Data Science Project

**Disagreements in Hate Interpretation**

Report

|  |  |  |
| --- | --- | --- |
| Students Name | | ID |
| Section #80657 | Shahad alotaibi | 444201168 |
| Raghad AL-Qwaider | 444200109 |
| Zuhd Ibrahim | 444201206 |
| Renad Alowais | 444200530 |
| Ghada AlQahtani | 444200371 |

Instructor: Dr. Abeer Aldayel

Contents

[Table of Contents 2](#_Toc209719391)

[1. Introduction 3](#_Toc209719392)

[2. Related Works 3](#_Toc209719393)

[3. Data Collection 4](#_Toc209719394)

[3.1 Task 4](#_Toc209719395)

[3.2 Logbook Entry 4](#_Toc209719396)

[4. Deliverables 5](#_Toc209719397)

[4.1 Introduction 5](#_Toc209719398)

[4.2 Data Sources 5](#_Toc209719399)

[4.3 Objectives 5](#_Toc209719400)

[4.4 Method 6](#_Toc209719401)

[4.5 Challenges and Recommendations 6](#_Toc209719402)

[5. Data Sources – Bias Evaluation 7](#_Toc209719403)

[6. Raw Data File(s) 7](#_Toc209719404)

[7. Jupyter Notebook 7](#_Toc209719405)

[8. Logbook 7](#_Toc209719406)

# 1. Introduction

# Hate speech is one of the most pressing challenges facing today’s societies, especially with the widespread use of social media platforms. These platforms have created vast spaces for interaction and discourse, but they also enable harmful and discriminatory speech. Despite significant research efforts, there is still no universal agreement on what constitutes hate speech. These disagreements arise from cultural, social, and linguistic differences (e.g., sarcasm, irony, regional slang). This project investigates how hate speech is expressed and interpreted across three distinct platforms: Twitter, Reddit, and YouTube. Each platform differs in format (short posts, threaded discussions, and video-based comments), moderation policies, and user demographics, all of which shape both the occurrence and interpretation of hate content.

# Research Question:

Do interpretations of hate speech vary across platforms (Twitter, Reddit, YouTube), and which platform provides the most balanced coverage of hate speech types?

# 2. Related Works

# 1. Hate Speech Detection Using Cross-Platform Social Media Data (Shahi & Majchrzak, 2024, arXiv) This study investigates bilingual hate speech detection using YouTube comments and evaluates the impact of integrating datasets from other platforms such as Twitter and Gab. The authors found that combining datasets based on content similarity, hate word overlap, and definition alignment significantly improves classification performance. The highest F1-score (0.74) was achieved by merging YouTube, Twitter, and Gab datasets, indicating that no single platform fully captures all hate speech interpretations, but cross-platform integration enhances coverage.

[Hate Speech Detection Using Cross-Platform Social Media Data In English and German Language](https://arxiv.org/html/2410.05287v1)

2. Hate Speech Dashboard – GitHub Repository (Dinesh052, GitHub)  
This interactive dashboard compiles hate speech datasets from Twitter and Reddit, allowing users to compare linguistic patterns and annotation strategies. The analysis reveals that Twitter tends to feature short, direct expressions of hate, while Reddit includes more nuanced and context-rich discussions. These differences suggest that each platform reflects distinct dimensions of hate speech, and a unified dataset must account for these variations to ensure balanced representation. [GitHub - Dinesh052/hate-speech-dashboard](https://github.com/Dinesh052/hate-speech-dashboard)  
  
3. Analysis of Online Hate Speech Detection for Multilingual and Multimodal Data (Springer, 2024)  
This comprehensive review of 23 studies emphasizes the challenges of generalizing hate speech detection across platforms and languages. It highlights that platforms like YouTube, which include both video and textual content, require multimodal approaches, while Twitter and Reddit rely primarily on text. The study concludes that platform-specific definitions and moderation practices influence how hate speech is labeled and understood, reinforcing the need for cross-platform datasets to achieve equitable coverage.

[Analysis of Online Hate Speech Detection for Multilingual and Multimodal Data Using Artificial Intelligence | SpringerLink](https://link.springer.com/chapter/10.1007/978-3-031-91331-0_2)

# 3. Data Collection

Data was collected from three platforms using their official APIs: Twitter, Reddit, and YouTube. The focus was on unstructured textual data, specifically user-generated posts and comments containing keywords related to hate speech and negative sentiment. By targeting multiple sources, the dataset ensures diversity in expression styles and contexts of hate speech across different online communities.

