Project 2: Toxic Comment Classification

October 3, 2025

Plan

Context

- **Moderation challenge**
  - Online platforms like Wikipedia struggle to detect and manage toxic comments
- **Project goal**
  - Explore NLP methods for binary classification: toxic (1) vs. non-toxic (0)
- **Compared approaches**
  - Bag-of-Words
  - Embeddings
  - Transformers
  - Prompt Engineering
- **Evaluation criteria**
  - Precision, Recall, F1-score, AUC, Runtime

Data team

Brayann A.
Evaluation Specialist

Romulus A.
NLP Research Engineer

Noélie K.
Data Scientist

Sonagnon K.
Data Scientist

Best Model from Project 1

Table: TF-IDF + SVM Performance

Model	Prec.	Rec.	F1	AUC	Time
TF-IDF + SVM	0.97 / 0.94	0.99 / 0.59	0.98 / 0.72	0.98	10s

- **TF-IDF limitations:** While it emphasizes rare words, it produces high-dimensional vectors and ignores semantic context
- **Embeddings as a solution:** We transitioned to embedding models to capture word meaning and contextual relationships

FastText and MiniLM

Table: Performance with Word and Sentence Embeddings

Embedding	Model	Prec.	Rec.	F1	AUC	Time
FastText	Logistic Reg.	0.98 / 0.48	0.90 / 0.84	0.94 / 0.61	0.942	N/A
FastText	SVM	0.95 / 0.85	0.97 / 0.47	0.96 / 0.60	0.945	N/A
MiniLM	Logistic Reg.	0.99 / 0.50	0.91 / 0.88	0.94 / 0.64	0.962	2.83
MiniLM	SVM	0.96 / 0.85	0.99 / 0.62	0.97 / 0.72	0.962	13.67

- **Embeddings improve toxic recall**, especially with FastText, but classifier choice affects precision and balance
- **Limitation:** These embeddings are static—they ignore word meaning in context. To overcome this, we explore transformers for dynamic, context-aware representations

Toxic-BERT Fine-Tuned

- **Transformers understand global context** through the attention mechanism, enabling them to weigh the relevance of each word in a sequence
- **Key advantage:** They generate contextual embeddings (dynamic embeddings) that capture complex relationships and nuanced meanings within sentences

Table: **Transformer Performance – Toxic-BERT (Best Overall)**

Model	Prec.	Rec.	F1	AUC	Time
Toxic-BERT	0.99 / 0.91	0.99 / 0.94	0.99 / 0.92	**0.997**	746.4s

- **Toxic-BERT delivers top performance**, with an F1-score of 0.92 and AUC of 0.997—making it the most robust model for toxic comment detection
- **Trade-off:** This performance comes at a high computational cost (746s), making it less practical for real-time moderation

Flan-T5 Prompting

Table: **Prompt Engineering with Flan-T5 (n=1000)**

Prompt	Prec.	Rec.	F1	AUC	Time
Zero-shot	0.99 / 0.19	0.62 / 0.94	0.76 / 0.31	0.50	757s
Role Prompt	0.99 / 0.16	0.55 / 0.96	0.70 / 0.28	0.50	783s
Few-shot	0.98 / 0.12	0.35 / 0.92	0.51 / 0.21	0.55	918s

- **High recall, low precision:** All prompt types flagged toxic comments well, but severely overclassified safe content, leading to poor F1 and AUC scores
- **Slow and unreliable:** Despite its flexibility, prompting was computationally expensive and inconsistent across variants
- **Few-shot underperformed:** Even with examples, the few-shot setup yielded the worst precision and F1-score

Deployment Trade-offs

- **Cost**
  - *Bag-of-Words* and embeddings are fast to train and infer
  - Toxic-BERT is accurate but requires over 12 minutes
  - Prompting is slow (750s for 1000 samples) and inaccurate
- **Bias** : Transformers don't have real meaning they just echo what they've seen during training.
  - Embedding models may misclassify neutral comments (e.g., "She is a woman") as toxic
  - This reflects biases in training corpora and societal stereotypes
- **Transparency**
  - *TF-IDF + SVM* is the most interpretable n grams directly influence decisions
  - Transformers are harder to interpret due to complex attention mechanisms
- **Privacy** :Where does your data go, and what is it used for?
  - Prompt engineering via external APIs sends user data to third parties
  - This poses privacy risks for sensitive platforms like Wikipedia
  - RGPD Violation