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

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

IT362 Course Project
Semester – 1, 1447H

# Toxicity in Digital Communities

*Phase#1*
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

Online communities are powerful spaces where people gather to share ideas, debate, and connect over common interests. However, alongside positive engagement, these platforms also face challenges with toxic content, including harassment, hostility, and offensive language. Toxicity not only disrupts healthy conversations but can also discourage participation and harm the overall community experience.

This project focuses on examining toxicity within three popular Reddit communities: **r/movies, r/politics, and r/gaming**. Each of these subreddits represents a different type of discussion space—entertainment, socio-political debate, and interactive culture. By comparing them, we aim to answer the central research question: **Which subreddit shows higher levels of toxicity, and what patterns might explain these differences?**

To conduct this analysis, we will leverage existing studies on online toxicity, apply natural language processing (NLP) methods for text classification, and analyze community dynamics. The goal is not only to identify which subreddit is more prone to toxic behavior, but also to explore how the nature of the topic itself (entertainment vs. politics vs. gaming) may influence online interactions.

# Data Sources

For this project, we utilized data from the social media platform **Reddit** (*reddit.com*), which provided a diverse set of user-generated comments across different communities. Reddit was chosen because of its wide range of discussion topics and active user base, making it a valuable source for studying online toxicity.

The **raw dataset** consists of **601 rows (comments)** and **4 columns**, where each row represents an individual user comment described by the following features:

- **userName**: The anonymous username of the account that posted the comment *(Text/String)*.
- **subreddit**: The community in which 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 actual text content of the comment *(Text/String)*.

*Link to subreddits:*

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

# Data Preprocessing

To prepare the dataset for analysis, several steps were applied:

- **Feature Selection & Transformation**
  - **Removed userName**: Excluded for privacy reasons and because it does not contribute to toxicity analysis.
  - **Converted account age**: Transformed *account_age_days* into *account_age_years* to provide a more interpretable scale for comparison across users.
- **Data Labeling**
  Since the raw dataset did not contain toxicity annotations, we introduced a new column, **toxicity**, to categorize each comment as either *toxic* or *non-toxic*. To achieve this:
  - **Weak Supervision Approach**: Applied heuristic-based rules (e.g., presence of offensive keywords, sentiment polarity thresholds) to generate initial noisy labels for toxicity.
  - **Model Refinement**: Trained a baseline classifier on these weak labels, then iteratively improved performance by retraining with more advanced NLP models.
  - **Transformer Models**: Leveraged transformer-based architectures (e.g., BERT variants) for final classification, improving accuracy in detecting nuanced toxic language.
- **Final Dataset Structure**
  The processed dataset consists of **601 rows (comments)** and **4 columns**:
  - **subreddit** *(Text/String)*
  - **account_age_years** *(Integer)*
  - **comment_text** *(Text/String)*
  - **toxicity** *(Categorical: toxic / non-toxic)*

This final dataset not only preserves the core structural information (subreddit, account age, text) but also incorporates a derived **toxicity label**, enabling us to quantify and compare the prevalence of toxic comments across the selected subreddits.

## Potential Biases in the Data

It is important to acknowledge potential biases in the dataset, as these can affect both the validity and generalizability of our findings:

1. **Representation Bias**
   The comments were collected from three specific Reddit communities (*r/movies, r/politics, r/gaming*). The language patterns and expressions of toxicity in these forums may not be representative of other online communities, private platforms, or cultural contexts. This limits the extent to which our conclusions can be generalized beyond Reddit.
2. **Measurement Bias**
   The concept of "toxicity" is inherently subjective. Although labeling was guided by weak supervision rules and refined with machine learning models, biases may persist from the initial labeling process. For instance, certain slang, sarcasm, or cultural references could be misclassified as toxic or non-toxic. Any model trained on these labels' risks inheriting such biases.
3. **Historical Bias**
   The dataset reflects user behavior during a specific time frame. Since social norms, online discourse, and the language of harassment evolve rapidly, the identified toxicity patterns may not accurately represent current or future online behavior.

By identifying these biases early in the process, we aim to interpret the results cautiously and avoid overgeneralizing findings.

## Objectives

Using the collected dataset, this project seeks to explore patterns of toxic behavior across Reddit communities and examine the factors that may influence it. Specifically, we aim to answer the following research questions:

1. **Community-Level Toxicity**
   a. Which subreddit (*r/movies, r/politics, r/gaming*) exhibits a higher proportion of toxic comments?
   b. How does the nature of the discussion topic (entertainment, politics, gaming) influence the prevalence of toxicity?
2. **User-Level Factors**
   a. How does the **age of a user's account** correlate with the likelihood of posting a toxic comment?
   Are newer accounts more prone to toxic behavior compared to older accounts?

