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#3

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 5

Method 5

Data Cleaning 6

Questions 6

Challenges & Recommendations 8

EDA Insights 9

Questions 9

Figures 11

Secondary Data 16

Summary Insights and Hypothesis 17

Modeling 18

Regression Models: 19

Comparing models based on regression scores: 22

Clustering Models: 24

Conclusion And Future Work 24

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.

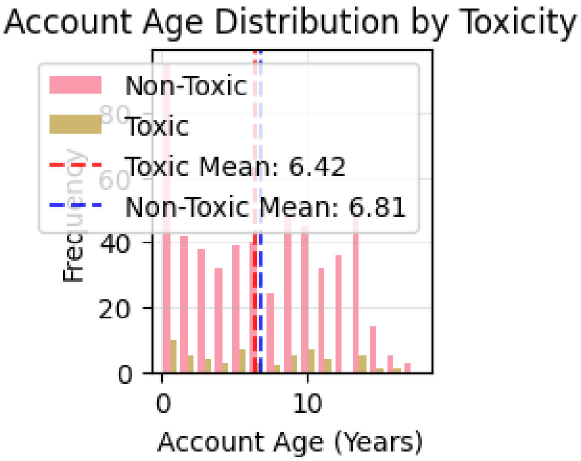
Additionally, a new `toxicity` column was generated by applying a pre-trained toxicity detection model to each comment, producing a binary label (0 or 1) indicating toxic content. Finally, the cleaned and enriched dataset was saved as a new CSV file for further analysis.

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

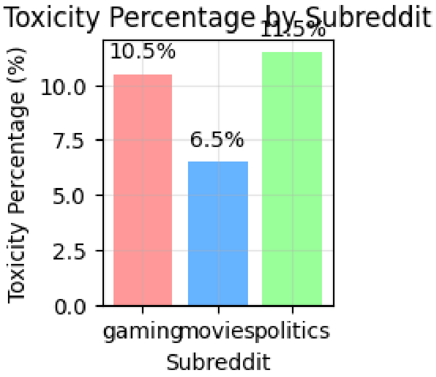
1. **How does the age of a user's account correlate with the likelihood of them posting a toxic comment?**



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?**

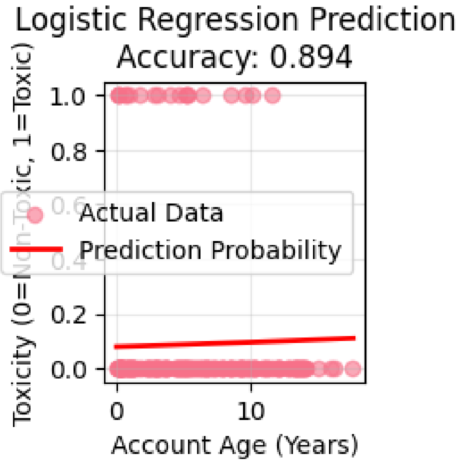


Answer: Yes, there are significant differences in toxicity rates across subreddits:

- 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.

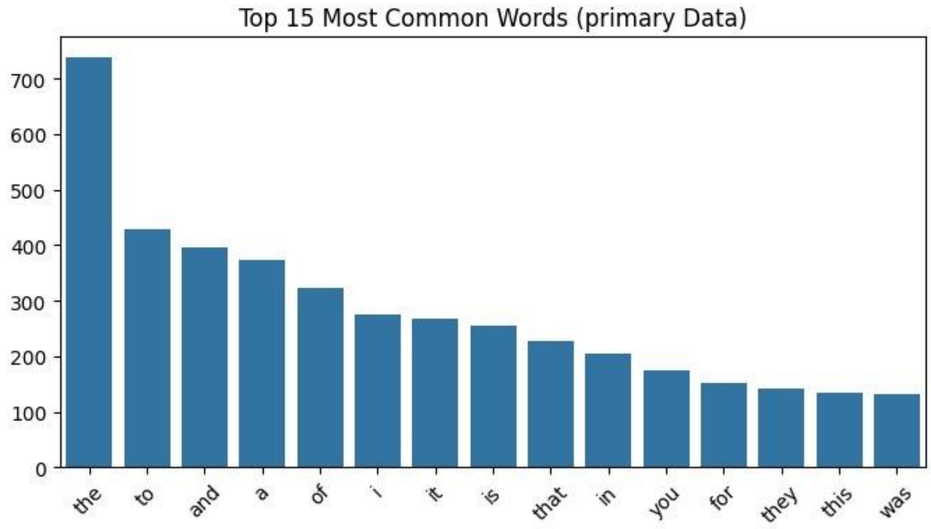
Can we accurately predict the toxicity of a comment based on account age? (Question 3)

- 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.

