# COSC4371 #1 DIFrauD Language Detection [Mascardo, Gupta] Bid

**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)*

**Project Manager:** Joseph Mascardo, Niket Gupta

**Team Members:**

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

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

**Submission Date:** September 25, 2025

**Ranking Position:** #1 (First Choice)

## 2. PERSONAL INTEREST STATEMENT

This project aligns directly with both of our academic interests in natural language processing within cybersecurity applications. As fraud detection systems increasingly operate in multilingual environments, it's crucial to understand how language diversity affects model performance. This understanding is key to developing robust, real-world security solutions.

Our background in cybersecurity specialization coursework, including Cybersecurity, Software Design, Database Systems, Algorithms, Data Structures, and a little bit of Data Science, has provided us with knowledge in text classification and preprocessing techniques. This project offers an opportunity to apply these skills to a practical cybersecurity problem while gaining experience with multilingual text analysis.

From a career perspective, this work directly relates to our goal of broadening our knowledge in a cybersecurity scope, with the hope of working as Cybersecurity Analysts in the future. Joseph's experience with full-stack development and enterprise IT infrastructure management provides a practical foundation for understanding how data quality issues impact real-world security systems. Niket has worked in IT roles and has developed a full-stack, database-driven website.

**Joseph Mascardo - Individual Motivation**: As a Computer Science student with a cybersecurity specialization and current IT Coordinator experience, I bring a practical enterprise-level perspective to this project. My technical skills in Python, JavaScript, and database systems, combined with my experience in developing multi-tenant applications and data visualization platforms, directly support the technical requirements of this research. This research opportunity allows me to bridge my practical IT experience with academic research in cybersecurity analytics.

**Niket Gupta - Individual Motivation**: This project excites me because it sits at the intersection of natural language processing, fraud detection, and multilingual data analysis—areas that align perfectly with my academic focus in data science and my personal background. As someone who speaks multiple languages (English, Hindi, and Urdu), I'm deeply interested in how language diversity affects machine learning model performance, especially in critical applications like fraud detection. The project's focus on domain-wise analysis also connects to my career goal of working in cybersecurity and fraud prevention. This research will deepen my understanding of both NLP techniques and the practical considerations of deploying ML models in real-world environments.

# 3. RESEARCH PROBLEM DESCRIPTION

### 3.1 Problem Significance

The DIFrauD dataset represents a comprehensive collection of fraud detection data compiled from multiple public sources, making it a valuable resource for training and evaluating fraud detection models. A critical data quality concern is the multilingual nature of the content. Non-English samples can significantly impact classification performance.

Language diversity in cybersecurity datasets poses unique challenges for fraud detection systems. Non-English content can introduce noise, reduce model effectiveness, and lead to unfairness towards certain language groups. Understanding the linguistic composition of training data is crucial for developing fair fraud detection systems.

### 3.2 Current State of Knowledge

Recent research in multilingual text classification has demonstrated that language mixing can substantially degrade model performance (Conneau et al., 2020). Studies have shown that transformer models like BERT, while powerful, can struggle with multilingual inputs when trained primarily on monolingual data (Devlin et al., 2019).

In the cybersecurity domain, most fraud detection research assumes monolingual datasets, with limited systematic investigation of language diversity impacts (Shu et al., 2017). The DIFrauD dataset documentation does not explicitly address language composition, creating uncertainty about potential multilingual effects on model performance.

### 3.3 Key Challenges

Several technical challenges remain unresolved:

- Detecting languages correctly in short, noisy fraud-related messages
- Managing text that mixes multiple languages together
- Measuring how different languages affect classifier performance across various fraud detection tasks
- Maintaining complete datasets while working with models designed for single languages

### 3.4 Proposed Approach

Our methodology consists of three integrated phases executed over the semester timeline. First, we will conduct a comprehensive linguistic analysis by implementing a multi-library language detection pipeline. We will configure both langdetect and spaCy's language identification models, then cross-validate their

outputs on a sample of 1,000 DIFrauD instances to establish detection accuracy baselines. Any disagreements between libraries will be manually reviewed and decided upon to create ground truth labels.

Second, we will perform systematic dataset partitioning and analysis. Using the validated language labels, we will create two experimental conditions: a filtered English-only dataset and the complete multilingual dataset. We will calculate exact class-wise and domain-wise language distribution percentages, then generate statistical significance tests to determine if language distribution varies meaningfully across fraud types.

Third, we will conduct rigorous classification experiments through a controlled evaluation framework. Our experimental design will compare model performance across both traditional machine learning approaches and modern transformer architectures. For traditional methods, we will implement Random Forest and Support Vector Machine classifiers with standard configurations. The transformer component will utilize DistilBERT models, which we will fine-tune using established best practices for text classification tasks.

Each classifier will undergo training and evaluation on both the filtered English-only dataset and the complete multilingual dataset using stratified cross-validation to ensure robust performance estimates. To establish statistical significance of our findings, we will apply paired t-tests comparing performance metrics between the two dataset conditions. Additionally, we will calculate effect sizes to quantify the practical magnitude of performance differences, distinguishing between statistically significant and practically meaningful results

## 4. EXPECTED OUTCOMES AND DELIVERABLES

Our primary deliverable will be a comprehensive multilingual analysis framework consisting of four integrated components. The language detection pipeline will process the entire DIFrauD dataset, generating statistical breakdowns of language distribution across all fraud domains. We anticipate finding measurable non-English content, with varying concentrations across different fraud categories based on their typical communication patterns.

