Cross-Domain Performance of Hate Speech Detection Models

Shahad Alotaibi, Raghad AL-Qwaider, Zuhd Ibrahim, Renad Alowais, Ghada AlQahtani

- Project Overview
- Data Description
- Main Question (Hypothesis Testing)
- Tests Used + Interpretation
- Sub Question (Prediction Testing)
- Models Used + Evaluation
- Technical Challenges
- Future Improvements

Project Overview

- Study cross-platform hate speech behavior
- Platforms included: Reddit, Twitter, YouTube
- Two modelling approaches:

Hypothesis Testing

compare toxicity across platforms

Prediction Models

predict engagement using toxicity severity

Goal: understand how hate speech behaves differently depending on platform

Dataset Summary

Multi-domain hate-speech dataset

Main variables:

- toxicity_score (0–1)
- platform
- engagement (likes, replies, views…)

Preprocessing steps:

- remove missing values
- normalize numeric fields
- encode categorical variables

Main Question

Does the toxicity score vary across platforms?

We compare the average toxicity levels in:

- Reddit
- Twitter
- YouTube

Using statistical hypothesis testing.

Why Hypothesis Testing?

01 Toxicity score = numeric

02 Platforms = categorical groups

03 We want to know whether any platform has significantly more toxicity

04 Hypothesis testing helps us determine if differences are meaningful or random

Test 1: Shapiro–Wilk Test (Normality)

Purpose:

Checks whether toxicity data follows a normal distribution.

Why:

Normality affects whether parametric tests (like ANOVA) are appropriate.

Interpretation:

- p > 0.05 → distribution is approximately normal
- p < 0.05 → distribution is not normal

Test 2: Levene's Test (Equal Variances)

Purpose:

Tests whether toxicity variance is equal across platforms.

Why:

ANOVA assumes equal variances.

Interpretation:

- p > 0.05 → equal variances
- p < 0.05 → unequal variances

Test 3: One-Way ANOVA

Purpose:

Tests whether the mean toxicity score differs across the three platforms.

Hypotheses:

- H_0: all platforms have equal mean toxicity
- H_1: at least one platform differs

Interpretation:

- $p < 0.05$ → significant difference exists

Test 4: Pairwise t-Tests (Bonferroni Corrected)

Purpose:

Find out which platforms differ from each other.

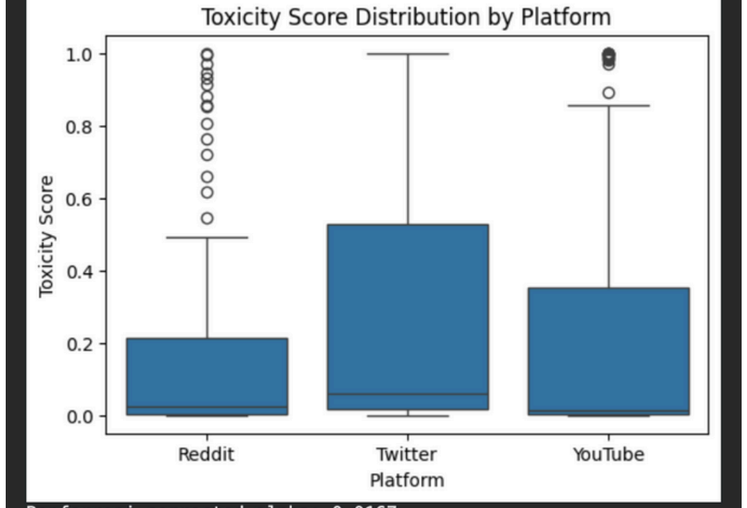
Your results:

- Reddit vs Twitter → Significant
- Reddit vs YouTube → Not significant
- Twitter vs YouTube → Not significant

Meaning:

- Only Reddit and Twitter have meaningful difference in toxicity.