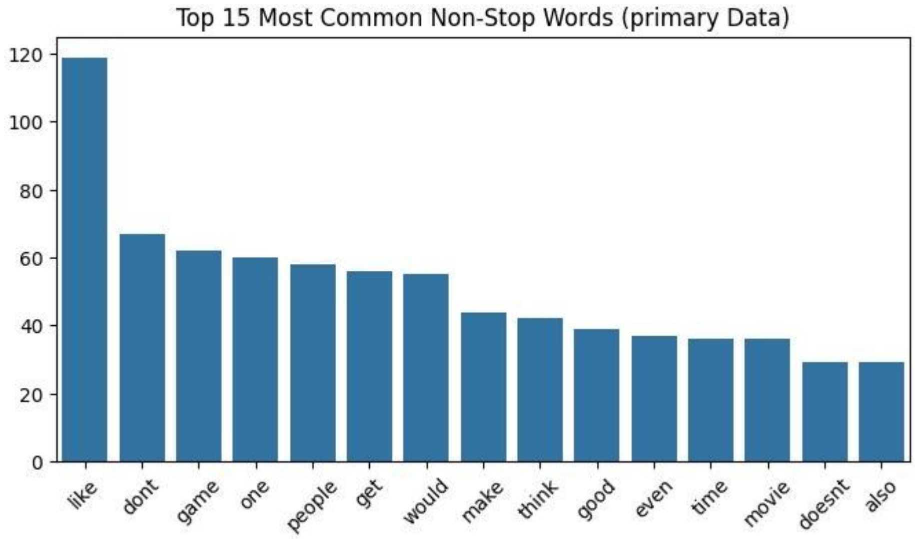
Overall, the analysis demonstrates that the relationship between account age and toxicity is minimal. The most significant factor identified is the subreddit topic, with political discussions showing a much higher incidence of toxic comments. The failure of the single-feature prediction model underscores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

Figures

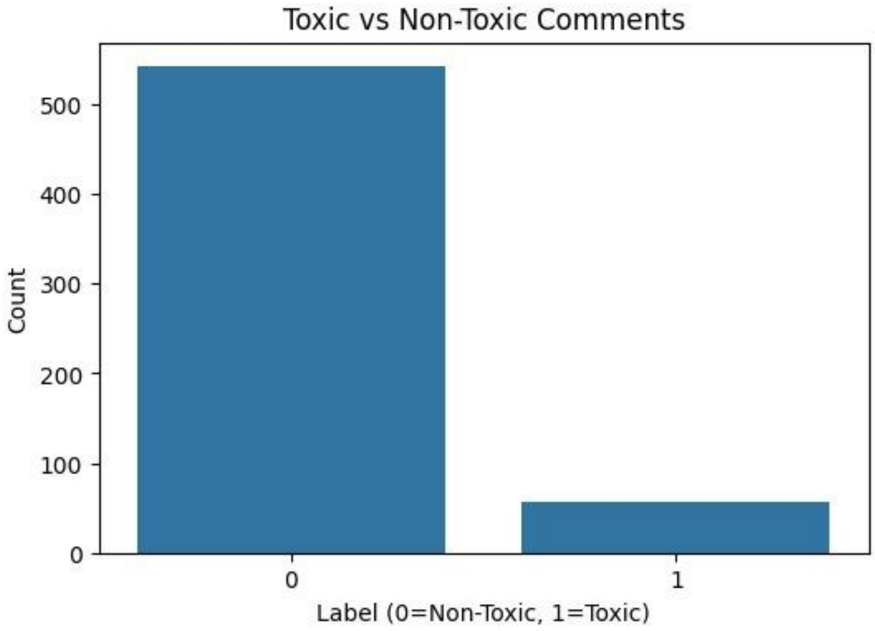
This chart shows the most frequent words used in all comments. Common English words like “the,” “to,” and “and” dominate, indicating they are filler or structural words rather than content specific.