## 3.1 Task

Collect unstructured textual data related to hate speech across platforms and prepare it for analysis.

## 3.2 Logbook Entry

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Platform | Date Collected | Method | Tool Used | Query Format | Notes |
| Twitter | Sep 22–23, 2025 | API | Tweepy (Twitter API v2) | ("hate OR slur") -is: retweet lang:en | Faced rate limits, missing metadata |
| Reddit | Sep 22–23, 2025 | API | |  | | --- | | PRAW (Python Reddit API Wrapper) |  |  | | --- | |  | | |  | | --- | | Keyword-based search across r/all |  |  | | --- | |  | | |  | | --- | | Required handling API limits; only post bodies (no comment text parsed). |  |  | | --- | |  | |
| YouTube | Sep 22–23, 2025 | API | YouTube Data API v3 | Search videos + extract comments | Pagination issues, quota restrictions |

# 4. Deliverables

## 4.1 Introduction

## Explores cross-platform differences in hate speech interpretation and seeks to identify which platform provides broader coverage

## 4.2 Data Sources

## • Twitter: Tweets containing selected hate-related keywords • Reddit: Posts from all subreddits using keyword search • YouTube: Video comments from content retrieved by keyword searches

## 4.3 Objectives

1. Identify differences in how hate speech is expressed across platforms.  
2. Detect the most frequent hate-related keywords across datasets.  
3. Determine which groups or communities are most frequently targeted.  
4. Compare engagement patterns between hateful and neutral content.  
5. Analyze co-occurrence of negative sentiment words with hate terms.  
6. Explore cultural and linguistic variations in framing hate speech.

## 4.4 Method

1. Data Preprocessing (cleaning, filtering duplicates, language restriction)

2. Data Transformation (unifying formats into CSVs with common fields)

3. Analysis Techniques (keyword frequency, co-occurrence analysis, engagement comparison)

4. Reporting Insights (cross-platform comparison, highlighting disagreements)

## 4.5 Challenges and Recommendations

Challenges:

• API rate limits and quota restrictions

• Data structure variations across platforms

• Dynamic nature of content

• High volume of retrieved data

Recommendations:

