Projet 1 Machine Learning

Toxic Comment Classification

Orlane Bochard, Pauline Dumas, Simon Labracherie, Emeline Lavaux

19 Septembre 2025

Part 1

Preprocessing

Part 1 — Preprocessing

Objective : Transform comments so the machine learning models can understand them *without destroying toxicity cues*.

Part 1 — Preprocessing : Core steps

- **Step 1 : Cleaning** — remove URLs, emails, @mentions, HTML
- Preserve **aggression signal** : keep **ALL CAPS** words and repeated punctuation (`!!`, `??`).
- **Step 2 : Tokenization** — split text into words
- **Step 3 : Removing stopwords** — *except* :
  - **Negations** : `not, no, never` (flip polarity).
  - **2nd person** : `you, your` (direct targeting).

Goal : clean the text *without destroying* toxicity cues that models need.

Part 1 — Preprocessing : Additional tweaks

Deobfuscation (leetspeak → plain text)

- Map symbols to letters : `@`→`a`, `1`→`i`, `$`→`s`, `0`→`o`, . . .
- Example : `$illy` → `silly`, `b!tch` → `bitch`.
- Purpose : prevent users from bypassing detection with obfuscation.

Normalize repeated characters (max 3 letters)

- Compress runs : `AAAAAAHHHHHHHH` → `AAAHHH`, `soooooooooo` → `sooo`.
- Reduces sparsity without losing meaning.

```
DEOB = str.maketrans({
    '@': 'a', '4': 'a',
    '1': 'i', '|': 'i',
    '3': 'e',
    '0': 'o',
    '$': 's', '5': 's',
    '7': 't',
    '€': 'e', '£': 'l'
})
```

Figure – Example mapping used in our code

Part 1 — Preprocessing : skipped step

Stemming/Lemmatization step.

For toxicity, exact word forms matter

Ex : `"fucked"`, `"fucking"`, `"fuck"`

Don't convey the same tone

Should not be reduced to the same root

Results of preprocessing & Class Imbalance

```
Preprocessing preview:
                                        comment_text  \
0  Explanation\nWhy the edits made under my usern...
1  D'aww! He matches this background colour I'm s...
2  Hey man, I'm really not trying to edit war. It...
3  "\nMore\nI can't make any real suggestions on ...
4  You, sir, are my hero. Any chance you remember...
5  "\n\nCongratulations from me as well, use the ...
6       COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK
7  Your vandalism to the Matt Shirvington article...

                                      processed_text
0  explanation edits made username hardcore metal...
1  daww matches background colour seemingly stuck...
2  hey man really not trying edit war guy constan...
3  ca nt make real suggestions improvement wonder...
4             you sir hero chance you remember page
5          congratulations well use tools well talk
6                     COCKSUCKER YOU PISS AROUND WORK
7  your vandalism matt shirvington article revert...
```

Class imbalance in dataset :

Class	Count	Distribution
0 (non-toxic)	144,277	90.42%
1 (toxic)	15,294	9.58%

- *Implication :* accuracy is misleading ; we will report Precision/Recall/F1 and PR-AUC in later parts.

Part 2
Feature Engineering

Part 2 — Feature Engineering : BoW dimentionality

- Compare different **Bag-of-Words** representations :

N-grams	**Shape**	**Features**
Uni-grams (1,1)	159k $\times$ 210k	210,067
Uni+Bi-grams (1,2)	159k $\times$ 2.9M	2,905,614
Uni+Bi+Tri-grams (1,3)	159k $\times$ 7.3M	7,344,371

- **Uni-grams only** : miss context (*"shut up"* *"go die"*).
- **Tri-grams** : capture more context but increase sparsity
- Best trade-off between context and dimensionality : **Uni+Bi-gram**

Part 2 — Feature Engineering : TF-IDF vectorization

- **TF-IDF :** gives more weight to discriminative terms
- We tested several configurations on a Linear SVM model

Config	Max feat.	Min df	Max df	F1-weighted	F1-macro
#1	50k	5	0.80	0.9520	0.8662
#2	100k	5	0.80	0.9561	0.8756
#3	150k	3	0.85	0.9574	0.8786
#4	150k	1	0.85	0.9577	0.8793
#5	1.5M	1	0.85	0.9615	0.8891
#6	2.0M	1	0.85	**0.9616**	**0.8891**

Best configuration : #6 (2M features, 1–2 grams, min_df=1, max_df=0.85)

- include rare toxic terms, filter very frequent terms.

