جامعة عجمان
AJMAN UNIVERSITY

SEMEVAL-2026 TASK 13: DETECTING MACHINE-GENERATED CODE WITH MULTIPLE PROGRAMMING LANGUAGES, GENERATORS, AND APPLICATION SCENARIOS

PROF. MAHMOUD ALSHBOUL

MARWA MARWAN ALNAJAR
202412504

SHAIKHA SHAMMA ALNUAIMI
202412586

Task Overview

Identifying machine-generated, hybrid, adversarial, and human-written code across multiple programming languages and LLM generator families.
Subtasks:

- A: Binary classification (Human vs AI)
- B: Multi-class authorship attribution (11 classes)
- C: Hybrid + adversarial detection (4 classes)

Dataset Summery

- Large multilingual dataset (.parquet format)
- Covers: Python, C++, Java, Go, PHP, C#, C, JavaScript
- Includes human-written and LLM-generated code
- Separate train/validation/test sets
- Labels:
- Subtask A → 0/1
- Subtask B → 11 classes
- Subtask C → {0,1,2,3} categories

Preprocessing

Subtask A

- Convert to binary labels
- Replace missing fields with empty strings
- Character-level TF–IDF

Subtask B

- Keep whitespace, symbols, indentation
- No comment removal
- Concatenate <language> + code

Subtask C

- Merge train + validation
- Use raw code
- Convert to character n-grams → TF–IDF features (sparse matrix X)

System Description

- Use TF–IDF (char-level) to extract stylistic patterns
- Convert code into 3–5 character n-grams
- Use SGDClassifier (linear model)
- Train on combined train + validation sets
- Bundle vectorizer + classifier into one model object

System Description

Subtask A System

- Binary classifier using TF–IDF → SGD (hinge/log loss)
- Targets OOD generalization across unseen languages & domains
- Tested multiple n-gram settings (3–4, 3–5) and feature sizes up to 200k
-

Subtask B System

- Multinomial SGD classifier
- TF–IDF with n-grams (3–4, 3–5)
- Vocabulary tested up to 200k
- Captures stylistic fingerprints of LLM families
- Best config: (3,4), 150k features, α=5e−5

Subtask C System

- Simple classical model
- TF–IDF (3–5 n-grams) → Linear SGD classifier
- Merged training + validation
- Predicts among 4 classes: Human, AI, Hybrid, Adversarial
- Heavy computational load caused kernel to stop; no implementation errors

RESULTS (SUBTASK A)

- Strong recall for class 1 (AI-generated)
- Difficulty distinguishing human code → slight bias
- Needs class balancing or language-specific models in future

RESULTS (SUBTASK A)

```
Using data folder: /Users/marwa/Desktop/Task_A
Files in folder: ['train.parquet', 'test.parquet', 'test_sample.parquet', 'validation.parquet', 'sample_submission.csv']
Train: (500000, 4)
Val: (100000, 4)
Test: (1000, 2)
Test_sample: (1000, 4)

================ HYPERPARAMETER SEARCH ==================

Training config: 3_5_200k_a1e-4
  Val Accuracy : 0.9326
  Val Macro F1 : 0.9326

Training config: 3_5_150k_a1e-4
  Val Accuracy : 0.9330
  Val Macro F1 : 0.9329

Training config: 3_4_150k_a5e-5
  Val Accuracy : 0.9409
  Val Macro F1 : 0.9408

Training config: 3_5_200k_a5e-4
  Val Accuracy : 0.8969
  Val Macro F1 : 0.8969

===== Validation Results (sorted by Macro F1) =====
3_4_150k_a5e-5: Macro F1=0.9408, Accuracy=0.9409, ngram=(3, 4), max_features=150000, alpha=5e-05
3_5_150k_a1e-4: Macro F1=0.9329, Accuracy=0.9330, ngram=(3, 5), max_features=150000, alpha=0.0001
3_5_200k_a1e-4: Macro F1=0.9326, Accuracy=0.9326, ngram=(3, 5), max_features=200000, alpha=0.0001
3_5_200k_a5e-4: Macro F1=0.8969, Accuracy=0.8969, ngram=(3, 5), max_features=200000, alpha=0.0005

Best config selected: {'name': '3_4_150k_a5e-5', 'ngram': (3, 4), 'max': 150000, 'alpha': 5e-05, 'val_acc': 0.94088, 'val_f1': 0.940840560034
1754}

================ FINAL TRAINING ON TRAIN+VAL ==================

Final model saved as: /Users/marwa/Desktop/Task_A/subtaskA_sgd_3_4_150k_a5e-5.joblib

================ TEST_SAMPLE EVALUATION ==================

Test_sample Accuracy : 0.3170
Test_sample Macro F1 : 0.3091
```

RESULTS (SUBTASK A)

```
================ TEST_SAMPLE EVALUATION ==================

Test_sample Accuracy : 0.3170
Test_sample Macro F1 : 0.3091

Classification report on test_sample:

              precision    recall  f1-score   support

           0       0.91      0.14      0.24       777
           1       0.24      0.95      0.38       223

    accuracy                           0.32      1000
   macro avg       0.57      0.54      0.31      1000
weighted avg       0.76      0.32      0.27      1000


================ GENERATING SUBMISSION FILE ==================

Submission file created: /Users/marwa/Desktop/Task_A/submission_subtaskA_3_4_150k_a5e-5.csv
```

RESULTS (SUBTASK B)

- Good performance for major LLM families
- Struggles with minority classes
- Best macro F1 (validation): 0.2746
- Test_sample macro F1: 0.2154

