# COSC4371 Project Plan

**COSC 4371 Security Analytics**

**Fall 2025**

**Project Title:** Multilingual Data Quality Assessment: Analyzing Language Diversity Impact on DIFrauD Classification Performance

*(Are all the samples in the DIFrauD dataset in English? Use a language detector to determine the number and percentage of samples (classwise) that are from other languages.)*

*(Grammarly AI was used to assist in the creation of this document for grammatical purposes)*

**Team Members:**

**[Joseph Mascardo] - [jamascar@cougarnet.uh.edu]**

**[Niket Gupta] - [ngupta21@cougarnet.uh.edu]**

## 1. INTRODUCTION AND HYPOTHESIS

The DIFrauD dataset represents a comprehensive collection of fraud detection data compiled from multiple public sources, making it valuable for training fraud detection models. However, there's a critical data quality concern that hasn't been explored that we plan to address, which is the multilingual nature of the content. As fraud detection systems increasingly operate in multilingual environments, understanding how language diversity impacts model performance is crucial for developing real-world security solutions.

This project addresses a fundamental gap by providing the first systematic evaluation of how multilingual content affects classification performance in DIFrauD. We hypothesize that non-English samples have a significant impact on classification performance when models are trained on the complete multilingual dataset compared to filtered English-only data.

**H1 (Data Composition Hypothesis):** We expect to find that DIFrauD contains measurable non-English content that is distributed unevenly across domains, with international fraud types, such as Nigerian prince scam and a fake online persona, containing significantly higher concentrations of multilingual content compared to domestic fraud categories. We'll validate this distribution pattern using chi-square tests to determine statistical significance.

**H2 (Performance Impact Hypothesis):** We anticipate that models trained on multilingual data will show lower F1-scores compared to models trained on English-only filtered data. Specifically, we expect transformer-based models like DistilBERT to demonstrate greater robustness to multilingual content compared to traditional classifiers like Random Forest and SVM.

Systematic experiments will be conducted to test these hypotheses. The results will be evaluated using paired t-tests and Cohen's d effect sizes to identify both statistically significant and practically meaningful differences.

## 2. DATA PLAN AND RELATED WORK

We're using the DIFrauD Dataset from Boumber et al.'s COLING 2024 paper, available on HuggingFace with approximately 50,000+ fraud-related text samples across multiple domains, including fake news, SMS, political statements, and phishing. The dataset provides structured text data with class labels and domain annotations.

Our data quality protocol involves cross-validation using both langdetect and spaCy's language identification models. We'll manually review a sample of 1,000 to 10,000 instances to establish ground truth labels. When langdetect and spaCy identify different languages for the same text, both us members will manually review those cases to determine the correct language.

We'll check that all samples contain valid text content and identify any missing or corrupted entries. We'll use balanced sampling to ensure each fraud domain is proportionally included in our analysis. Once we've validated the language labels, we'll create two versions of the dataset: one containing only English samples and another containing all samples regardless of language. Both versions will keep the same fraud domain categories and class distributions as the original dataset.

Conneau et al. (2020) show how language mixing can greatly degrade model performance, providing a theoretical basis for our hypothesis. Devlin et al. (2019) establish transformer architecture capabilities with multilingual data, informing our model selection. Verma et al. (2019) demonstrate the importance of data quality assessment in security datasets, motivating our focus on linguistic data quality.

## 3. METHODOLOGY

Our methodology consists of four integrated phases:

**Phase 1:** We'll set up language detection using both langdetect and spaCy, then test them on sample data to make sure they are performing accurately. We'll calculate what percentage of each fraud type and domain contains non-English content. Then we'll use chi-square statistical tests to determine if certain fraud categories have significantly more multilingual content than others.

**Phase 2:** We'll implement standardized text preprocessing identical for both datasets to ensure fair comparison. Using validated language labels, we'll partition data into two experimental conditions and implement cross-validation, ensuring representative samples from all domains.