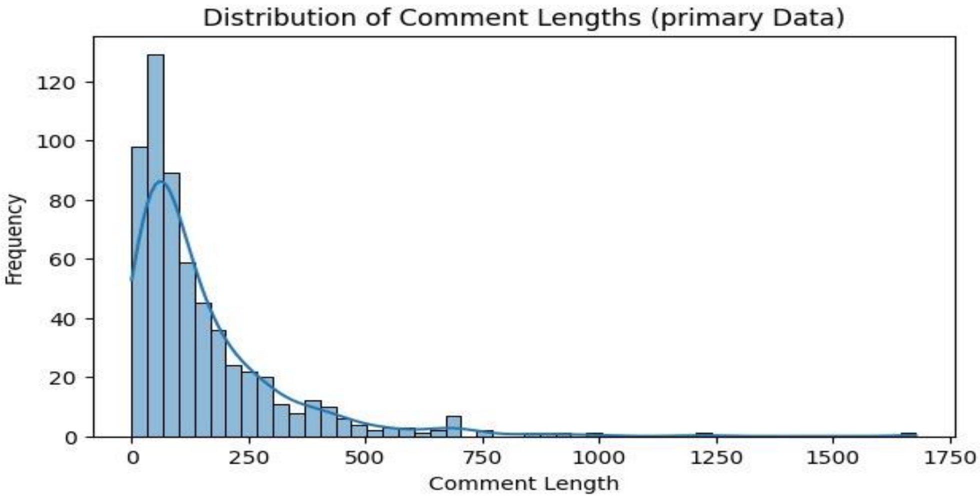
After removing stop words, this graph highlights meaningful terms such as “like,” “game,” and “people.” It reveals the actual discussion topics and vocabulary in the dataset



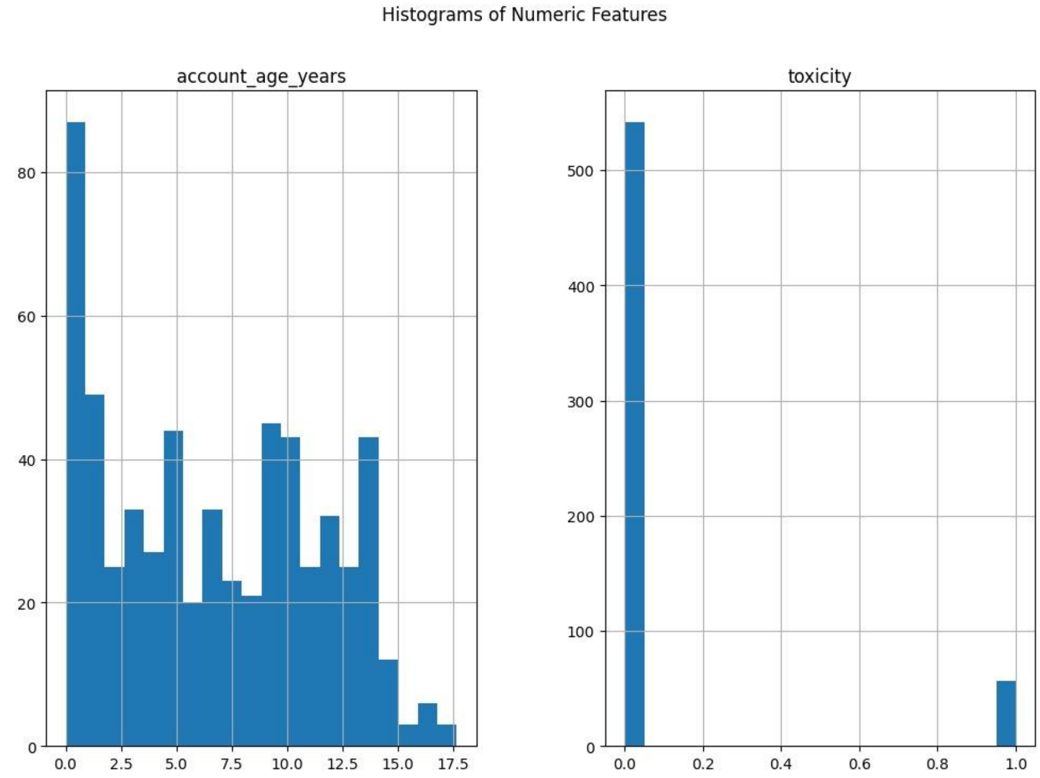
This bar chart compares the number of toxic and non-toxic comments. Most comments are non-toxic, showing that toxic language makes up a small portion of the data