• Store static CSV snapshots immediately after collection

• Standardize structures into unified schema

• Focus on representative samples

• Use batch collection to manage rate limits

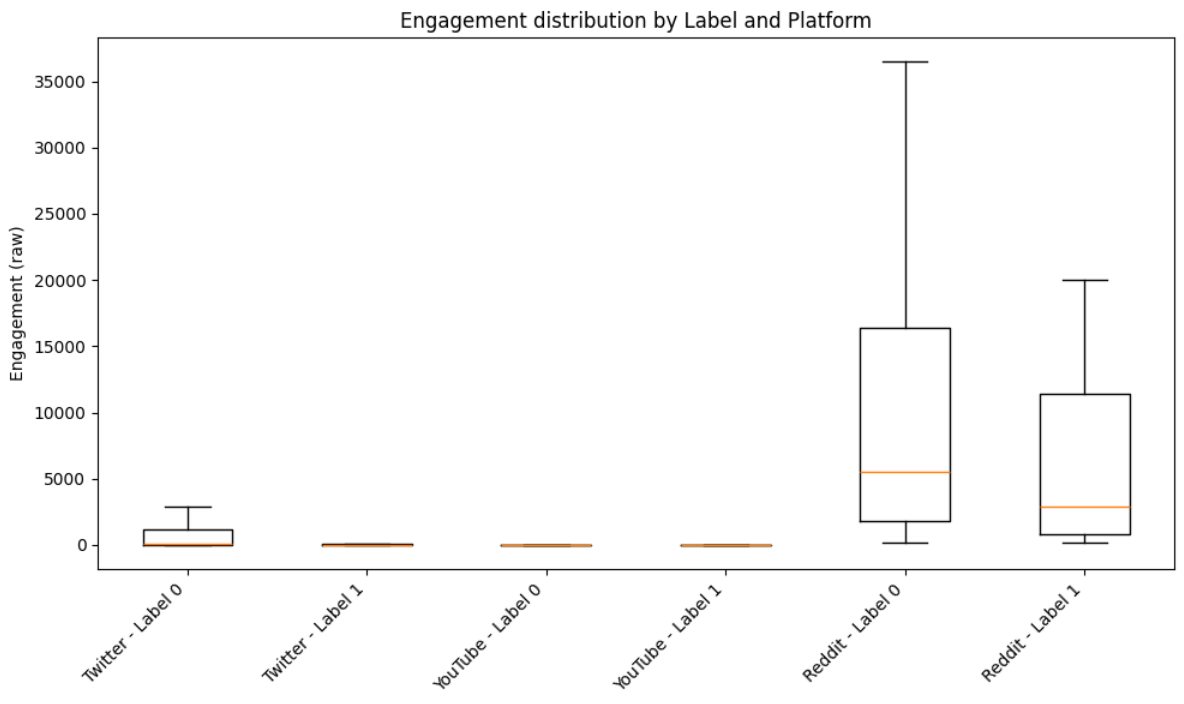
# 5. Data Sources – Bias Evaluation

1. Representation Bias: Twitter overrepresents younger users, Reddit focuses on niche discussions, YouTube has broader audiences. Some groups remain underrepresented.

A graph of social media

AI-generated content may be incorrect.

YouTube shows the highest share of content labeled as hate speech, with about half (50%) of its posts flagged. Twitter follows with roughly 28%, while Reddit has the lowest proportion at around 11%. This suggests that in your sampled data, hate-related or offensive language is most prevalent on YouTube, moderately present on Twitter, and comparatively rare on Reddit.

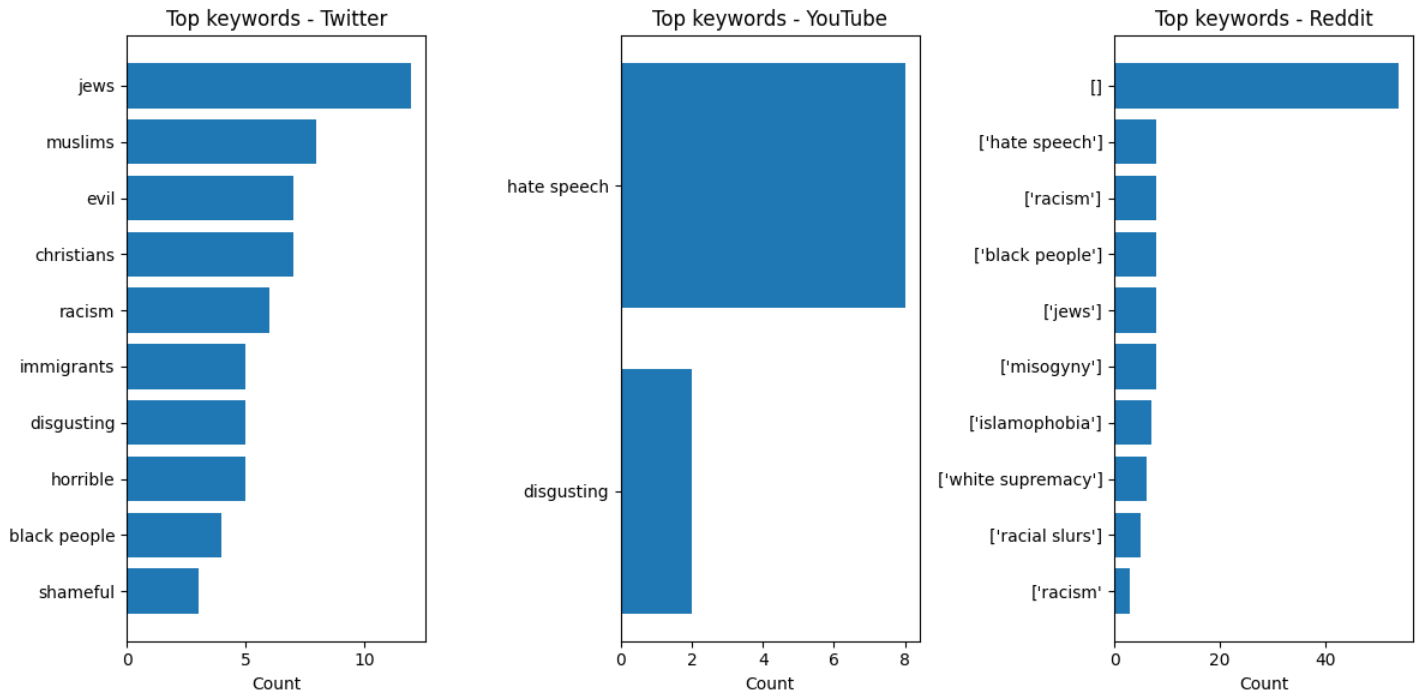


Reddit posts dominate engagement, with both hate (Label 1) and non-hate (Label 0) content reaching far higher interaction levels compared to Twitter and YouTube. The median Reddit engagement is thousands of interactions, while Twitter and YouTube posts stay mostly under a few hundred. The spread (box height) is also widest for Reddit, indicating some posts — hate or not — can go extremely viral. Twitter shows slightly higher engagement for non-hate content than hate, while YouTube engagement remains very low overall.

