**Final Project Submission – Michael MacLennan**

**Data Card   
Dataset: Jigsaw Toxic Comment Classification Challenge**

# Motivation

* **Purpose:** To benchmark toxic comment detection for online moderation.
* **Intended use:** Research, education, and decision-support in moderation workflows.
* **Not suitable for:** Fully automated moderation bans or law enforcement applications.

# Dataset Overview

* **Source:** Kaggle competition hosted by Jigsaw/Google, based on Wikipedia talk page comments.
* **Size:** ~160,000 annotated comments.
* **Labels:** Multi-label classification with six categories:
  + Toxic
  + Severe toxic
  + Obscene
  + Threat
  + Insult
  + Identity hate

# Composition

* **Input Modality:** Raw text (user comments).
* **Language:** Predominantly English. Limited representation of multilingual/variant English usage.
* **Demographics:** Not explicitly provided; comments are from open online communities (Wikipedia editors).
* **Label Distribution:** Highly imbalanced, e.g., “toxic” and “insult” labels more frequent; “threat” and “identity hate” rare.
* **Annotation:** Crowd-sourced labelling on a per-comment basis. Multiple annotators per sample; majority-vote aggregation.

# Collection Process

* **Method:** Harvested from Wikipedia talk pages.
* **Annotation Workforce:** Crowd workers (Mechanical Turk style).
* **Potential Bias:** Annotator demographics not disclosed; introduces cultural/linguistic bias. Label subjectivity in “toxic” interpretations.

# Preprocessing/Cleaning/Labelling

1. **Cleaning:** Removed URLs, special characters.
2. **Tokenisation:** Vocab capped at 30k.
3. **Sequence:** Padded/truncated to 256 tokens.
4. **Representations:** TF-IDF features, neural embeddings.

# Known Biases and Limitations

* **Cultural Bias:** Certain terms interpreted differently across cultural groups. Risk of over-flagging non-native English expressions.
* **Class Imbalance:** Under-representation of rare labels (e.g., threats), leading to weaker recall.
* **Domain Shift:** Dataset limited to Wikipedia; transferability to other platforms (Reddit, Twitter, forums) uncertain.
* **Linguistic Coverage:** Limited coverage of minority dialects, slang, or code-switching.

# Ethical Considerations

* **False Negatives:** Harmful comments slipping through moderation can expose communities to abuse.
* **False Positives:** Over-flagging risks silencing marginalised voices, especially those using non-standard English.
* **Annotation Transparency:** Lack of demographic info on annotators limits fairness auditing.
* **Use Constraints:** Not suitable for fully automated banning without human oversight. Intended for decision support.

# Data Splits

* **Training:** ~144k comments
* **Validation:** 16k (held-out, stratified 10%)
* **Cross-Validation:** 5-fold stratified CV used in experiments.
* **Test:** Kaggle hidden test set (labels withheld for leaderboard evaluation).

# Uses

* **Other tasks**: Could support research in toxicity lexicons, bias detection, NLP fairness studies, or benchmarking text classification pipelines.
* **Risks for reuse**: The dataset’s imbalanced labels and lack of annotator transparency may cause unfair treatment (e.g., disproportionate flagging of minority dialects). Reuse in production moderation must include bias mitigation (e.g., rebalancing, fairness-aware sampling, model calibration).
* **Not suitable for**:
  + High-stakes decision-making (e.g., law enforcement, employment screening).
  + Fully automated moderation without human-in-the-loop oversight.
  + Deployment in multilingual, non-Wikipedia platforms without adaptation.

# Distribution

* **Past distribution**: Released via Kaggle as part of the 2018 Jigsaw competition. Dataset remains available on Kaggle for research and education.
* **Copyright / IP / ToU**:
  + Distributed under Kaggle competition terms of use.
  + Derived from Wikipedia content (CC-BY-SA license applies to original text).
  + Permitted for research and educational use, not guaranteed for unrestricted commercial deployment.

# Maintenance & Updates

* **Original Dataset:** Static snapshot; no ongoing maintenance by Kaggle/Jigsaw.
* **Our Project:** No additional data collection. Dataset is used as-is with documented preprocessing.