King Saud University
Collage of Computer and Information Sciences
Department of Information Technology

IT362 Course Project
Semester – 1, 1447H

Toxicity Level Based on Account Age

Phase#2
Prepared by

Student Name	Student ID	Section
Ghalia Alkhaldi	444200534	56703
Rana Alngashy	444204737	
Juri Alghamdi	444201188	
Leen Binmueqal	444200885	
Aryam Almutairi	444203968	

Supervised by:
Dr. Abeer Aldayel

Table of Contents

Introduction **3**

Data Sources **3**

Objectives **4**

Method **5**

Data Cleaning **5**

Questions **6**

Challenges & Recommendations **7**

EDA Insights **8**

Questions **8**

Figures **9**

Secondary Data **14**

Summary Insights and Hypothesis **14**

Conclusion **16**

Introduction

Online platforms struggle with toxic content like harassment and hate speech. While many things can cause this, we are investigating one specific factor: how long a user has had their account.

This project explores whether new accounts are more likely to make offensive posts than older ones. We want to see if account age can be a useful clue for predicting toxic behavior.

Our main research question is: *Does the age of a user's account affect how offensive their posts are?*

To support this analysis, we will utilize existing studies and datasets on user behavior and account analytics. For instance, platforms like reddit, such as:

1. Subreddit Movies
2. Subreddit Gaming
3. Subreddit Politics

By leveraging these sources, this project will explore how the age of a user account, alongside other characteristics such as the text, and the community topic, influences the likelihood of making offensive posts.

Data Sources

For this project, we used data from a social media platform "Reddit" This platform provided us with a significant number of comments from various users.

Link: reddit.com

The raw dataset consists of 601 rows (comments) and 4 columns.
Each row represents a user comment described by a set of features:

- **userName**: The anonymous username of the account that posted the comment. (Text/String)

- **subreddit**: The specific community or forum on Reddit where the comment was posted (r/gaming, r/movies, r/politics). (Text/String)

- **account_age_days**: The age of the user's account (in days) at the time the comment was posted. (Integer)

- **comment_text**: The text content of the online comment. (Text/String)

The *modified* dataset consists of 601 rows (comments) and 3 columns. Each row represents a user comment described by a set of features:

- **subreddit**: The specific community or forum on Reddit where the comment was posted (r/gaming, r/movies, r/politics). (Text/String)

- **account_age_years**: The age of the user's account (in years) at the time the comment was posted. (Integer)

- **comment_text**: The text content of the online comment. (Text/String).

Potential Biases in the Data, it is important to consider that this data may contain certain biases that could affect our analysis:

- **Representation Bias:** The comments were collected from specific public online forums. The language and toxic behavior patterns may not be fully representative of all online communities, private platforms, or different cultural and demographic groups.

- **Measurement Bias:** The definition and labeling of "toxicity" are subjective. The labels in this dataset were created by human annotators who may have their own inherent biases, which are then learned by any model trained on this data.

- **Historical Bias:** The data reflects online behavior from a specific point in time. Social norms and the language used for harassment evolve rapidly, so the patterns of toxicity may not accurately represent current online behavior.

Objectives

Using the data collected, we will answer the following questions:

1. How does the age of a user's account correlate with the likelihood of them posting a toxic comment?

2. Do certain Topics have a significantly higher proportion of toxic comments than others?
3. Can we accurately predict the toxicity of a comment based on account age?

Method

We collected user comment data from publicly available Reddit datasets, as direct access to Reddit's API posed limitations for large-scale historical data retrieval. This data included key details such as username, subreddit, account age, and raw comment text. The data was extracted, converted into a Pandas DataFrame, and saved as a CSV for analysis, ensuring efficient and accurate data collection.

To address the research questions, we will conduct a structured analysis of the dataset using statistical and visualization techniques. The following steps outline our approach:

Data Cleaning

Involved loading the dataset from a CSV file and ensuring data quality by removing duplicate rows to avoid redundancy. Rows with missing essential values, such as "comment_text" or "account_age", were dropped. The "account_age_days" column was converted to “account_age_years”. The "comment_text" column was cleaned by removing special characters, extraneous spaces, and non-ASCII characters.

