Data Description for:

Human-in-the-loop Machine Learning for Real time feedback systems.

Why this project?

Online platforms are full of conversations, some constructive, some critical, and some that cross into uncomfortable territory. But where exactly is that line? What makes a comment problematic*,* offensive*,* or harmful*?* It turns out, even humans don't always agree.

This project explores how machine learning systems can work with human input to make better decisions when faced with these kinds of grey areas. The idea is to build a system where the model identifies comments it's not confident about and asks for human feedback. That feedback is then used to retrain the model and improve its future predictions.

To build and evaluate this approach, I'm using two publicly available datasets released by Jigsaw and Google:

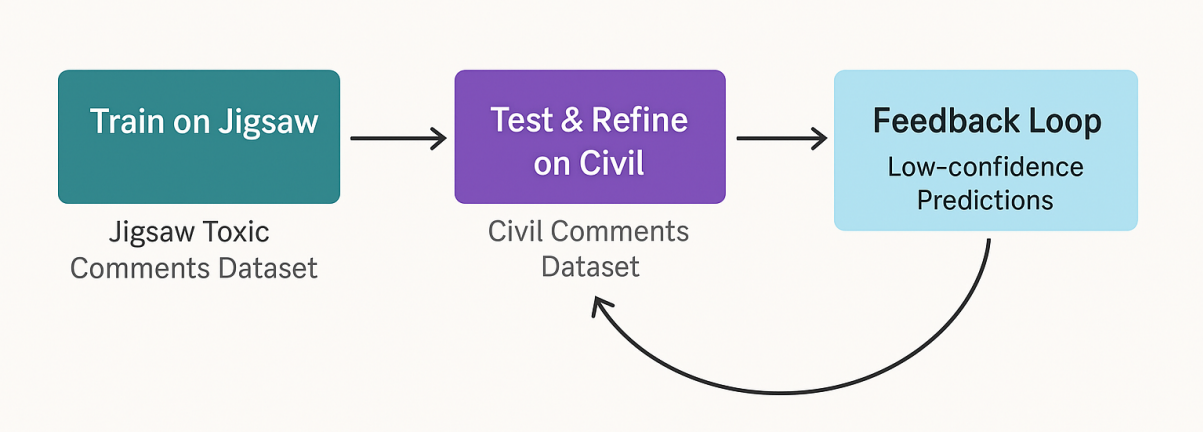
1. The first, drawn from Wikipedia talk pages, contains binary labels indicating whether a comment was flagged as offensive or disruptive by human raters.
2. The second, from the now-retired Civil Comments platform, is more nuanced. It contains averaged scores from multiple raters on how likely a comment is to be considered harmful, along with tags that identify whether the comment mentions certain groups (e.g., gender, religion, race). These allow for deeper analysis of how model behaviour may shift across different kinds of content.

I'm aware that terms like "toxic" can be controversial or overly broad. In this project, I use it only in the context of the original datasets and always with caution. The goal is not to define what is or isn't acceptable speech, but rather to study how human feedback can improve the reliability of a model in ambiguous or uncertain cases**.**

Later in the project, I'll also run a small study where people are asked to rate a few comments themselves. This will help compare how real-world human judgment aligns with the model's uncertainty and the original dataset labels.

At its core, this project is about teaching ML systems to recognise their limits and about letting humans step in when machines aren't sure.

How the Two Datasets Work Together?



This project uses two complementary datasets, each serving a distinct but connected purpose. The Jigsaw Toxic Comments dataset provides a straightforward, binary-labelled foundation to train a baseline text classification model. It helps establish the core learning pipeline, test early performance, and fine-tune the architecture in a relatively simple environment.

In contrast, the Civil Comments dataset is richer, more nuanced, and includes soft labels, subgroup identity tags, and annotator disagreement all of which are essential for evaluating model fairness, calibration, and bias. After training the initial model on Jigsaw, we will apply it to Civil Comments to identify low-confidence or ambiguous predictions, which are ideal candidates for human-in-the-loop feedback.

In short, Jigsaw helps us build and stabilize the model, while Civil Comments allows us to probe the model's ethical behaviour and adaptability in more complex real-world settings. This layered approach mirrors real deployment scenarios — where a model may be trained on one dataset but must adapt and evolve in response to diverse, shifting user inputs.

