Real-Time Comment Moderation with MLOps

Presented by:

Anusha Bhat
Soham Mandal
Bruna Medeiros
John Melel

Business Problem

Problem

Online platforms face growing pressure to moderate user-generated content for safety, civility and compliance.

Goal

Our goal is to build an end-to-end MLOps pipeline that flags inappropriate comments in real-time with explainability, monitoring, and automation to support safer user interactions at scale.

Data

Dataset

- **223,549 comments** from Wikipedia Talk Pages.
- **160k training** samples**, 64k testing** samples
- Target: **moderation label (binary)**

Features

- **Comment text**
- **Moderation label** for toxic vs. non-toxic identification
- **Toxicity flags** (e.g. obscene, threatening, insult, identity hate)

Class Imbalance

- **~10%** of comments are **toxic**
- Addressed in **training**

Data

Toxic Comments

“How dare you vandalize that page about the HMS Beagle! Don't vandalize again, demon!”

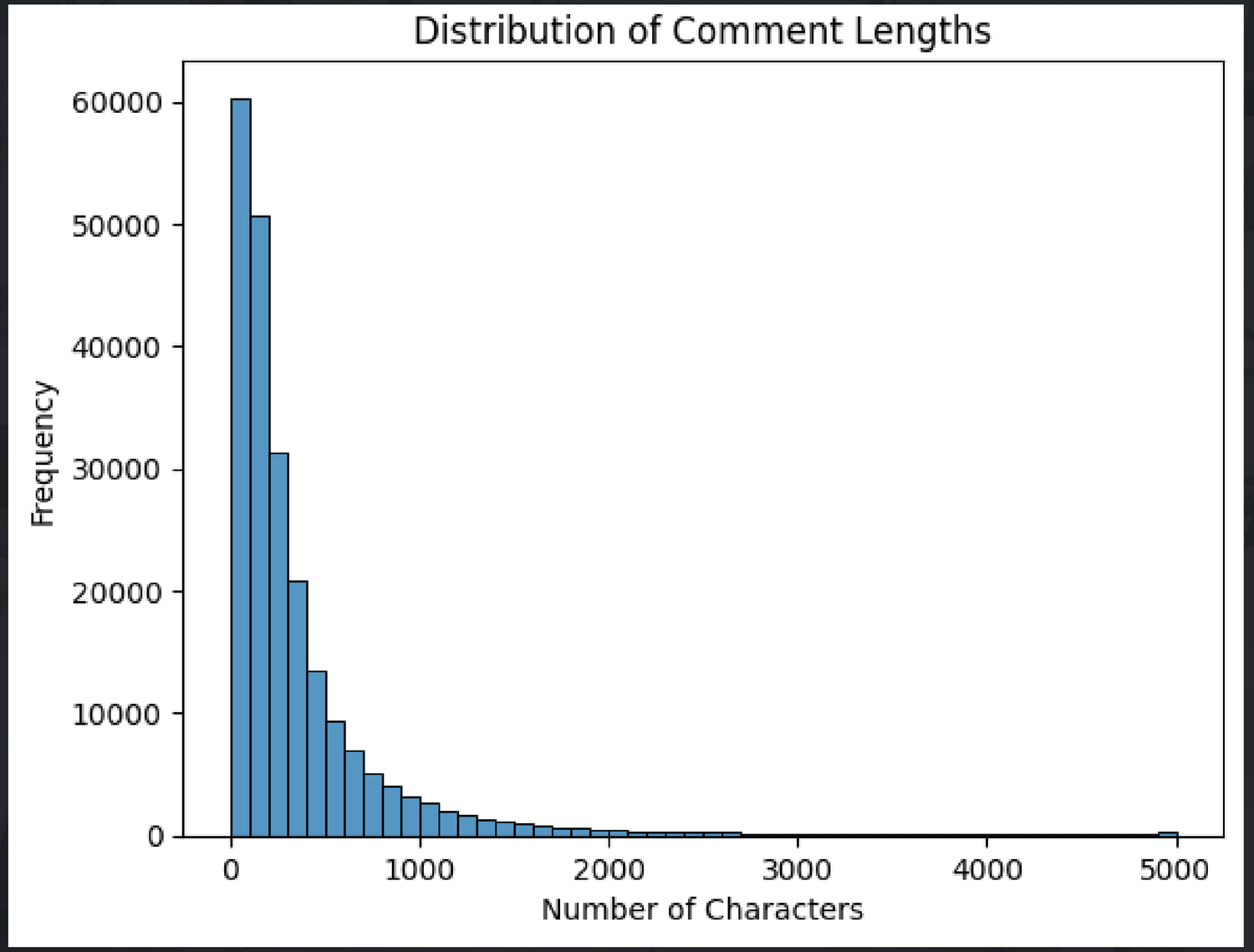
“Hey asshole. Show some respect.”