RESULTS (SUBTASK B)

```
Training config: char_3_5_200k_alpha1e-4
  Validation Accuracy : 0.8644
  Validation Macro F1 : 0.2443

Training config: char_3_5_150k_alpha1e-4
  Validation Accuracy : 0.8639
  Validation Macro F1 : 0.2452

Training config: char_3_4_150k_alpha5e-5
  Validation Accuracy : 0.8695
  Validation Macro F1 : 0.2746

Training config: char_3_5_200k_alpha5e-4
  Validation Accuracy : 0.8577
```

```
===== Validation Results (sorted by Macro F1) =====
char_3_4_150k_alpha5e-5: Macro F1=0.2746, Accuracy=0.8695, ngram=(3, 4), max_features=150000, alpha=5e-05
char_3_5_150k_alpha1e-4: Macro F1=0.2452, Accuracy=0.8639, ngram=(3, 5), max_features=150000, alpha=0.0001
char_3_5_200k_alpha1e-4: Macro F1=0.2443, Accuracy=0.8644, ngram=(3, 5), max_features=200000, alpha=0.0001
char_3_5_200k_alpha5e-4: Macro F1=0.2037, Accuracy=0.8577, ngram=(3, 5), max_features=200000, alpha=0.0005

Best config selected: char_3_4_150k_alpha5e-5
```

RESULTS (SUBTASK B)

```
================ TRAINING FINAL MODEL ON TRAIN+VAL ================

Final model saved as: subtaskB_sgd_char_3_4_150k_alpha5e-5.joblib

================ EVALUATION ON TEST_SAMPLE (LOCAL ONLY) ================

Test_sample Accuracy : 0.5300
Test_sample Macro F1 : 0.2154

Classification report on test_sample:

              precision    recall  f1-score   support

           0       0.77      0.95      0.85       474
           1       0.14      0.05      0.07        21
           2       0.50      0.05      0.10        73
           3       0.35      0.33      0.34        21
           4       0.11      0.20      0.14        10
           5       0.09      0.42      0.15        36
           6       0.25      0.30      0.27        54
           7       0.00      0.00      0.00        61
           8       0.09      0.39      0.15        18
           9       0.12      0.06      0.08        18
          10       0.57      0.14      0.22       214

    accuracy                           0.53      1000
   macro avg       0.27      0.26      0.22      1000
weighted avg       0.55      0.53      0.49      1000
```

RESULTS (SUBTASK C)

- Kernel froze due to extremely large dataset + high-dimensional TF–IDF
- Not an implementation error

Future runs require:

- Smaller dataset
- Fewer features (80k instead of 150k)
- Reduced n-gram range
- Machine with higher RAM

RESULTS (SUBTASK C)

```
Using data folder: /Users/marwa/Desktop/Task_C
Files: ['train.parquet', 'test.parquet', 'test_sample.parquet', 'validation.parquet', 'sample_submission.csv']
Train: (900000, 4)
Validation: (200000, 4)
Test: (1000, 2)

Unique labels: [1 0 2 3]
Number of classes: 4

Starting hyperparameter search on subsets...

Training subset size: 200000
Validation subset size: 80000

Testing configuration: 3_4_200k_a5e-5
python(79532) MallocStackLogging: can't turn off malloc stack logging because it was not enabled.
Training time (seconds): 335.8
Validation Accuracy: 0.7022
Validation Macro F1: 0.4915

Testing configuration: 3_4_150k_a1e-4
Training time (seconds): 302.4
Validation Accuracy: 0.6828
Validation Macro F1: 0.4528

Testing configuration: 3_5_150k_a5e-5
Training time (seconds): 2616.8
Validation Accuracy: 0.7051
Validation Macro F1: 0.4968


Hyperparameter search results (sorted):
3_5_150k_a5e-5 | F1 = 0.4968 | Accuracy = 0.7051 | ngram = (3, 5) | max_features = 150000 | alpha = 5e-05
3_4_200k_a5e-5 | F1 = 0.4915 | Accuracy = 0.7022 | ngram = (3, 4) | max_features = 200000 | alpha = 5e-05
3_4_150k_a1e-4 | F1 = 0.4528 | Accuracy = 0.6828 | ngram = (3, 4) | max_features = 150000 | alpha = 0.0001

Best configuration: {'name': '3_5_150k_a5e-5', 'ngram': (3, 5), 'max': 150000, 'alpha': 5e-05, 'val_acc': 0.7051125, 'val_f1': 0.4967940155690
02764, 'fit_time': 2616.839792728424}

Training final model on full dataset
```

Kaggle Submission

SemEval-2026-Task13-Subtask-B

Submit Prediction

Overview | Data | Code | Models | Discussion | Leaderboard | Rules | Team | Submissions

Submission and Description	Public Score	Select
submission_sgd_balanced_new.csv Complete · Sh.Shamma Alnuaimi · 11d ago	0.21204	☐
submission_sgd_char_3_4_150k_alpha5e-5.csv Complete · Sh.Shamma Alnuaimi · 11d ago	0.19284	☐
submission_subtaskB.csv Error · Sh.Shamma Alnuaimi · 11d ago		
submission1.csv Complete · Sh.Shamma Alnuaimi · 13d ago	0.23862	☐
submission.csv Complete · Sh.Shamma Alnuaimi · 13d ago	0.24379	☐
submission.csv Complete · Sh.Shamma Alnuaimi · 13d ago	0.23296	☐

submission_subtaskA_3_4_150k_a5e-5.csv Complete · Marwa Alnajjar · 11d ago	0.31771	☐
submission_subtaskA.csv Complete · Marwa Alnajjar · 1mo ago · First submission — TF-IDF + Logistic Regression (baseline binary m...	0.35459	☐

Future Work

Future research will focus on three areas:

- Focal loss variants and class rebalancing;
- Ensemble architectures that combine lexical and semantic features; and
- Language-specific modelling as opposed to a single multilingual vectorizer.

THANK YOU