Scam Scan

AI-Powered Fake Website Detection

Capstone Project

Text, Web & Social Media Analytics

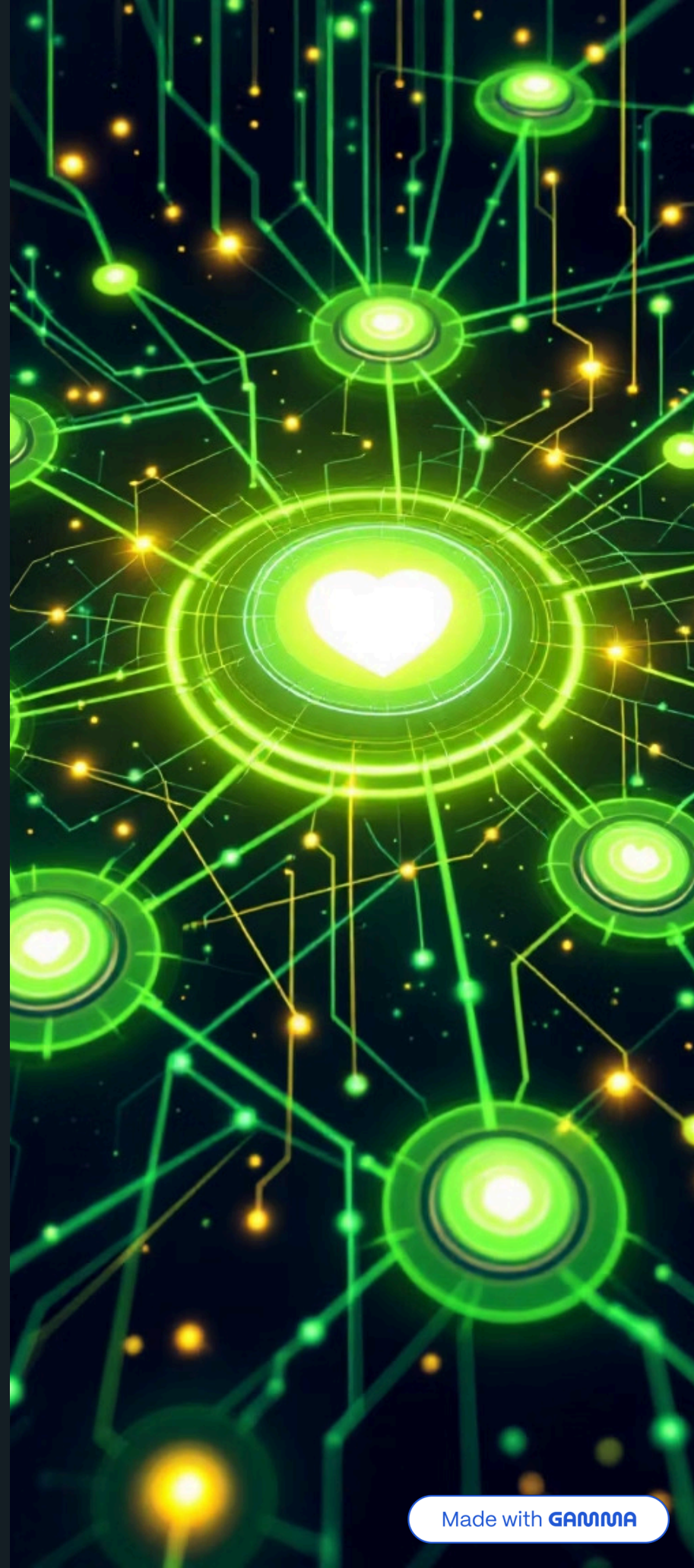
Domain: E-commerce & Delivery

Name: Gumpu UshaSri

Reg. Number: 24MBMB36

Course: MBA - Business Analytics

Year: 2024-2026



Project Overview

Scam Website Detection

AI-driven identification of fraudulent e-commerce platforms.

Multi-faceted Analysis

Combines Web Scraping, NLP, WHOIS lookup, and Machine Learning.

Real-time Integration

Instant detection capabilities delivered via FastAPI.

Made with GAMMA

Problem Statement

Rise in Fake Online Stores

Proliferation of fraudulent e-commerce platforms imitating trusted brands.

Users Fall Prey to Scams

Increasing instances of payment scams and phishing attacks targeting online shoppers.

Manual Verification is Difficult

Human-led processes struggle to keep pace with the scale and sophistication of new threats.

Need for Automated Detection

Critical demand for an accurate, real-time automated scam detection solution.

Made with GAMMA

Motivation

E-commerce fraud increasing rapidly

The digital landscape sees a constant surge in fraudulent online activities.

High financial losses & data theft

Consumers and businesses face significant monetary and data security risks.

Many scam sites look visually real

Sophisticated tactics make it difficult for users to distinguish legitimate from fake sites.

Users need a quick "trust score" tool

A reliable and immediate mechanism is essential for users to assess website legitimacy.

Made with GAMMA

Use Cases

Detecting Fake Online Stores

Identify and flag fraudulent e-commerce sites to protect shoppers.

Preventing Payment Fraud

Safeguard financial transactions by detecting and blocking suspicious activities.

Monitoring Deals & Promotions

Ensure legitimacy of offers and prevent consumers from falling for fake discounts.

Supporting Merchant Verification

Provide tools for users to verify the credibility of online merchants.