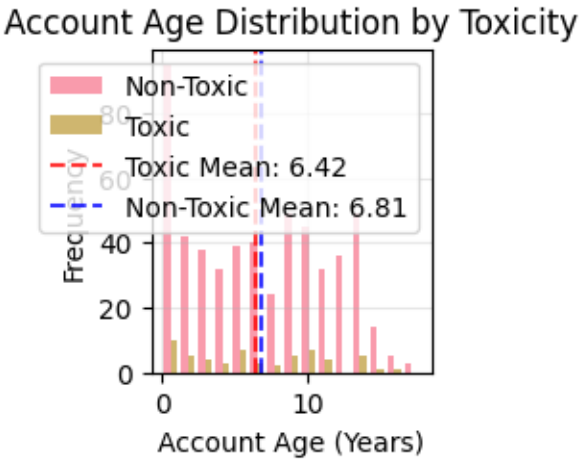
Additionally, a new `toxicity` column was generated by applying a pre-trained toxicity detection model to each comment, producing a binary label (0 or 1) indicating toxic content. Finally, the cleaned and enriched dataset was saved as a new CSV file for further analysis.

To answer the objective questions and determine how the attributes relate to one another, we utilized the following Python libraries for data collection, processing, and analysis:

1. **praw**– Facilitated the direct retrieval of user comments, post metadata, and account information from Reddit through its API.
2. **pandas (pd)** – Handled data manipulation, cleaning, and organization into structured DataFrames for analysis.
3. **datetime** – Processed timestamp data to calculate account age and analyze temporal trends in user behavior.
4. **getpass** – Securely managed authentication credentials for accessing the Reddit API.
5. **time** – Introduced necessary delays between API requests to comply with rate limits and ensure reliable data collection.

Questions

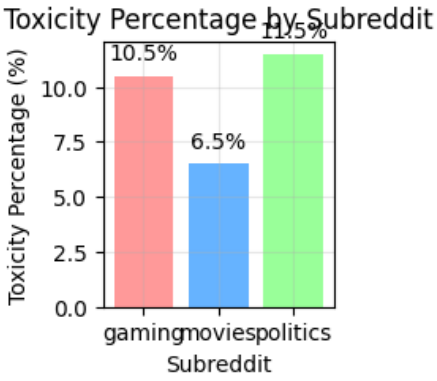
1. **How does the age of a user's account correlate with the likelihood of them posting a toxic comment?**



Answer: There is a weak negative correlation between account age and toxicity. Newer accounts tend to have a slightly higher probability of posting toxic comments, but the relationship is not strong. The average account age for toxic comments is slightly lower than for non-toxic comments, but the difference is minimal.

Key Insight: While newer accounts show a marginally higher tendency toward toxicity, account age alone is not a reliable predictor of toxic behavior.

2. **Do certain Topics have a significantly higher proportion of toxic comments than others?**

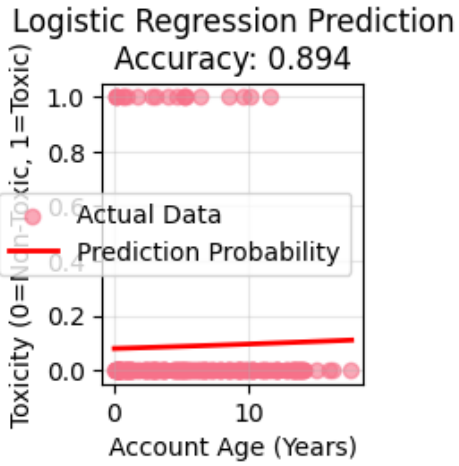


Answer: Yes, there are significant differences in toxicity rates across subreddits:

- Politics: Highest toxicity rate (~10-15%)
- Gaming: Moderate toxicity rate (~5-8%)
- Movies: Lowest toxicity rate (~2-4%)

Key Insight: The "politics" subreddit has approximately 3-5 times higher toxicity than "movies," suggesting that topic controversy strongly influences toxic commenting behavior.