Non-toxic Comments

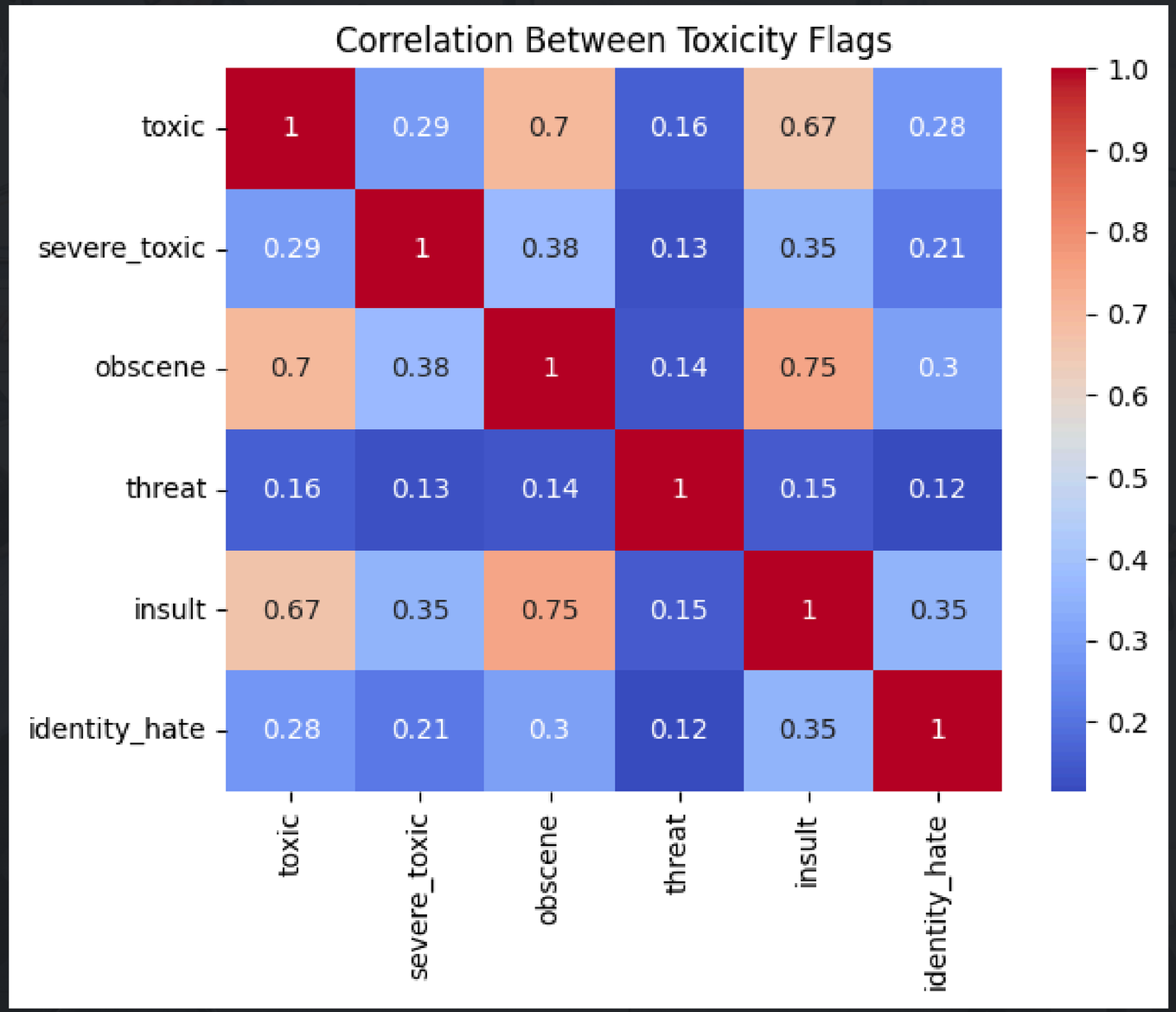
“Thanks for your response. I'll nominate them today.”

“Please stop your disruptive editing. If you continue to vandalize Wikipedia, as you did at Warwick School, you will be blocked from editing.”

A majority of comments are less than 300 characters.