Made with GAMMA

Technology Stack

Programming

Python

NLP

spaCy, TextBlob

ML

scikit-learn, RandomForest

Web

BeautifulSoup, Requests

Hosting / API

FastAPI, Uvicorn

Data Tools

pandas, numpy, tldextract, WHOIS

Made with GAMMA

System Architecture

User Input

URL Preprocessing

Web Scraping

NLP & WHOIS Analysis

TF-IDF

ML Classifier

Prediction

FastAPI Output

Made with GAMMA

Project Methodology

Phase 1: Data Creation & Feature Extraction

Phase 2: ML Model Training & Evaluation

Phase 3: API Creation & Deployment

Made with GAMMA

Feature Engineering

URL Features

Domain WHOIS Features

Web Content + NLP Features

TF-IDF Text Features

Made with GAMMA

NLP Techniques Used in ScamScan

Text Extraction & Cleaning

- Extracts website text
- Lowercasing, symbol removal, normalization

TF-IDF Vectorization

- Converts text to feature vectors
- Captures scam-related terms

Keyword-Based Lexical Analysis

- Detects scam-related words
- Computes keyword scores

Named Entity Recognition

- Counts entities (ORG, PRODUCT, PERSON)
- Identifies lack of proper entities

Made with GAMMA

Why Advanced NLP Techniques Were Not Used

Topic Detection (LDA / Topic Modeling)

- Inconsistent topics on fake sites
- Minimal/random text on scam pages
- Yields noisy, meaningless clusters

Sentiment / Emotion Analysis

- Scam sites use neutral/promotional text
- Sentiment scores don't indicate legitimacy
- "Positive" scam wording misleads models

Behavioral NLP Analysis

- Requires user interaction logs
- System analyzes websites, not user behavior
- Not applicable to our approach

Relevance to Scam Detection

- Relies on URL patterns, domain age, keywords
- Offers higher accuracy, faster inference

Made with GAMMA

Machine Learning Model Used

Algorithm: Random Forest Classifier

- Suitable for mixed feature types (URL + text + domain features)
- Handles high-dimensional TF-IDF vectors efficiently
- Resistant to overfitting and noise
- Produces stable and reliable predictions

Input Feature Vector: 47 Dimensions

- URL structural features
- Domain and WHOIS features
- Text/NLP features
- TF-IDF vector

Training Strategy

- Train-test split for evaluation
- Balanced dataset (real + synthetic scam URLs)
- Evaluated using accuracy, precision, recall, and F1 score

Features

Random Forest

Output: Legit / Fake

Made with GAMMA

Model Performance & Evaluation

Accuracy

100%

Precision

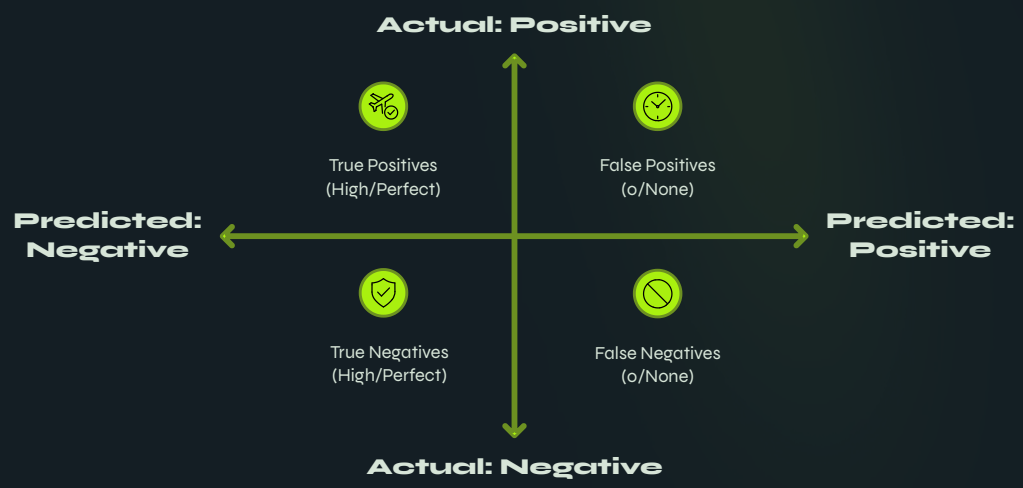
100%

Recall

100%

F1 Score

100%



Made with GAMMA

API Deployment - FastAPI

Backend Technology: FastAPI

- Lightweight and high-performance Python web framework
- Ideal for real-time ML predictions
- Auto-generated Swagger documentation

Key Endpoints Implemented

- /predict → Fake vs. Legit website classification
- /payment-check → Detect risky payment patterns
- /scan-promotions → Identify suspicious discounts or offers
- /merchant-verification → WHOIS-based legitimacy check
- /risk-score → Combined scam score (0–100)
- /full-analysis → Complete scam intelligence report

Deployment Features

- Fast JSON responses
- Cross-platform use (web/app/browser extension)
- Modular feature extraction + ML pipeline

Simple API Flow:

User Input URL → FastAPI → Feature Extraction → ML Model → Output Prediction

Made with GAMMA

Challenges & Future Scope

Challenges Faced

- WHOIS lookup failures or inconsistent responses
- Some websites block scraping or load via JavaScript
- Limited publicly available scam datasets
- Many scam websites have very little text or auto-generated content
- Balancing real vs. synthetic data

Future Scope

- Integrate deep learning models (BERT, RoBERTa) for richer text understanding
- Multi-language website analysis (Hindi, Telugu, Tamil, etc.)
- Browser extension for real-time scam detection
- Social media scam post detection (Instagram, WhatsApp, Telegram)
- Mobile app for URL scanning and fraud prevention
- Deploy API using AWS / GCP

Made with GAMMA

Thank You

Made with GAMMA