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

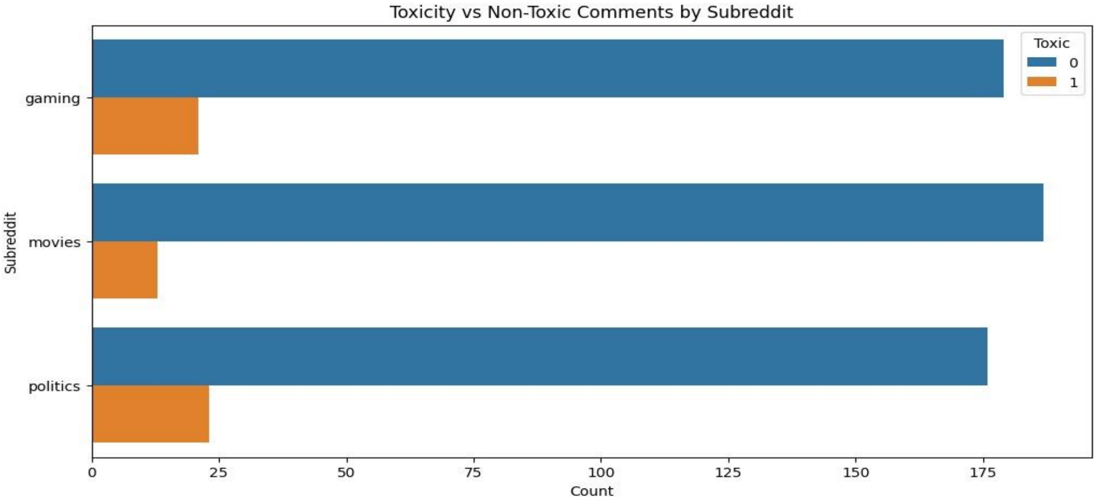


This figure shows the distributions of two numeric variables — account age (years) and toxicity. Most accounts are relatively new, and the toxicity histogram shows that most comments are non-toxic (0) with only a few toxic ones (1)

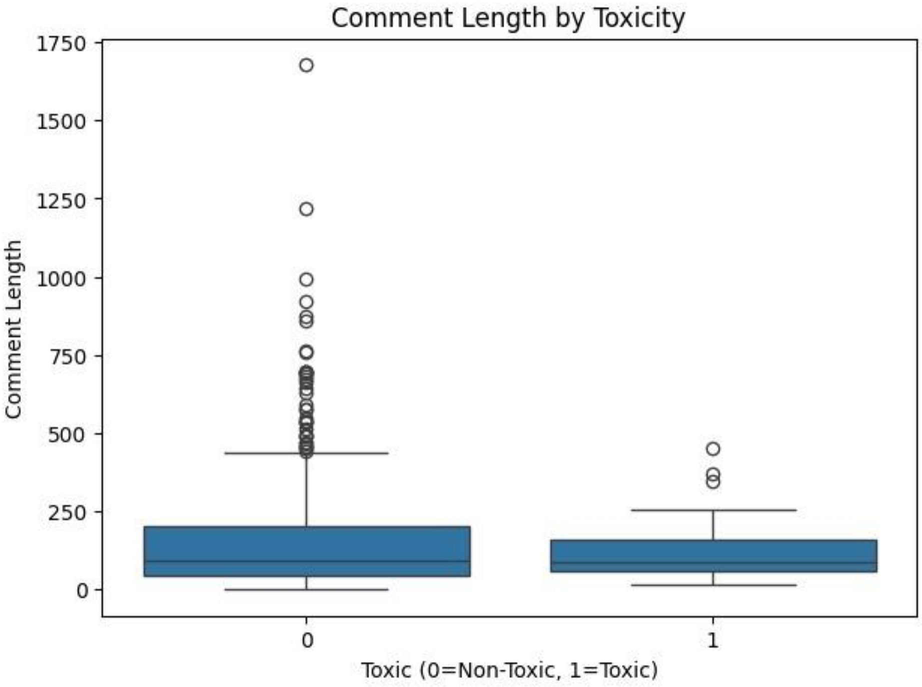


This chart displays the number of toxic and non-toxic comments in each subreddit.

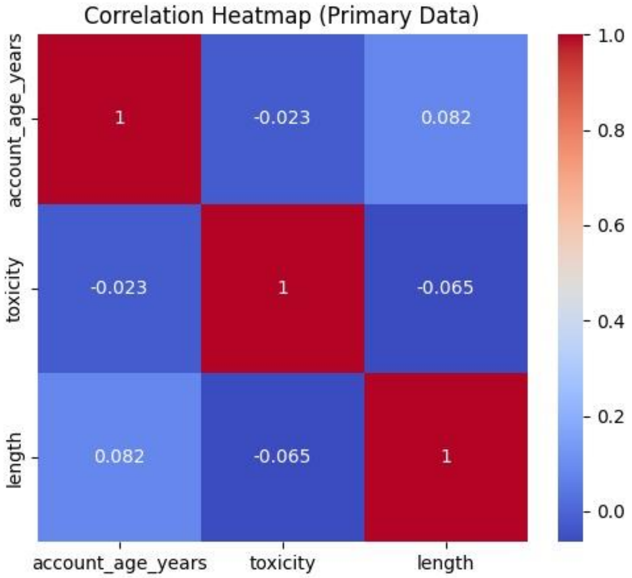
All three subreddits (gaming, movies, politics) have a much higher number of non-toxic comments, indicating that positive or neutral discussions dominate



