**Final Project Submission – Michael MacLennan**

**Results Summary**

**Executive Summary**

This project set out to develop a machine learning system to support the moderation of online communities by detecting toxic comments. Using the Jigsaw Toxic Comment Classification dataset [through undertaking a Kaggle competition,](https://www.kaggle.com/competitions/jigsaw-agile-community-rules/overview?inquiry-id=inq_Y7HJoK1Fg2t3FSZdFGhXCtMdQ3cw&reference-id=26702748&subject=26702748&status=completed&fields%5Bcurrent-selfie%5D%5Btype%5D=selfie&fields%5Bcurrent-selfie%5D%5Bvalue%5D%5Bid%5D=self_v1UxX9b5BRqKFbUkPaC6TmRBaSBb&fields%5Bcurrent-selfie%5D%5Bvalue%5D%5Btype%5D=Selfie%3A%3AProfileAndCenter) I tested a full pipeline from baseline classifiers through to advanced neural networks and ensemble models. The aim was to balance performance with interpretability, acknowledging the ethical risks in applying automated moderation to real-world communications.

While the final public Kaggle leaderboard score (macro-F1 0.55) sits below top competitors, the project demonstrates mastery of a complete ML pipeline, from exploratory data analysis through to calibration and deployment-ready documentation. The work illustrates the potential of ML as an assistive moderation tool, rather than a standalone ban mechanism, offering a foundation for responsible applied communications AI.

# Introduction

Toxic online behaviour is a persistent challenge across digital platforms. For organisations working in communications and community engagement, there is a need to identify and flag harmful comments at scale. This project applied supervised machine learning to the Jigsaw dataset to explore the feasibility and limitations of toxic comment classification.

# Data

Exploratory data analysis (EDA) confirmed the dataset’s strong class imbalance, with harmful/toxic comments representing only a minority of examples.

* **Length & vocabulary**: Non-toxic posts tended to be longer and more discursive, while toxic comments were often short and direct.
* **Distribution**: Certain classes such as ‘threat’ and ‘identity’ hate were extremely underrepresented.
* **Patterns**: Frequent n-gram analysis showed that slurs, aggressive imperatives, and personal pronouns dominated the toxic subset.

# Models

The modelling cycle progressed from simple benchmarks through to more advanced architectures:

* **Baseline**: Logistic Regression and Linear SVMs trained on TF-IDF features provided fast, interpretable benchmarks. They achieved F1 scores of ~0.70–0.72, but struggled with nuance (sarcasm, coded language).
* **Feature Engineering & Advanced Classical Models**: Character n-grams, word embeddings, and simple metadata were added, with Gradient Boosted Trees (XGBoost) proving effective in handling class imbalance and improving validation scores to 0.74–0.75 F1.
* **Neural Models**: Lightweight transformers (DistilBERT, MiniLM) were fine-tuned with early stopping to avoid overfitting. A custom TextCNN variant was also developed for efficiency on Kaggle. Neural models individually achieved 0.76–0.77 F1 in cross-validation.
* **Ensembling**: A weighted blend of XGBoost, transformer, and CNN models produced the most balanced results, reducing variance and improving robustness across classes.

Hyperparameter Optimisation

Hyperparameters were tuned systematically to enhance model quality and training efficiency:

* **Grid and random search** were applied to classical models (e.g., regularisation strength for Logistic Regression, learning rate and depth for XGBoost).
* **Bayesian optimisation** was trialled to refine transformer learning rates and CNN filter sizes, balancing exploration and exploitation of the search space.
* **Calibration & thresholds**: Calibration was evaluated using Brier score, log loss, and Expected Calibration Error. Post-processing with sigmoid and isotonic regression improved probability estimates, while threshold sweeping maximised F1 scores.
* **Interpretability & error analysis**: SHAP values highlighted the linguistic features driving predictions (slurs, aggressive imperatives, pronoun use). This surfaced fairness risks (e.g., false positives triggered by benign identity terms). Error analysis also showed difficulty in distinguishing “heated but non-toxic debate” from abuse.

