SEC Filings MD&A Extraction & Multimodal Credit Rating Dataset Creation

1. Project Overview

Objective

The primary goal of this project is to create a multimodal dataset for corporate credit rating prediction by integrating:

- **Tabular financial data:** 25 financial ratios sourced from Kaggle.
- **Textual disclosure data:** Extracted MD&A sections from SEC filings.
- **NLP-derived features:** 11 metrics capturing sentiment and risk.

Final Output

The final dataset, credit_ratings_multimodal.csv, will contain 2029 entries, including:

- Original financial ratios (25 features)
- Company metadata (Ticker, Sector, Industry)
- Credit ratings (multiclass and binary encodings)
- Extracted MD&A text
- 11 NLP-derived scores
- Temporal metadata (rating date, fiscal quarter)

2. Step-by-Step Process Documentation

Step 1: Base Dataset Acquisition & Cleaning

Source & Features

- **Source:** Kaggle Corporate Credit Ratings Dataset
- **Tabular Features:**
  - 25 Financial Ratios: Metrics include liquidity, profitability, debt, operating performance, and cash flow.
  - Categorical Metadata: Company ticker, industry sector, rating agency
  - Target Variable: Credit rating (ranging from AAA to D)

Preprocessing Operations

- **Data Loading:** Imported corporate_rating.csv with 2029 entries and 31 columns.
- **Column Renaming:** Standardized Symbol to Ticker.

- **Rating Class Consolidation:** Consolidated ratings into 8 classes (e.g., AAA → AA+, CC → CCC–).
- **Target Encoding:** Encoded as multiclass and binary (investment grade vs. non-investment grade).
- **Missing Value Strategy:**
  - Created indicators for features with missing values.
  - Used median imputation for low-missing features (<20%).
  - Applied KNN imputation for high-missing features (≥20%).
- **Output:** Saved as credit_ratings_tabular_clean.csv.

Step 2: SEC Filings Retrieval

Mapping Strategy

- **Temporal Mapping:** Converted rating dates to fiscal quarters in year_Q{quarter} format.
- **File Path Structure:** Saved as sec_filings/{ticker}/{year_qtr}.html.

Extraction Pipeline

- **EDGAR API Integration:** Used sec-edgar-downloader with polite crawling, retry logic, and SEC-compliant rate limiting.
- **Filing Selection Logic:** Prioritized 10-Q filings, fallback to most recent if necessary.
- **Document Processing:** Extracted text using HTML parsing and pdfminer for PDFs.
- **File Organization:** Stored in ticker-specific folders.

Step 3: MD&A Section Extraction

Extraction Methodology

- **Table of Contents Navigation:** Identified MD&A links using regex patterns.
- **Section Heading Detection:** Searched for common MD&A headings to determine boundaries.
- **Pattern-Based Extraction:** Used regex to capture text between sections.
- **Anchor-Based Parsing:** For HTML with named anchors, extracted text between #mda anchors.

Quality Control

- **Success Rate:** Approximately 60-70% successful extractions.
- **Failure Handling:** Logged failures for manual review.
- **Validation:** Ensured extracted text length exceeded 500 characters.

Step 4: NLP Feature Computation

Analytical Framework

Implemented a FinancialTextAnalyzer class for 11 metrics, categorized as follows:

- **Sentiment:** Positivity, Negativity, Sentiment (VADER + TextBlob).

- **Risk & Safety:** Risk, Safety, Fraud (custom lexicon frequency analysis).
- **Legal & Certainty:** Litigiousness, Certainty, Uncertainty (domain-specific word lists).
- **Readability:** Flesch Reading Ease.
- **Composite:** Polarity (combined sentiment score).

Lexicon Development

Developed specialized word lists for financial terms, including risk, fraud, safety, certainty, uncertainty, and litigiousness.

Score Calculation

- **Frequency Normalization:** Term counts divided by total word count.
- **Sentiment Aggregation:** Weighted average of VADER (60%) and TextBlob (40%).
- **Readability Scaling:** Normalized Flesch scores to 0-1 range.

Step 5: Multimodal Dataset Assembly

Integration Process

- **Temporal Alignment:** Matched MD&A text to corresponding rating dates.
- **Feature Concatenation:** Combined all features into a single dataset.
- **Data Validation:** Ensured consistent row counts across all features.