The performance evaluation system will generate detailed classification reports comparing multilingual versus English-only training conditions. We will document accuracy, precision, recall, and F1-score differences for each classifier-dataset combination, with statistical significance testing results. Our analysis will quantify the performance impact caused by multilingual content, providing empirical evidence of how language diversity affects fraud detection effectiveness.

The experimental framework will be fully reproducible through standardized development environments and automated processing pipelines. Our implementation will include streamlined data preprocessing, consistent train-test procedures, and standardized evaluation metrics. The complete pipeline will execute systematically from raw data to final results with minimal manual intervention.

Finally, we will produce actionable recommendations for practitioners, including optimal preprocessing strategies for multilingual fraud datasets, guidance on language filtering approaches, and practical trade-off analyses for different implementation strategies. These guidelines will be grounded in our experimental findings and supported by rigorous statistical validation.

# 5. TEAM QUALIFICATIONS

*Joseph Mascardo*

### 5.1 Technical Skills

**Programming Languages:**

- Python: Experience with data analysis and web development
- JavaScript/TypeScript: Full-stack development experience with React and Node.js
- Database technologies: MySQL, SQLite, Azure database services
- Version control (Git) and collaborative development through GitHub Actions

**Data Analysis and Visualization:**

- Power BI for data visualization and performance metrics analysis
- Database design and management (MySQL, SQLAlchemy)
- Experience in processing and analyzing data for enterprise-level systems
- Statistical analysis tools (R programming language)

**Web Development and System Integration:**

- Full-stack development with React, Node.js, Express, and Flask
- Multi-tenant application architecture and role-based access control
- API integration and workflow automation (Power Automate, Office 365 API)
- Cloud platforms (Azure) and containerization (Docker)

### 5.2 Relevant Experience

**Academic Background:**

- Cybersecurity Specialization - Core understanding of security principles and threat analysis
- Database Systems - Experience with data storage, retrieval, and analysis methodologies
- Algorithms and Data Structures - Foundation for understanding classification algorithms and performance optimization
- Software Design - System architecture and development lifecycle management
- Digital Image Processing - Experience with data processing and analysis techniques

**Project Experience:**

- **GIS Municipal Utility Management Platform** - Developed a multi-tenant platform with geospatial data processing, demonstrating the ability to handle large datasets and performance optimization
- **CoogMusic Platform** - Team-based development project showcasing collaborative software development and data management skills
- **Student Services Management Platform** - Flask-based application with automated workflows and data processing, relevant to systematic data analysis
- **Enterprise IT Coordination** - Experience managing data systems, creating performance metrics, and troubleshooting technical issues across diverse user bases

**5.3 Collaborative Abilities**

Strong communication skills demonstrated through coordinating IT issues with consultants and providing technical support to users of all skill levels at RG Miller. Extensive experience with collaborative software development practices through team-based projects, including the five-person CoogMusic development team and multi-department coordination for enterprise web development projects. Proven ability to work effectively in team environments through both academic group projects and professional experience managing technical solutions for 100+ employees across multiple departments.

*Niket Gupta*

**Programming Languages:**

- **Python**: Extensive experience with pandas for data manipulation, scikit-learn for machine learning implementation, numpy for numerical computing, and matplotlib for data visualization
- **Full-stack development**: Proficient with React, Node.js, Express, Django, and database technologies (MySQL, PostgreSQL), demonstrated through collaborative team projects
- Strong familiarity with version control (Git) and collaborative development workflows

**Machine Learning Expertise:**

- **Text classification experience**: Hands-on work with traditional algorithms (KNN, Linear Regression, SVM)
- **Model evaluation and statistical analysis**: Comprehensive experience with evaluation metrics, including confusion matrices, classification reports, accuracy, precision, recall, F1-scores, and cross-validation techniques
- **Data preprocessing proficiency**: Skilled in feature scaling using StandardScaler, train-test splitting, handling missing values with dropna(), and feature engineering techniques

**Natural Language Processing:**

- **Multilingual processing awareness**: Personal experience with multiple languages (Hindi, Gujarati, Urdu) provides valuable insight into language detection challenges and cross-linguistic model performance

- **Text preprocessing experience**: Familiar with data cleaning, normalization techniques, and the importance of handling linguistic diversity in datasets

**Relevant Experience: Academic Background:**

- **COSC Data Science and Database coursework** - Gained hands-on experience with classification algorithms, statistical analysis, and model evaluation methodologies through practical projects, including Iris dataset classification and KNN optimization
- **Statistical analysis skills** - Understanding of overfitting/underfitting concepts, gradient descent optimization, and appropriate metric selection for imbalanced datasets

**Project Experience:**

- **Large-scale collaborative development**: Led development of Full-Stack Zoo Management System with Team-8-Uma2025, managinga  complex codebase with multiple contributors
- **Agile project management**: Successfully delivered the Academic Approval System using GitHub project management, issue tracking, and maintained high code quality standards
- **Data quality assessment**: Experience identifying optimal model performance through error rate analysis and cross-validation techniques

**Collaborative Abilities:** Proven track record of effective teamwork through multiple large-scale group projects, including the Full-Stack Zoo Management System and Academic Approval System, demonstrating collaborative software development using GitHub project management, issue tracking, and code review processes. Strong communication skills from my teaching experience at ICode and experience coordinating with diverse team members while maintaining code quality standards and meeting project deadlines.

## 6. REFERENCES


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.