Project 1 : Toxic Comment Classification

Orlane BOCHARD, Pauline DUMAS, Simon LABRACHERIE, Emeline LAVAUX

1) Preprocessing

- For data pretraining, we performed: *cleaning* (URL's, emails), *deobfuscation*, *normalisation*, *tokenization*, *remove stop words* (except negation and second-person pronouns ("you").
- *Intensity preserved* : ALL-CAPS words, repeated punctuation (!!!, !?, ??), no stemming.

2) Feature engineering

- Use 3 different *Bags of Word* (*BoW*).

BoW Representation

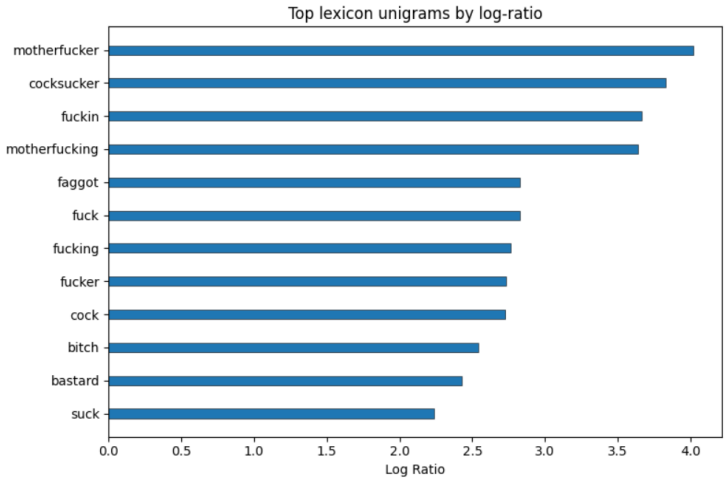
N-grams	Shape
Unigram (1,1)	(159571, 210067)
Uni + Bigram (1,2)	**(159571, 2905614)**
Uni + Bi + Trigram (1,3)	(159571, 7344371)

- **Unigrams only:** fail to capture *context/phrases* (e.g., "shut up", "go die") **Trigrams:** *dimensionality* and *sparsity* explode
- **Best trade-off:** uni+bi-grams (1,2)
- Instead of BoW, we use TF-IDF with N-gram = (1,2) to consider the value of words and discriminative terms.
- **Best Config 6:** `"max features"=2000000, "ngram range"=(1, 2), "min df"=1, "max df"=0.85`
  Improved the F1-score of SVM.

3) Modeling

0.1 Baseline

- We consider both Unigram and Bigram to compare with our models builds using TF-IDF(1,2) which yielded the best config.
- Dictionnary of toxic word used : "asshole", "cocksucker", "fucking cunt"...
- Method : compute a log-ratio of relative frequencies to identify n-grams overrepresented in the toxic class.



0.2 Models

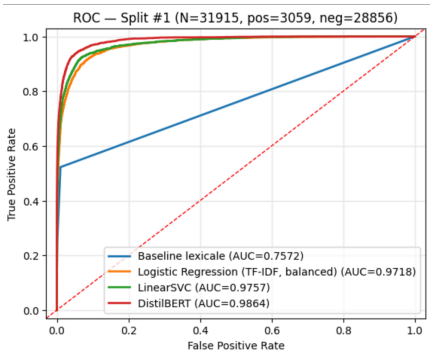
- 3 Models : LR, SVM and DistilBERT, a Transformer model that understands context, and therefore detects toxicity.
- To mitigates class imbalance, we use $class\_weight =$ "*balanced*"

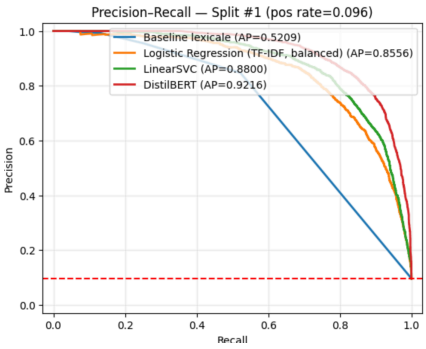
4) Evaluation

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**

- **Best performing Model:** *DistilBERT* – strong ability to detect toxicity with highest F1-score.





5) Reflection

- If Wikipedia deploys our model :

  Accuracy : less informative and misleading with imbalanced data as toxic comments.

  Recall : matter the most to catch toxic comments.

  Precision : also important to avoid misclassifying non-toxic ones.

- Wikipedia would probably favor precision because the platform is based on voluntary contributions from a lot of users so misclassifying too many innocent comments could frustrate contributors and discourage participation.

- Best metric : F1-score : Wikipedia must balance recall and precision.

Consequencies of each error type

Error type	Who is harmed ?	Consequenses
False Positive (non toxic classified as toxic)	Commentor	Unfairly censored, lose of confidence, frustration
	Users	Lose of information and useful content
	Platform	Risk to free expression, lower participation
False Negative (toxic classified as non toxic)	Commenter (toxic)	No impact
	Users	Exposed to harmful, schocking content
	Moderators	Extra workload to catch error manually
	Platform	Loss of credibility and quality with unsafe platform.

Error consequencies and based metric depending on the context

Context	False Positive	False Negative	Best Metric
Automatic blocking	More harmful : unfair censorship of non-toxic comment	Harmful for users but the comment can still be signaled manually	Precision
Moderator Support Tool	Less harmful, human can correct the error	More harmful : if a toxic comment isn't detected	Recall

In health, false negatives are most critical since disease spreads undetected, whereas on Wikipedia the trade-off involves both protecting users and safeguarding free speech. Age-based moderation (e.g., stricter filters for minors) could help balance these risks.

Conclusion

- **Findings**: DistilBERT was great for binary. It appears to be the best performing model.
- **Limits**: Transformers took a lot of time. Class imbalance.
- **Threshold as a lever:** the decision threshold directly shapes the model's behavior.
  - A higher threshold → favors **precision**, fewer false positives, protects **freedom of expression**.
  - A lower threshold → favors **recall**, fewer false negatives, protects **users from toxicity**.
- **Takeaway:** Moderation is not only technical but also a **policy choice**. Automatic blocking should prioritize precision, while moderator support should prioritize recall.

[1] we used AI/tool assistance for the code and Hugging Face website for the transformer model