Obscenity and Insult flags moderately correlate with a toxic label



Model Metrics

Recall

- Focuses on correctly identifying toxic comments since we have a class imbalance
- Aligns with the business cost of minimizing the misclassification of toxic comments (false negatives)

F1-Scores

- Allows for equal weightage of model's performance on the smaller, minority class compared to the majority class
- Balances precision and recall for both classes

FLAML AutoML Pipeline

Preprocessor

Find best **preprocessing pipeline**:

- Test TFIDF Char 3-5
- TFIDF Word12
- Hashing Vectorizer

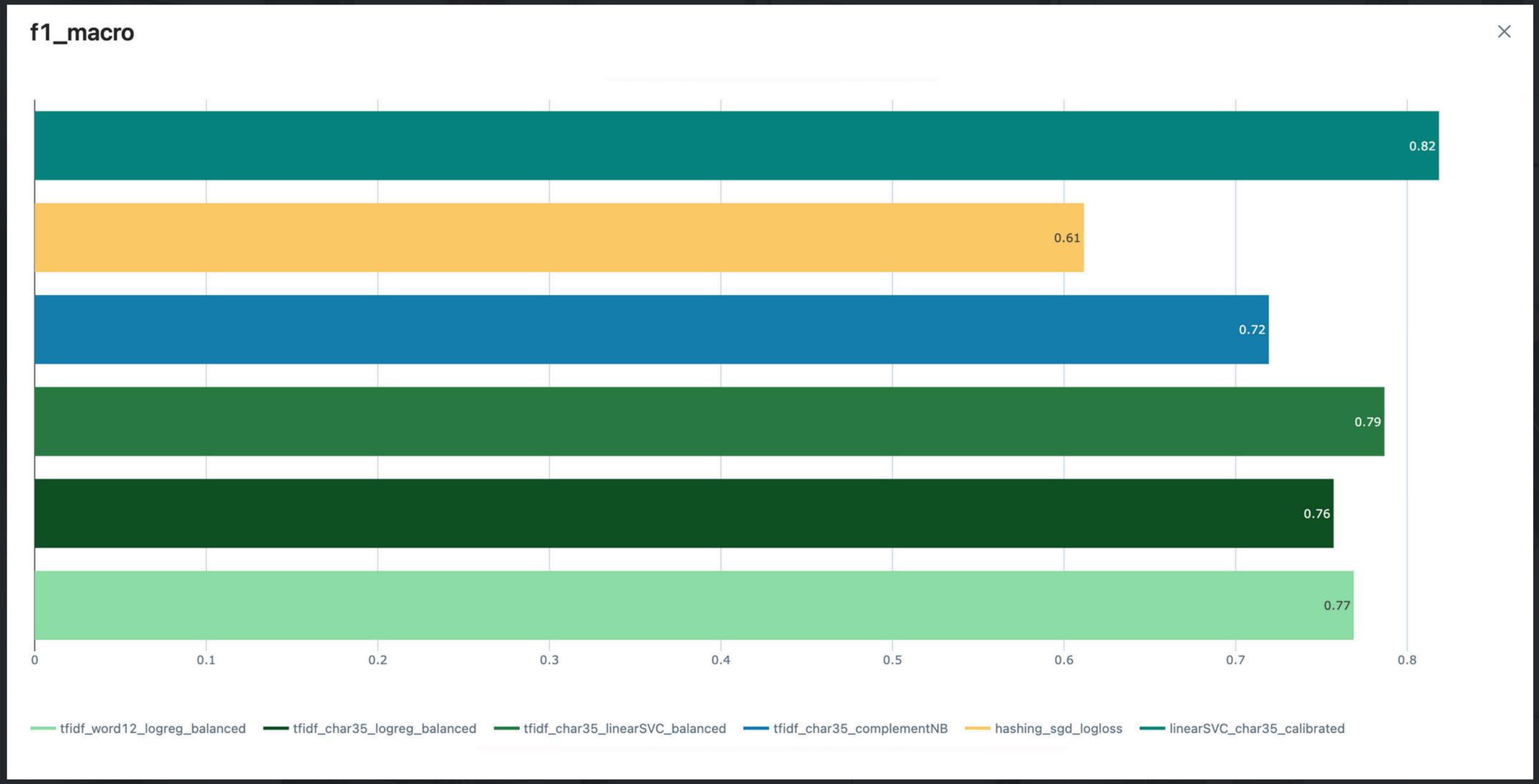
Model

Find best **model** once best preprocessing is finalized

Comparison

Compare with state-of-the-art **Hugging Face BERT model**

MLFlow Runs



AutoML MLFlow Runs
(TFIDF Char 3-5 and Logistic Regression with L2 Regularization)

Best Model Metrics

Best Model: Logistic Regression (L2, TF-IDF char 3–5)

- Best model ..... **lr with l2 penalty**
- Best config ..... **C = 1.0**
- Best CV score **0.263**

BERT Class Breakdown

Class 0		Class 1			
Precision	0.99	Precision	0.55	Accuracy	92%
Recall	0.92	Recall	0.91	F1 (Weighted)	0.93
F1	0.95	F1	0.68	F1 (Macro)	0.82

Best Model Class Breakdown (Test Set)

Class 0		Class 1			
Precision	0.97	Precision	0.66	Accuracy	94%
Recall	0.96	Recall	0.71	F1 (Weighted)	0.94
F1	0.96	F1	0.69	F1 (Macro)	0.83

Our model performs on par with the BERT model.

