**Consumer Complaint Analysis (NLP + ML Project)**

**Overview**

This project performs an end-to-end analysis of **consumer complaints** data using **Python, NLP, and Machine Learning**. The dataset contains thousands of real-world complaint narratives from consumers regarding financial institutions.  
The goal is to clean, analyze, and model this data to uncover hidden insights - such as sentiment, complaint clusters, and potential fraud indicators.

**Key Objectives**

1. **Data Cleaning & Preparation**
   * Removed missing narratives and standardized column names
   * Converted date columns into proper datetime formats
   * Sampled 20,000 rows for efficient processing
2. **Exploratory Data Analysis (EDA)**
   * Identified companies and states with the highest fraud-related complaints
   * Analyzed most disputed products and issues
   * Visualized complaint trends over time
   * Evaluated company response times and timely resolution rates
3. **Sentiment Analysis (VADER)**
   * Applied NLTK’s **VADER sentiment analyzer** on complaint narratives
   * Categorized complaints into *positive*, *negative*, and *neutral* tones
   * Visualized sentiment distribution using Seaborn
4. **Topic Modeling (LDA)**
   * Implemented **Latent Dirichlet Allocation (LDA)** to cluster similar complaints
   * Extracted key complaint topics such as:
     + *Credit Report Disputes*
     + *Debt Collection Practices*
     + *Loan and Mortgage Payment Issues*
     + *Bank Account or Credit Card Problems*
   * Visualized topic distribution using Seaborn
5. **Fraud Detection Model (Machine Learning)**
   * Labeled complaint texts using keyword-based fraud indicators (e.g., “identity theft”, “scam”, “unauthorized charge”)
   * Cleaned and tokenized complaint narratives using **TF-IDF Vectorization**
   * Trained a **Random Forest Classifier** to auto-flag potential fraud cases
   * Achieved **98% accuracy** and **ROC-AUC = 0.995**
   * Validated model predictions and probability distributions

**Tools & Libraries Used**

* **Python** (Pandas, NumPy, Matplotlib, Seaborn)
* **NLTK** (VADER, Lemmatization, Stopwords)
* **Gensim** (LDA Topic Modeling)
* **scikit-learn** (TF-IDF, RandomForest, Metrics)
* **PyLDAvis** (Topic visualization)
* **TQDM** (progress tracking)

**Major Insights**

* **Top Companies with Fraud Complaints:** Experian, Equifax, TransUnion, PayPal, and Chase
* **Most Fraud Cases by State:** California, Florida, Texas, and New York
* **Timely Response Rate:** 98.76% of companies responded within time
* **Sentiment Distribution:** Majority of complaints were **negative**, indicating strong consumer dissatisfaction

**Model Evaluation**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 98% |
| F1-score (Fraud Class) | 0.96 |
| ROC-AUC | 0.996 |

Confusion matrix and fraud probability distribution visualizations were also generated to assess model reliability.