

Behavioral AI in Fraud Detection



How Sentiment, Bias & Ethics Shape the
Fight Against Deception

Unified AI-Powered Fraud Detection

Objective: Set the stage for an interdisciplinary approach to detecting fraud.

Deep Learning (DL), a subset of machine learning powered by neural networks that mimic the human brain, rose to prominence in the 2010s and now drives advanced applications like image recognition, language translation, cancer detection, and autonomous vehicles.

Supervised learning uses labeled data to train models that can predict either discrete categories (classification) or continuous values (regression) for new, unseen inputs.

Natural Language Processing (NLP), a branch of AI focused on understanding human language, powers financial applications like sentiment analysis and SEC filing reviews and serves as the foundation for Large Language Models (LLMs) like GPT and BERT, which use deep learning to perform tasks such as text generation, summarization, and translation at scale.

"We aim to combine behavioral insights with ML/NLP to preempt and detect fraud, not just react to it."



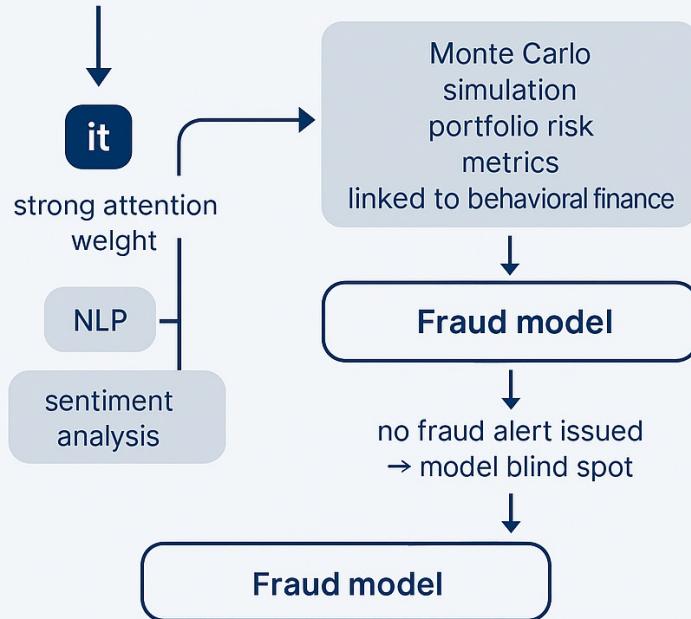
The Role of Self-Attention & Behavioral Finance

GPTs can be fine-tuned on financial texts to detect fraud by identifying suspicious language patterns, anomalies, and deceptive communication, making them powerful tools for early fraud detection and automated risk alerts.

Self-attention allows GPTs to trace contextual dependencies—like what "it" refers to—making them effective in decoding ambiguous or deceptive language often found in fraud-related communications.

Self-Attention

"The wire wasn't flagged because *it* looked routine."



Types Of Fraud

Objective: Categorize major fraud risks.

Types: Identity theft, insurance fraud, cyberfraud, loan fraud, insider fraud

Behavioral Angle: Overconfidence in systems leads individuals to blindly trust automated processes, ignoring subtle signs of fraud simply because "the system didn't flag it."

Authority bias causes employees to accept decisions from senior figures without question, enabling unethical actions to go unchecked.

Example: Text-based red flags in SEC filings—like exaggerated positivity, evasive wording, or sudden shifts in tone—can signal potential fraud and are effectively detected through advanced NLP techniques.



