NTA** tags (250 each). Posts were retrieved with Reddit's *top* search algorithm, which returns content with the highest upvotes and overall engagement. For each post we stored, in a `pandas` DataFrame, a rich set of attributes, including `submission_id`, author, tag, creation time and date, `upvote_ratio`, score, title, body text, number of comments, and so on.

The resulting dataset was filtered to keep only submissions from 2019–2023—the period with the greatest posting activity. Figure 1a shows the distribution of posts across these years, while Figure 1b reports the number of comments per verdict (YTA vs. NTA).



(a) Posts per year



(b) Comments by tag

Figure 1: (a) Yearly distribution of posts. (b) Comments by tag.

2.2 Comments Dataset

After creating the main dataset, we extended it to include comments up to three levels of depth. The purpose of this dataset is to perform sentiment analysis; accordingly, information about who replied to whom was discarded, and we retained only the year, the depth level of each comment within a submission, the comment text, the author, and the post tag (YTA/NTA). Figure 2 shows histograms of the distribution of comment counts at each of the three depth levels in a single combined image.

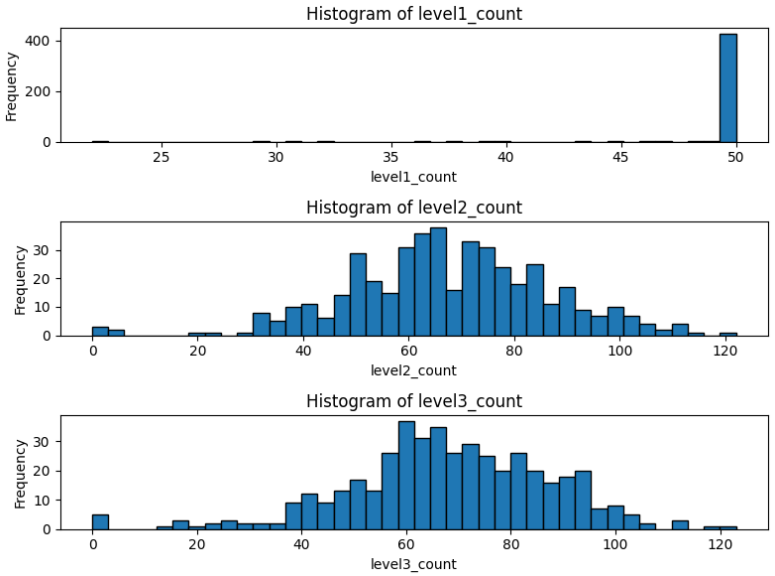


Figure 2: Distribution of the number of comments at depth levels 1, 2, and 3.

2.3 Reply Graph Dataset

To construct the reply graph for a submission, we need to retrieve the full comment tree to the greatest possible depth. However, doing so for all 450 posts would require an excessive number of Reddit API calls. Therefore, we randomly selected 40 posts from our main dataset (20 YTA and 20 NTA) and collected the complete comment trees only for this subset. During this process, we recorded each comment's ID and author to establish the reply relationships.

2.4 User-Interaction Graph Dataset

For the user–interaction graph, we used all 450 submissions from the main dataset. For each submission, we extracted comments up to the third depth level, excluding the first level (OP → top-level comments) to focus solely on community interactions and avoid over-emphasizing OP activity. For every reply pair (responder → original commenter), we recorded the comment score (number of up-votes), any vote tags (presence of "YTA" or "NTA" in the comment), and the authors' usernames.

The raw dataset contained multiple interactions between the same user pairs; we therefore aggregated these to create weighted edges, where each edge weight equals the total number of interactions in that direction between the two users. Likewise, we summed upvote counts per user to identify those with the most appreciated responses. Finally, for each author, we computed a vote-tendency metric defined as:

$$\text{Vote Tendency} = (\text{Total YTA votes}) - (\text{Total NTA votes}).$$

Figure 3 shows histograms (both linear and logarithmic scales) of the distributions of vote tendency and total up-vote counts, combined into a single image.

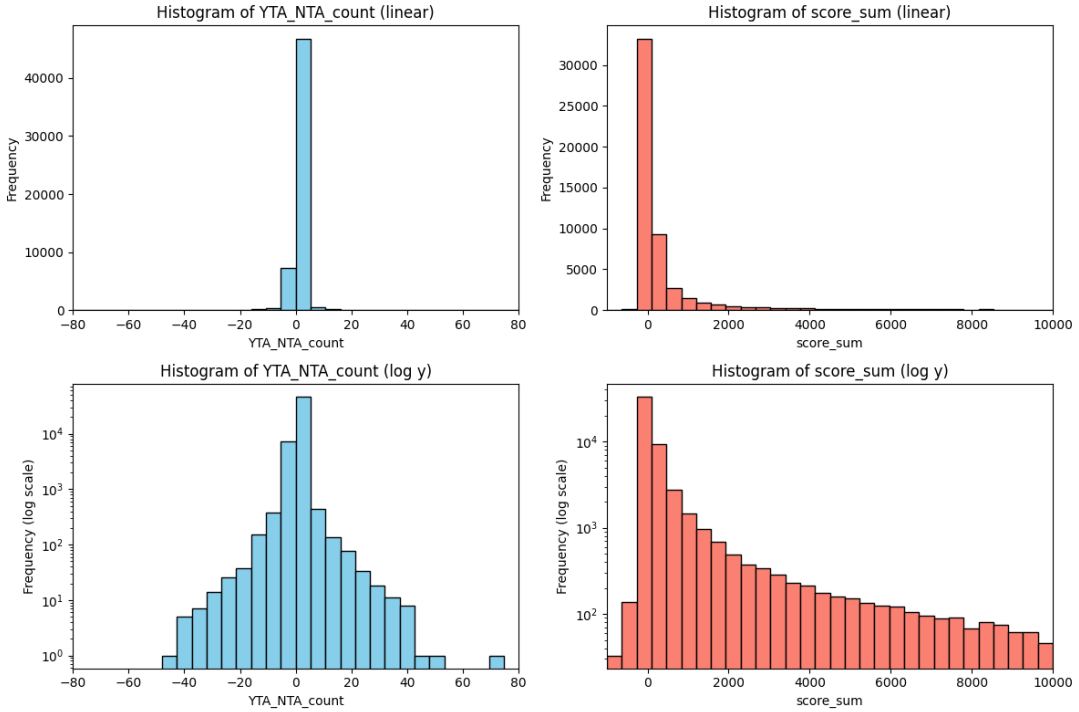


Figure 3: Distribution of author vote tendency and total up-vote counts (linear and log scales).

3 Methodologies and Models

3.1 Topology Analysis

3.1.1 Reply Graphs Model

To understand the differences between conversations where the OP is judged YTA rather than NTA, we used `NetworkX` to construct a separate reply graph for each of the 40 selected submissions. From each graph, we extracted the following metrics:

- **Number of nodes:** total comments in the tree.