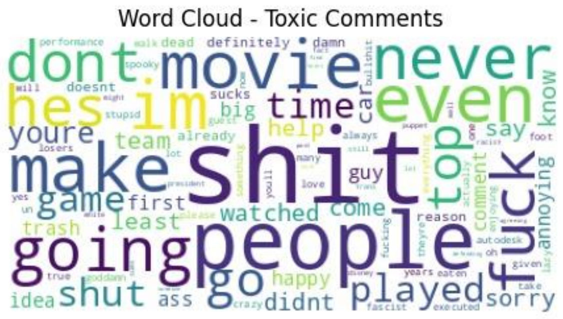
This boxplot compares comment lengths between toxic and non-toxic comments. Non-toxic comments tend to be longer on average, while toxic comments are generally shorter, with fewer extreme outliers.

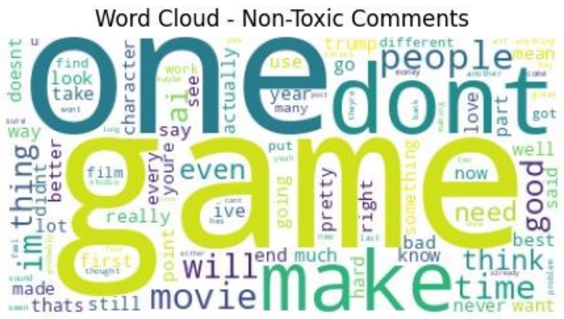


This heatmap shows the correlation between numerical features — account age, toxicity, and comment length. All correlations are very weak, meaning there’s no strong linear relationship among these
variables.

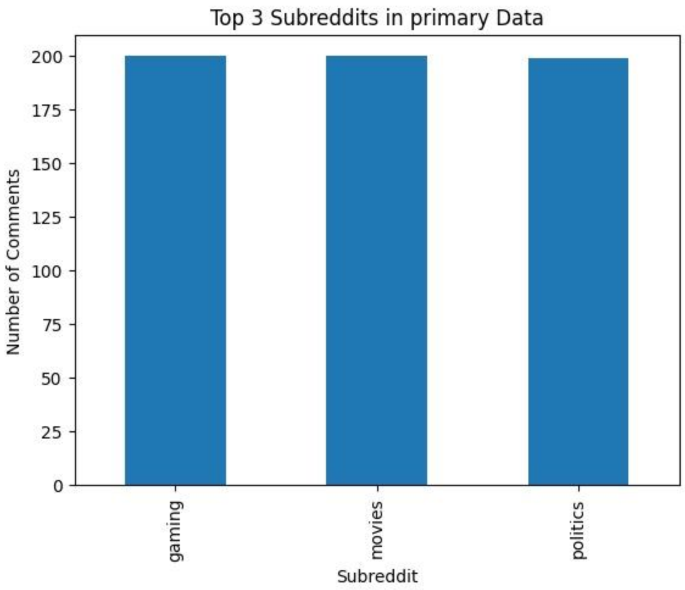


These two-word clouds visualize the most frequent words in toxic and non-toxic comments. Toxic comments contain negative or aggressive terms (e.g., “shit,” “fuck,” “trash”), while non-toxic ones use neutral or positive language (e.g., “game,” “make,” “good,” “movie”).





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

Modeling

At the beginning of our project, our main goal was to understand whether the **age of a user's account** had any influence on their likelihood of posting toxic comments.

This led us to formulate an initial descriptive question:

"How does account age affect the toxicity level of a user's comment?"

This question helped us to:

- Explore the dataset and understand the distribution of account ages.
- Compare toxicity rates between newer and older accounts.
- Examine whether account age could serve as a meaningful indicator of toxic behavior.

However, as we progressed through our exploratory analysis, we discovered that **account age had only a very weak relationship with toxicity**. The distributions of toxic and non-toxic comments overlapped heavily, and early models built using account age alone performed poorly. At the same time, our analysis revealed a **much stronger pattern** related to the context of the comment. Specifically, different subreddits showed **very different toxicity levels**, with political discussions exhibiting significantly higher toxicity than gaming or movie communities.