Identity Theft



Insurance Fraud



Cyberfraud



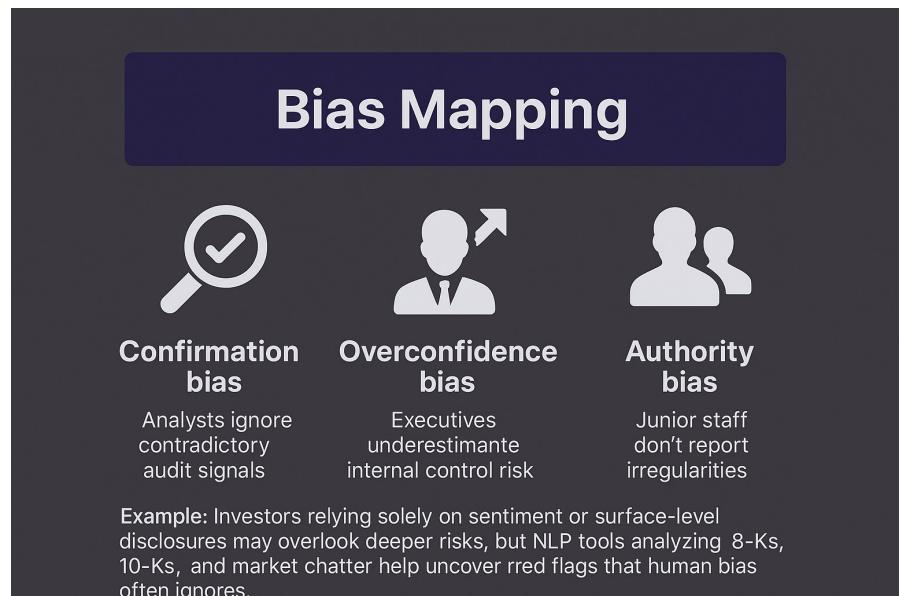
Loan Fraud



Insider Fraud

Behavioral Finance + Fraud

Objective: Show how human biases fuel fraud or miss it



ML Algorithms in Fraud Detection

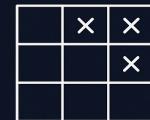
Objective: Link ML with Fraud Flags

| Model | Fraud Use Case | Strength | Limitation |
|---------------------|---|---------------------------------------|---|
| Logistic regression | Predicts probability of fraud in credit card transactions based on structured data | Interpretable and fast | May miss complex or nonlinear fraud relations |
| KNN | Flags geo-temporal anomalies in payment networks by comparing to nearby events | Simple logic, good for local patterns | Slower with large data sets; no model memory |
| SVM | Classifies fraudulent invoices or claims via text and numeric fields pinning behavior | Effective in high-dimensional space | Requires tuning, hard to interpret |

ML Algorithms in Fraud Detection

Fraud Use Case:

- SVM + TF-IDF for textual red flags
- KNN on geo-temporal anomalies



SVM + TF-IDF



KNN

Supervised learning

ML Algorithms in Fraud Detection

Logistic Regression



Baseline fraud classifier with probability output

K-Nearest Neighbors (KNN)



Classifies new transactions based on proximity to known frauds

Support Vector Machine (SVM)



SVM separates complex fraud vs. non-fraud utilizing hyperplanes

K-Means Clustering



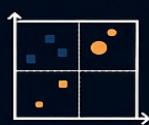
Detects subtle nonlinear behavior over sequences

PayPal uses hybrid ML for fraud prevention

ML Algorithms in Fraud Detection

Fraud Detection Tactics Using ML Models

Use Case Strategy



SVM + TF-IDF
Textural red flags



KNN -GTIDF
Geo-temporal anomalies

Flagging suspicious patterns using proximity, frequency, and contextual indicators

ML Model Mechanics in Action

Real-World Deployment



Logistic Regression



Support Vector Machine



K-Means Clustering



PayPal uses hybrid ML

combining multiple algorithms for real-time fraud prevention

Comparative Overview: ML Models for Fraud Flag

Model Comparison Matrix

| Interpretability | Speed | Fraud complexity |
|------------------|-------------------|--------------------------|
| ✓ | Fast | Basic fraud patterns |
| ✗ | Fast | Local anomalies |
| ✗ | Text 3. numerical | Telecommunications fraud |

Mapping each model to its strengths and application boundaries

NLP Techniques in Fraud Detection

Objective: Show how text reveals deception.

TF-IDF is a word scoring method that highlights informative terms by balancing their frequency within a document against their rarity across documents, and when combined with Naïve Bayes—a simple yet powerful probabilistic classifier that assumes feature independence—it enables efficient and scalable text classification despite the naïve assumption that textual features (tokens) are uncorrelated.

In fraud detection, NLP techniques can uncover behavioral red flags by identifying frequent use of emotionally charged or deceptive language patterns—such as exaggeration or evasion—within communications or disclosures, helping flag potential fraud when combined with traditional data signals.

Evaluation Metrics

- High recall is critical to minimize undetected fraud

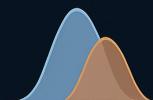
Validating Fraud Models

| | | Confusion Matrix | |
|--------|----------|------------------|----------------|
| | | Positive | Negative |
| Actual | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |
| | | Predicted | |

Accuracy
Accuracy = True Predictions
All Predictions
Measures the proportion of correct predictions out of all predictions.