Dataset Overview:

1. Jigsaw Toxic Comment Classification:

This dataset was released as part of a Kaggle competition in 2018, focusing on detecting offensive, abusive, or disruptive comments on Wikipedia talk pages. Each comment in the dataset has been annotated by multiple human reviewers, who decided whether the comment contains one or more types of problematic content.

The labels are binary: a value of 1 indicates that the comment was marked as containing that type of content by at least one rater, and 0 otherwise. This makes it useful for initial model development, training, and simulation of a human-in-the-loop feedback loop using binary ground truth.

Dataset shape: (159571, 8)

Insights:

Column names: ['id', 'comment_text', 'toxic', 'severe_toxic', 'obscene', 'threat', 'insult', 'identity_hate']

Label distribution:
toxic 15294
obscene 8449
insult 7877
severe_toxic 1595
identity_hate 1405
threat 478
dtype: int64

Comments with multiple labels:
num_labels
0 143346
1 6360
2 3480
3 4209
4 1760
5 385
6 31

Column Name	Description
ID	Unique identifier for the comment
Comment_text	The raw text of the user comment
toxic	1 if the comment is considered generally offensive
Severe_toxic	1 if the comment contains severe toxicity (e.g., violent insults)
obscene	1 if the comment contains obscene language
threat	1 if the comment contains threats (e.g., of violence)
Insult	1 if the comment contains insults
identity_hate	1 if the comment targets people based on identity (e.g., race, gender, religion)

Class Imbalance and Its Implications:

One important characteristic of this dataset is that it is highly imbalanced. Out of ~160,000 comments, the vast majority (over 140,000) are labelled as non-problematic, while only a small fraction are marked with any kind of harmful content. For example, just 9.5% of comments are labelled as "toxic", and labels like "threat" or "identity hate" appear in less than 1% of the data.

This imbalance can create challenges during training. A model that simply predicts "clean" for everything would still achieve very high accuracy — but it wouldn't actually be useful. More importantly, rare but serious cases (like threats) might be missed entirely. To address this, the project will rely on metrics like precision, recall, F1-score, and AUC, rather than just accuracy. If needed, resampling techniques or class weighting will be applied to help the model learn from minority classes effectively.

2. Civil Comments:

This dataset was released by Jigsaw and Google as part of the "Unintended Bias in Toxicity Classification" competition. It contains over 1.8 million public comments originally posted on the Civil Comments platform (which was shut down in 2017). The goal of this dataset is not just toxicity classification, but also to explore fairness**,** bias**,** and group-level performance across demographic identities.

Each comment in this dataset has been annotated by multiple human raters, and their votes were aggregated into soft scores continuous values between 0 and 1 — indicating the proportion of raters who believed the comment had a certain property (e.g., "toxic" or "obscene").

This dataset provides both:

1. Soft target labels (e.g., target, severe_toxicity, insult)
2. Demographic identity tags (e.g., male, black, jewish)
3. Extra metadata about the comment (e.g., article_id, rating, likes)

Dataset Size

1. Total rows: 1,804,874
2. Columns: 45
   This is a large-scale, real-world dataset, great for bias, calibration, and HITL simulation.

Missing Values

1. Many identity columns (e.g., muslim, transgender, white, etc.) have ~400,000+ missing values
2. One metadata column (parent_id) has ~1M missing

Interpretation:

1. Identity columns are optional annotations: most comments don't reference any group
2. NaN simply means "this group was not mentioned" : we can safely treat as 0 or "not present"
3. This is expected and not a problem

Category	Columns	Description
Main Input	comment_text	The raw user comment
Soft Labels	target, severe_toxicity, obscene, identity_attack, insult, threat	Scores between 0–1 showing how many annotators flagged the comment with this label
Identity Attributes (22 columns)	male, female, black, white, muslim, jewish, etc.	Indicate whether the comment references a specific identity group; value ∈ [0, 1]
Engagement Features	likes, disagree, wow, funny, sad	User reaction metrics (less relevant for model training)
Metadata	id, article_id, created_date, rating, publication_id, etc.	Extra info about the source and user reactions
Annotation Stats	identity_annotator_count, toxicity_annotator_count	Number of human raters who annotated this comment for identity/toxicity