This insight motivated us to refine our question into a more analytical and meaningful one: **“Which subreddit influences toxicity level the most, and can we predict toxicity using subreddit and comment text?”**

The reason for this shift was:

- Our data clearly showed noticeable differences in toxicity across communities (e.g., politics vs. movies).
- We identified that **subreddit and comment text** carried far stronger predictive power than account age.
- Transitioning from a descriptive question to a **predictive modelling question** allowed us to build stronger machine learning models that leveraged NLP and classification techniques.
- This refined question aligned better with the patterns observed during EDA and allowed us to evaluate the performance of multiple models meaningfully.

Ultimately, this shift strengthened our project by focusing on the factors that truly influence online toxicity rather than those with minimal impact.

To evaluate whether toxic comments can be predicted from user and text attributes, we developed two supervised classification models and compared them to a simple baseline. Our goal is to determine which model performs best on our imbalanced dataset (~10% toxic comments). Since F1-score is the most important metric in imbalanced classification, we used it to select the final model.

Regression Models:

We used the following features:

- **Independent Variables:**
  • Subreddit (categorical)
  • Account age (numeric)
  • Comment text (TF-IDF features)

- **Dependent Variable:**

  • Toxicity (0 = non-toxic, 1 = toxic)

The dataset was split into an **80% training set and 20% testing set** using stratified sampling to preserve the imbalance ratio. After preprocessing and feature engineering, we trained and evaluated three models:

1-Baseline Regression Model:

This model acts as our minimum performance model, taking the least affective factor the account_age in relation with the value of toxicity. This model acts as our benchmark and helps us provide a reference point for evaluating the complexity
and effectiveness of more sophisticated models.

2-Logistic Regression + TF-IDF:

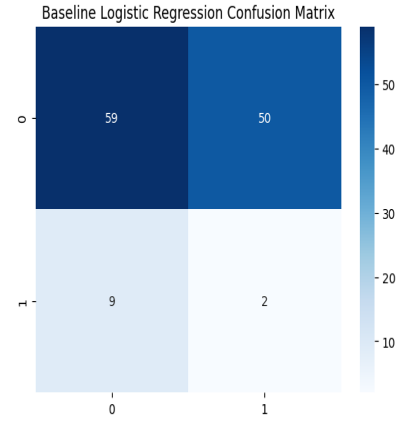
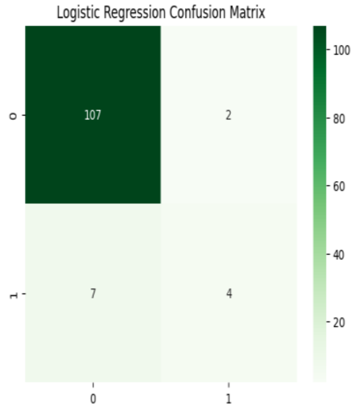
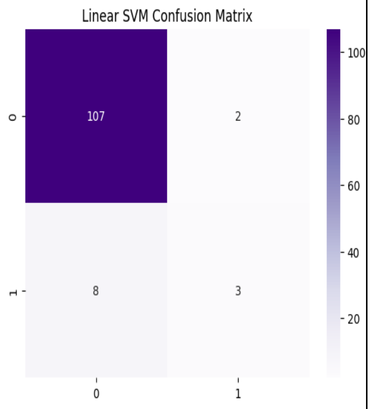
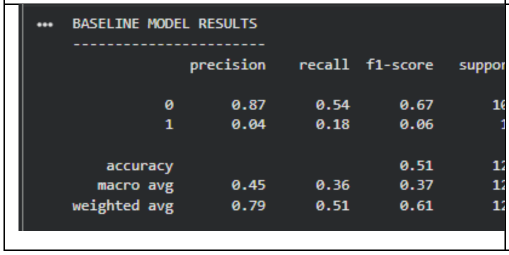
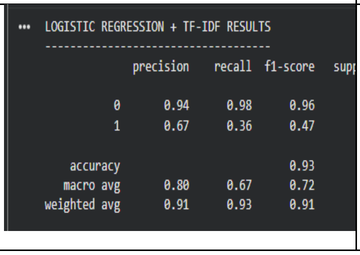
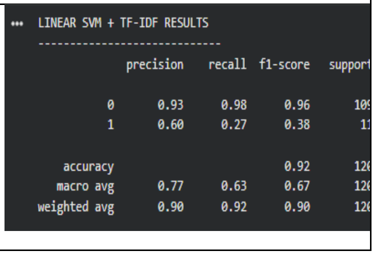
This model combines Logistic Regression with TF-IDF text vectorization to predict the toxicity of Reddit comments. Logistic Regression is a widely used classification algorithm that works well with high-dimensional, sparse data—making it ideal for text classification tasks.