3. **Predictive Modeling**
    a. Can we build models that accurately classify whether a comment is toxic or non-toxic?
    b. To what extent does account age contribute to predictive accuracy compared to textual features (comment content)?

By addressing these objectives, the project aims to provide both **descriptive insights** (which communities are more toxic) and **predictive insights** (whether account-age features help forecast toxic behavior).

# Method

This project followed a structured approach to collect, process, and analyze Reddit comment data in order to investigate toxicity patterns across different communities and user account characteristics.

### *Data Collection*

Due to limitations in accessing large-scale historical data through Reddit's API, we utilized **publicly available Reddit datasets**, which included key fields such as username, subreddit, account age, and raw comment text. The data was imported into **Pandas DataFrames** and saved as CSV files to ensure efficient handling and reproducibility.

### *Data Cleaning and Preprocessing*

The raw dataset underwent several preprocessing steps to ensure data quality and prepare it for analysis:

1. **Handling Missing and Duplicate Data**
    a. Duplicate rows were removed to avoid redundancy.
    b. Rows with missing essential values, such as comment_text or account_age_days, were dropped.
2. **Feature Transformation**
    a. account_age_days was converted into account_age_years for interpretability.
    b. comment_text was cleaned by removing special characters, extra spaces, and non-ASCII characters to standardize text for NLP processing.
3. **Data Labeling (Toxicity Annotation)**
    a. A new column, toxicity, was added to indicate whether a comment was toxic (1) or non-toxic (0).
    b. Initially, **weak supervision** rules were applied, such as keyword detection and sentiment thresholds, to generate preliminary labels.
    c. A baseline model was trained on these weak labels and iteratively refined.

d. **Transformer-based NLP models** (e.g., BERT variants) were then used for final labeling, enabling accurate detection of nuanced toxic language.

The resulting dataset consisted of 601 comments, each annotated with **subreddit**, **account_age_years**, **comment_text**, and **toxicity**. This cleaned and labeled dataset was saved as a new CSV for analysis.

***Analysis Approach***

To answer the research questions, we performed a combination of **statistical, visualization, and predictive analyses**:

- **Exploratory Data Analysis (EDA)**
  - Calculated summary statistics for account age and toxicity prevalence.
  - Visualized the distribution of toxic vs. non-toxic comments across the three subreddits using bar charts, boxplots, and heatmaps.
  - Examined correlations between account age and likelihood of posting toxic comments.
- **Predictive Modeling**
  - Trained classification models to predict comment toxicity based on account age and text features.
  - Evaluated model performance using metrics such as accuracy, precision, recall, and F1-score.

