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

COSC 4371: Security Analytics — Fall 2025

Joseph Mascardo (jamascar@cougarnet.uh.edu), Niket Gupta (ngupta21@cougarnet.uh.edu)

1. Introduction

Fraud detection systems in the current day and age increasingly operate in multilingual environments, but the language composition of training datasets does not get as much attention as one would expect. The DIFrauD dataset for training fraud detection models, showcases itself as a valuable resource for security analytics research, which will be explained further. However, we were presented with a specific data quality concern, which is the extent to which non-English content affects classification performance.

Our project is the first to look closely at how different languages in the DIFrauD dataset affect how accurately fraud is detected. We first investigated the language distribution across classes and domains in the DIFrauD dataset.

After gaining familiarity with the domain through our review process, we developed two targeted hypotheses that would advance understanding in the space of security analytics. Our first hypothesis is that 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 further hypothesize that modern transformers like DistilBERT will demonstrate greater robustness to multilingual content compared to traditional classifiers like Random Forest and SVM.

We tested our hypotheses through systematic experiments comparing model performance on our original DIFrauD dataset and our synthetic multilingual dataset. We assessed the performance using the chi-squared test, but mainly the F1-score due to class imbalances.

2. Related Work

We found that prior research in fraud detection has mainly been driven on algorithm development without systematically looking at data quality issues related to language diversity. Boumber et al. (2024) introduced the DIFrauD dataset at COLING 2024, compiling together fraudulent text from seven domains, some of which were: fake news, phishing, SMS spam, and job scams. While this dataset provides substantial scale and domain diversity, the original work did not focus on the language composition of the dataset.

Research in cross-lingual natural language processing gave us the background theory for our study. Conneau et al. (2020) demonstrated that mixing together languages can substantially degrade model performance. Their work found that even the best models don't work equally well across all languages

Verma et al. (2019) emphasized the importance of data quality assessment in security datasets, showcasing that unexamined data characteristics can lead to misleading evaluation. Their framework for systematic data quality evaluation motivates our projects focus on linguistic data quality in fraud detection. The work that was done by Devlin et al. (2019) on how transformers use context to understand text led us to our previous hypothesis that BERT-based models would be more robust to multilingual content than traditional classifiers.

3. Methodology

3.1 Dataset and Language Detection

We utilized the DIFrauD dataset from HuggingFace, comprising 95,854 text samples across seven fraud domains: fake news (20,456), job scams (14,295), phishing (15,272), political statements (12,497), product reviews (20,971), SMS (6,574), and Twitter rumours (5,789). Language detection was performed using two complementary tools: the langdetect library and spaCy's language identification pipeline. This dual-method approach enabled cross-validation of detection results.

For each sample, we extracted the detected language label and confidence score. We used a minimum text length threshold of 20 characters to ensure reliable detection. Agreement between the two detection methods reached 99.79%, with only 203 disagreements flagged for potential manual review.

3.2 Synthetic Multilingual Dataset Generation

To systematically evaluate the impact of multilingual content, we created a synthetic multilingual dataset by translating English samples from the original dataset using Meta's NLLB-200 model. The synthetic dataset comprised 20,000 samples with the following language distribution: 40% English (original), 30% Spanish, 20% French, and 10% Arabic. Stratified sampling ensured domain and class balance remained proportional to the original dataset, with a deceptive class ratio of 38.9%.

3.3 Classification Experiments

We implemented three classifier architectures: Random Forest and LinearSVC using TF-IDF vectorization (scikit-learn), and DistilBERT fine-tuned for sequence classification. Each classifier was trained and evaluated on both the original (English-dominated) and synthetic (multilingual) datasets using an 80/20 train-test split with stratified sampling. Evaluation metrics included accuracy, balanced accuracy, weighted F1-score, and macro F1-score to account for class imbalance.