- **Maximum depth:** longest distance from the root (OP) to any leaf.
- **Longest reply chain:** the maximum shortest-path length from the root.
- **Average reply depth:** mean distance of all comments from the root. Higher values indicating more sustained, detailed back-and-forth exchanges.
- **Average branching factor:** average number of direct replies per node, indicating community engagement and diversity of parallel conversations.
- **Maximum branching factor:** the highest number of replies to any single comment.
- **Shape imbalance:** measured by the normalized Sackin index, computed as $I = \frac{S}{n \log_2 n}$, where $S = \sum_{v \in \text{leaves}} \text{depth}(v)$ is the total sum of distances from the root and n is the total number of nodes. Higher values indicate greater skew. This metric helps distinguish between "viral" discussions—characterized by broad branching and shallow depth—and more "polarized" or "technical" threads, which exhibit few branches but greater depth.

As illustrative examples of the resulting reply graphs, see Figure 4 in Section 4.

3.1.2 OP Presence

We examined how the OP interacts with comments depending on the discussion outcome (YTA vs. NTA). To this end, we computed several new measures on the reply graph. After identifying the node corresponding to the original poster, we calculated:

- The number of OP replies (all OP nodes excluding the root node),
- The ratio of OP replies to total comments,
- The average depth at which the OP replies.

3.1.3 User-Interaction Graph

Using the dataset described above, we constructed a **global user–interaction graph** in which each node represents a user who replied to an OP or another comment, and each directed edge is *weighted* by the number of interactions between those two users. Each node is annotated with its **voting tendency** (YTA vote count minus NTA vote count) and its **score** (sum of upvotes minus downvotes received).

We then examined the relationship between **out-degree** (the number of distinct comments a given user replied to), **in-degree** (the number of distinct reply received by other users), and **voting tendency** (YTA – NTA count) via a scatter plot. A similar analysis compared both in- and out-degree against user *score*. This exploration can help identify particular *social interaction patterns* based on these factors.

Next, we calculated **assortativity** and the **clustering coefficient**. Assortativity—which measures the tendency of similar users to interact more with each other than with dissimilar users—was computed using three similarity measures: *score*, *voting tendency*, and *total degree* (in-degree + out-degree). To obtain the *score* assortativity and *voting tendency* assortativity measures, we used a random subset of the graph containing 40% of the edges, due to computational resource constraints. For clustering, we used the **average clustering coefficient**—defined as the mean of the local clustering coefficients—to explore the propensity of users to form *tightly connected groups*.

Finally, we analyzed three different **centrality measures** for each user and investigated their correlations with **score** and **voting tendency**. These metrics were computed only on the **giant**

component (the largest weakly connected component). Due to computational constraints, *betweenness centrality* was estimated using a 5,000-sample approximation per node.

Visualizations of the two graphs (colored by voting tendency or score), the scatter plots, and the centrality measures are presented in Section 4.

3.2 Sentiment Analysis

To investigate whether there is a correlation between the final judgment and the average sentiment of a post's comments, we conducted sentiment analysis.
The analysis was based on the comments dataset described in Section 2.2.
We applied four sentiment analysis techniques: three dictionary-based methods (Afinn, NRCLex, and Vader) and one neural network-based model (Bert). We wanted to explore both the difference between two very different approaches, lexicon and neural, and between distinct lexicons, each with its strengths and weaknesses.

- **Afinn** assigns a score between -5 and +5 to words to represent their sentiment and intensity. However, it does not handle negations and nuances in the language.
- **NCRLex** is a more refined lexicon that includes emotions. For the sake of our analysis, we computed the score using only the polarity provided by NCRLex, disregarding the emotion scores. The score for this tool was thus computed as the positive score minus the negative score. It should be noted that each word has a binary score for the different emotions and polarities, so NCRLex does not account for intensity. Moreover, like Afinn, this lexicon does not take into consideration negations and context.
- **Vader** integrates a rule-based system together with the lexicon to handle negations, slang, emojis, capitalization, and intensifiers. It should be able to better capture the context of the text.
- The **Bert**-based model we used is the `cardiffnlp/twitter-roberta-base-sentiment`. The model has been fine-tuned on tweets for sentiment analysis. Although we are working in a different domain, we were not able to find such a robust and reliable model made for Reddit. The final score is again computed as the difference between the positive and negative scores.

3.2.1 Flair Analysis

After obtaining the 4 sentiment scores for each comment, we started by grouping the comments by submission, so that each comment has the same weight on the final submission sentiment score, regardless of the level. The total number of entries is equal to the starting number of posts (450).

3.2.2 Level and Flair Analysis

This time, we grouped and averaged the comment scores by level and submission. In this way, we obtained 3 data points for most submissions, totaling 1346 entries.

3.2.3 Word Clouds

We generated word clouds for all comments, comments from submissions labeled as 'Not the A-hole', and comments from submissions labeled as 'Asshole'. To do so, first we cleaned the comment texts by lowercasing, removing punctuation, removing stopwords, some uninformative words that are not usually considered stopwords (like 'even' and 'one'), and other not meaningful words specific to the context like 'YTA', 'NTA', and 'op'. Then, we used the library `wordcloud` to display the word clouds of the 50 most common words in each set.

3.3 Temporal Analysis

To analyze the AITA dataset and uncover community trends, a three-step methodological approach was employed. The goal was to identify temporal patterns, detect significant events, and perform both time series and statistical analysis.

3.3.1 Post Distribution Analysis

Post frequency was examined across different time frames for the YTA and NTA flairs. This included:

- **Hourly Distribution:** Identifying peak activity hours.
- **Weekly Trends:** Analyzing post frequency by each day of the week.
- **Seasonal Trends:** Aggregating post counts by month.
- **Yearly Trends:** Observing changes in post volume from 2019 to 2023-11.

These insights provided a foundation for detecting peak activity periods and fluctuations in engagement, which are discussed in a later section.

3.3.2 Comment Activity Analysis

Comment data from non-OP was analyzed in two ways:

- **Temporal Trends (2019-2023)** Comment counts for YTA and NTA flairs across three depth levels. Average comment lengths for the same period and depth levels.
- **Overall Comparison** Mean comment count and mean length compared between the two flairs without depth-level separation.

3.3.3 Sentiment Score Analysis

Sentiment scores were calculated using four models—AFINN, VADER, NRClex, and BERT. Time series analysis was then conducted for each score type, comparing YTA and NTA posts across the same 58-month period.

4 Results

4.1 Topology Analysis

4.1.1 Reply Graphs

Figure 4 shows two example reply trees: the **left** tree represents an NTA discussion, while the **right** tree represents a YTA discussion.