This dataset contains over 1.8 million comments originally posted on the Civil Comments platform, and was released as part of a Kaggle competition focused on understanding unintended bias in toxicity classification. Each comment is accompanied by multiple soft label scores, such as target, severe toxicity, obscene, insult, and others, which represent the average response from a group of human annotators, rather than a binary label. A key feature of this dataset is the inclusion of identity attributes**,** which indicate whether a comment mentions specific demographic groups such as gender, religion, or race. These columns are sparsely populated, and a missing value typically means the group wasn't mentioned. This design allows for deeper analysis of fairness and bias — for example, testing whether the model behaves differently on comments that reference marginalized identities. The dataset also includes metadata like article IDs, reaction counts (likes, sad, wow, etc.), and the number of annotators per comment. These features make the Civil Comments dataset particularly well-suited for studying model calibration, ethical decision-making, and subgroup-level performance, especially within a human-in-the-loop learning framework.

Missing values:

Many of the identity-related columns in the Civil Comments dataset (e.g., male, muslim, black) appear to contain missing values — but these are not traditional nulls. A missing value simply means that the identity was not mentioned in the comment. These columns were intentionally left blank by the annotators unless the comment referred to that specific group. For the purposes of training and fairness evaluation, these values will be safely filled with 0.0, indicating "not referenced." This is consistent with the dataset's documentation and ensures that identity-related analysis can proceed without data leakage or misinterpretation.

How was toxicity labelled and by whom?

The labels in both the Toxic Comments and Civil Comments datasets were not assigned by a single person or by an algorithm. They were created through crowdsourced human annotation, where multiple independent raters were shown each comment and asked to decide whether it contained certain types of problematic content such as insults, threats, or identity-based attacks.

For the Toxic Comments dataset, each comment was labelled by several individuals through a managed human-review process. The labels are binary (0 or 1), indicating whether *any* of the annotators believed the comment should be flagged. These labels cover six specific categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate.

In contrast, the Civil Comments dataset uses a probabilistic scoring approach. Each comment was rated by several crowd workers, and the labels (like target, obscene, threat, etc.) represent the fraction of annotators who agreed that the comment fits that label. So a target score of 0.87 means 87% of reviewers marked that comment as problematic. The identity columns were also crowd-annotated, indicating how likely it is that a given identity group was mentioned or referenced in the comment.

These ratings were collected on platforms like Figure Eight (formerly CrowdFlower) a crowd annotation service using basic training and quality checks to ensure some consistency among raters. However, like any crowdsourced labeling process, it reflects human subjectivity**,** including cultural, personal, and social bias.

This is important to acknowledge: a comment labelled as "toxic" might not be seen that way by everyone and what one annotator sees as threatening, another might consider sarcastic or harmless. This variability in human judgment is precisely what makes Human-in-the-Loop learning so valuable: it creates an opportunity to refine the system by incorporating fresh, contextual, or domain-specific feedback especially in borderline cases.

Label Ambiguity and Disagreement:

Disagreement among annotators is common, especially in borderline or sarcastic comments. This reinforces the idea that context and subjectivity matter and supports the case for Human-in-the-Loop systems where human judgment continues to play a role even after training.

Why Human-in-the-Loop Is Needed?

While the datasets used in this project are large and carefully labelled, they represent a snapshot of public judgment from a particular time and context. Over time, language, norms, and expectations change. A Human-in-the-Loop system is designed to adapt to these changes by involving real people when the model is unsure, and learning from that evolving input.

Planned Use of Real Human Feedback (via Google Form):

In addition to using the dataset labels, a small-scale feedback exercise will be conducted using a Google Form. A subset of low-confidence comments will be presented to human reviewers (friends/classmates), who will label whether the comments seem offensive or problematic. This real-time feedback will help compare human judgment with both model predictions and dataset labels.

Next Steps:

The next phase will focus on building the initial model pipeline and testing the human-in-the-loop feedback loop. The plan is to:

1. Train a baseline classifier using the Jigsaw dataset.
2. Identify low-confidence predictions using thresholding.
3. Simulate human feedback using ground truth.
4. Retrain and evaluate the model.
5. Transfer the model to Civil Comments for fairness and calibration analysis.
6. Collect real human feedback via a Google Form to compare with model predictions.

Final Note:

Together, these two datasets offer a powerful foundation: one to train the base model and another to explore fairness, uncertainty, and human judgment. With a clear understanding of the data, its origins, and its limitations, we're now ready to begin building the initial model pipeline, define low-confidence thresholds, and simulate human feedback in a way that's grounded in both ethics and real-world relevance.