



Adverse News Detection and Classification System for Anti-Money Laundering Investigations

Author: Fabiola Rivera Rodríguez
Advisor: Alfredo Cruz, Ph.D.
Master in Computer Science
Graduate Project EXPO, May 2026

Abstract

Adverse media screening supports Anti-Money Laundering (AML) compliance by identifying entities potentially associated with financial crime indicators. This project presents an automated adverse media detection system that collects news articles from RSS feeds and applies Natural Language Processing techniques for entity extraction and risk analysis. The system uses spaCy models, rule-based filtering, and keyword-based risk scoring to classify article-level risk and aggregate entity-level exposure across multiple articles. A Streamlit-based web application was developed to support exploration of ranked entities and adverse media indicators. The system achieved 90.66% precision in entity extraction, with a 95% confidence interval of 85.55% to 94.09%, demonstrating the feasibility of an interpretable NLP-based approach for adverse media analysis.

Introduction

Adverse media screening is an important component of Anti-Money Laundering (AML) compliance and Enhanced Due Diligence processes. Financial institutions use publicly available information, including news articles, to identify entities potentially associated with financial crime indicators [1], [2]. However, the increasing volume of unstructured news content makes manual review difficult to scale efficiently. Natural Language Processing (NLP) techniques provide opportunities to automate adverse media analysis through entity extraction, risk scoring, and structured risk evaluation [3].

Problem

- Manual adverse media review is time-consuming and difficult to scale efficiently.
- Keyword-only approaches may generate false positives and limited contextual understanding.
- Multilingual news sources introduce additional extraction and interpretation challenges [3].
- AML investigators require interpretable outputs that support entity prioritization and risk-focused review.

Methodology

- Collection of multilingual news articles from RSS feeds
- Data preprocessing and text normalization
- Named Entity Recognition using spaCy English and Spanish models [4]
- Rule-based entity filtering and validation
- Keyword-based article risk scoring
- Entity-level aggregation and ranking
- Interactive visualization through a Streamlit web application [5]

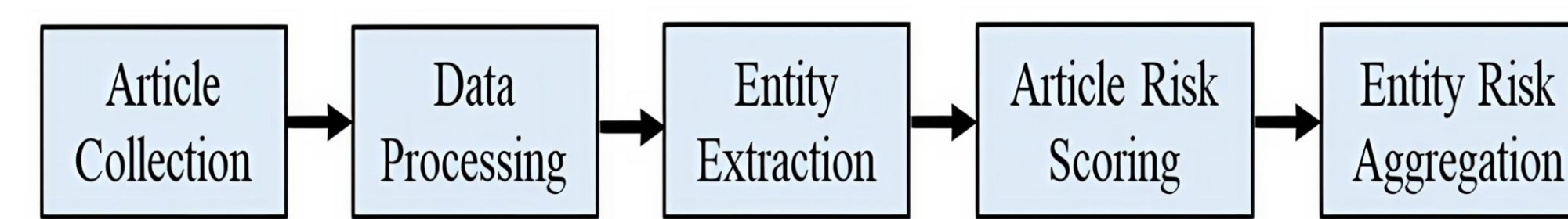


Figure 1
Overall Pipeline of the Proposed Adverse Media Screening System

Web Application

A Streamlit-based web application was developed to support interactive exploration of ranked entities and article-level risk indicators. Users can filter entities, search results, and review associated adverse media articles.