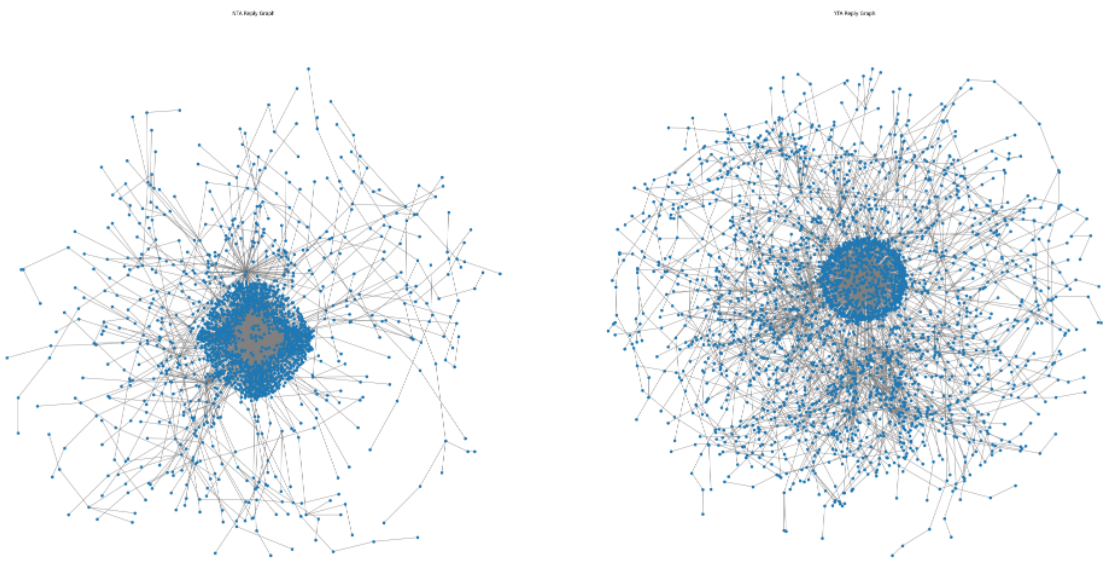


Figure 4: Comparison of reply graph structures: left shows a NTA discussion, right shows an YTA discussion.

Figure 5 summarizes key reply-graph metrics for NTA and YTA threads:

- **YTA** discussions are *larger and deeper*, with a median of ≈ 4,700 nodes (IQR 2,900–6,200) versus ≈ 3,600 nodes (IQR 2,100–4,600) in **NTA**.
- The median maximum depth increases from 18 (IQR 15–21) in **NTA** to 23 (IQR 15–30) in **YTA**, although the median average depth remains *similar* at ≈ 2.5 (IQR 2.1–3.1) for **NTA** and ≈ 2.4 (IQR 2.2–2.6) for **YTA**.
- **NTA** threads exhibit a *slightly higher* median branching factor (2.15; IQR 2.0–2.45) compared to **YTA** (2.00; IQR 1.8–2.15), while maximum branching is identical at a median of 89 in both.
- The shape imbalance shows *minimal difference*, with a median of 0.18 in **NTA** (IQR 0.15–0.20) and 0.18 in **YTA** (IQR 0.15–0.26).

These boxplots highlight that *YTA* discussions tend to be *larger and deeper* on average, whereas *NTA* threads exhibit *slightly higher branching*.

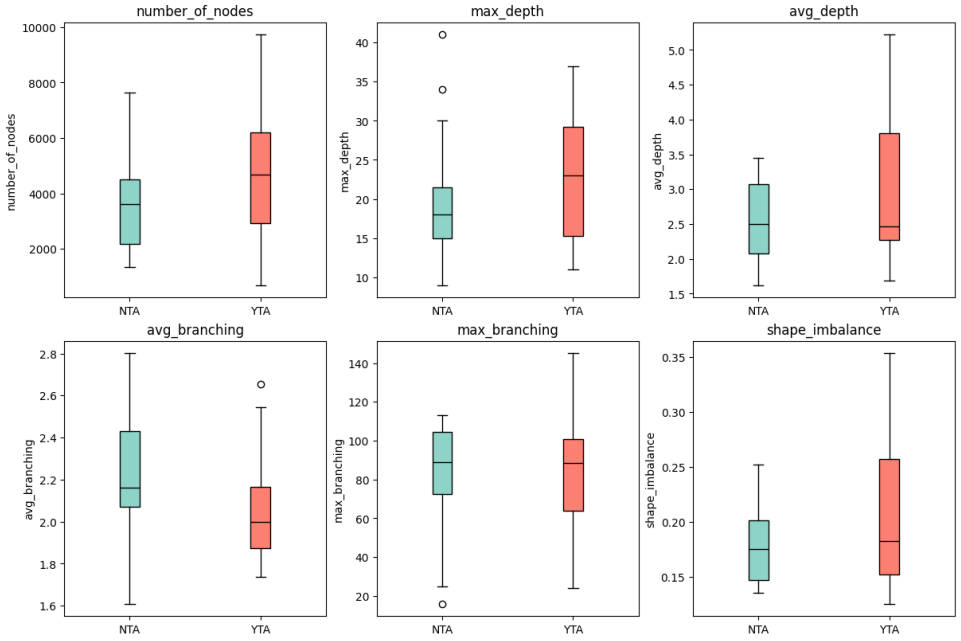


Figure 5: Distributions of reply graph metrics for NTA vs YTA threads.

4.1.2 OP Presence

Of the 40 submissions analyzed (20 with final verdict **NTA** and 20 with final verdict **YTA**), the OP replied at least once in *18* NTA threads, whereas only *11* YTA threads include an OP reply. The data presented in the boxplots in Figure 6—calculated only for posts with at least one OP reply—reveal that when OPs do engage:

- **NTA** threads have a *slightly higher* median number of replies (6; IQR 4–17) than **YTA** threads (5; IQR 2–14).
- OP replies also make up a *larger share* of the overall discussion in **NTA** threads (median ratio ≈ 0.0023; IQR 0.001–0.005) compared to **YTA** threads (median ratio ≈ 0.0007; IQR 0.0003–0.005).
- Finally, OP responses occur at *similar depths* in **NTA** (median ≈ 2.35; IQR 2.0–2.74) and **YTA** threads (median ≈ 2.0; IQR 2.0–3.44), though the *variance* is notably higher in **YTA** threads (std. $= 1.11$ vs. 0.72).

Note: The sample size for these analyses is relatively small, so these descriptive comparisons should be interpreted with caution.