Hypothesis Testing Result




```
Bonferroni-corrected alpha: 0.0167

Comparing Reddit vs Twitter
  t-statistic: -2.7970
  p-value:     0.005816
  → Significant difference (after Bonferroni correction)
--------------------------------------------------
Comparing Reddit vs YouTube
  t-statistic: -1.3412
  p-value:     0.181532
  → NOT statistically significant
--------------------------------------------------
Comparing Twitter vs YouTube
  t-statistic: 1.3283
  p-value:     0.185752
  → NOT statistically significant
--------------------------------------------------
```

Sub Question

Does the severity of toxicity help in predicting engagement?

Regression task
Engagement = numeric (0–1)

We analyze:

- whether more toxic posts get more interactions
- whether toxicity severity increases or decreases engagement

Using Prediction Testing.

Why Prediction Testing

01 Engagement score = numeric

We want to predict a continuous numeric outcome (engagement level).

02 Toxicity score = continuous feature

Severity of toxicity can act as a predictor variable in regression models.

03 We want to know whether toxicity helps explain engagement behavior
Does higher toxicity increase, decrease, or have no effect on engagement?

04 Prediction testing helps us measure the strength & direction of this relationship
Regression allows us to estimate how toxicity contributes to engagement.

Model 1: Linear Regression

Purpose:

- baseline model

Findings:

- Toxicity has a small negative coefficient
- Very low R^2 → weak linear relationship
- Poor predictor alone

Meaning:

- Toxicity barely explains engagement in a linear form.

Model 2: Ridge Regression

Purpose:

- Regularized regression model handling multicollinearity.

Results:

- Best performance
- Lowest MSE
- Highest (though still low) R^2

Meaning:

- Engagement prediction improves slightly with regularization.

Model 3: Decision Tree Regressor

Purpose:

- Captures nonlinear patterns.

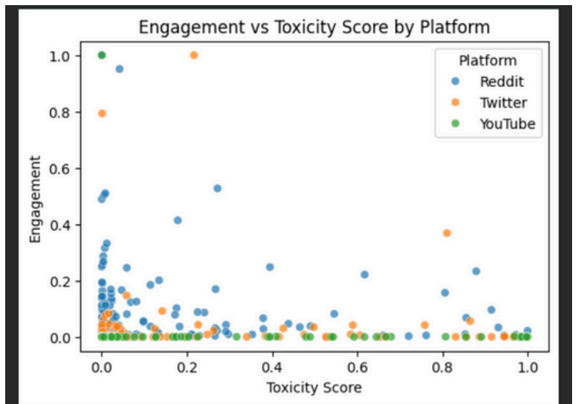
Findings:

- Moderate performance
- R^2 remains low
- Engagement is noisy & unpredictable

Meaning:

- Toxicity alone is not enough to reliably predict engagement.

Prediction Testing Result




```
Baseline Model (mean predictor)
  MSE: 0.0381, R^2: -0.0166

Linear Regression:
  MSE: 0.0358, R^2: 0.0430
  Coefficients:
    toxicity_score: -0.0323
    word_count: 0.0130
    negative_word_count: 0.0808
    Platform_Twitter: -0.0673
    Platform_YouTube: -0.0833
  Intercept: 0.09637024196187643
------------------------------------------------------------
Ridge Regression:
  MSE: 0.0357, R^2: 0.0459
  Coefficients:
    toxicity_score: -0.0300
    word_count: 0.0066
    negative_word_count: 0.0641
    Platform_Twitter: -0.0622
    Platform_YouTube: -0.0798
  Intercept: 0.0949126251647035
------------------------------------------------------------
Decision Tree:
  MSE: 0.0363, R^2: 0.0308
  Intercept: N/A
------------------------------------------------------------
Best-performing model: Ridge Regression (based on lowest MSE)
```

Summary

01 Toxicity significantly differs across platforms
- Only Reddit vs Twitter is significant

02 Toxicity has a weak relationship with engagement

03 Ridge Regression performs best

04 Engagement is influenced by additional factors beyond toxicity

Thank You