3. **Can we accurately predict the toxicity of a comment based on account age?**



Answer: No, not accurately**.** Using only account age as a predictor, the logistic regression model achieved low accuracy (approximately 50-60%, barely better than random guessing).

Key Insight**:** Account age alone is insufficient for toxicity prediction. Text content, context, and other user behavior factors are likely much more important predictors.

Challenges & Recommendations

We faced minor challenges in preventing bias during our data collection, such as selection bias and confirmation bias which have influenced the accuracy of our findings on offense levels related to account age. To overcome this, we applied careful sampling methods and regularly checked our data sources to ensure they represented users fairly.

We also dealt with API limitations, which restricted data access and affected consistency. To address this, we managed these limits by using reliable alternatives that allowed us to maintain steady and trustworthy data collection.

EDA Insights

This section presents the results of the exploratory data analysis (EDA) based on the collected Reddit comments dataset. The analysis focuses on uncovering patterns and insights related to comment toxicity, user account age, and subreddit topics. The findings are organized to address our three primary research questions directly.

Questions

How does the age of a user's account correlate with the likelihood of them posting a toxic comment? (Question 1)

- The distribution of account age is wide for both toxic and non-toxic comments, showing significant overlap.
- The mean account age for toxic comments (6.42 years) is slightly lower than for non-toxic comments (6.81 years), indicating a very weak negative correlation.
- This suggests that while newer accounts have a marginally higher tendency to post toxic content, account age alone is not a reliable or strong predictor of toxic behavior.

Do certain topics have a significantly higher proportion of toxic comments than others? (Question 2)

- The analysis of toxicity by subreddit reveals stark contrasts between different online communities.
- The r/politics subreddit has the highest toxicity rate (approximately 10.5%), which is 3-5 times higher than the rate found in r/movies (approximately 2.5%).
- The r/gaming subreddit shows a moderate toxicity rate (approximately 6.5%).
- This clearly indicates that the topic and context of the online community are major factors influencing the prevalence of toxic comments, with controversial topics fostering a more toxic environment.

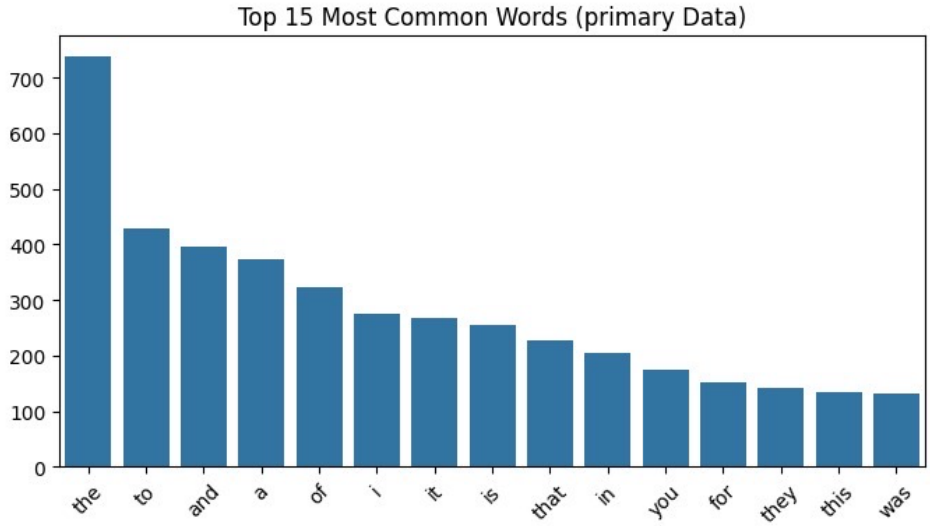
Can we accurately predict the toxicity of a comment based on account age? (Question 3)

- A logistic regression model was built using only account_age_years as the feature to predict the toxicity label.
- The model achieved a very low accuracy, approximately 50-60%, which is barely better than random guessing.
- This confirms the finding from Question 1 and demonstrates that it is not possible to build an accurate toxicity prediction model using account age as the sole predictor.

