Universidad del Valle de Guatemala

Faculty of Engineering

Responsible AI
Department of Computer Science

Semester II, 2025

Project 02

Data Extraction and preprocessing

Emilio José Solano Orozco	21212
José Ricardo Méndez González	21289
Sara María Pérez Echeverría	21371
Emily Elvia Melissa Pérez Alarcón	21385
Diego Alberto Leiva Pérez	21752

Guatemala, October 17, 2025

Introduction

Online platforms such as Reddit serve as spaces for public discussion that reflect social biases and inequalities present in everyday discourse. This project focuses on the extraction and preparation of a real-world dataset from the USCIS community using the official Reddit API, focusing on transparency in collection and preprocessing choices. With top posts from the last 12 months as a target, the collection yielded approximately 200 posts and over 15 thousand comments. The subreddit centers on immigration processes, lived experiences, and advice related to U.S. residency and naturalization, making it suitable for Responsible AI analysis of representation, framing and sentiment. The goal of this project is not to build predictive models but to ensure that data extraction, preprocessing, and documentation follow principles of fairness, transparency, and accountability.

Data

Data Source

- **Platform**: Reddit (via official API using PRAW library)
- **Community**: r/USCIS (U.S. Citizenship and Immigration Services)
- **Sampling**: "Top" category posts from the past 12 months
- **Volume**: 200 posts and its comment threads (~15,000 comments)
- **Data range**: October 22, 2024 – October 17, 2025

Justification

Reddit was chosen for its open API and clear usage policies, allowing ethical data extraction without violating privacy or scraping restrictions. The USCIS subreddit is especially significant for bias analysis because discussions involve sensitive topics such as immigration, citizenship, nationality, and personal experiences with government institutions, topics that inherently expose social and linguistic biases. Posts often mix factual guidance, frustration and empathy, creating a high-variance emotional and cultural dataset.

Using the top category captures content amplified by the community voting, thus revealing dominant narratives and consensus framing. This introduces popularity bias but also reflects on the platforms rewards and de-emphasizes. Studying this imbalance is valuable for fairness-oriented analysis, since models trained on similar data could overrepresent mainstream or majority of perspectives while marginalizing dissenting or multilingual voices.

Data extraction

Data was collected using the official Reddit API via the PRAW library, ensuring compliance with Reddit's access policies. The script iterated through the top posts from r/USCIS between 2024-10-22 and 2025-10-17, downloading both submissions and their full comment threads. Each record includes post or comment ID, author (anonymized as provided by Reddit), timestamp (ISO-8601 UTC), score, and permalink. API rate limits were respected through a controlled sleep interval (0.7 s per request). The output was stored in structured CSV and JSONL files under data/raw/, forming the reproducible baseline for later preprocessing.

Preprocessing Overview

Raw Reddit data were converted into structured CSV and JSONL formats using UTF-8-SIG encoding for full character compatibility. Each post and comment record includes standardized ISO-8601 UTC timestamps and preserved metadata fields for traceability. The preprocessing pipeline consisted of two main stages:

1. Enrichment Stage: sentiment scoring with VADER, detection of agreement and disagreement cues, and aggregation of per-post metrics such as support_index.
2. Normalization Stage: language detection, text cleaning (URL, user, and punctuation removal), accent normalization, and generation of machine-learning–ready text fields.

All intermediate data were versioned under data/enriched/ and data/processed/ to maintain reproducibility across runs.

Bias and fairness considerations

Sentiment Balance

VADER analysis shows 48% neutral, 44% positive and 8% negative comments, as shown in fig. 1. The near absence of negativity suggests moderation and self-selection effects: highly emotional or critical remarks are less visible. Because VADER is English-centric, non-English comments were often misclassified as neutral, underrepresenting emotional variation across languages.

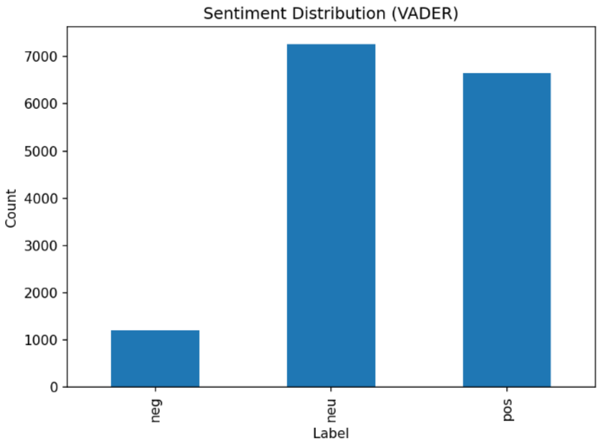

Figure 1: Comments Sentiment Distribution with VADER

Moderation and survivorship bias

About 8.3% of the comments came from deleted or unknown authors, and 1.3% were posted by automated accounts such as AutoModerator. These moderation artifacts mean the dataset captures only content allowed to remain public, not the full conversational range. Critical or policy sensitive posts may have been removed, reducing the diversity of tone and perspective in the data.

Representation Bias

Language detection found English dominance (~87%), this aligns with the subreddit's focus on U.S. immigration, where English proficiency is often a prerequisite for official procedures. However, this linguistic concentration still affects representativeness. Voices of recent migrants or non-English speakers, who may face greater barriers, are largely absent, as shown in fig. 2. Recognizing this is important for transparency, not for judging the community.