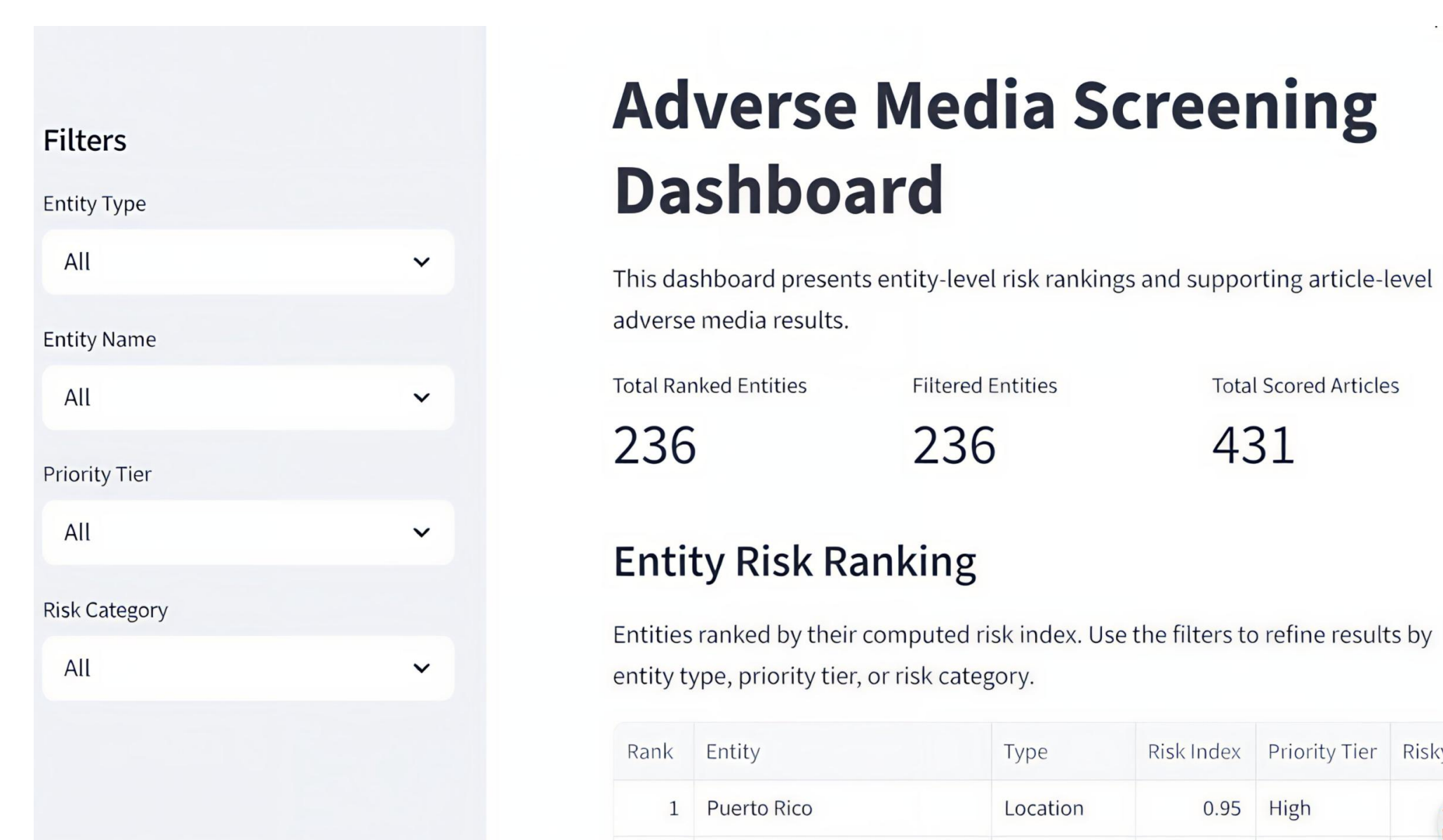


Figure 2
Adverse Media Screening Dashboard with Entity Ranking and Sidebar Filters

Results and Discussion

- The system achieved 90.66% precision in entity extraction.
- Most articles were classified as Low or Medium Risk, while fewer articles were categorized as High Risk.
- Entity-level aggregation enabled prioritization of individuals and organizations associated with repeated adverse media exposure.