3.4 Statistical Analysis

We used chi-square tests to check if the language differences across classes and domains were statistically significant. We then compared how each classifier performed on both datasets by calculating percentage change.

4. Results

4.1 Language Distribution in DIFrauD

Language detection revealed that the DIFrauD dataset is predominantly English, with 95,090 samples (99.20%) classified as English. The remaining 764 samples (0.80%) were distributed across 28 non-English languages, with the most common being Afrikaans (78, 0.08%), German (54, 0.06%), French (48, 0.05%), and Dutch (44, 0.05%). An additional 208 samples (0.22%) could not be reliably classified due to insufficient text length or mixed content.

Table 1: Language Distribution by Domain

Domain	Total	English	Non-Eng %	Top Non-English
SMS	6,574	5,979	8.05%	unknown, af, cy
Twitter Rumours	5,789	5,733	0.97%	de, af, da
Political Statements	12,497	12,417	0.64%	fr, da, nl
Phishing	15,272	15,241	0.20%	unknown, it, vi
Fake News	20,456	20,456	0.00%	—

The SMS domain exhibited the highest concentration of non-English content at 9.05%, substantially higher than all other domains. This concentration is consistent with the international nature of SMS spam campaigns, which often target multilingual populations. Chi-square analysis revealed a statistically significant association between domain and language ($\chi^2 = 223.52$, $p < 0.001$).

4.2 Language Distribution by Class

Analysis by class (deceptive versus non-deceptive) revealed an asymmetric distribution: 99.74% of deceptive samples were English compared to 98.86% of non-deceptive samples. The chi-square test confirmed this difference as statistically significant ($\chi^2 = 223.52$, $p = 1.55 \times 10^{-50}$). Non-deceptive samples contained 668 non-English instances versus only 96 in the deceptive class, suggesting that legitimate content in the dataset exhibits greater linguistic diversity than fraudulent content.

4.3 Classification Performance Comparison

Table 2 presents classification results comparing model performance on the original English-dominated dataset versus the synthetic multilingual dataset. All models exhibited performance degradation when trained on multilingual data.

Table 2: Classification Performance Comparison

Model	Dataset	Accuracy	Bal. Acc.	F1 (W)	F1 (M)	Change
Random Forest	Original	0.769	0.732	0.760	0.741	—
	Synthetic	0.732	0.686	0.717	0.693	-4.8%
LinearSVC	Original	0.769	0.752	0.768	0.754	—
	Synthetic	0.732	0.711	0.729	0.713	-4.9%
DistilBERT	Original	0.833	0.825	0.833	0.825	—
	Synthetic	0.771	0.755	0.770	0.757	-7.4%

DistilBERT achieved the highest absolute performance on both datasets but exhibited the largest relative degradation (7.4% accuracy drop versus 4.8–4.9% for traditional classifiers). The macro F1-score, which equally weights performance across classes, showed DistilBERT declining from 0.825 to 0.757 (8.2% decrease), compared to Random Forest's 6.5% decrease and LinearSVC's 5.5% decrease.

5. Discussion

Our findings partially support the first hypothesis regarding data composition: non-English content exists in DIFrauD but at lower levels than anticipated (0.80% versus expected higher concentrations). However, the distribution is notably uneven, with the SMS domain containing 9.05% non-English content compared to 0% in fake news. This asymmetric distribution has implications for domain-specific fraud detection applications.

The second hypothesis regarding transformer robustness was not supported. Contrary to expectations, DistilBERT exhibited greater sensitivity to multilingual content than traditional classifiers, with the largest performance degradation across all metrics. This counterintuitive result may stem from the model's reliance on English-specific tokenization patterns and pre-training that, despite theoretical multilingual capabilities, does not transfer effectively to the mixed-language fraud detection context. The monolingual DistilBERT architecture (distilbert-base-uncased) lacks explicit cross-lingual training, limiting its multilingual generalization.

