**Final Project Submission – Michael MacLennan**

**Model Card   
Community Crisis Prevention System – Calibrated Ensemble for Toxicity Detection**

# Model Description

* **Model name**: Toxic Comment Classifier (Capstone Project)
* **Developer**: Michael MacLennan (for Imperial College PCMLAI Capstone Project).
* **Frameworks & libraries**: PyTorch, scikit-learn, XGBoost.
* **Architectures**: Logistic Regression (TF-IDF), Ridge, NB-SVM, Gradient Boosted Trees (XGBoost), TextCNN, lightweight transformers (DistilBERT, MiniLM).
* **Version/date**: 3 October 2025 (educational project, not production release).
* **Intended purpose**: Assistive moderation of online communities; triage toxic comments for human review.
* **Status**: Research/educational model (not production-hardened).

# Intended Use

* **Users:** Moderators, community managers, researchers, students of ML/NLP.
* **Goal:** Provide probability estimates and ranked lists of potentially toxic comments to support decision-making.
* **Scope:** English-language online comments (Wikipedia-style discourse).
* **Deployment:** Decision-support only. Always requires human-in-the-loop oversight.

# Model Architecture

* **Baselines**: Logistic Regression, Multinomial Naive Bayes
* **Advanced Model**: XGBoost with engineered features (n-grams, sentiment, length, punctuation)
* Neural Models:
  + TextCNN trained on tokenised input (vocab cap 30k, max length 256)
  + DistilBERT (lightweight transformer) fine-tuned with early stopping
* **Ensemble**: Weighted averaging across models (weights selected via CV)
* **Calibration**: Post-hoc sigmoid and isotonic regression to improve probability reliability

# Training Procedure

* **Dataset:** Jigsaw Toxic Comment Classification (≈160k comments, 6 toxicity labels)
* **Pre-processing:** Text cleaning, tokenisation, truncation/padding to 256 tokens
* **Training Strategy:** 5-fold stratified cross-validation, with early stopping
* **Hyperparameters:** Tuned via grid/random search (XGBoost) and learning-rate/filters (TextCNN)
* **Thresholding:** Decision threshold optimised per model using F1
* Evaluation Metrics: F1, Brier score, log loss, Expected Calibration Error (ECE)

# Results Summary

|  |  |  |
| --- | --- | --- |
| Model | F1 Score | Notes |
| Logistic Regression | ~0.72 | Baseline |
| Naïve Bayes | ~0.70 | Baseline |
| XGBoost (default) | ~0.73 | Untuned |
| XGBoost (tuned) | ~0.74 | Improved via feature engineering |
| TextCNN | ~0.75 | Convolutional neural net |
| DistilBERT | ~0.75 | Transformer, compute-constrained |
| Ensemble | ~0.76 | Weighted average |
| Calibrated Ensemble | 0.765 | Reliability improved (ECE ↓ 0.116 → 0.045) |

Evaluation Across Dimensions:

* **Predictive Power**: Incremental gains from baseline → ensemble → calibration.
* **Reliability**: Calibrated ensemble significantly reduced overconfidence, enabling trustworthy probability scores.
* **Efficiency**: Lightweight models (XGBoost, TextCNN) provided a strong performance/compute trade-off.
* **Fairness**: Evaluated subgroup risk (e.g. identity-related comments). While calibration helped, disparities remained.

# Fairness & Bias Considerations

* **Bias in Labels**: Dataset annotations are crowd-sourced, potentially reflecting annotator bias (e.g. cultural subjectivity in defining toxicity).
* **Group Risk**: Risk of over-flagging comments from non-native English speakers or dialects.
* **Mitigations**: Threshold tuning, calibration, and ensemble weighting reduced (but did not eliminate) disparities.
* **Residual Risks**: Misclassifications could silence marginalised voices or fail to protect communities from subtle harassment.

Trade-offs & Limitations

* **Class imbalance:** Performs poorly on rare classes (threat, identity hate).
* **Domain generalisation:** Effective on Wikipedia-like text but weaker on social media slang or multimodal content.
* **Resource efficiency:** Neural/transformer models outperform classical baselines, but require more compute and risk overfitting with small data.
* **Interpretability vs. accuracy:** Linear models offer transparency but lower performance; neural models provide higher F1 but act as “black boxes.”

# Ethical Considerations

* **False Negatives**: Allow harmful content to persist, potentially escalating crises.
* **False Positives**: Risk silencing users, especially minority voices.
* Deployment Guidance:
  + Always keep human moderation in the loop
  + Use calibrated scores for prioritisation, not final judgement
  + Communicate limitations transparently to end-users and stakeholders

# Not Suitable Uses

* Automated bans or penalties without human oversight.
* High-stakes decision-making (e.g. law enforcement, employment, financial risk assessments).
* Multilingual or highly informal/short-text platforms without further retraining.

# Caveats & Recommendations

* Retraining is required for new domains or languages.
* Fairness-aware resampling or augmentation is needed for underrepresented labels.
* Should be deployed as part of an integrated moderation workflow combining automation with human judgment.