Precision
$$\text{Precision} = \frac{\text{True Positives}}{\text{Predicted Positives}}$$

Calculates the accuracy of positive predictions.

Recall

Assesses the model's ability to identify actual positives

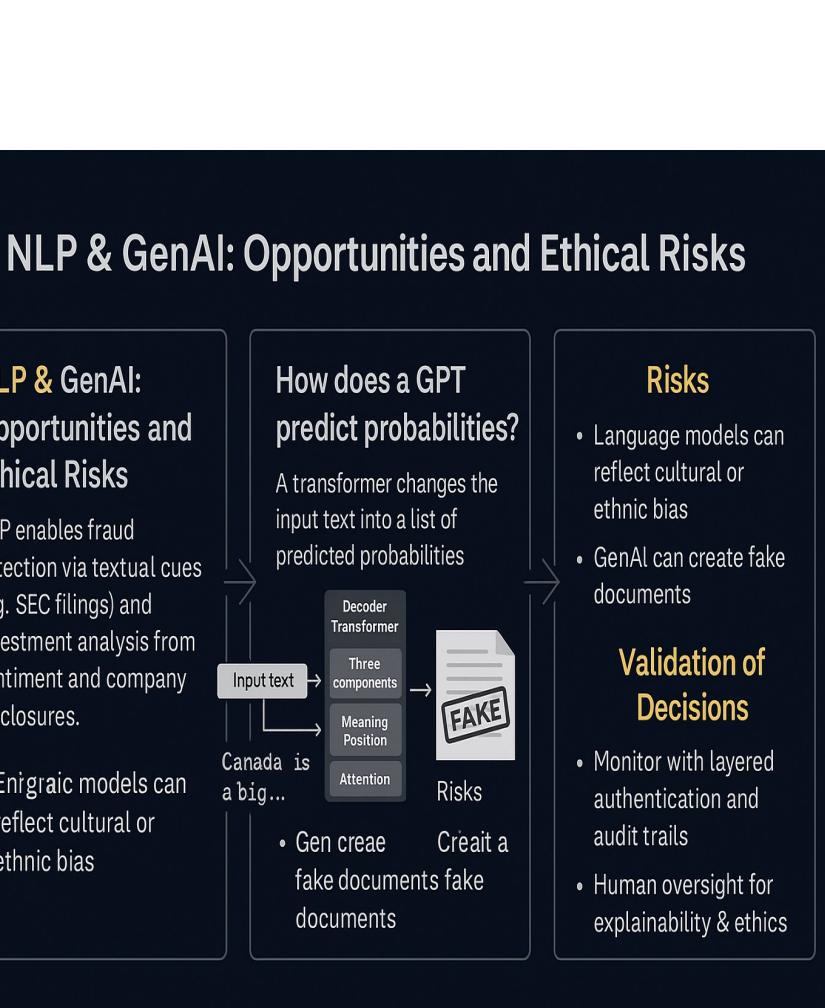
Ethics and Explainability

NLP & GenAI: Opportunities and Ethical Risks

NLP & GenAI: Opportunities and Ethical Risks

NLP enables fraud detection via textual cues (e.g. SEC filings) and investment analysis from sentiment and company disclosures.

- Enigmaic models can reflect cultural or ethnic bias



Confusion Matrix

How the Confusion Matrix Helps Optimize Fraud Models



Model Tuning

- Adjust threshold to balance false alerts and missed fraud
- Evaluate metrics like precision, recall, and F1 score



Behavioral Biases

- Loss aversion favors high recall
- Overconfidence skews judgment
- Misclassifications reveal blind spots

Visual Diagnosis of Fraud Prediction Errors

| | | Predicted | |
|--------|---------------|-------------------------------------|-------------------------------------|
| | | Not Fraud | Fraud (1) |
| Actual | Not Fraud (0) | True Negative (TN) CLEAR | False Positive (FP) False alerts |
| | Fraud (1) | False Negative (FN) Fraud missed | True Positive (TP) Fraud caught |

TF-IDF Heatmap

TF-IDF in Fraud Detection

Spotting Red Flags in Text

TF-IDF in Fraud Detection

- TF (Term Frequency): How often a term appears in a document
- IDF (inverse Document Frequency): Down-scales terms common in many documents; boosts rare, informative terms

$$TF-IDF_{t,d} = TF_{t,d} \times IDF_t$$

$$TF_{t,d} = \frac{f_{t,d}}{\sum_{t' \in e} \sum_{d' \in D} f_{t',d'}}$$

$$IDF_t = \frac{N}{|\{d \in D : t \in d\}|}$$

N = total number of documents
 $|\{d \in D : t \in d\}|$ = number of documents with term