A graph of a bar chart

AI-generated content may be incorrect.

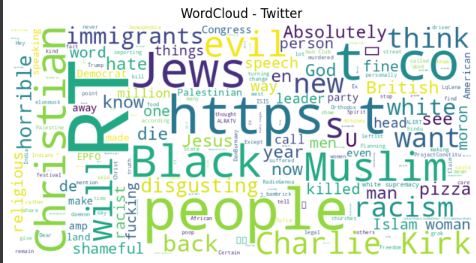
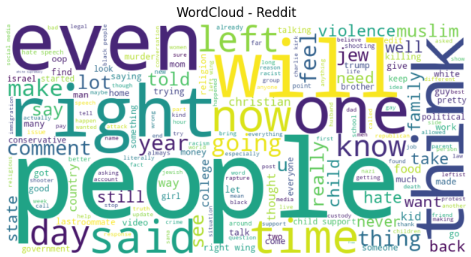
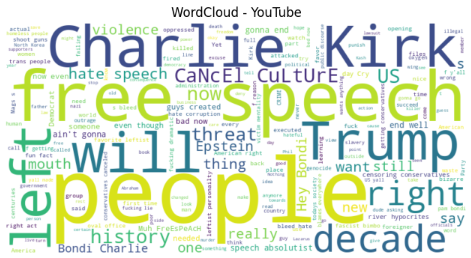
YouTube has the largest total number of posts in the dataset and also the highest count of hate-labeled content — nearly half of its entries. Twitter has far fewer posts overall and a moderate portion of hate. Reddit has the second-largest dataset size but only a small fraction of hate-labeled posts, showing that while Reddit is well-represented, most of its content here is non-hate.



Reddit: Most rows appear empty ([]), but where terms exist, they mention “hate speech,” “racism,” “misogyny,” “islamophobia,” and “white supremacy.” This indicates that many Reddit entries either lacked matched keywords or were not automatically tagged.

YouTube: The keyword set is very small and dominated by the phrase “hate speech,” with “disgusting” appearing as a distant second, suggesting less diversity of flagged terms

Twitter: The most frequent hate-related terms revolve around religion and ethnicity — especially “jews”, “muslims”, and “christians” — along with general insults like “evil”, “horrible”, and “disgusting.”



Twitter WordCloud

Shows strong focus on identity and religion: “jews, christian, muslim, black, immigrants” alongside hate-related adjectives (“evil, disgusting, horrible”). Also includes links (https, t.co) typical of Twitter and some popular figures (Charlie Kirk). This suggests a mix of targeted hate terms and news-sharing behavior.

YouTube WordCloud

Dominated by “free speech,” “Charlie Kirk,” “Trump,” “people,” and “left/right.” YouTube discussions here are centered on free speech debates, political figures, and culture-war themes like cancel culture and censorship.

Reddit WordCloud

Most frequent terms include “people,” “right,” “left,” “even,” and “one.”The conversation seems highly political and polarized, with frequent references to sides (left/right), time-based words (now, time, days), and some religion- or identity-related mentions (e.g., jew, christian).

2. Measurement Bias: Content types differ (short tweets vs. long Reddit posts vs. YouTube comments), leading to inconsistency in interpretation of hate speech.

3. Historical Bias: Moderation policies and content removal affect data availability, shaping the dataset based only on what remains visible.

# 6. Raw Data File(s)

- twitter\_data.csv

- youtube\_comments.csv

- reddit\_data.csv

# 7. Jupyter Notebook

Contains scripts for API access, preprocessing, and analysis.

* Data collection and Extraction Notebooks

Twitter: <https://colab.research.google.com/drive/1s0sNmPXVtMY8irGXOQDVrngU9Oe0Lduj?usp=sharing>

Reddit: <https://colab.research.google.com/drive/1BhgKkYaotBkS9T5eT8P_UluZSuYlbD3P?usp=sharing>

YouTube:

<https://colab.research.google.com/drive/1fgCA-nw5JJhUMUNDNkZtpMeKJrLCCGN5?usp=sharing>

* Graphical Representation Notebook

<https://colab.research.google.com/drive/1v4x2-973h_SoMU5AFARobp6X7hIyY_bj?usp=sharing>

# 8. Logbook

Provided in a separate document.