In this model, the comment text is converted into numerical features using TF-IDF, while
the subreddit and account age attributes are processed using one-hot encoding and standardization. We also applied class_weight="balanced" to address the significant
imbalance in our dataset, ensuring that the model pays equal attention to both toxic and non-toxic comments.

3- Linear SVM (LinearSVC) + TF-IDF:

For this model, we implemented a Linear Support Vector Machine (LinearSVC) to classify Reddit comments as toxic or non-toxic. Linear SVMs are commonly used for text classification because they perform well with high-dimensional, sparse input features such as TF-IDF vectors. The goal of this model is to learn a separating hyperplane that distinguishes toxic comments from non-toxic ones based on linguistic patterns.

In this method, the comment text is transformed into numerical features using TF-IDF, while the *subreddit* attribute is encoded using OneHotEncoder and the *account age* feature is scaled through the preprocessing pipeline. As with the previous model, we applied class_weight="balanced" to reduce the impact of class imbalance and ensure the toxic class receives appropriate weight during training.

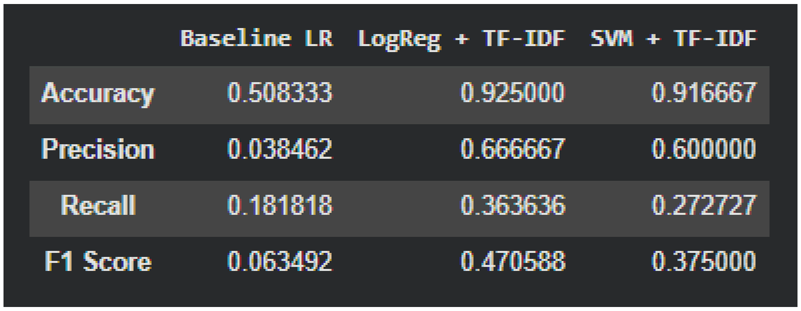
Baseline Model — Logistic Regression	**Logistic Regression + TF-IDF**	**Linear SVM (LinearSVC) + TF-IDF+ OneHotEncoder**
		
		

```
BASELINE MODEL RESULTS
-----------------------
              precision    recall  f1-score   suppor

           0       0.87      0.54      0.67       10
           1       0.04      0.18      0.06        1

    accuracy                           0.51       12
   macro avg       0.45      0.36      0.37       12
weighted avg       0.79      0.51      0.61       12
```

```
LOGISTIC REGRESSION + TF-IDF RESULTS
-----------------------------------
              precision    recall  f1-score   supp

           0       0.94      0.98      0.96
           1       0.67      0.36      0.47

    accuracy                           0.93
   macro avg       0.80      0.67      0.72
weighted avg       0.91      0.93      0.91
```

```
LINEAR SVM + TF-IDF RESULTS
----------------------------
              precision    recall  f1-score   support

           0       0.93      0.98      0.96       109
           1       0.60      0.27      0.38        11

    accuracy                           0.92       120
   macro avg       0.77      0.63      0.67       120
weighted avg       0.90      0.92      0.90       120
```

Comparing models based on regression scores:



	Baseline LR	LogReg + TF-IDF	SVM + TF-IDF
Accuracy	0.508333	0.925000	0.916667
Precision	0.038462	0.666667	0.600000
Recall	0.181818	0.363636	0.272727
F1 Score	0.063492	0.470588	0.375000

Key Findings: Logistic Regression + TF-IDF is the Best Performing Model

- Achieved the highest F1-score (0.47), indicating the best balance between precision and recall for detecting toxic comments.

- Outperformed both the baseline Logistic Regression and the SVM model in precision, recall, and overall consistency.

- Delivered the strongest results in the Precision–Recall Curve, with an Average Precision (AP) of 0.415 — significantly higher than what would be expected from a random classifier in an imbalanced dataset.

- Demonstrated more stable performance across evaluation metrics, showing that TF-IDF features combined with Logistic Regression are highly effective in capturing meaningful patterns in toxic language.

- SVM performed reasonably well but showed lower recall and F1-score, indicating it missed more toxic comments compared to the Logistic Regression model.