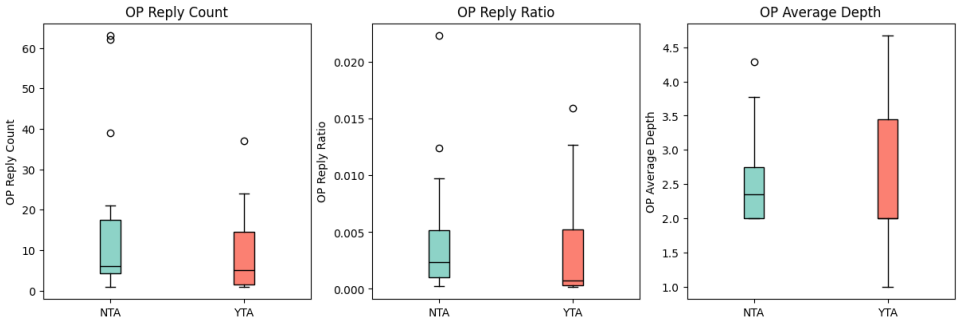
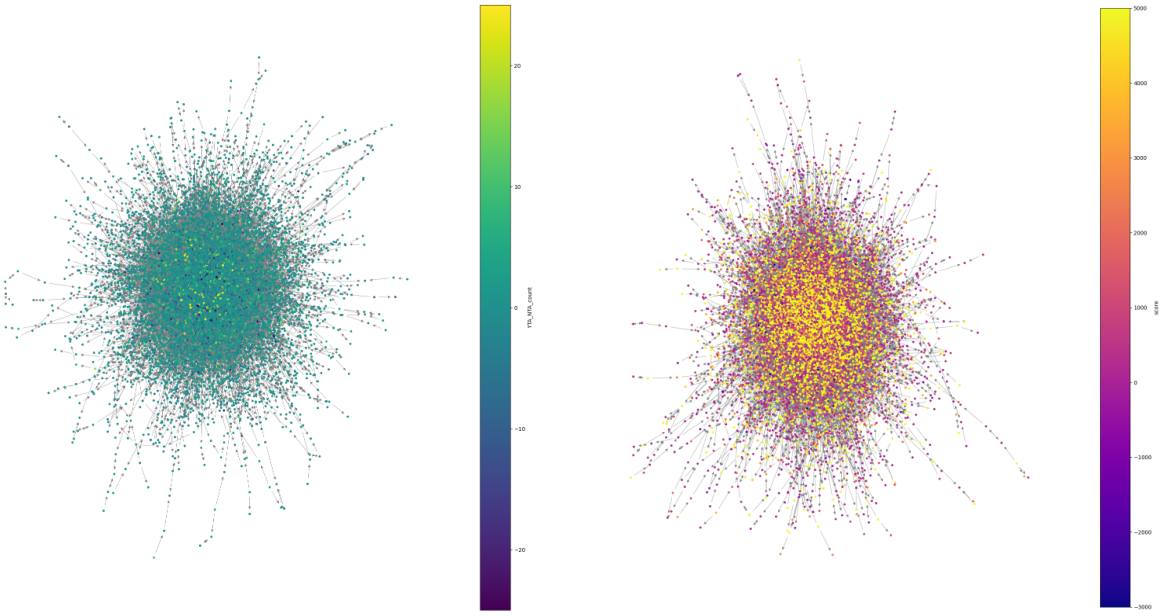


Figure 6: OP reply count, reply ratio, and average reply depth for NTA vs. YTA threads.

4.1.3 User Interaction Graphs: Visualization

We generated two *visualizations* of the resulting graphs (see Figure 7). In both, only the maximum weakly connected component is shown, self-loops have been removed, and only nodes with a total degree (in-degree + out-degree) greater than 2 are displayed, both to accommodate our time constraints and to improve readability.



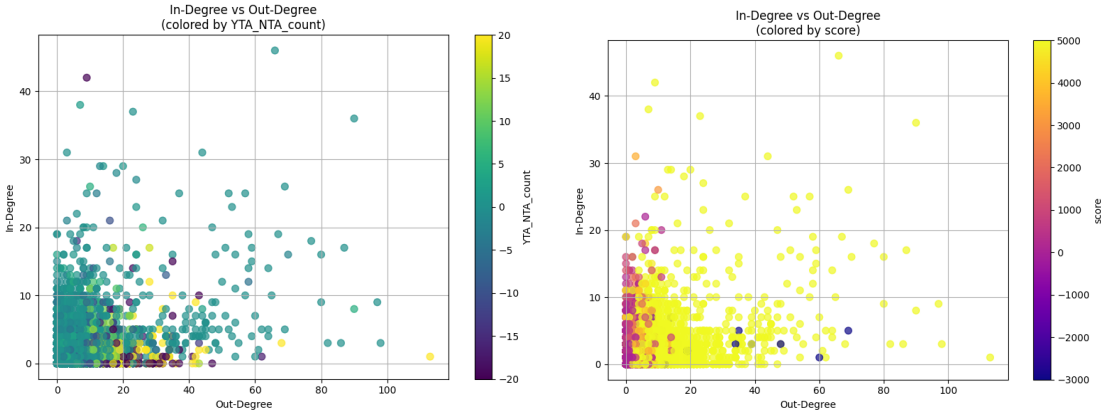
(a) Nodes colored by *voting tendency*

(b) Nodes colored by *user score*

Figure 7: User interaction graphs (largest weakly connected component, self-loops removed) with nodes filtered by degree > 2. (a) Nodes colored by **voting tendency**. (b) Nodes colored by **user score**. Voting tendency values are shown on a scale from -20 to $+20$, and user score values range from -3000 to $+5000$, with any values outside these ranges clipped to the nearest bound.

4.1.4 User Interaction Graphs: Vote Tendency vs. Score

The following two graphs (Figure 8a and Figure 8b) depict, for each user, the number of distinct replies they made to others (x-axis) versus the number of distinct replies they received (y-axis). The two graphs differ in the metric shown by the color scale: the first illustrates voting tendency, while the second represents user score.



(a) Voting Tendency (b) User Score

Figure 8: User out-degree vs. in-degree colored by (a) voting tendency and (b) score.

In Figure 8a, neutral vote counts are distributed across the full range of in- and out-degrees, whereas YTA-leaning and NTA-leaning votes cluster around out-degrees of 10–40 and in-degrees of 0–10.

Figure 8b shows that the most active users (high out-degree) also attain the highest scores. Conversely, users with the lowest scores are few: they post many replies but receive few responses (in-degree < 10). Since low-score comments may be deleted, this potential bias should be taken into account. Users with neutral scores tend to be the least active (low out-degree), and in-degree appears to have minimal impact on overall score.

4.1.5 User Interaction Graphs: Assortativity, Clustering Coefficient, and Centrality Measures

Assortativity and Clustering Coefficient We computed three different assortativity coefficients: **score assortativity** of *-0.0068*, **vote-tendency assortativity** of *0.0698*, and **degree assortativity** of *-0.0515*. The **average clustering coefficient** is *0.0021*. All assortativity coefficients are *negligible* (absolute value < 0.1) and the mean clustering coefficient is almost zero, indicating the **absence of any systematic connection patterns** among nodes. Note that *subsampling* a portion of the edges tends to attenuate assortativity, so these values should be interpreted with **caution**.

Centrality Measures We calculated three centrality metrics for every node in the giant component.

- **Closeness centrality**
  Range: 0.000000–0.013277; mean: 0.002793; median: 0.000054; SD: 0.003729.
- **Betweenness centrality**
  Range: 0.000000–0.007377; mean: 0.000016; median: 0.000000; SD: 0.000157.
- **Degree centrality**
  Range: 0.000024–0.002986; mean: 0.000063; median: 0.000024; SD: 0.000109.

Figure 9 displays the joint distribution of *Vote Tendency* (first row) and *Score* (second row) with each of the three centrality measures.

- **Vote Tendency vs. Closeness.** Users whose verdicts are most balanced occupy the most central positions in the network, whereas many users with extreme verdicts have closeness values near 0. Notably, there is an absence of observations with closeness between 0.001 and 0.004.

- **Vote Tendency vs. Betweenness.** The users who connect the community most strongly (high betweenness) also exhibit the most neutral vote tendencies; users with extreme tendencies cluster at low betweenness values.
- **Vote Tendency vs. Degree Centrality.** The scatterplot forms a distinctive triangular shape with three main linear clusters. As degree centrality increases, vote tendency appears to diverge linearly in two opposite directions, while neutral tendencies occur across the entire degree-centrality range.
- **Score Plots.** The score-based plots show very similar patterns, except that the values are not symmetric around zero: almost all individuals have positive scores.

