AI-Driven Entity Intelligence Risk Analysis

Team: Game Changers

Abstract

The Entity Risk Assessment and Classification System is designed to assess the risk associated with companies or entities by analyzing multiple data sources and performing complex evaluations. This system combines machine learning models, fuzzy matching algorithms, sentiment analysis, and external data sources to classify entities, assign risk scores, and generate summaries with a confidence level. The process integrates data from multiple sources, including SEC-EDGAR, GLEIF, and OFAC Sanctions Lists, and applies advanced techniques such as Spacy-based Named Entity Recognition (NER) and sentiment analysis to enhance entity classification and risk scoring.

Contents

1 Introduction

The Entity Risk Assessment and Classification System evaluates entities (e.g., companies or individuals) based on various criteria, including financial data, regulatory sanctions, and external reports. The system leverages multiple data sources, fuzzy matching techniques, and sentiment analysis to generate risk scores, confidence levels, and classifications. A final risk report is provided, giving a comprehensive risk assessment of the entity.

2 Data Sources

2.1 SEC-EDGAR Data

The SEC-EDGAR (Securities and Exchange Commission - Electronic Data Gathering, Analysis, and Retrieval) system provides financial and corporate filings. The input company name is matched against company names in SEC filings using fuzzy matching techniques. The retrieved SEC data includes:

- SIC Code (Standard Industrial Classification)
- CIK Code (Central Index Key)
- Company Names
- Financial Reports

These are obtained from two sources on SEC, one from the SEC API's itself and second, is web scraping the search results of the EDGAR website.

A fuzzy matching score is calculated for each company based on the input name's similarity to the company names in the dataset. The fuzzy match uses a combination of metrics:

- 40% Ratio
- 30% Token Sort
- 30% Partial Ratio

This scoring helps rank companies in the SEC-EDGAR database by their similarity to the input entity.

2.2 GLEIF Data

The Global Legal Entity Identifier Foundation (GLEIF) maintains a registry of global legal entity identifiers (LEIs), which are unique identifiers for entities involved in financial transactions. The GLEIF dataset includes:

- Company Information
- Policy Conformity Status
- Corroborating Status
- Ultimate Parent Information

Data from GLEIF is similarly processed using fuzzy matching to score the similarity between the input company and the companies listed in the GLEIF database. In cases where multiple companies share the same parent entity, a weighted average is used to rank them, considering the match score and parent information.

2.3 OFAC and Other Sanctions Lists

The OFAC (Office of Foreign Assets Control) maintains a list of sanctioned entities. Other regulatory bodies, including SDN (Specially Designated Nationals), Debarred Lists, and PEP (Politically Exposed Persons) Lists, are also used for assessing potential risks. Fuzzy matching is applied to these lists to identify entities that may have sanctions or other regulatory restrictions.

3 Model Development

3.1 Fine-tuning of Spacy Model

The system uses a Spacy-based Named Entity Recognition (NER) model that was fine-tuned using data from FiNER-ORD (Fine-grained Named Entity Recognition for Organizations, Locations, and Persons) over 50 epochs. The training was conducted on BIO format data, which is a popular format for NER tasks. The entities targeted for fine-tuning were:

- PER (Person)

- ORG (Organization)
- LOC (Location)

By fine-tuning the model, it achieved improved accuracy for recognizing these entities compared to a generic model.

3.2 Model Metrics Comparison

Before Fine-Tuning:

Entity Type	Precision	Recall	F1-Score
PER	0.1309	0.2536	0.1726
LOC	0.5000	0.0635	0.1128
ORG	0.4461	0.4944	0.4690

After Fine-Tuning:

Entity Type	Precision	Recall	F1-Score
PER	0.9078	0.8982	0.9030
LOC	0.8833	0.9365	0.9091
ORG	0.7661	0.9104	0.8321

The fine-tuning process led to significant improvements in the model's precision, recall, and F1-Score, especially for PER and LOC entities.

4 Data Processing and Merging

4.1 Data Normalization

Before performing any further analysis, company names from all sources (SEC-EDGAR, GLEIF, OFAC, etc.) are normalized to ensure consistent entity name matching. This is done by applying a normalize_name function that processes name variations (e.g., "IBM" vs "International Business Machines").

4.2 Fuzzy Matching and Scoring

Fuzzy matching is employed to compute the similarity score between the input name and the names in the company datasets. A combination of different metrics is used:

- Ratio: A string comparison metric to measure overall similarity.
- Token Sort: Compares words in the names after sorting them.
- Partial Ratio: Measures the similarity of partial tokens within the names.

Each match is assigned a score that reflects how closely the input name matches the dataset entries.

4.3 Weighted Average Calculation

A new column, average, is created based on weighted averages using the following criteria:

- 25% of parent_percentage
- 50% of match_score
- 25% of equality_score

For rows where the match score exceeds 70, a different weighting is applied:

- 70% of match_score
- 30% of parent_percentage

This weighted average helps rank entities based on their relevance.

5 Sentiment Analysis

Sentiment analysis is performed on news headlines scraped from Yahoo Finance using the FinBERT model, which is a sentiment analysis model fine-tuned for financial texts. The sentiment score derived from these analyses adds another layer of assessment to the company's risk evaluation.

6 Gemini AI Integration

The data processed from the various sources is fed into Gemini AI, a large language model, which generates the final output. The AI model:

- Computes a risk score based on all available data.
- Provides a confidence score reflecting the certainty of the risk assessment.
- Categorizes the entity (e.g., Organization, Person, or Location).
- Generates a summary explaining the rationale behind the categorization and risk score.

7 Summarization Integration

The summaries generated are a combination of summary for all 3 entities so there were a lot of overlapping ideas. Using a summarizer transformer model, the summmary was consolidated into a small summary.

8 API Endpoints

The system exposes a FastAPI endpoint, allowing users to submit company names and receive detailed risk reports. The endpoint processes the input data, performs entity risk assessment, and returns the response in the following format:

9 Confidence Scoring

The confidence score is calculated based on the match score and the final average score. The score is assigned as follows:

- 1.0 for a match score above 80
- 0.8 for a match score between 60 and 80
- 0.6 for a match score below 60

This confidence score is crucial for determining how reliable the risk assessment is.

10 Conclusion

This system provides an advanced and multi-layered approach to entity risk assessment by combining machine learning models, fuzzy matching, sentiment analysis, and data from regulatory sources. The integration of a fine-tuned Spacy model enhances the accuracy of the named entity recognition, while the Gemini AI model ensures a reliable and transparent risk evaluation. The system's ability to process multiple data sources and generate actionable insights makes it a valuable tool for financial institutions, regulatory bodies, and businesses seeking to assess the risk associated with various entities.