The statistically significant association between language and class ($p < 0.001$) suggests that language may serve as an unintended proxy feature in classification. Non-deceptive content exhibits greater linguistic diversity, potentially because legitimate international communications appear in the dataset while fraudulent content is more consistently English-based. This pattern merits consideration when deploying fraud detection systems in multilingual environments.

Several limitations affect interpretation. The synthetic multilingual dataset was created through machine translation, which may not perfectly replicate naturally occurring multilingual fraud. Additionally, the 60% non-English proportion in the synthetic dataset represents an extreme scenario; real-world applications may encounter smaller multilingual proportions. Future work should incorporate naturally multilingual fraud datasets and evaluate multilingual transformer variants such as mBERT or XLM-RoBERTa.

6. Conclusion

This study provides the first systematic analysis of language diversity in the DIFrauD fraud detection dataset. We found that while the dataset is predominantly English (99.20%), non-English content is concentrated in specific domains, particularly SMS (9.05%), with statistically significant differences in language distribution between deceptive and non-deceptive classes.

Classification experiments demonstrated that multilingual content degrades model performance across all tested architectures. Notably, DistilBERT exhibited the largest performance drop (7.4% accuracy decrease) despite achieving the highest absolute performance, contradicting our hypothesis that transformer architectures would prove more robust to multilingual content. Traditional classifiers (Random Forest and LinearSVC) showed more stable performance degradation in the 4.8–4.9% range.

These findings have practical implications for fraud detection system deployment. Organizations operating in multilingual environments should evaluate dataset composition and consider language-specific preprocessing or multilingual model architectures. The cross-validation framework developed in this study, achieving 99.77% agreement between langdetect and spaCy, provides a reproducible methodology for language quality assessment in security datasets. Future research should explore multilingual transformer variants and naturally occurring multilingual fraud data to develop more robust cross-lingual fraud detection systems.

7. What Did We Learn from the Project?

This project provided valuable hands-on experience in security analytics research methodology. We learned to systematically evaluate data quality issues that are often overlooked in machine learning pipelines, recognizing that dataset characteristics can significantly impact model behavior. The process of implementing cross-validation between multiple language detection tools taught us the importance of verification in automated analysis pipelines.

Working with transformer models alongside traditional classifiers highlighted the trade-offs between model complexity and robustness. Our counterintuitive finding that DistilBERT was less robust to multilingual content than simpler models reinforced the principle that more sophisticated architectures do not guarantee better real-world performance across all conditions. This experience will inform our future approach to model selection, emphasizing empirical evaluation over architectural assumptions.

Finally, synthesizing a controlled experimental dataset through translation enabled rigorous hypothesis testing while teaching us about the challenges of creating representative multilingual data. The project demonstrated how combining NLP techniques (language detection, machine translation) with statistical analysis (chi-square tests) and machine learning evaluation creates a comprehensive framework for security analytics research.

References

Boumber, D., et al. (2024). Domain-agnostic adapter architecture for deception detection. *Proceedings of LREC-COLING 2024*. Retrieved from https://huggingface.co/datasets/difraud/difraud (Accessed: October 2025).

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 NAACL-HLT 2019*, 4171–4186.

langdetect (v1.0.9). (2025). Python language detection library. Retrieved from https://pypi.org/project/langdetect/ (Accessed: October 2025).

Meta AI. (2025). NLLB-200 distilled 600M. Retrieved from https://huggingface.co/facebook/nllb-200-distilled-600M (Accessed: October 2025).

scikit-learn (v1.7.2). (2025). Machine learning in Python. Retrieved from https://scikit-learn.org/ (Accessed: October 2025).

spaCy (v3.7). (2025). Industrial-strength natural language processing. Retrieved from https://spacy.io/ (Accessed: October 2025).

Transformers (v4.57.0). (2025). HuggingFace Transformers library. Retrieved from https://huggingface.co/docs/transformers/ (Accessed: October 2025).

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.