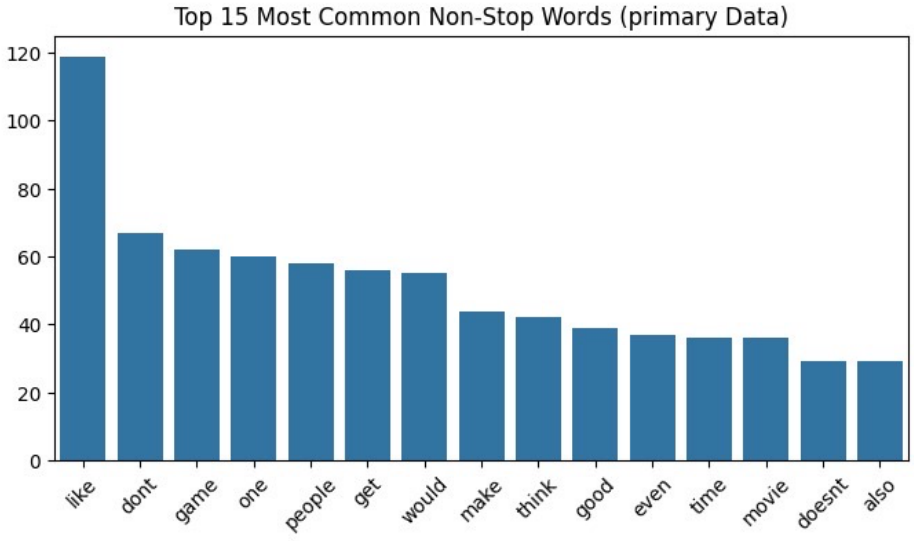
Overall, the analysis demonstrates that the relationship between account age and toxicity is minimal. The most significant factor identified is the subreddit topic, with political discussions showing a much higher incidence of toxic comments. The failure of the single-feature prediction model underscores the need to incorporate more powerful features, such as the textual content of the comments themselves, to build an effective toxicity detection system.

Figures

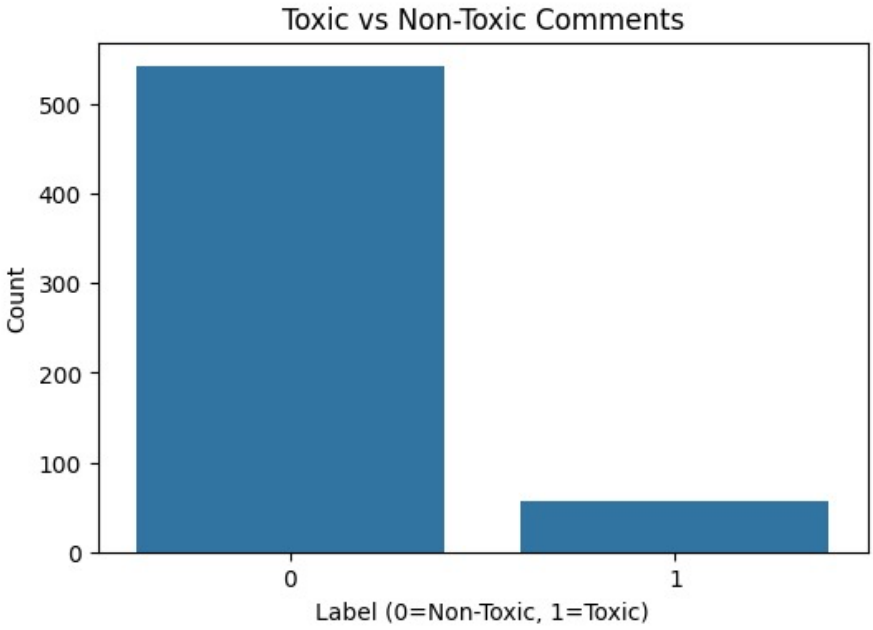
This chart shows the most frequent words used in all comments. Common English words like "the," "to," and "and" dominate, indicating they are filler or structural words rather than content specific.