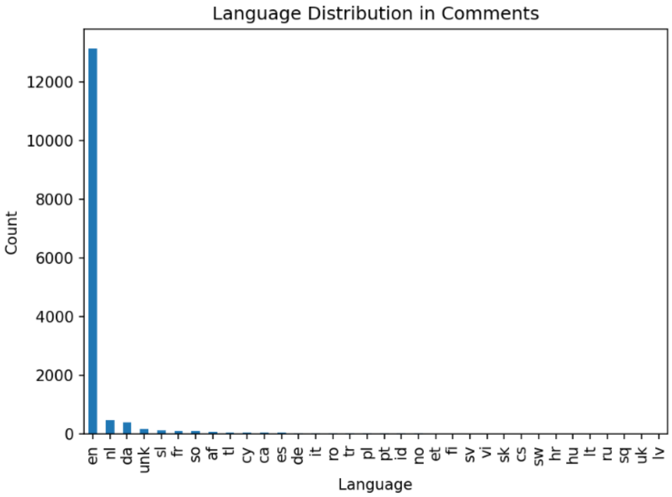

Figure 2: Language distribution in comments

Technical Constraints

The Reddit API omits shadow-banned and removed threads, and VADER cannot reliably interpret sarcasm, code-switching, or idiomatic Spanish. Cleaning operations (URL, username, and punctuation removal) improved consistency but stripped context about who or what users' reference, which slightly reduces interpretive fidelity.

Support Index

The support index, an aggregate of positive sentiment and agreement cues, fig. 3. shows that most posts receive predominantly supportive or agreeable responses. This pattern indicates consensus reinforcement rather than open debate. On Reddit, highly upvoted or positively framed posts remain visible longer, while contentious or critical ones are buried or removed. As a result, the dataset overrepresents cooperative or affirming narratives, underrepresenting disagreement and frustration. This does not invalidate the data but highlights a community-level bias toward conformity and civility shaped by platform mechanics and moderation norms.

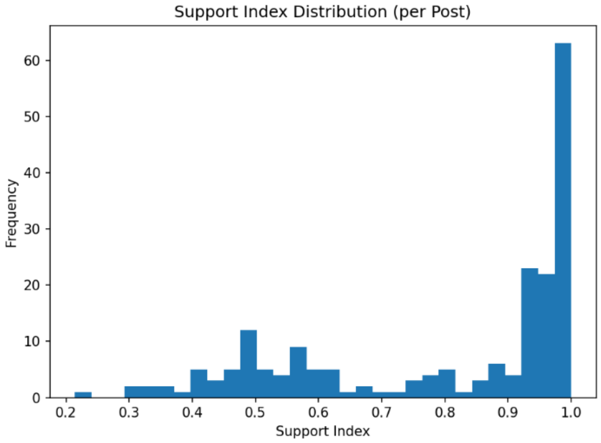

Figure 3: Support index value per post

Reflection

Every preprocessing step involves a trade-off between cleanliness and authenticity. Removing deleted comments or normalizing text improves technical quality but erases emotional and critical content, shifting the dataset's social balance. Similarly, prioritizing English-only discussions may reduce linguistic noise yet silences multilingual or recently arrived voices.

Such transformations affect interpretability. Simplifying text, removing usernames, aggregating sentiment into single scores make the data easier to model but harder to trace back to real conversational context. These steps hide nuance, sarcasm, or dissent-factors essential for understanding public opinion around immigration to the U.S.

Sentiment enrichment using VADER added interpretability, numerical emotion metrics, but imported its own lexical biases. Because VADER is optimized for English internet slang, it interprets certain words or emojis differently across cultures. Later models trained on these scores may inherit those value judgments. Similarly, the support index provides a compact signal of agreement but compresses complex social interactions into a single number, which can obscure dissent or minority viewpoints.

From a Responsible AI perspective, this project illustrates that bias control begins long before model training. Fairness depends on transparent documentation of data origins, missing segments, and preprocessing assumptions. Ethical data pipelines should record each transformation, enable reversion to raw text, and avoid using derived sentiment or support scores for individual-level inference.

Conclusions

The Reddit-based dataset offers a rich textual foundation for studying discourse around immigration and citizenship in the U.S. However, even with a transparent, reproducible pipeline, representational and selection biases remain, rooted in who participates, what content is rewarded and how moderation shapes the narrative. Preprocessing might improve structure and data readability but cannot eliminate these structural imbalances. This project demonstrates that Responsible AI begins well before model training, in the ethical awareness of how data are collected, filtered and documented.

References

Cjhutto. (2021). *VADER Sentiment Analysis*. GitHub. https://github.com/cjhutto/vaderSentiment

Herkewitz, W. (2013). Upvotes, Downvotes, and the Science of the Reddit Hivemind. *Popular Mechanics*. https://www.popularmechanics.com/science/health/a9335/upvotes-downvotes-and-the-science-of-the-reddit-hivemind-15784871/

Reddit. (s. f.-a). *Reddit API Documentation*. https://www.reddit.com/dev/api/

Reddit. (s. f.-b). *U.S. Citizenship and Immigration Services (USCIS) Subreddit*. https://www.reddit.com/r/USCIS/