Frequent use of terms like „urgent,” ‘exception’ or manual override

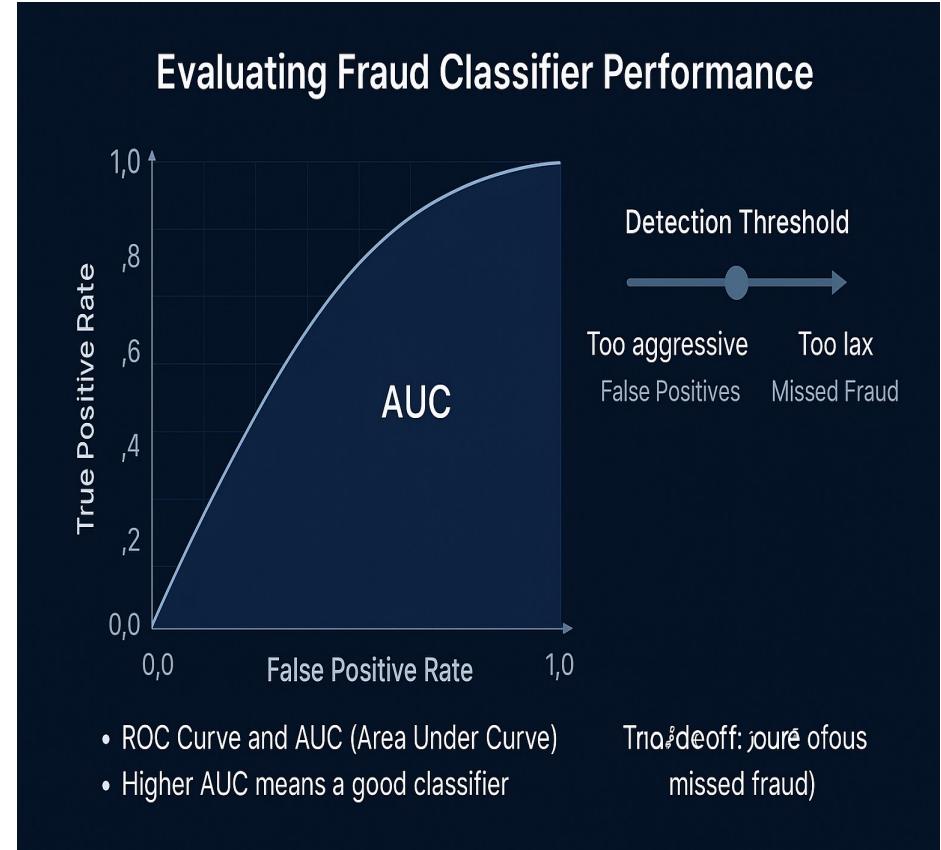
Rare but Informative Words Signal Fraud



- urgent
- exception
- manual override

Frequent use of terms like “urgent,” ‘exception.’ or ‘manual override’ could be red flags in internal reports

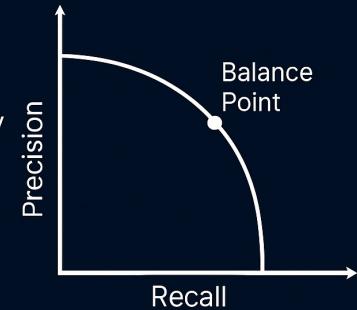
ROC Curve



Precision - Recall Curve

Precision-Recall Analysis

- Use in class-imbalanced fraud settings
- Precision: How trustworthy are the fraud alerts?
- Recall: Are we catching enough actual frauds?



Message: Choose a balance point based on operational resources (investigators, audit bandwidth)

Key Takeaways

ML Fundamentals

Covers the basics of supervised learning in the context of fraud detection

Model Development

Explains metric evaluation, optimization, and fine-tuning of fraud classifiers

NLP Pipeline

Describes text processing and analysis steps to flag potential fraud in free-form documents

AI Risks & Safeguards

Addresses emerging risks of AI-generated fraud, emphasizing monitoring and control

Behavioral Insights

Examines human biases in decision-making and ethics in deploying fraud detection models