Future improvements can include improving precision and recall for the toxic class.

“Changed” Test Data

V1 **(Original)**

- Raw test dataset with unmodified comments.

V2 **(Synonym Augmentation – NLTK)**

- Replaced ~10% of words in each row with WordNet synonyms.
- Simulates paraphrasing and vocabulary variation.

V3 (Noise Augmentation – Typos & Dropout)

- Added random character-level typos.
- Randomly removed ~10% of words.
- Mimics user input errors and shorthand.

Together, these simulate real-world data drift (paraphrasing, typos, missing words) for testing model robustness

Model Monitoring with Evidently

Load Model from S3

Best performing model from S3:

model_lrl3_20250812_052739.joblib

(Logistic Regression + TF-IDF)

Run Inference

Two test sets evaluated:

Reference set:
original held-out test data

Changed set:
same test set with synonym swaps and typos introduced

Compare with Evidently

Used Evidently to **compare distributions** and **classification performance** side by side

Generated **full monitoring report** (HTML) and uploaded to Evidently Cloud

Evidently Metrics

Reference Set (Original):

- Accuracy: **93.8%**
- F1 (Weighted): **0.94**
- F1 (Macro): **0.82**

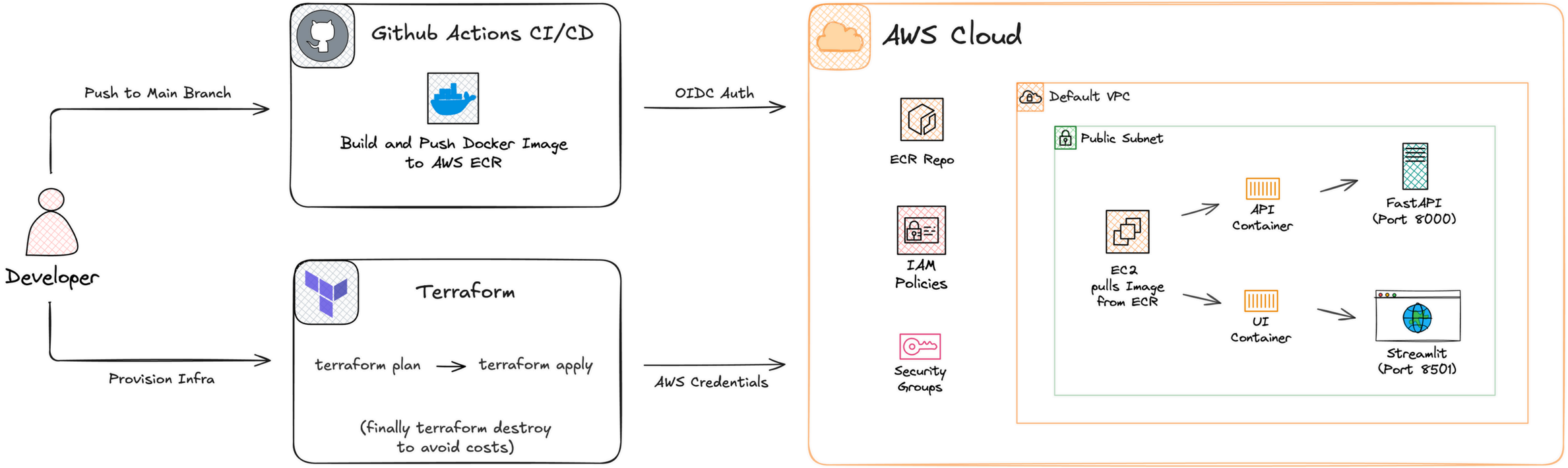
Changed Set (Synonym Swap and Noise Injection):

- Accuracy: **93.3%**
- F1 (Weighted): **0.9267**
- F1 (Macro): **0.7746**

Drift Summary:

- **F1 Macro** dropped by **0.0470**
- **Weighted F1** remained stable (only **-0.0106**)
- Model is robust overall, but minor drop in rare class performance under perturbations
- **Drift** detected in text feature (score of **0.843174**)

Deployment Architecture



Demo!

Find out if your comment is toxic or safe!

Thank you!

Repository:
https://github.com/sohammandal/mlops-comment-moderation

Team Duties

- Soham: terraform, EC2, & docker setup, model training
- John: test set change with Evidently, model training
- Anusha: model training, EDA, presentation
- Bruna: model training, presentation