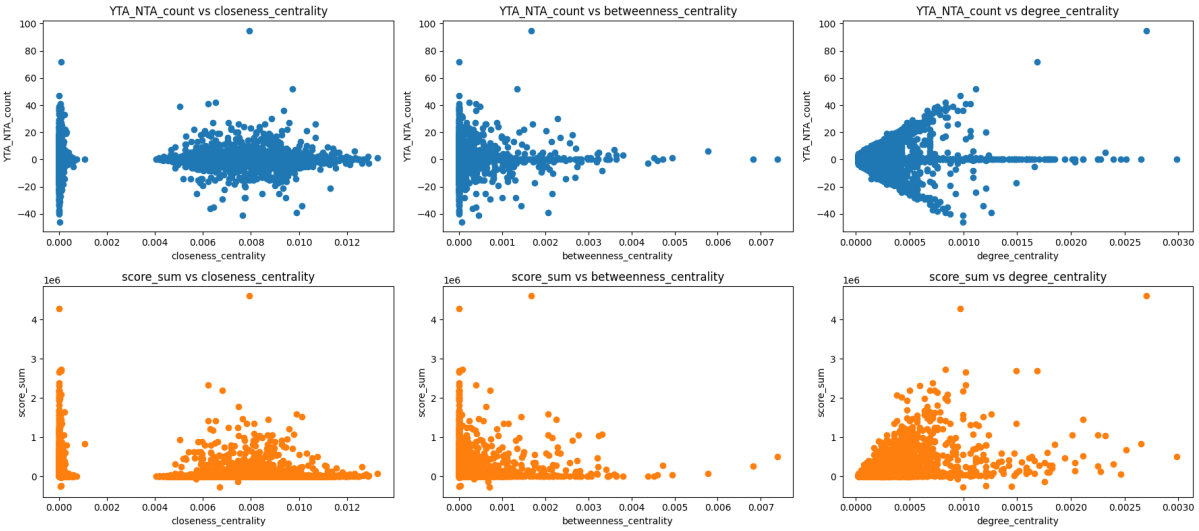
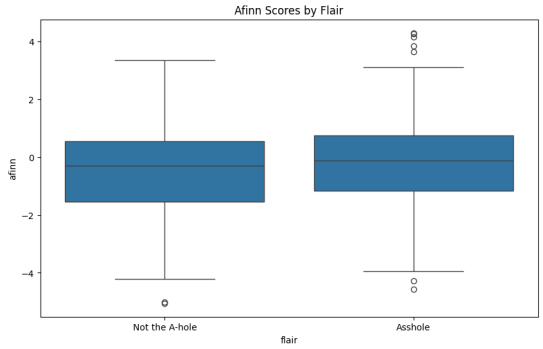


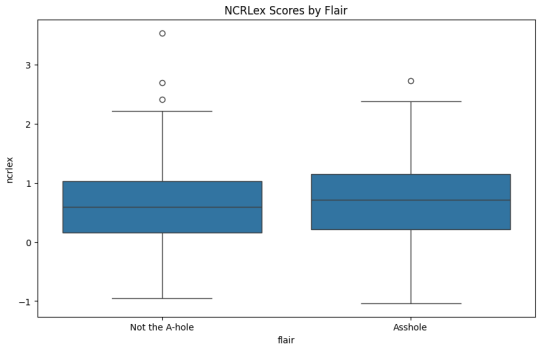
Figure 9: Scatterplots of `Vote Tendency` (top row) and `Score` (bottom row) against closeness, betweenness, and degree centrality, respectively.

4.2 Sentiment Analysis

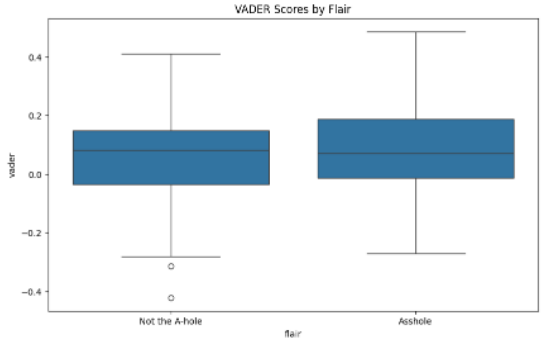
4.2.1 Sentiment scores analysis



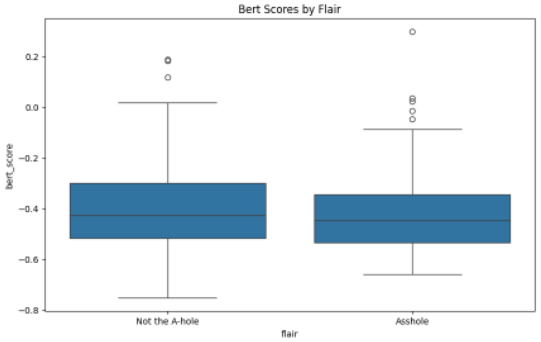
(a) Afinn score



(b) NCRLex score

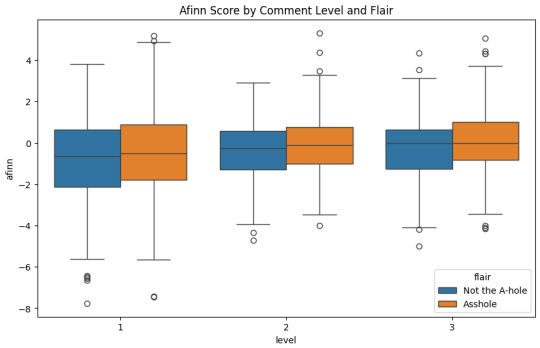


(c) Vader score

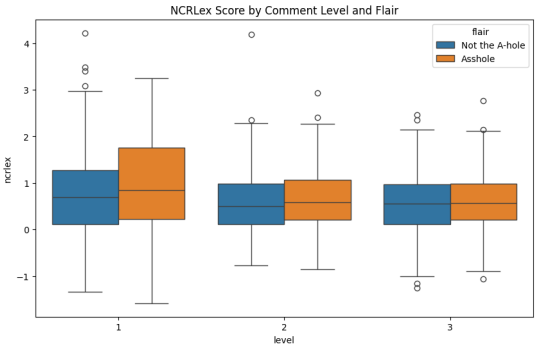


(d) Bert score

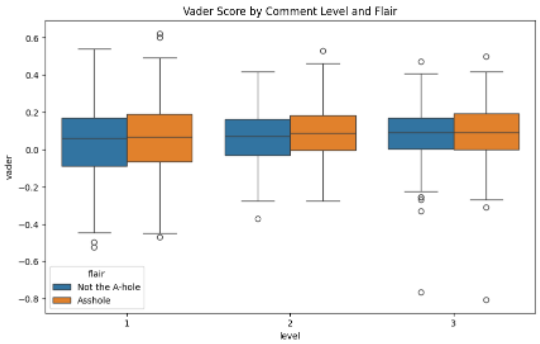
Figure 10: Scores of the different sentiment analysis techniques in the flair analysis.