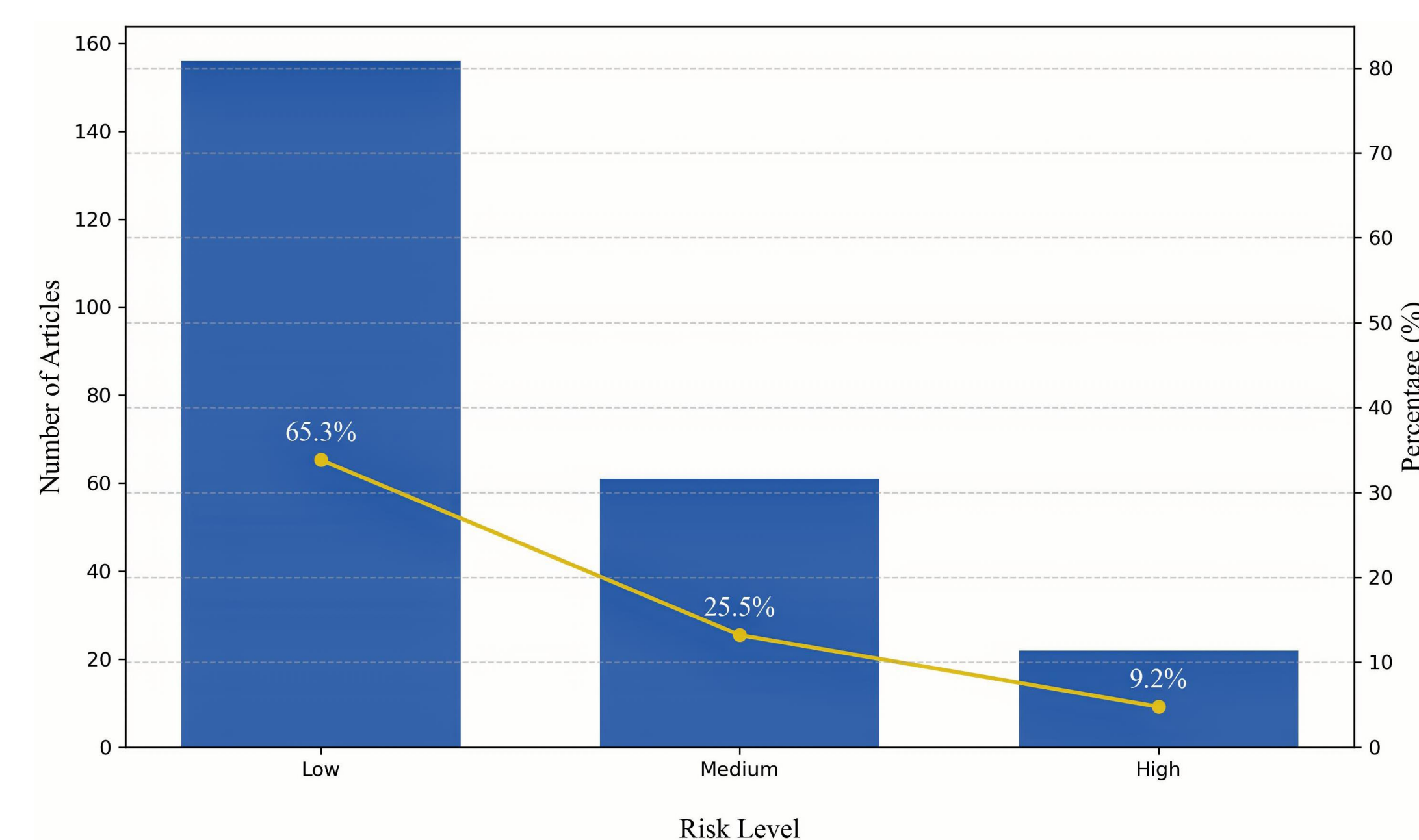


Figure 3
Distribution of Articles by Risk Level

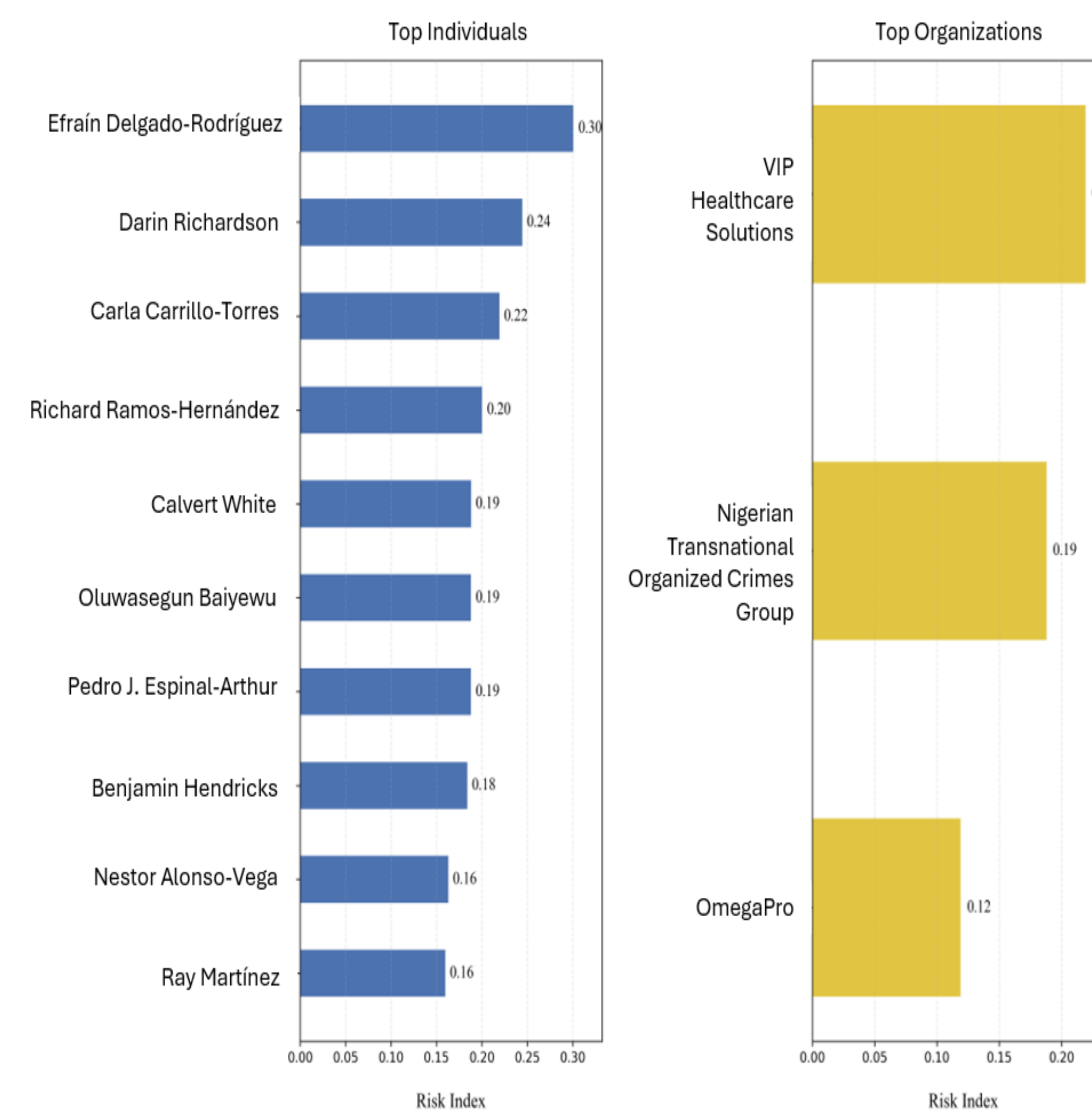


Figure 4
Highest-Ranked Entities by Risk Index

Conclusions

The proposed system demonstrated the feasibility of applying multilingual Natural Language Processing techniques to automate adverse media analysis for AML-focused investigations. By combining multilingual entity extraction, rule-based refinement, risk scoring, and entity-level aggregation, the methodology transformed unstructured news content into interpretable risk indicators. The system achieved 90.66% precision in entity extraction and supported exploration of ranked entities and article-level risk indicators through a Streamlit web application.

Future Work

- Expand multilingual source coverage and entity resolution capabilities.
- Incorporate incremental article processing and automated updates.
- Evaluate additional NLP and machine learning techniques for adverse media classification.
- Compare system performance against baseline keyword-based approaches.

References

- [1] Financial Action Task Force. "FATF Recommendations." FATF-GAFI, 2024.
- [2] Financial Crimes Enforcement Network (FinCEN). "Anti-Money Laundering Act of 2020." U.S. Department of the Treasury, 2020.
- [3] S. Roy. "AI Adoption to Combat Financial Crime: Study on Natural Language Processing in Adverse Media Screening of Financial Services in English and Bangla Multilingual Interpretation." arXiv (Cornell University), 2024.
- [4] spaCy. "Industrial-Strength Natural Language Processing in Python." spaCy, 2015.
- [5] Streamlit, Inc. "Streamlit: A faster way to build and share data apps." Streamlit.io, 2025.