Team ID: 3

Team Name: Data Verse

Multilingual AI Generated Text Detection using Comparative Classification Models

UNIVERSITÉ Concordia UNIVERSITY

Abstract

This project explores the task of detecting AI-generated text in a multilingual setting, focusing on English and Chinese. We evaluate three models:

1. TF-IDF + Naive Bayes — statistical baseline
2. FastText + FNN — embedding-based model
3. mBERT — contextual transformer model

All models use the same multilingual setup. Naive Bayes provides the baseline, FastText improves representation, and mBERT achieves the best accuracy through stronger contextual understanding.

Goal of the Project

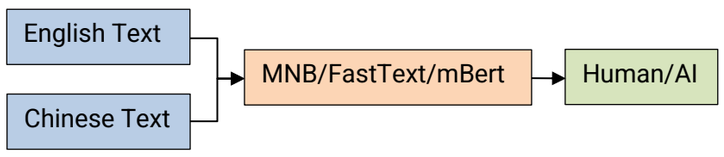
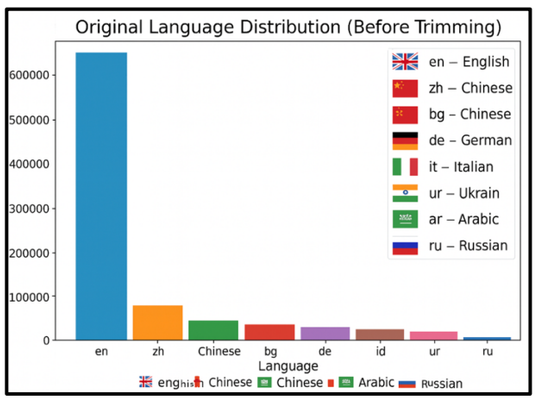
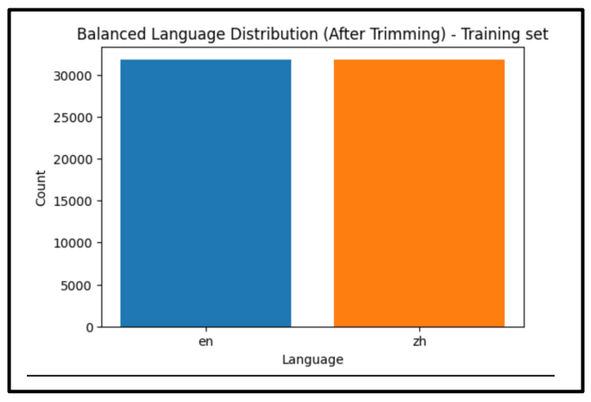


Figure 1. Goal of the project

The goal of this project is to build a multilingual system capable of distinguishing AI-generated text from human-written text and to compare three different modeling approaches to understand how linguistic complexity, context modeling, and representation learning impact detection performance.

Dataset





- Filtered the dataset to EN and ZH samples from the COLING 2025 MGTD dataset.[1]
- Cleaned text using:
  - Lowercase for EN
  - Jieba segmentation for ZH [2]
- Applied consistent sampling to ensure balanced multilingual training, inspired by best practices in multilingual model evaluation.

Methodology

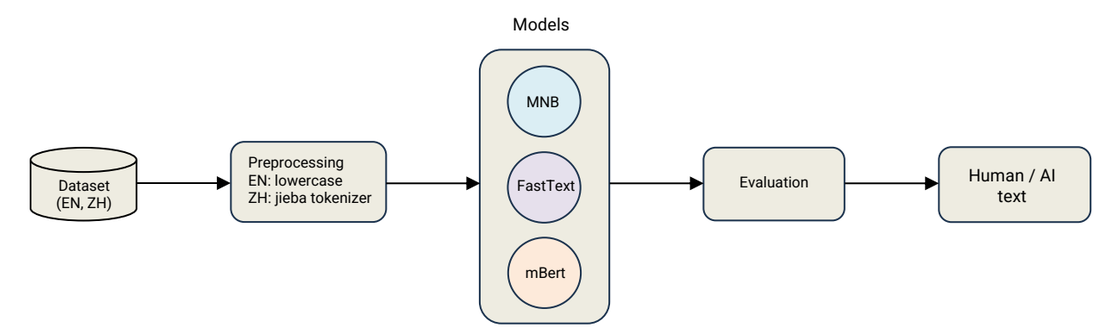


Figure 2. Architecture of the multilingual AI-generated text detection

Model 1 – Multinomial Naïve Bayes (MNB)

- Used TF-IDF vectorisation with unigrams and bigrams.
- Trained MNB classifiers separately for EN and ZH

Model 2 – FastText Embeddings + Feedforward Neural Network

- Loaded pretrained multilingual FastText word vectors.
- Produced sentence embeddings by average word embeddings, a method frequently applied in shallow neural text classifiers.[3]
- Trained an FNN for binary classification.

Model 3 – Multilingual BERT (mBERT)

- Tokenized text with the mBERT tokenizer and padded to 128 tokens.
- Fined-tuned BERT-based-multilingual-cased, a transformer model trained on 104 languages.[4]
- Leveraged contextual embeddings to capture semantic nuance, syntax, and writing style.

Results

Metrics	Lang	Accuracy	Precision	Recall	F1-score
MNB	*EN*	0.6639	0.7588	0.6746	0.7142
	ZH	0.8033	0.8823	0.6525	0.7502
FastText	*EN*	0.7591	0.7553	0.7591	0.7544
	ZH	0.6129	0.6899	0.6129	0.5389
	Overall	0.6860	0.7842	0.5738	0.6627
mBERT	*EN*	0.8572	0.8379	0.9554	0.8928
	ZH	0.9515	0.9102	0.9906	0.9487
	Overall	0.9044	0.8675	0.9702	0.9160

Figure 3. Evaluation metrics for all 3 models on EN and ZH.