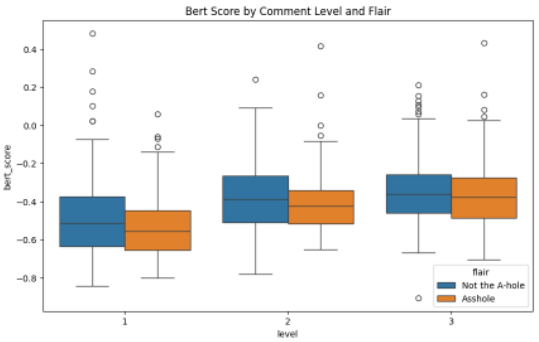
(a) Afinn score



(b) NCRLex score



(c) Vader score

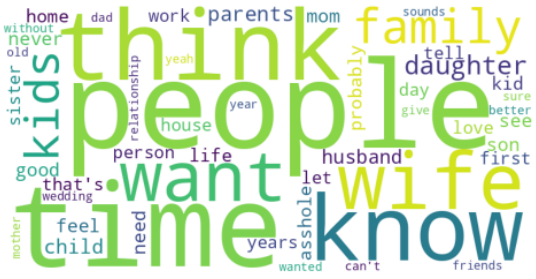


(d) Bert score

Figure 11: Scores of the different sentiment analysis techniques in the level analysis.

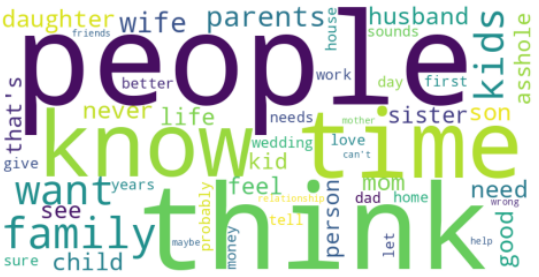
From the data it clearly emerges a disagreement between the lexicon-based approaches and the neural-based approach. The lexicon-based tools all agree to give higher sentiment scores to comments from submissions labeled as 'Asshole'. On the other hand, Bert assigns higher sentiment scores to comments from submissions labeled as 'Not the A-hole'.

4.2.2 Word clouds



(a) Word cloud of comments from submissions labeled as 'Asshole'.

(b) Word cloud of comments from submissions labeled as 'Not the A-hole'.

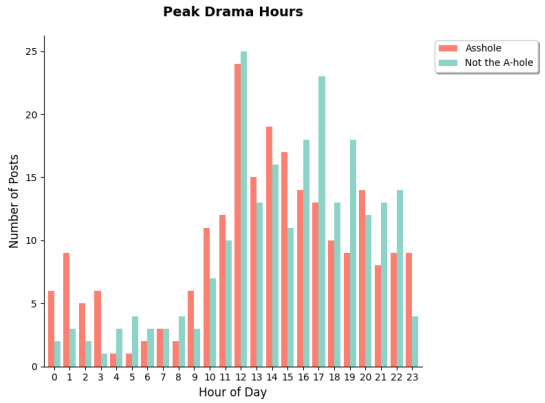


(c) Word cloud of comments from all submissions.

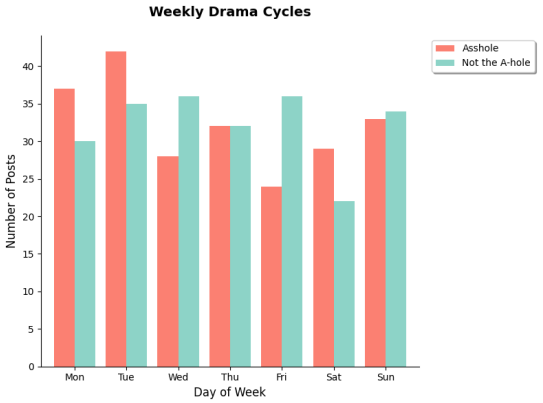
Figure 12: Word clouds for different sets of comments.

4.3 Temporal Analysis

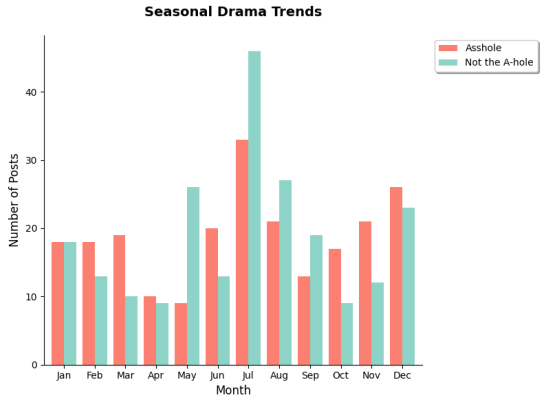
4.3.1 Temporal Post Distributions



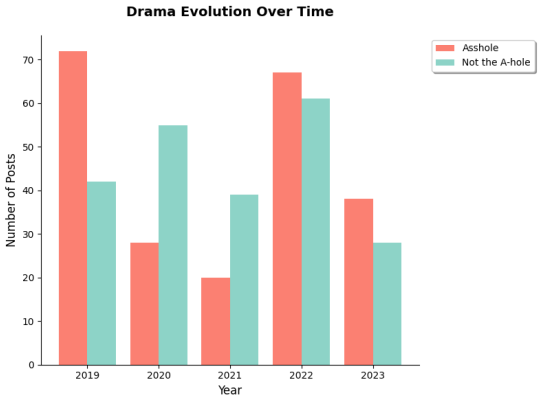
(a) Hours Per Day



(b) Weeks Per Month



(c) Months Per Year



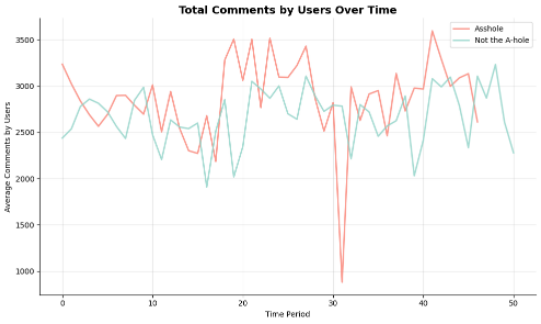
(d) Years

Figure 13: Temporal post patterns.

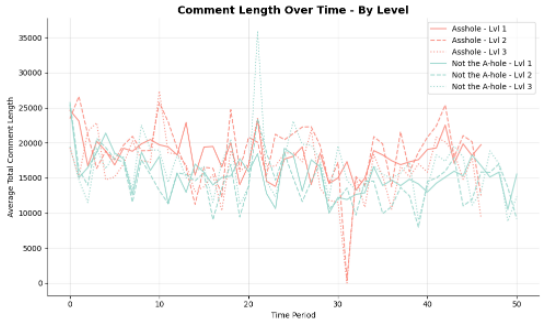
- **(a) Peak Drama Hours** Activity peaks sharply at 12:00 with 49 posts, while 4:00 sees the lowest at 4 posts. The average hourly post count is 18.8, with 17:00 showing the highest variance. NTA posts slightly outnumber YTA at peak, with NTA maintaining a lead across multiple later hours. Minimal posting from 00:00–06:00 reflects a typical circadian rhythm.
- **(b) Weekly Drama Cycles** Tuesdays lead with 77 posts, while Saturdays lag at 51. The weekday vs weekend ratio is 2.78:1, suggesting users favor judgment-heavy posts earlier in the week. The average daily count stands at 64.3 posts.
- **(c) Seasonal Drama Trends** July is the most active month (79 posts), far surpassing April (19 posts). Posting rises in summer, as shown by a summer-to-winter ratio of 1.38, and again in December, likely tied to seasonal events.
- **(d) Drama Evolution Over Time** Post volume peaked in 2022 (128 posts) and dipped in 2021 (59 posts). The overall year-over-year growth rate is –42.1%, indicating volatility. YTA dominated in 2019, while NTA spiked in 2020, likely due to the pandemic's influence. Both saw partial recovery in 2022, followed by a decline in 2023.