Part 3

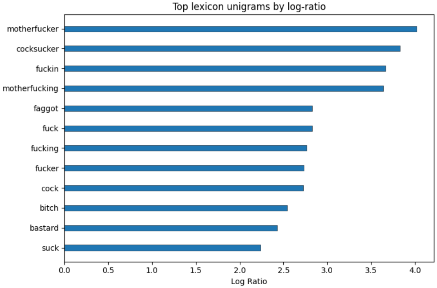
Modeling

Part 3 — Modeling : Overview

Objective : Train models to detect toxic comments.

- **Baseline (lexicon-based) :** predict toxic if a toxic word/phrase appears.
- **Classical ML (TF-IDF features) :**
  - Logistic Regression
  - Linear SVM
- **Deep learning :** DistilBERT fine-tuned on our dataset.

Part 3 — Modeling : Lexicon baseline



- Bars = **log-ratio** (toxic vs. non-toxic) ; values > 0 = more typical of toxic comments.
- Used to **seed the lexical baseline** (cue words).
- Simple but **limited** : ignores context/sarcasm, easy to bypass with obfuscation.

Pros :

- Simple, transparent, interpretable.

Cons :

- Misses subtle toxicity (sarcasm, paraphrases).
- Fragile : bypassed by spelling variations.

Part 3 — Modeling : Classical ML

Features : TF-IDF with 1–2 grams, ∼2M features (fitted on TRAIN to avoid leakage).

Models :

- **Logistic Regression (LR) :** robust with sparse high-dim data ; improved with `class_weight=balanced`.
- **Linear SVM :** strong text classifier, finds maximum-margin separation.

Results (TEST set) :

Model	Toxic F1	Macro F1	ROC-AUC
Logistic Regression	0.74	0.86	0.972
Linear SVM	0.80	0.89	0.976

LR reaches higher recall (0.85) but lower precision, while Linear SVM achieves a stronger balance overall.

Part 3 — Modeling : DistilBERT

Why use DistilBERT ?

- **Context-aware :** captures meaning beyond keywords.
- **Robust :** subword tokenization handles misspellings/obfuscation.
- **Data-efficient :** pretrained on large data sets, adapts well to imbalanced data.
- **Practical :** lighter and faster than BERT-base, suitable for this project.

Training setup :

- Re-weighted the loss to balance toxic vs non-toxic.
- Tuned the treshold on the validation set to maximize F1.

Part 3 — Modeling : Threshold as a product lever

Why threshold matters :

- **High threshold** (strict) : prioritize Precision, fewer false positives.
- **Low threshold** (lenient) : prioritize Recall, fewer false negatives.

Use cases :

- **Automatic blocking :** high threshold → protect free speech.
- **Moderator support :** low threshold → catch more toxicity.

Takeaway : The threshold is not just a technical detail, but a real-world **policy choice**.

Part 3 — Modeling : Threshold comparison (DistilBERT)

Policy = balanced (tuned on PR)

Class	P	R
non-toxic	0.98	0.98
toxic	0.82	0.85

Acc = 0.97, Macro-F1 = 0.91

Confusion (counts)
TN=28 287, FP=569

FN=452, TP=2 607

Threshold = 0.5 (reference)

Class	P	R
non-toxic	0.98	0.98
toxic	0.85	0.83

Acc = 0.97, Macro-F1 = 0.91

Confusion (counts)
TN=28 389, FP=467

FN=512, TP=2 547

ROC-AUC = 0.9864 (invariant to threshold) — PR-AUC (toxic) = 0.9216

- **Balanced vs 0.5 :** toxic precision ↓ (0.85→0.82), toxic recall ↑ (0.83→0.85).
- **Trade-off :** FP ↑ (467→569) but FN ↓ (512→452).
- **Takeaway :** A lower threshold means the model catches more toxic comments but also risks blocking more innocent ones. For auto-blocking we prefer precision, for moderator support we prefer recall.

Part 4

Evaluation

Part 4 — Evaluation : Metrics under class imbalance

Context : $\sim$90% non-toxic vs $\sim$10% toxic.

- **Accuracy is misleading** : a trivial "always non-toxic" model $\approx$ 90%.
- We report per-class **Precision**, **Recall**, **F1**, and **Macro-F1**.
- **ROC-AUC** (ranking quality) and **PR-AUC (toxic)** :
  - ROC-AUC can look high under imbalance.
  - **PR-AUC for the toxic class** is more informative (rare positives).