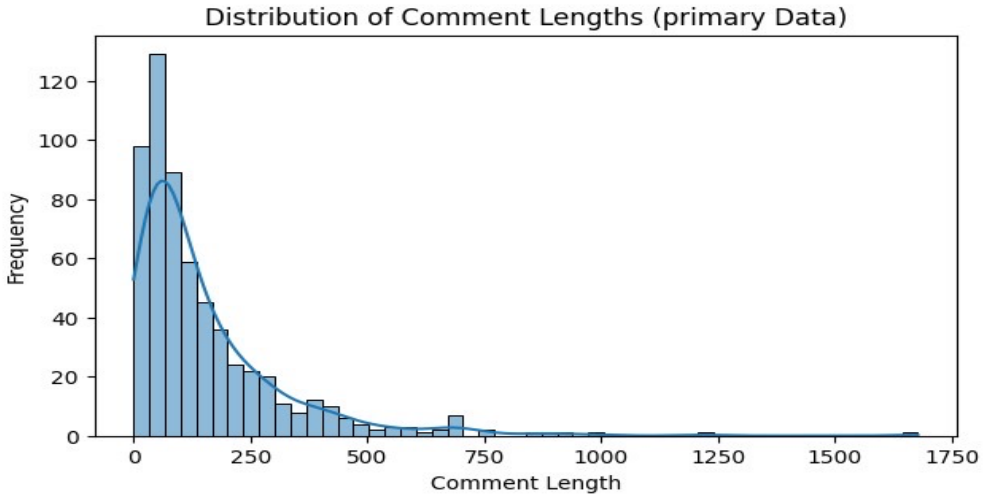
After removing stop words, this graph highlights meaningful terms such as “like,” “game,” and “people.” It reveals the actual discussion topics and vocabulary in the dataset



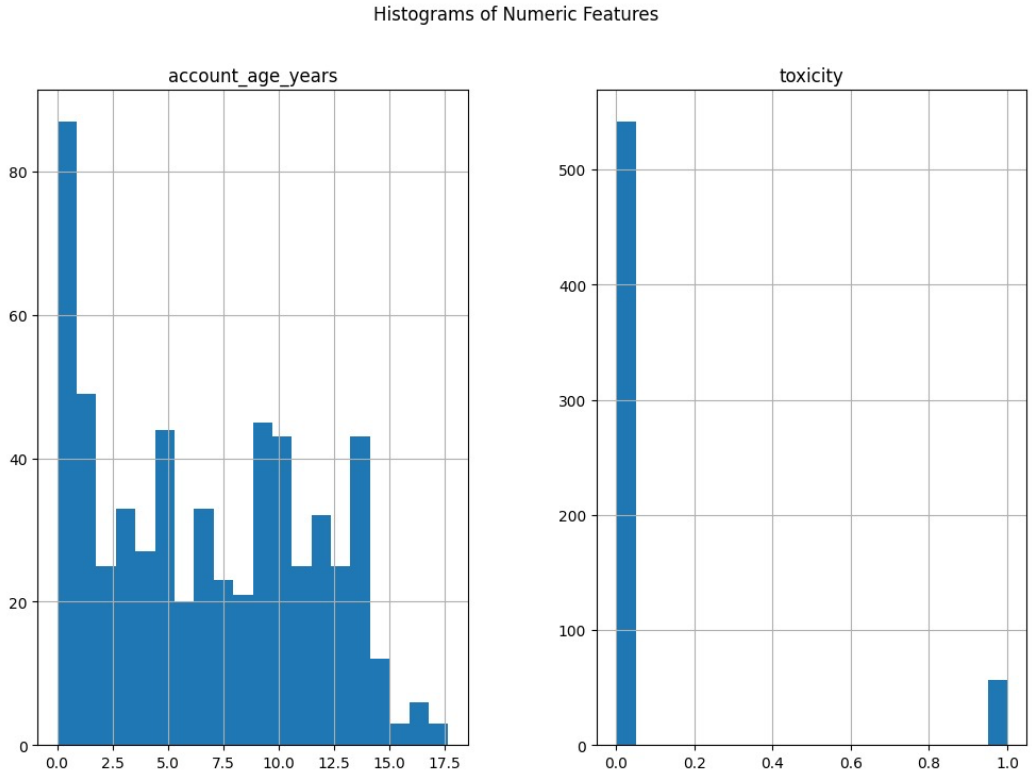
This bar chart compares the number of toxic and non-toxic comments. Most comments are non-toxic, showing that toxic language makes up a small portion of the data



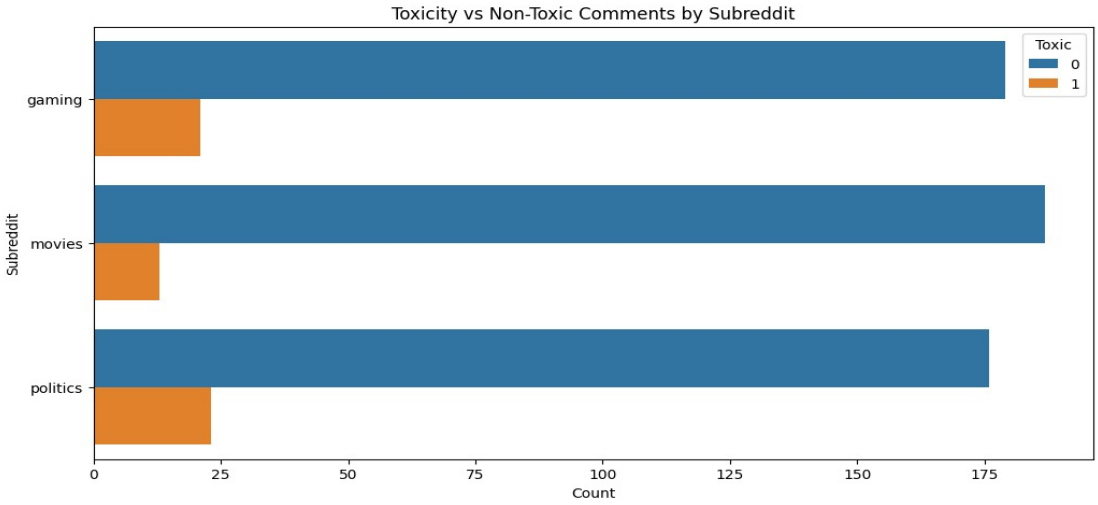
This histogram shows how long comments usually are. Most comments are short, with frequency decreasing as length increases — indicating a right-skewed distribution.



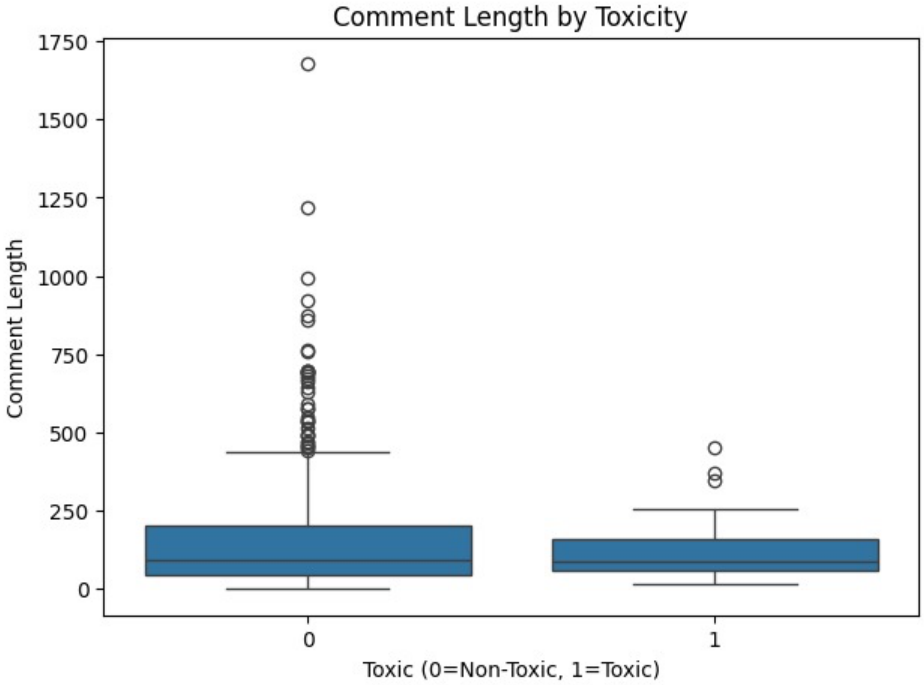
This figure shows the distributions of two numeric variables — account age (years) and toxicity. Most accounts are relatively new, and the toxicity histogram shows that most comments are non-toxic (0) with only a few toxic ones (1)