**Phase 3:** Our experimental design will compare model performance across traditional machine learning (Random Forest, SVM) and transformer architectures (DistilBERT). Each classifier undergoes training and evaluation on both filtered English-only and complete multilingual datasets. We'll record accuracy, precision, recall, and F1-scores, document training time, and generate confusion matrices for error analysis.

**Phase 4:** We will apply paired t-tests that will compare performance metrics between dataset conditions for each classifier. We'll also perform domain-specific analysis, generate all visualizations, provide an accessible replication method, and complete final documentation.

Our framework will generate statistical breakdowns of language distribution across all DIFrauD domains, and reports comparing multilingual versus English-only conditions, and will document the differences and results. We'll also create code and documentation that other people can use to repeat our experiments, along with guidelines for applying this analysis to other fraud datasets.

## 4. RISKS, MITIGATION, AND TEAM ROLES

One significant challenge we face is ensuring language detection accuracy. We'll test results between multiple language detection tools, manually check cases where they disagree, keep track of confidence scores, and remove unclear samples if necessary. Another concern is computational limitations; to address this, we will process data in efficient batches and optimize memory usage. Finally, class imbalance in the dataset could skew our performance metrics and make models appear more accurate than they actually are. We'll address this by using multiple evaluation metrics beyond just accuracy, ensuring balanced sampling across fraud types, and weighting classes appropriately during evaluation.

**Joseph Mascardo - Technical Implementation:** Build the language detection system, set up data preprocessing, automate experiments, package everything in Docker for reproducibility, and create visualizations for results.

**Niket Gupta - Analysis and Validation:** ML model implementation and tuning, statistical analysis and hypothesis testing, cross-validation framework, and documentation preparation.

**Shared Responsibilities:** Literature review, experiment design, results interpretation, report writing, weekly progress meetings.

## 5. TIMELINE

**Week 1 (Oct 12-14):** Download DIFrauD dataset, set up development environments, implement language detection pipeline using langdetect and spaCy, create and begin manual validation of 1,000-sample subset, calculate preliminary language distribution statistics.

**Week 2 (Oct 15-21):** Complete manual validation process, finalize language distribution analysis, perform chi-square significance testing, implement standardized preprocessing, create multilingual and English-only dataset partitions, set up cross-validation framework.

**Week 3 (Oct 22-28):** Implement and train Random Forest and SVM classifiers on both datasets, implement and fine-tune the DistilBERT model on both datasets, record all performance metrics, generate confusion matrices, and document preliminary results.

**Week 4 (Oct 29-Nov 6):** Conduct paired t-tests and calculate effect sizes, perform domain-specific analysis, generate all visualizations and charts, compile practitioner recommendations, finalize project report and documentation, prepare final presentation, submit all deliverables.

## 6. TOOLS AND DATASETS

**Language Detection:** langdetect (v1.0.9) - https://pypi.org/project/langdetect/ (Accessed: Oct 2025); spaCy (v3.7) - https://spacy.io/ (Accessed: Oct 2025)

**Machine Learning:** scikit-learn (v1.7.2) - https://scikit-learn.org/ (Accessed: Oct 2025); transformers (v4.57.0) - https://huggingface.co/docs/transformers/ (Accessed: Oct 2025)

**Data Processing:** pandas (v2.3.3) - https://pandas.pydata.org/ (Accessed: Oct 2025); numpy (v2.3.0) https://numpy.org/ (Accessed: Oct 2025)

**Visualization:** matplotlib (v3.10) - https://matplotlib.org/ (Accessed: Oct 2025); seaborn (v0.13.2) - https://seaborn.pydata.org/ (Accessed: Oct 2025)

**Environment:** Docker (v2.39.4) - https://www.docker.com/ (Accessed: Oct 2025)

**Dataset:** DIFrauD Dataset - https://huggingface.co/datasets/difraud/difraud (Accessed: Oct 2025)

---

## REFERENCES


Boumber, D., et al. (2024). DIFrauD: A comprehensive fraud detection dataset. *Proceedings of COLING 2024*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

Verma, R. M., Zeng, V., & Faridi, H. (2019). Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2605-2607.

Verma, R. M., & Marchette, D. J. (2019). *Cybersecurity analytics*. CRC Press.