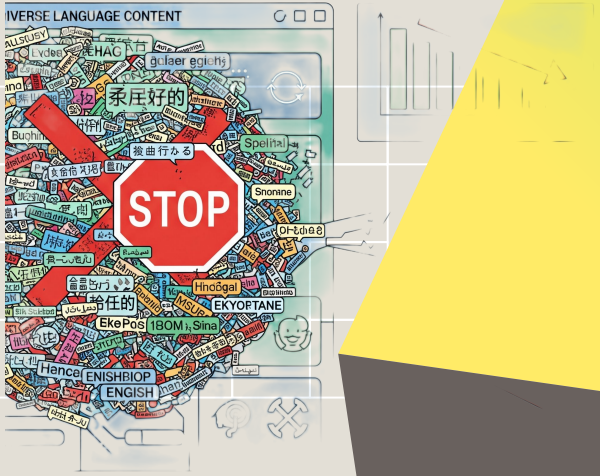
Multilingual Data Quality Assessment

Analyzing Language Diversity Impact on DIFrauD Classification Performance

Team Members: Joseph Mascardo, Niket Gupta

Problem Statement & Hypothesis



Problem

DIFrauD dataset assumed to be English-only, but likely contains multilingual content

Non-English samples can degrade fraud detection performance

No systematic evaluation exists of language diversity impact

Our Hypotheses

H1: DIFrauD contains non-English content unevenly distributed across fraud domains

H2: Models trained on multilingual data show lower F1-scores than English-only models

We expect transformer models (like BERT) to handle this better than old-school ML (Random Forest, SVM)

Project roadmap

Phase 1: Language Detection

Use langdetect and spaCy to identify non-English samples across fraud domains with manual validation

Phase 2: Data Preparation

Create two dataset versions: English-only and complete multilingual with standardized preprocessing

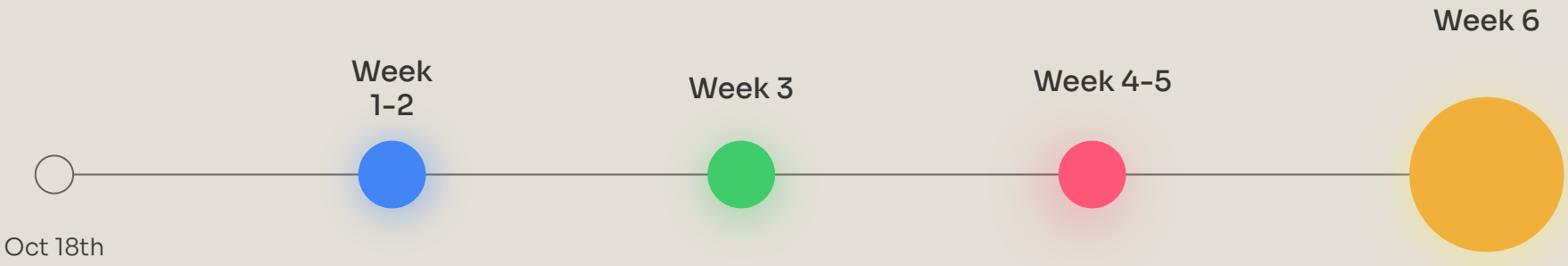
Phase 3: Model Training

Train and evaluate Random Forest, SVM, and DistilBERT on both datasets

Phase 4: Statistical Analysis

Compare performance using paired t-tests and generate comprehensive documentation

Timeline



Oct 18-31

Download dataset, setup language detection, start validation

Nov 1-10

Finish validation and create English/multilingual splits

Nov 11-24

Train all models, record metrics, generate confusion matrices

Nov 25-Dec 1

Statistical analysis, make visualizations and write the report!

Joseph Mascardo

- Build language detection system
- Data preprocessing
- Docker setup
- Make visualizations

Niket Gupta

https://github.com/XYZorro0

- Implement ML models
- Statistical testing
- Cross-validation
- Documentation

Who does what?