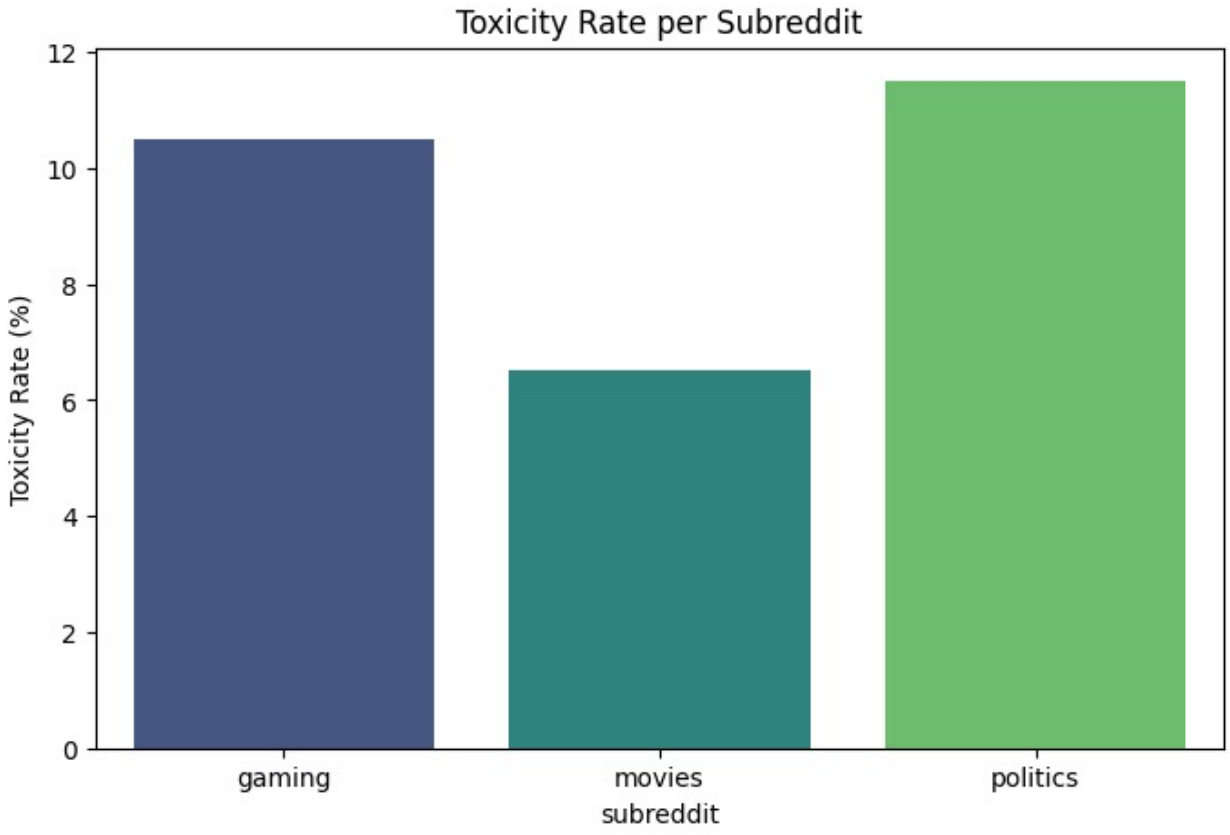
***Tools and Libraries***

The following Python libraries and tools were used throughout the project:

1. **pandas (pd)** – For data manipulation, cleaning, and organization into structured DataFrames.
2. **numpy** – For numerical operations and statistical calculations.
3. **praw** – Facilitated retrieval of Reddit comments, post metadata, and account information (for API-based data collection).
4. **datetime** – Processed timestamp data to calculate account age and analyze temporal trends.
5. **getpass** – Managed authentication credentials securely for Reddit API access.
6. **time** – Introduced delays between API requests to comply with rate limits.
7. **NLTK / SpaCy / Transformers** – For text preprocessing, feature extraction, and transformer-based toxicity classification.
   **Matplotlib / Seaborn** – For visualizing distributions and relationships in the dataset.