Final Dataset Structure

- **Metadata (3 columns):** Ticker, Sector, Industry
- **Temporal (2 columns):** Rating date, year_qtr
- **Target Variables (3 columns):** Rating_Merged, Rating_Encoded_Multiclass, Rating_Encoded_Binary
- **Financial Ratios (25 columns):** Various liquidity, profitability, debt, and cash flow metrics
- **Text Data (1 column):** Extracted MD&A text
- **NLP Features (11 columns):** Various sentiment, risk, and readability scores

3. Technical Implementation Details

Key Libraries Used

- **Data Processing:** pandas, numpy
- **Web Scraping & Text Extraction:** beautifulsoup4, requests, pdfminer.six
- **NLP Processing:** nltk, textblob, vaderSentiment, textstat, spacy
- **Machine Learning:** scikit-learn (LabelEncoder, KNNImputer)

File Structure

```
project/
├── data/
│   ├── raw/
```

```
│   │   └── corporate_rating.csv
│   └── processed/
│       ├── credit_ratings_tabular_clean.csv
│       ├── credit_rating_with_filing_periods.csv
│       └── credit_ratings_multimodal.csv
├── sec_filings/
│   ├── AAPL/
│   │   ├── 2023_Q1.html
│   │   └── 2023_Q2.html
│   └── MSFT/
│       ├── 2023_Q1.html
│       └── 2023_Q2.html
├── src/
│   ├── lexicons/
│   │   ├── risk.txt
│   │   ├── fraud.txt
│   │   └── ...
│   └── extraction_functions.py
├── notebooks/
│   ├── 01_data_cleaning.ipynb
│   ├── 02_sec_filing_extraction.ipynb
│   ├── 03_mda_extraction.ipynb
│   └── 04_nlp_scoring.ipynb
└── documentation.md
```

Performance Metrics

- **Initial Dataset:** 2029 companies, 31 features
- **MD&A Extraction Success Rate:** ~65%
- **Processing Time:** ~3-4 hours for the complete pipeline
- **Final Dataset Size:** 2029 rows × ~45 columns

4. Challenges & Solutions

Challenge 1: Inconsistent SEC Filing Formats

- **Problem:** Varying structures (HTML, XML, PDFs)
- **Solution:** Multi-strategy extraction with fallback mechanisms

Challenge 2: MD&A Section Identification

- **Problem:** Varying section headings
- **Solution:** Regex patterns, heading hierarchy, and table of contents navigation

Challenge 3: Missing Financial Data

- **Problem:** Incomplete financial ratios
- **Solution:** KNN imputation with missing indicators

Challenge 4: Computational Efficiency

- **Problem:** Processing large volumes of text with NLP metrics
- **Solution:** Batch processing and parallelization

5. Potential Applications

Research Applications

- **Credit Rating Prediction:** Multimodal models using combined data
- **Early Warning Systems:** Detecting language changes indicating credit risk
- **Regulatory Analysis:** Studying industry-specific disclosure patterns
- **Natural Language Processing:** Financial NLP research

Model Development

- **Traditional ML:** Logistic regression, Random Forests
- **Deep Learning:** LSTM/Transformer models
- **Multimodal Fusion:** Combining tabular and text data

6. Limitations & Future Improvements

Current Limitations

- **Coverage Gap:** ~35% of companies lack MD&A data
- **Temporal Alignment:** Potential misalignment between ratings and filings
- **Lexicon Completeness:** Limited by custom word lists
- **International Companies:** Different disclosure formats

Recommended Enhancements

- **Enhanced Extraction:** Use transformer-based detection
- **Additional Text Sources:** Include earnings call transcripts
- **Temporal Features:** Add time-series patterns
- **Cross-validation:** Prevent data leakage with proper splits
- **Feature Engineering:** Develop interaction terms

7. Usage Instructions

Quick Start

```
import pandas as pd

# Load the multimodal dataset
df = pd.read_csv('credit_ratings_multimodal.csv')
```

```
# View dataset structure
print(df.info())
print(df[['Ticker', 'Rating_Merged', 'nlp_sentiment', 'nlp_risk']].head())
```

For Model Training

```
from sklearn.model_selection import train_test_split

# Separate features and target
X_tabular = df.drop(columns=['Rating_Merged', 'md&a'])
X_text = df['md&a']
y_multiclass = df['Rating_Encoded_Multiclass']
y_binary = df['Rating_Encoded_Binary']

# Split data
X_train, X_test, y_train, y_test = train_test_split(
    X_tabular, y_binary, test_size=0.2, random_state=42
)
```

8. Conclusion

This pipeline effectively creates a comprehensive multimodal dataset for credit rating analysis by:

- Integrating financial and textual data sources
- Utilizing robust extraction methods for SEC filings
- Developing domain-specific NLP metrics
- Producing a clean dataset ready for machine learning applications

The dataset facilitates the development of more accurate credit rating models by leveraging both quantitative and qualitative data.

Author: Manish Kujur

Date: 15/12/2025

Version: 1.0

License: