KYIV AVIATION INSTITUTE

Comparing Methods for Multi-Label Classification of Manipulation Techniques in Ukrainian Telegram Content

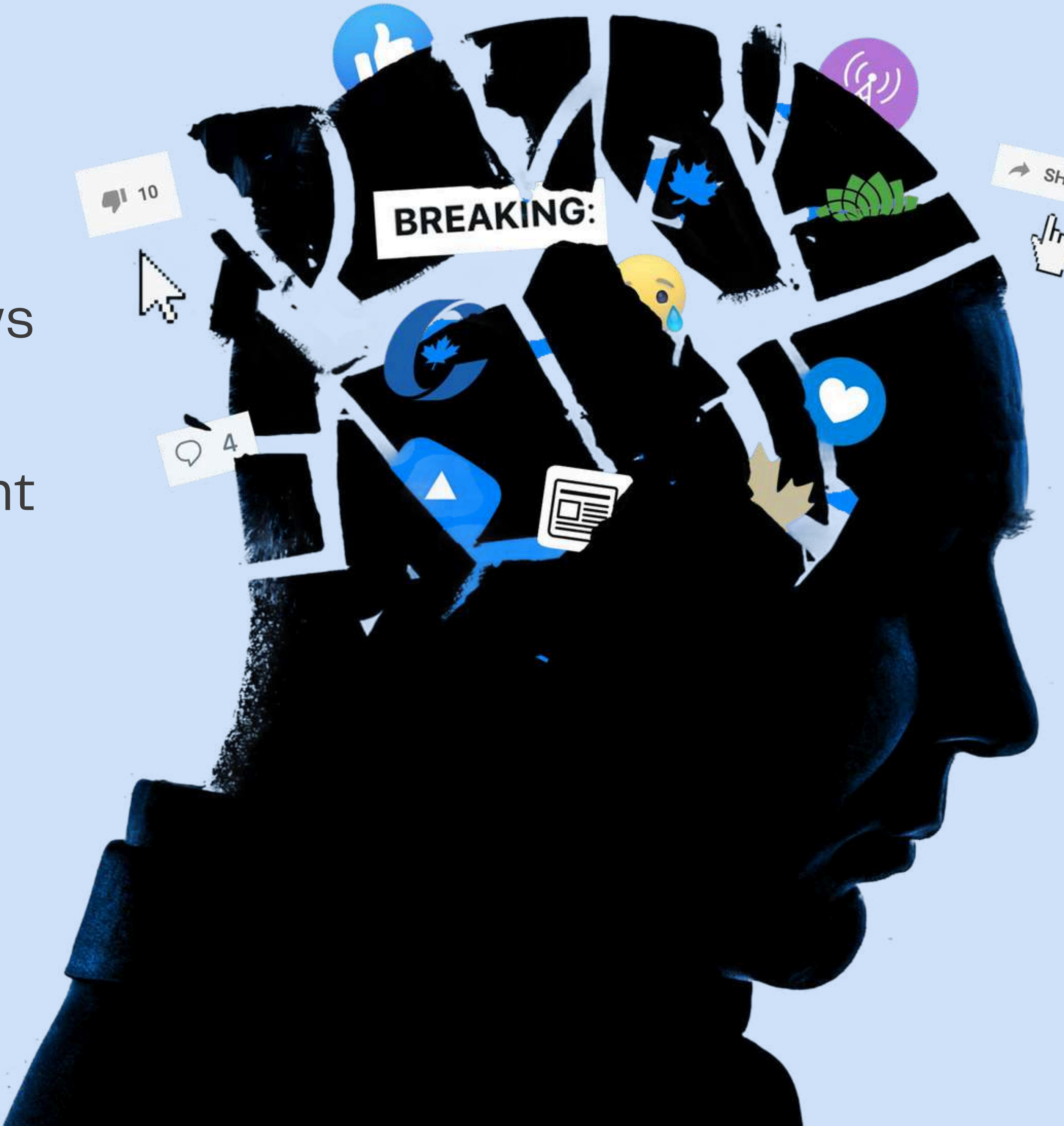
Author: Oleh Melnychuk

Motivation & Context

Motivation:

- 74% of Ukrainians use social media as their primary news source. Telegram is the dominant platform. (USAID-Internews study 2024)
- AI-generated propaganda can receive 37% more engagement than human-written content and is significantly harder to detect.