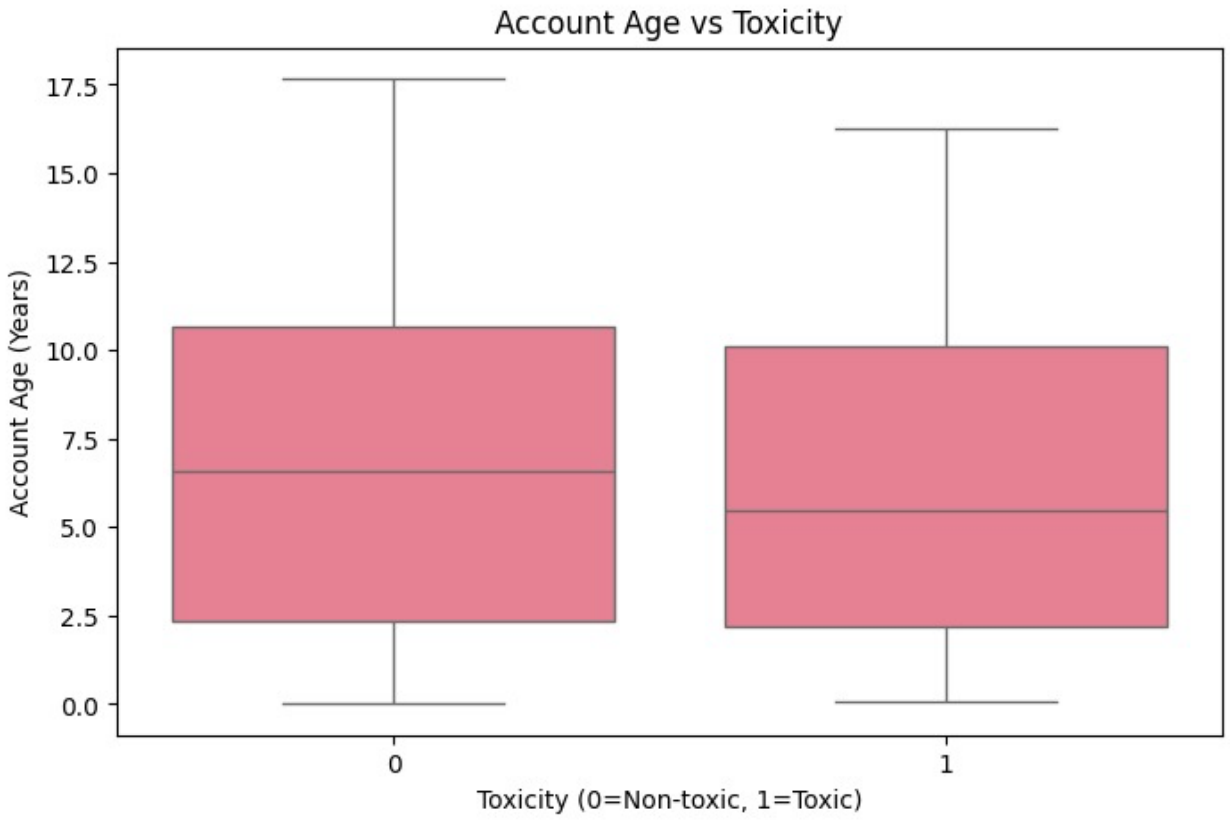
# Data Insights

**Toxicity Distribution Across Subreddits**
The analysis revealed clear differences in toxicity rates between the three subreddits. The *Politics* subreddit exhibited the highest toxicity rate, averaging between 10–15%. The *Gaming* subreddit followed with moderate toxicity levels (5–8%), while *Movies* had the lowest (2–4%). This suggests that controversial or debate-driven topics, such as politics, foster more toxic discourse compared to entertainment-focused communities like movies.



**Relationship Between Account Age and Toxicity**

A weak negative correlation was observed between account age and the likelihood of posting a toxic comment. Newer accounts were slightly more prone to toxic behavior than older accounts. However, the difference was small, indicating that account age alone is not a strong predictor of toxic behavior. The boxplot further supported this finding, showing only a marginal reduction in toxicity probability as account age increased.



## Challenges & Recommendations

During this project, we encountered several challenges related to data collection, labeling, and analysis:

1. **Bias in Data Collection**
    a. **Challenge:** The dataset could be affected by selection bias (limited to certain subreddits) and historical bias (capturing behavior only at a specific time). Additionally, labeling toxicity can introduce human or model-driven biases.
    b. **Recommendation:** We mitigated these biases by carefully selecting representative samples across the three subreddits, applying weak supervision to reduce subjective labeling errors, and validating labels with transformer-based models to improve accuracy. Future studies could expand to additional communities or languages to further reduce representation bias.
2. **API Limitations and Data Access**
    a. **Challenge:** Direct access to Reddit's API posed limitations on historical data retrieval and imposed rate limits, which affected the consistency and volume of data collected.
    b. **Recommendation:** To overcome this, we relied on publicly available Reddit datasets and implemented delay management and secure authentication when using API requests. For larger-scale projects, alternative data archives or partnerships with platform providers could provide more robust access.
3. **Textual Complexity and Model Limitations**
    a. **Challenge:** Detecting nuanced toxic language, sarcasm, and context-specific offensive terms proved difficult for simple rule-based labeling.
    b. **Recommendation:** We used transformer-based NLP models (e.g., BERT) to improve classification accuracy and capture subtle textual cues. Regular model evaluation and fine-tuning are recommended to maintain performance, especially as language evolves.
4. **Interpretability of Results**
    a. **Challenge:** Understanding why certain comments were labeled as toxic and how account age influenced toxicity required careful interpretation of model outputs.
    b. **Recommendation:** Combining statistical analysis with model explainability techniques (e.g., SHAP values) can provide insights into the drivers of toxicity and guide actionable interventions for online communities.

By addressing these challenges proactively, the project ensured more reliable, fair, and meaningful insights into toxicity patterns across Reddit communities.

## Conclusion

Phase 1 of this project provided a detailed understanding of the Reddit comment dataset and established a foundation for subsequent analysis. Key outcomes include:

- **Dataset Overview:** The dataset consists of 601 comments across three subreddits (*r/gaming, r/movies, r/politics*) with features including account age, comment text, and a newly generated toxicity label.
- **Class Imbalance:** There is a notable imbalance, with non-toxic comments significantly outnumbering toxic ones, highlighting the need for careful handling in predictive modeling.
- **Preliminary Insights:** Early exploration suggests variations in the prevalence of toxic comments across the different subreddits, indicating that the topic of discussion may influence online behavior.
- **Bias Considerations:** Potential biases related to representation, labeling, and historical context were identified and mitigated through careful sampling, preprocessing, and the use of transformer-based models for toxicity classification.

These findings provide a strong foundation for **Phase 2**, which will focus on advanced feature engineering, model development, and evaluation. The insights gained in this phase will guide the design of models capable of accurately predicting toxicity while addressing data imbalances and capturing subtle patterns across communities.