4.3.2 Temporal Comment Activity

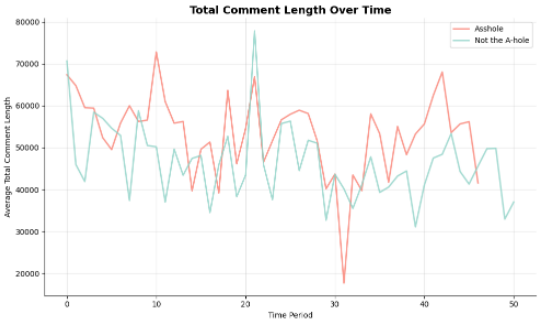
(a) Comments by level



(b) Comments averaged



(c) Comment length by level



(d) Comment length averaged

Figure 14: Temporal comment activity.

Comment behavior shows consistent depth-layer trends and content engagement differences.

- **(a) Comments by Level:** "Level 1" comments quantity are more consistent and lower than the other levels, while "Level 2" and "Level 3" fluctuate and spike upwards, this is an indicator that after the initial engagement, if the level depth increases, density of the argument increases. It is also clear that, the YTA counts are greater than NTA counts in the deeper levels.
- **(c) Comments Length by Level:** Same fashion follows, as in comments count by level, except for the "Level 1", here the graph demonstrates that the length of the text is not directly correlated with the depth of the comment. However, as mentioned before YTA category wins over the other, meaning that YTA's are most likely to comment longer.
- **(b)-(d) Overall** YTA posts attract more total comments and longer cumulative responses, suggesting broader community empathy. NTA posts maintain slightly lower volume.

4.3.3 Temporal Sentiment Score

(a) Afinn Score Over Time

(b) Vader Score Over Time

(c) NCRLex Score Over Time

(d) Bert Score Over Time

Figure 15: Sentiment Score Trends.

- **(a) Afinn:** The chart indicates a fluctuating dynamic between the two categories. The YTA sentiment initially dips, but, shows a gradual upward trend over time. In contrast, NTA sentiment remains variable without a clear long-term trend.
- **(b) Vader:** The pattern closely mirrors the Afinn results, reinforcing a cyclical trend and their inverse relationship is evident.
- **(c) NCRLex:** The inverse, cyclical pattern continues, but with greater crossovers. In contrast with other models, at the end of 2023, sentiment score for YTA decreases.
- **(d) Bert:** YTA discussions are inherently conflict-driven and carry a more negative tone. Within this context, NTA posts consistently show less negative sentiment. The model clearly differentiates between the two, with YTA exhibiting sharp, cyclical drops in sentiment.

5 Discussion

The obtained dataset represents a significant portion of the AITA subreddit, with a good temporal distribution that allows for the analysis of possible trends. Regarding the number of comments per post, Figure 1b shows that, on average, discussions with a YTA outcome generate more comments compared to those with an NTA outcome. This may reflect greater community engagement when the OP is judged negatively. Since the first level of replies typically consists of direct interactions between users and the OP, it remains unclear whether the increased number of comments is driven by disagreement with the OP or by the nature of the topic itself. Further analysis would be required to clarify this aspect.

5.1 Topology Analysis

Reply Graphs The two reply graphs shown in Figure 4 highlight the main differences we observed between YTA and NTA discussions. YTA graphs tend to be larger and deeper, while NTA graphs display higher branching (Figure 5). This may indicate that when the OP is judged to be in the wrong, the topic sparks more controversial or argumentative discussions, often involving direct replies and clarifications. In contrast, NTA threads show broader but shallower engagement, possibly due to general agreement and supportive comments from the community. As for OP engagement, although based on a limited number of samples and with only a slight difference, we observe that NTA discussions contain more replies from the OP compared to YTA discussions. While this observation is not conclusive due to the small sample size, it may suggest that a more supportive or agreeable environment encourages the OP to participate further, whereas in YTA threads, the critical tone or disagreement may discourage them from engaging (see Figure 6).

User–interaction graph. The global user–interaction network confirms that `r/AmITheAsshole` behaves more like an open agora than a set of insulated camps. All three assortativity coefficients are essentially zero (score $r = -0.0068$, vote-tendency $r = 0.0698$, degree $r = -0.0515$) and the mean clustering coefficient is almost nil ($\bar{C} = 0.0021$), indicating that users neither preferentially address like-minded interlocutors nor form tightly knit triangles. Visual inspection of the giant component (Fig. 7) and of the in- vs. out-degree scatterplots (Fig. 7a; Fig. 7b) reinforces this picture: highly polarised YTA- and NTA-leaning accounts lie on a sparsely connected periphery, while the central backbone is occupied by users whose verdict counts balance out. Centrality measures tell the same story. The nodes with the highest closeness or betweenness—i.e. the brokers who bridge otherwise disconnected threads—display vote tendencies close to zero, whereas outspoken partisans cluster at low centrality values, giving rise to the characteristic "polarisation funnel" in Fig. 9. score, by contrast, rewards productivity rather than influence: user score rises almost linearly with out-degree but shows little dependence on in-degree (Fig. 7b), suggesting that the community values talking more than being talked to. Taken together, these findings portray a debating floor where loud partisan voices are plentiful yet weakly connected, and where a small set of neutral "connectors" weaves the conversation into a single, albeit extremely sparse, whole.

5.2 Sentiment Analysis

Sentiment Scores Our initial analysis using boxplots (Figure 10 and 11) showed some visual differences in sentiment scores between comments on 'Asshole' (YTA) and 'Not the A-hole' (NTA) posts. The next logical step was to determine if these differences were statistically significant or simply due to random chance. Our null hypothesis was that there is no significant difference in the mean sentiment scores between the two flair categories. Some statistics about the results are reported in Table 1, while the results of the statistical test are reported in Table 2.

We first checked the assumptions for each data split using the Shapiro-Wilk test (for normality) and Levene's test (for equal variances). Afinn and Vader scores were normally distributed, making the t-test appropriate. The NCRLex and Bert scores were not, so we relied on the Mann-Whitney U test for those. We used a standard alpha level of 0.05 to determine statistical significance.

According to our analysis, only Afinn detected a statistically significant difference in sentiment between the two flairs. Both Vader and NCRLex found no significant difference. The Bert score showed a p-value of 0.064. While not significant at the 0.05 level, this is considered marginally significant.