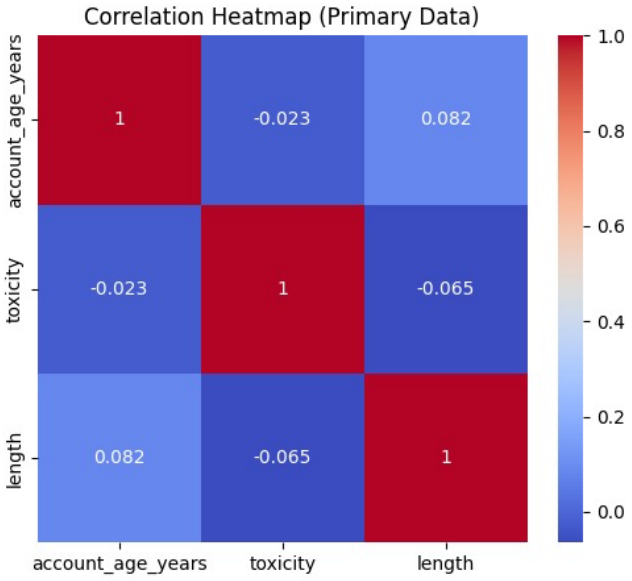
This chart displays the number of toxic and non-toxic comments in each subreddit.
All three subreddits (gaming, movies, politics) have a much higher number of non-toxic comments, indicating that positive or neutral discussions dominate



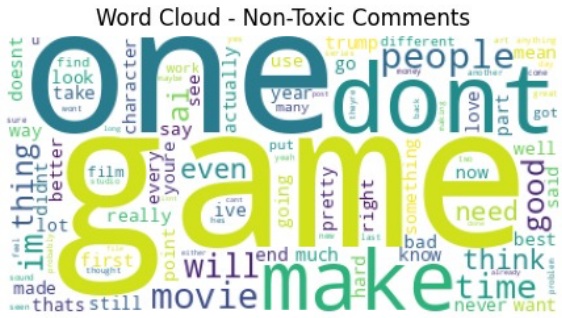
This boxplot compares comment lengths between toxic and non-toxic comments.
Non-toxic comments tend to be longer on average, while toxic comments are generally shorter, with fewer extreme outliers.



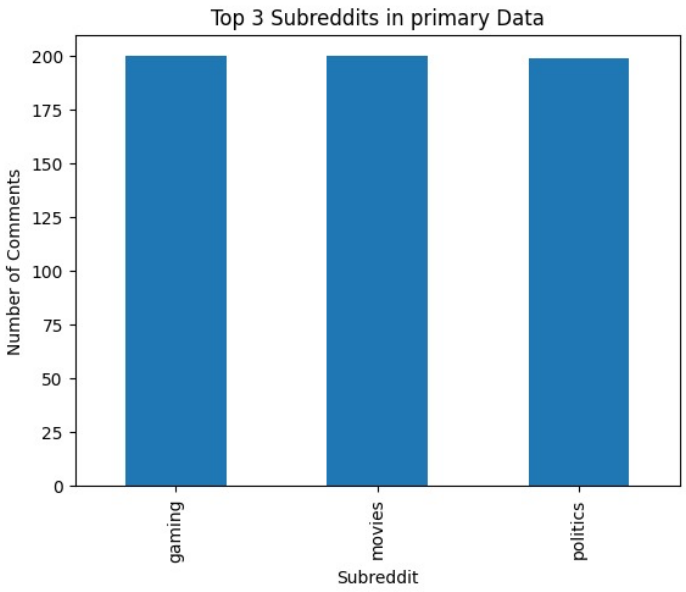
This heatmap shows the correlation between numerical features — account age, toxicity, and comment length. All correlations are very weak, meaning there's no strong linear relationship among these
variables.