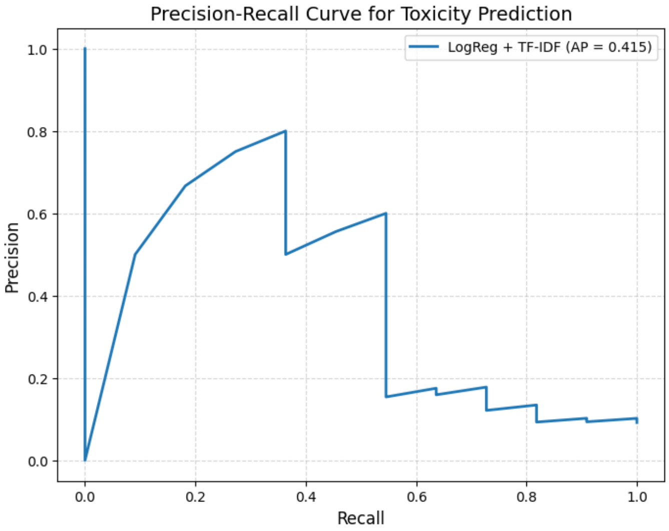
Overall, Logistic Regression with TF-IDF provided the most reliable and interpretable results, making it the optimal choice for our toxicity prediction task.

For further investigation we used The Precision–Recall curve for the Logistic Regression + TF-IDF model which shows that the classifier performs moderately well on the toxic (minority) class. The model achieves an Average Precision (AP) score of 0.415, which is substantially higher than the baseline AP of 0.10 expected from random guessing, given the class imbalance.

At low recall levels, the model reaches very high precision (up to 1.0), indicating that when it predicts a comment as toxic, it is highly confident and usually correct. As recall increases to the 0.3–0.4 range, precision remains relatively high (0.7–0.8), showing that the model can correctly detect a fair portion of toxic comments while maintaining reasonable accuracy.

However, beyond a recall of about 0.5, precision drops significantly, which reflects the difficulty of identifying more toxic comments without introducing false positives. This is expected in imbalanced text classification tasks.

Overall, the PR curve reveals that the model captures meaningful linguistic patterns associated with toxicity, performs far better than chance, and presents a reasonable balance between precision and recall for the minority class



Clustering Models:

Since our dataset includes a labeled target variable (toxicity), our modelling task is a supervised classification problem. Therefore, clustering algorithms are not appropriate for our dataset. Instead, we implemented three classification models — a baseline Logistic Regression model, a Logistic Regression model with TF-IDF model, and a Linear SVM with TF-IDF — and selected the best-performing model based on precision, recall, and F1-score.

Conclusion And Future Work

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.

In this phase, we developed, trained, and evaluated multiple machine-learning models to classify toxic comments using a combination of numeric, categorical, and text-based features. We began with a simple baseline Logistic Regression model that used only the numeric feature account_age_years. As expected, the baseline showed very poor performance due to the limited input information and the strong class imbalance in the dataset.

We then built two advanced models that incorporated full preprocessing using TF-IDF for text and One-Hot Encoding for categorical features:

1. Logistic Regression + TF-IDF + OneHotEncoder
2. Linear SVM (LinearSVC) + TF-IDF + OneHotEncoder

Both advanced models significantly outperformed the baseline, confirming the importance of using text-based features when predicting toxicity. Between the two advanced models, Logistic Regression consistently achieved the best balance between precision, recall, and F1-score, especially for the minority (toxic) class.

- Logistic Regression achieved an accuracy of 0.925, with a much higher precision and F1-score for class 1 compared to the baseline.
- SVM performed slightly worse, particularly in recall for the toxic class.

To further validate our results, we generated a Precision–Recall Curve for the best-performing model (Logistic Regression + TF-IDF). The PR curve confirmed that the model performs reasonably well on the minority class despite class imbalance, achieving an Average Precision

(AP) score of 0.415, which is much higher than random guessing. The curve also clearly demonstrated the trade-off between precision and recall and showed that the model is capable of detecting toxic comments while maintaining acceptable precision levels.

Overall, the results show that Logistic Regression with TF-IDF is the most effective model for our toxicity prediction task. It successfully learns meaningful patterns from text and provides the best balance between sensitivity to toxic comments and minimizing false positives. This model was therefore saved as the final deployed model.

Future Work

Future work on this project could further strengthen the performance and reliability of toxicity prediction. Expanding the dataset to include a wider variety of subreddits and a larger number of toxic comments would allow the models to generalize more effectively across different online communities. Incorporating more advanced text-representation methods, such as contextual embeddings (e.g., BERT), could also provide deeper linguistic understanding compared to TF-IDF.

Additionally, exploring specialized imbalance-handling techniques—such as SMOTE, may help improve recall for the minority toxic class. More extensive experimentation with different machine learning models, hyperparameter tuning, and cross-validation strategies could further enhance predictive performance.

Finally, future work could involve deploying the final model into an interactive application or dashboard, allowing real-time toxicity detection and offering practical value for platform moderation.