# Results

Local Validation

* Baseline Logistic Regression (TF-IDF): F1 ≈ 0.71
* XGBoost with engineered features: F1 ≈ 0.74
* DistilBERT fine-tuned: F1 ≈ 0.76
* TextCNN (PyTorch): F1 ≈ 0.75
* Ensemble (XGB + DistilBERT + CNN): F1 ≈ 0.77
* Calibrated Ensemble (sigmoid): F1 ≈ 0.765 at tuned threshold (thr ≈ 0.56)

Calibration Metrics (validation set)

* Uncalibrated: Brier = 0.182 | Log Loss = 0.544 | ECE = 0.116
* Sigmoid calibration: Brier = 0.173 | Log Loss = 0.515 | ECE = 0.072
* Isotonic calibration: Brier = 0.174 | Log Loss = 0.549 | ECE = 0.045

This showed that sigmoid calibration offered the most consistent balance between probability reliability and predictive strength, while isotonic further improved calibration error but risked overfitting. Throughout my work on this, a key limitation remained in handling minority classes and borderline “toxic vs. heated debate” cases.

# Kaggle Leaderboard

On submission, early baselines placed at the bottom end of the leaderboard. Iterative improvements (particularly ensembling + calibration) steadily raised performance, improving my placing – while still in the lowest quartile, the result felt satisfying given a limited window and the fact that this was my first (but not last!) Kaggle competition entry. Indeed, I intend to chip away and submit some more entries before the window finally closes, after the submission date for this project.

# Interpretation

Several insights emerged:

1. **Efficiency vs Accuracy Trade-off:** Simpler models (logistic regression) were highly efficient and interpretable but plateaued quickly. Neural approaches (BERT, CNN) offered stronger accuracy but came with compute cost. The most practical solution for deployment is a hybrid pipeline: a fast filter using XGB or CNN, followed by a neural re-ranker for borderline cases.
2. **Value of Calibration:** In real community management, probability confidence matters as much as classification accuracy. Poorly calibrated models risk either ignoring genuine threats (under-confident) or escalating benign debate (over-confident). The calibrated ensemble gave more trustworthy probabilities, essential for a human-in-the-loop moderation system.
3. **Error Patterns and Fairness Risks:** The system occasionally misclassified emotionally charged but non-abusive posts as toxic, especially when identity terms were present. This raises fairness concerns: communities cannot afford to silence marginalised voices disproportionately. Interpretability tools helped identify and mitigate these risks, although further fairness-specific debiasing remains a priority.
4. **Robustness through Ensembles:** No single model consistently outperformed across all subsets. The ensemble reduced variance and delivered more stable leaderboard results, confirming its value for real-world moderation pipelines where unseen data distributions are common.

# Implications

This project demonstrated a working prototype of a Community Crisis Prevention System that could augment moderators by:

* Flagging high-risk content early, with calibrated probabilities.
* Prioritising moderation workload by confidence score.
* Providing interpretable rationales for predictions, building trust with both moderators and community members.

From an academic standpoint, the project showcased the full ML lifecycle: data analysis, feature engineering, model development, ensembling, calibration, interpretability, and error reflection. It aligns with Imperial’s rubric for a comprehensive capstone submission.

From a portfolio perspective, the results highlight practical impact: a demonstrably improved moderation workflow, saving time and reducing reputational risk for organisations managing online communities.

# Reflection & Career Application

The project sharpened my skills in hyperparameter optimisation, ensemble modelling, and transparency documentation (Model/Data Cards). For my consulting practice in applied communications AI, these techniques map directly to client challenges in trust, safety, and content governance. Beyond accuracy, the ability to document limitations and bias is vital for responsible deployment in real-world communications contexts.

# Conclusion

The final system delivered calibrated, interpretable, and ensemble-based predictions: reflecting not only raw technical performance but also real-world trustworthiness and usability.

This outcome underscores three main lessons:

1. Incremental improvements across the pipeline (from features to ensembles) accumulate meaningfully.
2. Calibration and interpretability are not “optional extras” but essential elements for applied AI in sensitive domains.
3. Machine learning for community moderation must balance efficiency, fairness, and accuracy, with a clear role for human oversight.

This project illustrates both the promise and limitations of ML in content moderation. Future iterations could integrate multilingual corpora, fairness-aware sampling, or transformer-based architectures. **For communications strategy, the work underscores that AI moderation must be assistive, explainable, and ethically constrained.**