Part 4 — Evaluation : Model comparison

Table 1: Comparison of Model Performance

Model	**Toxic Precision**	**Toxic Recal**	**Toxic F1-score**	**Macro F1-score**	ROC-AUC	PR-AUC (toxic)
Lexicon Baseline	**0.85**	0.52	0.67	0.81	0.757	0.521
Logistic Regression (TF-IDF)	0.66	**0.85**	0.74	0.86	0.971	0.856
Linear SVM (TD-IDF)	0.80	0.79	0.80	0.89	0.976	0.880
DistilBERT	0.85	0.83	**0.84**	**0.91**	**0.986**	**0.922**

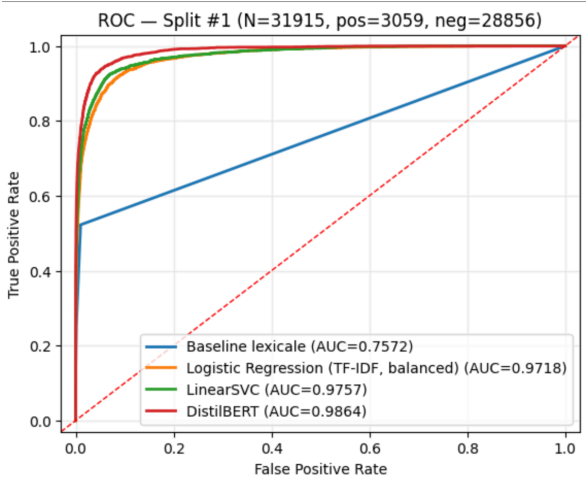
Part 4 — Evaluation : Best model (DistilBERT)

Classification report (Test)

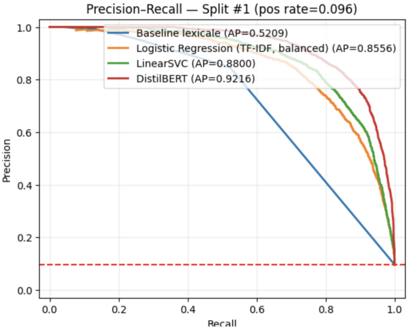
Class	Precision	Recall	F1-score	Support
non-toxic	0.984	0.980	0.982	28 856
toxic	0.821	0.852	0.836	3 059
accuracy			0.968	31 915
macro avg	0.903	0.916	0.909	31 915
weighted avg	0.969	0.968	0.968	31 915

ROC-AUC : 0.9864

Part 4 — Evaluation : ROC curves (Test)



Part 4 — Evaluation : Precision–Recall curves (Test)



Under imbalance, PR is more revealing : DistilBERT reaches **PR-AUC** $\approx$ **0.92**, while the baseline collapses quickly.

Part 5

Reflexion

Which metric should matter most ?

- **Accuracy** is misleading with imbalanced data.
- **Recall** is important : reduce toxic comments left online.
- **Precision** is also crucial : avoid unfairly censoring users.
- On Wikipedia, too many false positives risk frustrating volunteers and discouraging contributions.
- **Best compromise : F1-score** (balance between precision and recall).

Consequences of each error type

- **False Positive** (non-toxic $\rightarrow$ toxic) :
  - Commenter : unfair censorship, frustration.
  - Other users : lose access to useful content.
  - Platform : discourages contributions, harms trust.
- **False Negative** (toxic $\rightarrow$ non-toxic) :
  - Users : exposed to harmful/shocking content.
  - Moderators : extra workload to correct errors.
  - Platform : loses credibility and safety.

Context matters : deployment mode

- **Automatic blocking :**
  - False positives = unjust censorship.
  - Precision becomes the key metric.
- **Moderator support tool :**
  - False negatives = missed toxic content.
  - Recall becomes more important.
- **Summary :** Blocking favors precision, support tools favor recall.

Analogy with COVID testing

- False positive test : healthy person isolated unnecessarily.
- False negative test : infected person spreads the disease.
- Both Wikipedia moderation and medical testing involve **asymmetric error costs**.
- **Difference :** moderation also touches on **free speech**.
  - False positive = silencing a legitimate contributor.
  - False negative = harmful content remains visible.

Conclusion

- Moderation is always a trade-off.
- **Automatic decisions :** protecting free speech (precision) comes first.
- **With human moderators :** maximizing recall protects the community.
- As in COVID testing, errors have asymmetric costs — here, between community safety and free expression.