Our main finding is the a substantial disagreement between the various sentiment analysis mod-

els. This prevents us from confirming our initial hypothesis relating to the question outlined in Section 1. Our hypothesis was that comments from submissions labeled as 'Asshole' have an overall more negative sentiment than ones from submissions labeled as 'Not the A-hole'. This hypothesis was based on the premise that in general comments from 'Asshole' submissions focus on explaining to the OP why they are labeled like that, even using an abrupt tone. On the other hand, comments from 'Not the A-hole' submissions focus on reassurance and sympathy towards OP. For the sake of this discussion, we analyzed the data according to our hypothesis, discussing where this disagreement may originate from.

In this light, our interpretation of these results is that they reflect the varying ability of each sentiment analysis tool to navigate nuanced texts. The comments in the AITA subreddit are highly complex: they may be argumentative, sympathetic with the wronged side, full of sarcasm, validating, or a mix of any of these features.

Afinn confidently makes the wrong interpretation because it is not equipped to understand how the words are used in a long and articulate text. It identifies a significantly more positive sentiment in 'Asshole' threads. This is likely because it cannot distinguish context and is heavily influenced by the positive language used to validate the person wronged in the story, which often outweighs the negative words directed at the OP.

NCRLex and Vader are a step higher on the complexity scale. They are able to detect the mixed sentiment in the comments, and so they don't detect any statistically significant difference between the two categories. While this prevents them from being drawn to the wrong conclusion like Afinn, it also makes them unable to differentiate the nuances of the two categories.

Lastly, the Bert model has a superior ability to navigate complicated texts. Consequently, it identifies a trend consistent with our original hypothesis: recognizing a more negative sentiment in comments from submissions labeled as 'Asshole', albeit in a marginally significant way. It is plausible that a robust Bert model fine-tuned on Reddit posts and comments would confidently detect a difference in the sentiment, with 'Asshole' comments having a more negative sentiment.

An alternative interpretation, however, is that the results we obtained are due to chance. It is possible that no model we used is equipped to understand complex human generated text. The statistically insignificant difference between scores found in NCRLex, Vader, and Bert should be interpreted exactly as they are. This would suggest that we were not able to capture the underlying trend, if one exists.

Flair	**Afinn** μ / σ	**NCRLex** μ / σ	**Vader** μ / σ	**Bert** μ / σ
Asshole	-0.141 / 1.575	0.703 / 0.686	0.083 / 0.146	-0.427 / 0.146
Not the A-hole	-0.477 / 1.484	0.637 / 0.652	0.064 / 0.144	-0.400 / 0.163

Table 1: Mean and standard deviation of sentiment scores by flair

Table 2: Statistical Comparison of Sentiment Scores Between "Asshole" and "Not the A-hole" Flairs

Sentiment Method	**Statistical Test**		**p-value**	**Significant? (α=0.05)**
	Test Used	*Statistic*		
Afinn	t-test	t = -2.326	**0.021***	Yes
NCRLex	Mann-Whitney U	U = 23487.0	0.186	No
Vader	t-test	t = -1.442	0.150	No
Bert	Mann-Whitney U	U = 27866.0	0.064	No (Marginal)

Note: The most appropriate test was chosen based on data distribution. The negative t-statistic for Afinn indicates the mean score for 'Not the A-hole' comments was significantly **lower** than for 'Asshole' comments. An asterisk (*) denotes a statistically significant result.

Level Analysis Another interesting insight that derives from the boxplots is that the sentiment tends to slightly increase the deeper the comment is. This would suggest that deeper comments contain softer tones or simply agreement with the previous comments.

Word Clouds The word cloud visualizations provide two key insights into the word patterns of the AITA subreddit.
First, as we can see in Figure 12 the content of the AITA subreddit revolves mostly around familial conflicts. This is evidenced by the high frequency of words belonging to the family semantic field, like parents, husband, wife, son, ... Interestingly, wife appears more often in the 'Asshole' word cloud. This hints that most of the time, when there is a conflict between OP and their wife, OP is the 'Asshole'.
Second, there is a substantial vocabulary overlap between comments from submission labeled 'Asshole' or 'Not the A-hole'. So the key point is not about what words are used but how they are contextualized.

5.3 Temporal Analysis

The temporal analysis of the AITA dataset reveals complex yet interpretable behavioral patterns that reflect both platform dynamics and broader social trends. By decomposing the data across time units — hourly, daily, monthly, and yearly — alongside comment engagement and sentiment metrics, several important insights emerged regarding user activity, community interaction, and the emotional tone of posts.

Posting Patterns An exploration of posting activity on the AITA subreddit from early 2019 to late 2023 reveals some clear rhythms in how and when users engage with the community. Posts tend to follow daily and weekly routines, with more activity during typical waking hours and workdays, and less engagement during early mornings and weekends. These patterns suggest that users are more inclined to share their moral dilemmas during structured, routine-driven times rather than during periods of rest or downtime.
Seasonal trends also emerged, with increased participation during certain times of the year, likely influenced by school breaks, holidays, and changes in social interaction. These shifts may reflect moments when interpersonal tensions are more likely to surface or when users have more time to reflect and share their experiences.
On a broader scale, yearly posting trends fluctuated, hinting at how external events and societal changes may impact the community's activity. Interestingly, the balance between the two main judgment outcomes — "Asshole" and "Not the A-hole" — remained steady throughout, offering a solid foundation for further exploration.

As a result of these patterns, we were able to investigate how engagement timing might relate to the tone of discussions, the nature of moral conflicts presented, and how sentiment within the community evolves over time.

Comment Engagement Patterns Comment activity analysis, segmented by depth (Levels 1–3), shows that deeper engagement (Levels 2 and 3) is not only more variable but also more indicative of sustained discourse. YTA posts tend to accumulate more comments, particularly at deeper levels, suggesting that these posts spark extended debate and engagement. Interestingly, while Level 1 comments are more consistent in volume, deeper threads tend to expand significantly in both comment count and length, particularly under YTA posts. This suggests that morally contentious posts provoke more elaborate discussions and may elicit stronger community responses. On average, comment lengths follow a similar trend: YTA posts attract longer responses, even at the same depth, implying that users invest more time and detail when responding to morally ambiguous or controversial content.

Sentiment Score Dynamics Sentiment analysis using Afinn, Vader, NRClex, and Bert models provides a multi-dimensional view of emotional tone over time. Across all models, an inverse relation between 'Asshole and 'Not the A-hole' scores emerges. When the sentiment for one rises, the sentiment for the other tends to fall. The interpretation of this trend is difficult and it would require a much deeper analysis.

Overall Implications Taken together, these temporal insights contribute to a deeper understanding of how moral discourse unfolds and evolves in online communities. The time-based patterns in posting and engagement suggest that external social rhythms (e.g., workweeks, holidays, global events) influence both when users seek judgment and how the community responds. The disparity in sentiment and comment behavior between YTA and NTA posts highlights the community's greater emotional and conversational investment in morally ambiguous or negatively judged scenarios. These findings offer a rich framework for further inquiry into digital moral evaluation, public empathy dynamics, and the sociology of online judgment spaces.

6 Conclusion

In this project we explored multiple facets of the AmITheAsshole subreddit, from community topology to temporal patters, passing through the sentiment expressed within its discussions. We successfully ran the whole pipeline of data science: we gathered and processed data, deriving and interpreting insights from them. The whole process has been an invaluable learning experience.