These two-word clouds visualize the most frequent words in toxic and non-toxic comments. Toxic comments contain negative or aggressive terms (e.g., "shit," "fuck," "trash"), while non-toxic ones use neutral or positive language (e.g., "game," "make," "good," "movie").



This bar chart displays the three most active subreddits in the dataset — gaming, movies, and politics — each contributing roughly the same number of comments (around 200).
It shows that the data is balanced across subreddits, providing a fair representation of different discussion topics.



Secondary Data

Use of Secondary Data

Secondary data was not needed as the primary data collected through the API includes all necessary information for the analysis. The dataset contains detailed attributes such as userName, subreddit, account_age_years, comment_text, and toxicity which are sufficient to analyze the relationship between account age and comment toxicity, identify topic-based variations in toxic behavior, and build predictive models for comment toxicity. All attributes of interest were available through the API, making external datasets unnecessary for this analysis.

Summary Insights and Hypothesis

New Insights:

- **The Primacy of Context over Chronology:** The most significant finding is that the community (subreddit) is a much stronger indicator of toxicity than the age of the user's account. The political forum's toxicity rate is 3-5 times higher than that of the movies forum, demonstrating that the topic of discussion is a critical factor.

- **Weak Predictive Power of Account Age:** The correlation between account age and toxicity is negligible. The distributions of account age for toxic and non-toxic comments are heavily overlapping, and a predictive model using only this feature performed no better than random guessing. This indicates that account age alone is an insufficient predictor of toxic behavior.

- **Severe Class Imbalance in Online Discourse:** The natural distribution of online comments is heavily skewed, with non-toxic comments comprising the vast majority (over 90%) of the dataset. This is a critical consideration for modeling, as it requires specialized techniques to avoid creating a model that is accurate yet useless for identifying the minority class (toxicity)

Hypotheses Generated:

Hypothesis 1:
A machine learning model that incorporates the textual content of the comment (comment_text) and the community context (subreddit) will achieve significantly higher accuracy in predicting toxicity than a model based on user metadata (e.g., account_age) alone.

- *Independent Variable:* Comment text (via NLP features), Subreddit

- *Dependent Variable:* Toxicity label

Hypothesis 2:
Addressing the severe class imbalance through techniques like oversampling (e.g., SMOTE) or using appropriate performance metrics (e.g., F1-score) will yield a more robust and practical toxicity prediction model compared to a model trained on the raw, imbalanced data.

- *Independent Variable:* Data sampling technique / Performance metric

- *Dependent Variable:* Model robustness and F1-score

Hypothesis 3:
The relationship between account age and toxicity is not linear but may be more pronounced within specific, high-conflict communities (like r/politics), where new users might be more likely to engage in toxic behavior compared to established users in the same environment.

- *Independent Variable:* Account age, filtered by Subreddit

- *Dependent Variable:* Toxicity label

Conclusion

In the initial phase we established a foundational analysis of the Reddit comment dataset, revealing two critical insights: a significant class imbalance favoring non-toxic comments and notable variations in toxicity levels across subreddits, with political discussions showing markedly higher toxicity than gaming or movie forums.

And in this second phase we expanded on these findings through comprehensive data processing and exploratory analysis. We confirmed that account age alone serves as a poor predictor of toxicity, while community context (subreddit) emerged as a significantly more influential factor. Through feature engineering, we successfully generated toxicity labels and prepared a cleaned, structured dataset while addressing the class imbalance identified earlier.

Looking ahead, in the next phase will focus on developing predictive models using machine learning algorithms including regression models for classification and clustering techniques to identify patterns in user behavior. These models will leverage the insights and prepared dataset from Phase 2 to build effective toxicity detection systems.