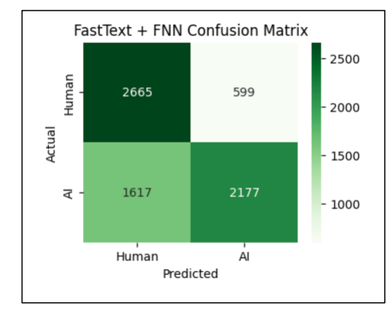


Figure 4. Confusion matrix FastText

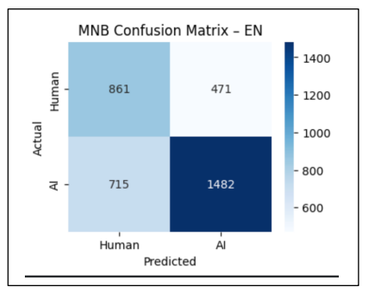


Figure 5. Confusion matrix MNB EN

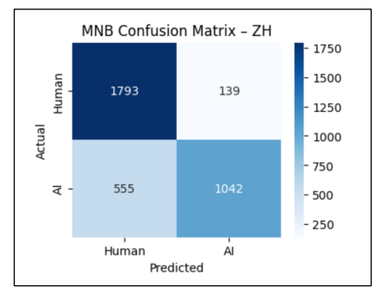


Figure 6. Confusion matrix MNB ZH

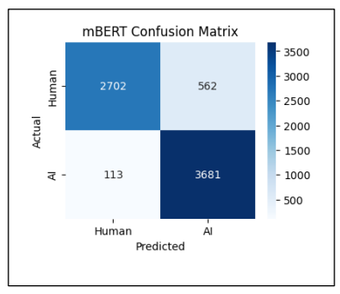


Figure 7. Confusion matrix mBERT

Sample Prediction in EN

```
Text: It was a really cold winter morning and my friends had a stupid idea. They filled water balloons with water and left the...
→ Actual: AI | Pred: AI
--------------------------------------------------------------------------------
Text: This paper reviews the economic and theoretical foundations of insolvency
risk measurement and capital adequacy rules. T...
→ Actual: Human | Pred: AI
--------------------------------------------------------------------------------
Text: What the Rose did to the Cypress is a Persian fairy tale. Andrew Lang included it in The Brown Fairy Book (1904), with t...
→ Actual: Human | Pred: AI
--------------------------------------------------------------------------------
Text: Eh. For someone who was on their first visit to Vegas, I sure wasn't impressed. The outside and inside are completely ta...
→ Actual: Human | Pred: Human
--------------------------------------------------------------------------------
Text: Gitane is a French manufacturer of bicycles based in Machecoul, France; the name "Gitane" means gypsy woman. The brand w...
→ Actual: Human | Pred: Human
--------------------------------------------------------------------------------
```

Sample Prediction in ZH

```
Text: 在生完小孩后，乳房胀痛是常见的现象。这是由于乳汁的分泌引起的，通常会在出生后的第三天到第五天出现。如果乳房胀痛严重，可以尝试以下几种方法来缓解症状：\
→ Actual: AI | Pred: Human
--------------------------------------------------------------------------------
Text: 您好这位朋友，导致手臂粗壮的原因有很多种，无克制的饮食是造成手臂粗的一种原因，甚是是造玉成身肥胖的原因，还有就是手臂不经常运动，所以轻易手臂粗，再者，
→ Actual: Human | Pred: Human
--------------------------------------------------------------------------------
Text: 如何设置锁电脑桌面

    先打开控制面版 再打开 用户帐户 选中你的用户名就可以设置密码了
    到BIOS里设置,找到USER PRASSWORD 点击设置,或者在ADRANCED BIOS FEATURES 里找到S...
→ Actual: Human | Pred: Human
--------------------------------------------------------------------------------
Text: 01 赫尔城 vs 博尔顿
02 桑德兰 vs 朴茨茅
03 西汉姆 vs 埃弗顿
04 利物浦 vs 西布朗 == 31  西布朗发挥出色可能拿到一分
05 布莱克 vs 切尔西 == 30  注意：没有平，切尔西一旦拿不下比赛，...
→ Actual: Human | Pred: Human
--------------------------------------------------------------------------------
Text: 仿瓷又称瓷釉涂料，是一种装饰效果酷似瓷釉饰面的建筑涂料。...
→ Actual: Human | Pred: Human
```

Analysis

- MNB and FastText show similar performance, with mBERT clearly achieving the highest accuracy and F1-score.
- Naive Bayes struggles with English due to limited context modeling.
- FastText helps capture semantics but loses sentence structure when averaging embeddings.
- mBERT achieves the highest scores because it models full context and multilingual patterns.
- Overall, mBERT is the most reliable model for multilingual AI-generated text detection.

Limitations and Future Work

- **Limitations:** Computational constraints limited how much multilingual data we were able to process, which restricted how much we could train and test our models.
- **Future work:** Adding more diverse language datasets to improve multilingual coverage and reduce bias. Including languages with different grammatical structures would help test how well the models generalize beyond English and Chinese.

References

[1] Jinyan1, COLING 2025 Multilingual Machine-Generated Text Dataset. Hugging Face, Accessed: Nov. 20, 2025. [Online]. Available: https://huggingface.co/datasets/Jinyan1/COLING_2025_MGT_multingual
[2] PyPI, "jieba: Chinese text segmentation," Accessed: Nov. 20, 2025. [Online]. Available: https://pypi.org/project/jieba/
[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," arXiv preprint arXiv:1607.04606, 2016. [Online]. Available: https://arxiv.org/abs/1310.4546
[4] T. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019. [Online]. Available: https://arxiv.org/abs/2005.09093