Main Focus of Research: AI—based methods for detecting mis/disinformation in social media directly on a user's device.

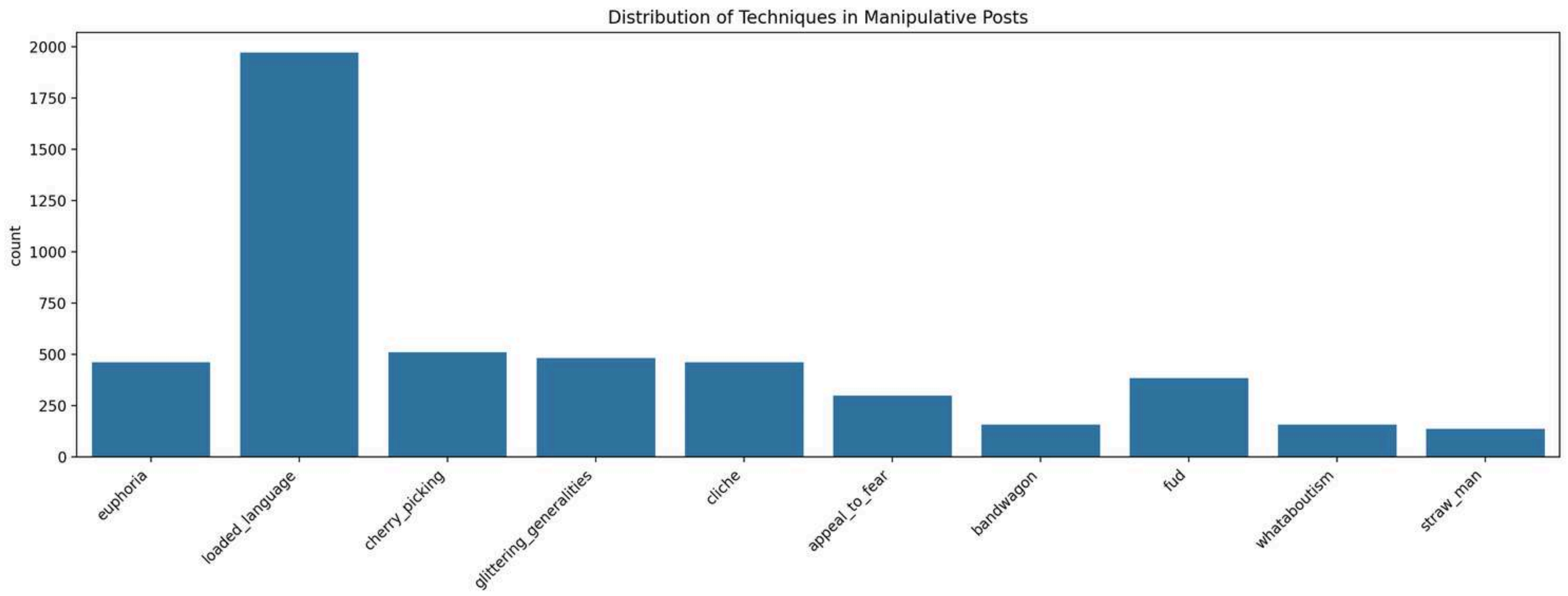


Dataset

Problem Statement

1) Multi-Label Complexity: A single post can contain multiple manipulation techniques, complicating classification.

2) High Class Imbalance across manipulation techniques.



Research Questions

1) How do LLM-based approaches (RAG, fine-tuning) compare to traditional approaches (TF-IDF, fine-tuned Transformers) for this multi-label classification task for running on device?

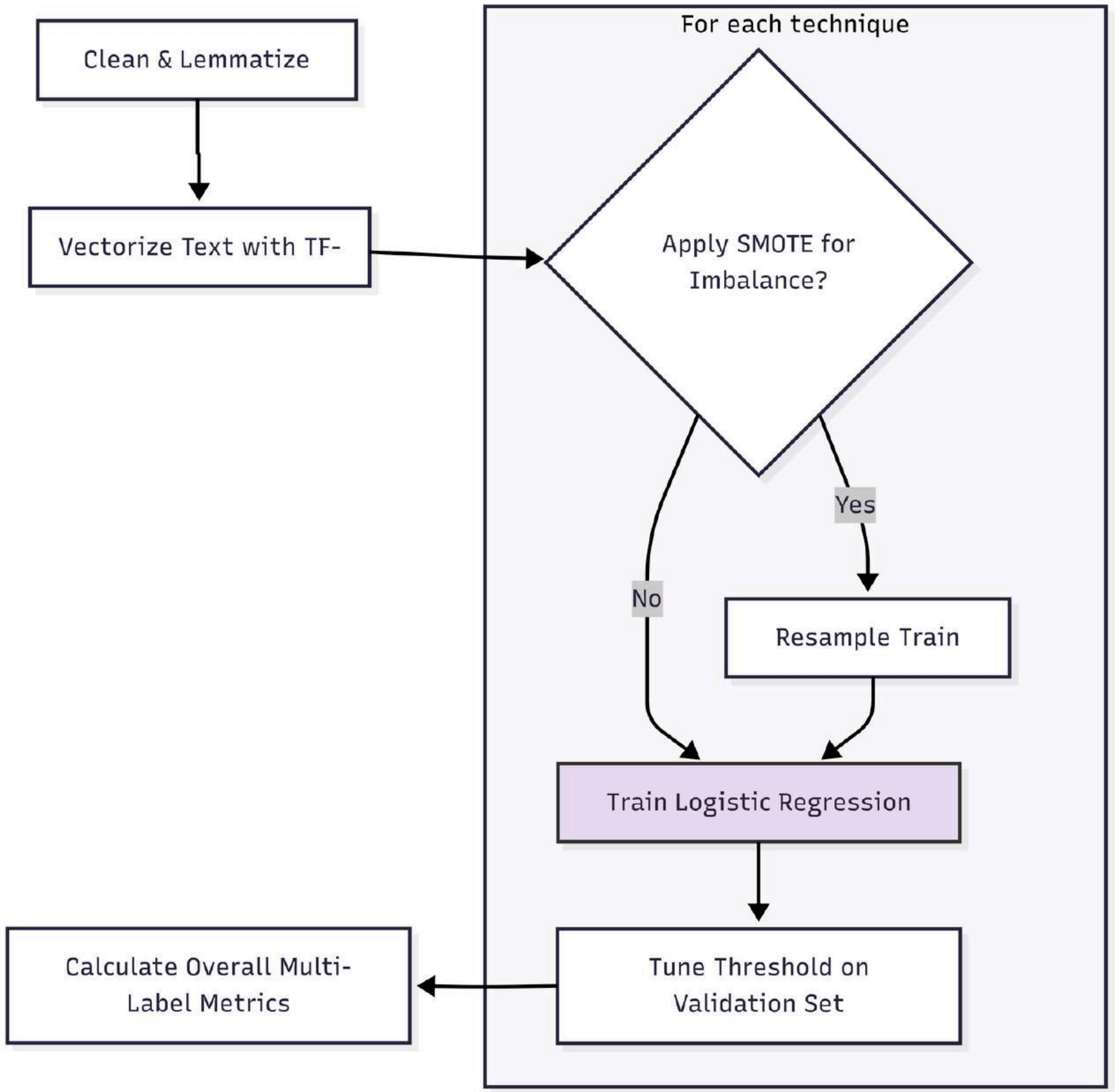
2) What is the impact of using LLM-generated synthetic data to address class imbalance?

Methods Compared

KYIV AVIATION INSTITUTE

TF-IDF (baseline)

Logistic Regression classifiers with SMOTE to handle imbalance.


Clean & Lemmatize

Vectorize Text with TF-

For each technique

Apply SMOTE for Imbalance?

Resample Train

Train Logistic Regression

Tune Threshold on Validation Set

Calculate Overall Multi-Label Metrics

KYIV AVIATION INSTITUTE

XML-RoBERTa-Large

Weights:

straw_man: Positives=128, Negatives=3311, PosWeight=25.87

appeal_to_fear: Positives=270, Negatives=3169, PosWeight=11.74

fud: Positives=348, Negatives=3091, PosWeight=8.88

bandwagon: Positives=138, Negatives=3301, PosWeight=23.92

whataboutism: Positives=146, Negatives=3293, PosWeight=22.55

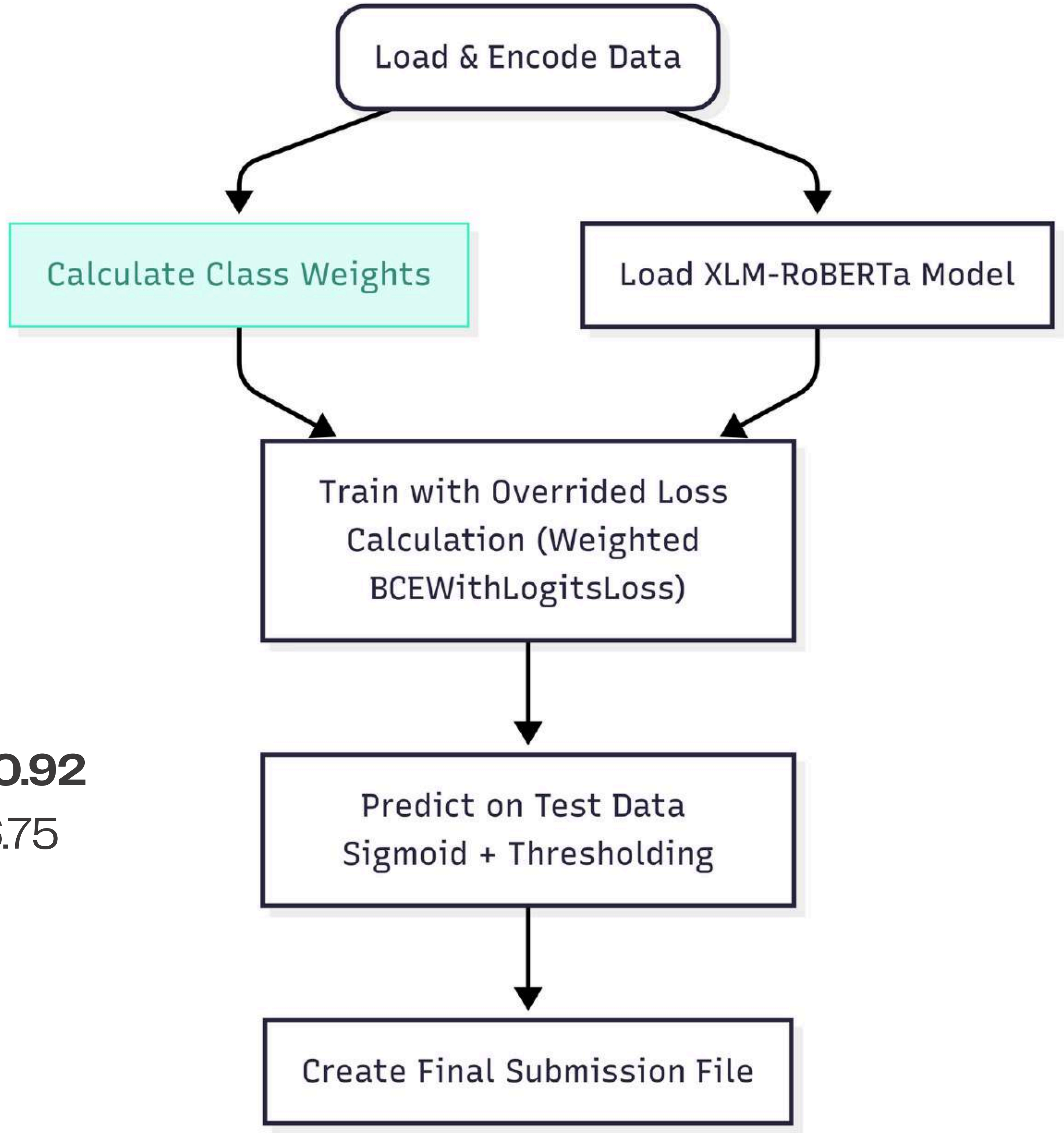
loaded_language: Positives=1788, Negatives=1651, PosWeight=0.92

glittering_generalities: Positives=444, Negatives=2995, PosWeight=6.75

euphoria: Positives=418, Negatives=3021, PosWeight=7.23

cherry_picking: Positives=463, Negatives=2976, PosWeight=6.43

cliche: Positives=418, Negatives=3021, PosWeight=7.23

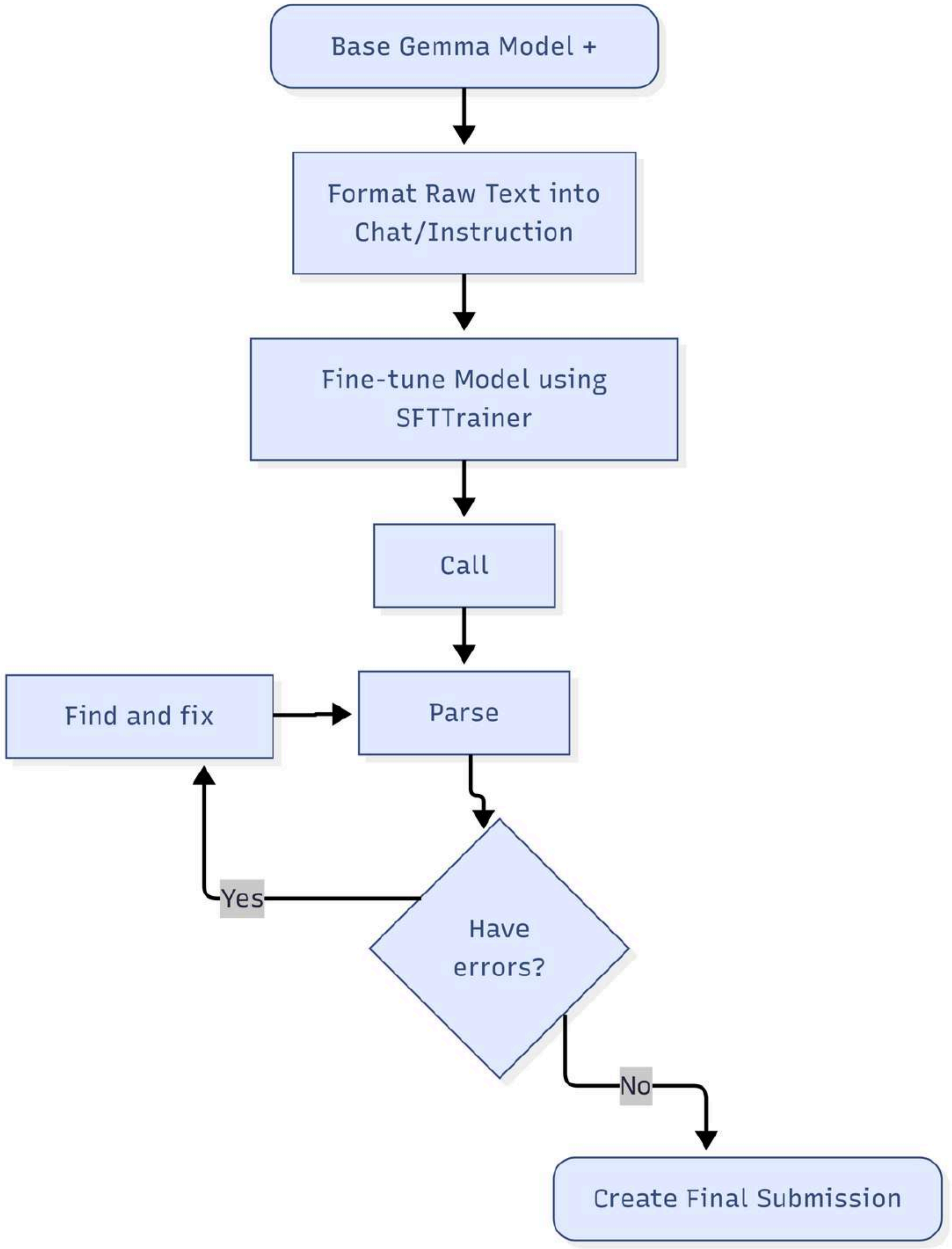

Load & Encode Data
Calculate Class Weights
Load XLM-RoBERTa Model
Train with Overrided Loss Calculation (Weighted BCEWithLogitsLoss)
Predict on Test Data Sigmoid + Thresholding
Create Final Submission File

KYIV AVIATION INSTITUTE

Fine-tuned Gemma 3-1B

Model: unsloth/gemma-3-1b-it-unsloth-bnb-4bit

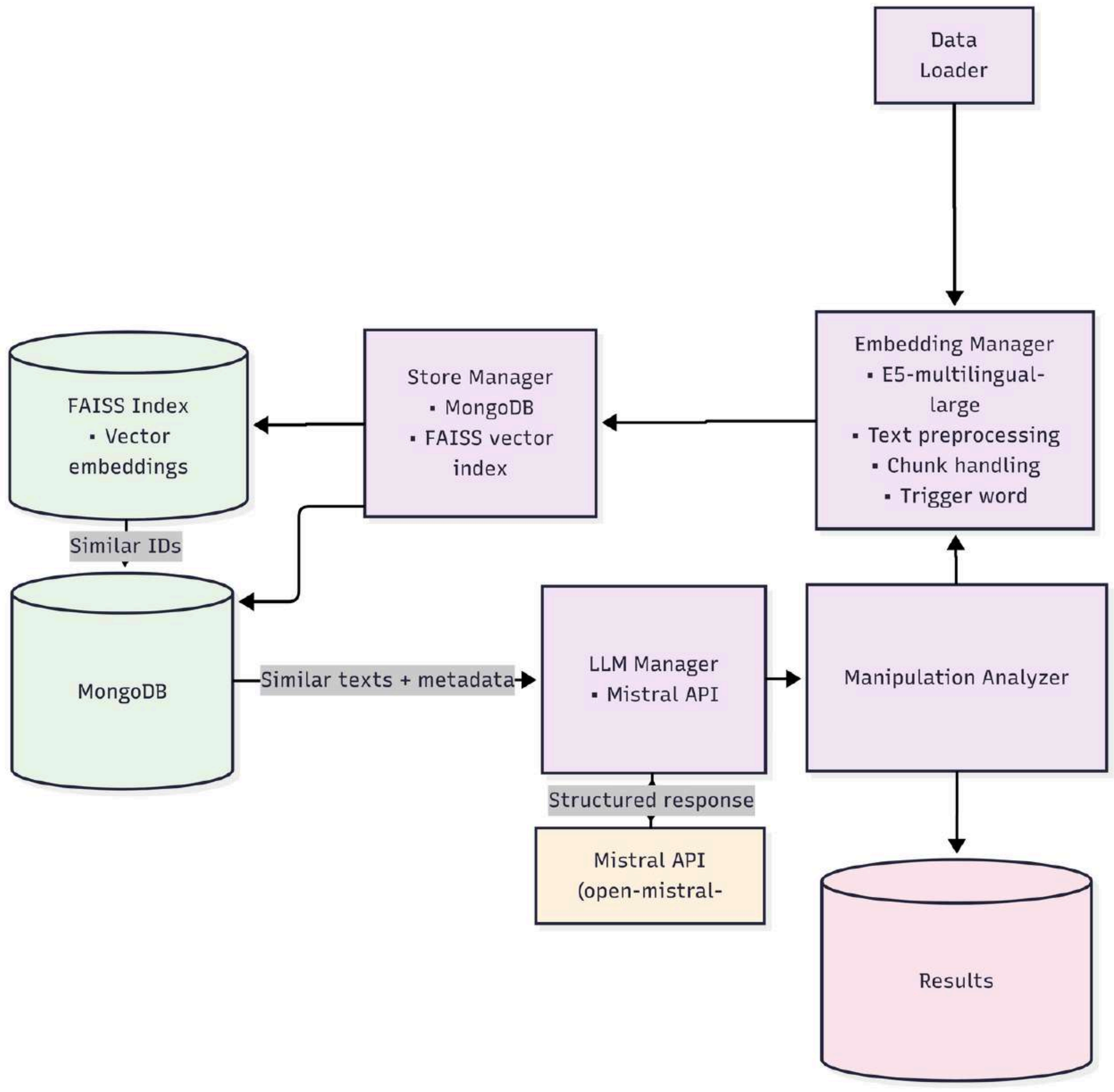
Subset: 100 examples per class



KYIV AVIATION INSTITUTE

RAG approach

RAG system: E5–large embeddings, a FAISS vector index, and a Mistral Nemo generator.



Synthetic Data Generation

```
Your task is to generate 50 plausible examples
that use the [Technique] technique – [brief
description].

The messages should resemble real social media
posts, comments, or tweets. Avoid repetition
or formulaic structure. Keep the tone
realistic and diverse.

Do not add explanations, labels, or numbering
– output just 50 lines, one message per line.

If input is insufficient, refuse and explain
why. Otherwise, proceed with generation.

Target language: Russian

Examples:
 1. …
 2. …
 3. …
```

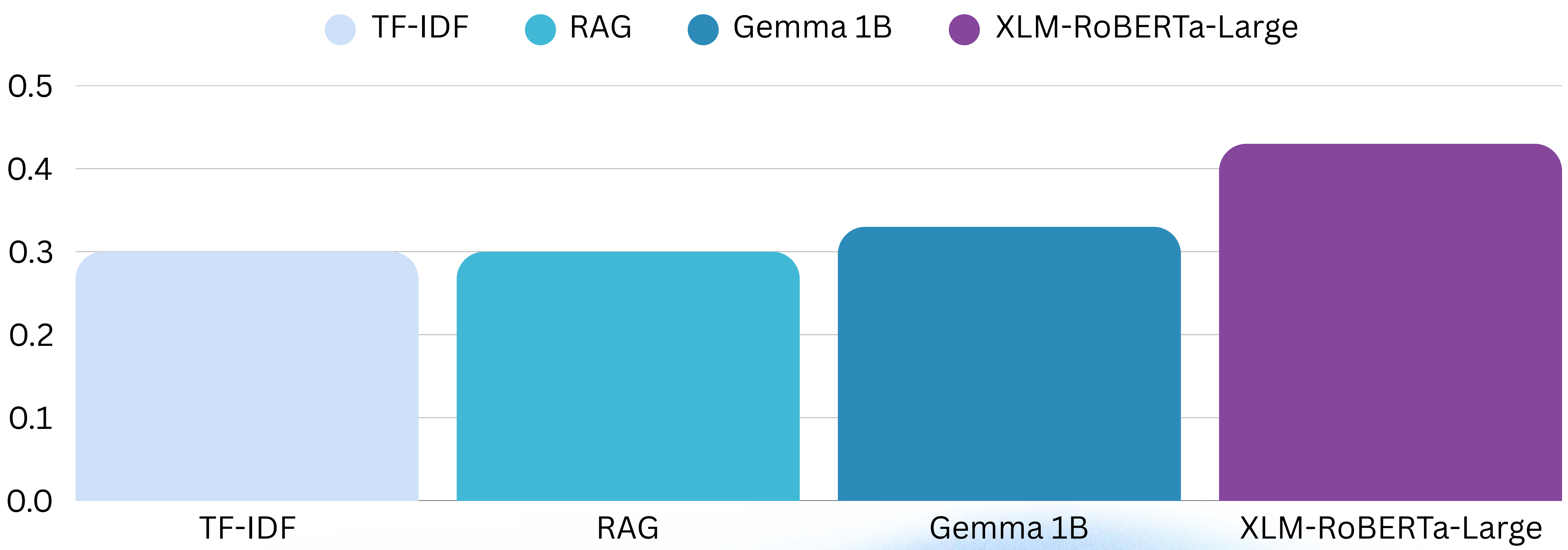
Model: Mistral Large

Strategy: Few-shot prompting for each manipulation class. Separate for UA and RU languages.

Goal: Increase amount of small classes in train dataset

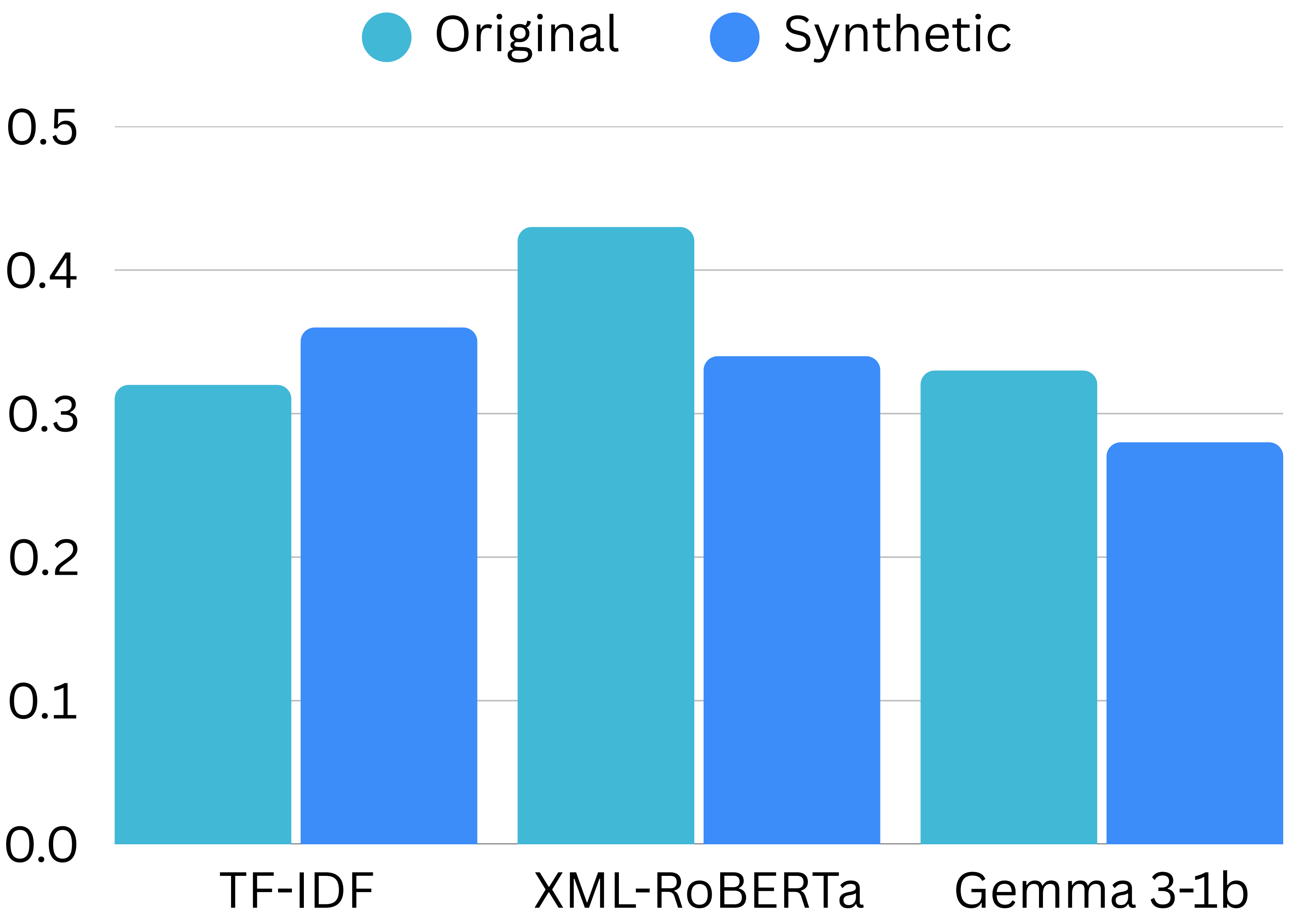
KAI
KYIV AVIATION INSTITUTE

Results Comparison (original dataset)



KYIV AVIATION INSTITUTE

Results Comparison (synthetic)



Positive Impact

Synthetic data improved the TF-IDF + Logistic Regression Macro F1 score (from ~0.30 to 0.36)

Negative Impact

10-20% drop for transformer-based models (XLM-ROBERTa and PEFT-LLMs).

KAI
KYIV AVIATION INSTITUTE

Key findings

XLM-ROBERTa is the Top Performer

Fine-tuned XLM-ROBERTa-Large with a weighted loss function achieved the highest score. This shows that well-configured BERT-like transformers outperform LLMs.

Synthetic Data Quality

Simple, prompt-based data generation helps bag-of-words models, but harms the performance of transformer models. The generated data likely lacked the required semantic richness.

PEFT Small LLMs Underperformed

The fine-tuned LLMs underperformed not just due to the limited data, but also due to hallucinations—Mistral inserted French articles, and Gemma added random symbols, which degraded performance

RAG Approach Complexity

The RAG approach struggled because semantic similarity from vector search did not correlate well with the distinct manipulation technique categories.

KAI
KYIV AVIATION INSTITUTE

Future Work

Enhance RAG

Investigate strategies beyond vector similarity to better align retrieved examples with specific manipulation techniques

New Small LLMs

Conduct fine-tuning of Gemma 3n on the complete dataset to assess their potential to outperform XLM-ROBERTa

Practical usage

Build a PoC browser extension that combines the best-performing classifier with a local RAG system for real-time analysis at the edge.

KYIV AVIATION INSTITUTE

Thank you!

My LinkedIn

Github