Credit Rating Prediction: Multimodal Approach with Structured and Unstructured Data

1. Problem Statement

The objective of this project is to build a credit rating classification model that leverages both structured financial features and unstructured textual data (short financial narratives) to predict a company's rating into 8 distinct credit rating categories.

Detailed approach and EDA and conclusion are in notebooks.

2. Approach

2.1 Data Loading & Preprocessing

The original dataset consisted of 8 distinct credit rating categories (e.g., AAA, A+, BBB, BB, etc.). However, given the limited dataset size and class imbalance, training a robust multi-class model was impractical.

To address this, we grouped the ratings into two broader categories:

- High: Ratings considered stable or strong (e.g., AAA, AA+, AA, A+)
- Low: Ratings that reflect higher risk (e.g., BBB, BBB+, BB)

This binary classification approach improves model stability, simplifies interpretation, and aligns well with real-world credit screening tasks, where analysts often need a quick "safe vs risky" segmentation.

- Structured features: Included financial ratios, index returns, and macroeconomic indicators. These were numerically clean, with minor missing values handled using imputation or column exclusion.
- Unstructured text: Found in the string_values column, consisting of short company-related statements (e.g., "Strategic expansion continues in international markets").

Text Preprocessing included:

- Lowercasing all text
- Removing punctuation and digits
- Stopword removal (NLTK)

- Lemmatization (WordNetLemmatizer)

2.2 NLP Feature Engineering

I engineered several types of features from string_values:

- TF-IDF (Term Frequency-Inverse Document Frequency): Created sparse representations capturing the uniqueness of each word.
- Word2Vec: Used pretrained embeddings to capture semantic similarity between words. Each sentence vector was the average of its word vectors.
- BERT Embeddings: Captured contextual meaning using sentence-transformers (all-MiniLM-L6-v2). These were especially powerful for differentiating subtle sentiment changes.
- Sentiment Analysis: Used TextBlob to assign polarity scores (-1 to 1) reflecting positive/negative tone.
- Topic Modeling (LDA): Extracted latent themes from the text using Gensim's LDA. Each document was converted to a vector of topic probabilities.

2.3 Feature Integration

We constructed three versions of the dataset for modeling:

- Structured-only model: Only used numeric financial variables.
- Text-only model: Used TF-IDF, sentiment, topic, and embedding features.
- Combined model: Fused both structured and unstructured data into a single feature set.

2.4 Modeling & Evaluation

We trained RandomForestClassifier models for each feature variant.

To improve accuracy, we tuned hyperparameters using GridSearchCV on the combined model.

We evaluated models using:

- Accuracy, Precision, Recall, F1-score
- Confusion matrix
- ROC-AUC curve

3. Key Insights

3.1 NLP Findings

3.1.1 Sentiment Score

- Both High and Low rating groups cluster sentiment scores around 0.0 to 0.05, but:
- The Low group shows a heavier density near zero (neutral/negative), indicating more cautious or negative sentiment.
- The High group appears slightly more skewed toward higher sentiment (positive zone around 0.3).

This supports the idea that sentiment scores can act as a predictive signal. Companies with consistently lower sentiment in their disclosures may be more likely to receive a low rating or even face a downgrade.

3.1.2 Topics

- Topic 1 appears much more strongly in Low-rated companies, suggesting this theme may represent financial concerns, market instability, or negative business outlooks.
- Topic 2 dominates more in High-rated firms, potentially aligned with themes of growth, investment, or operational strength.

3.2 Model Performance

• Structured model accuracy: ~50–60%
• Text-only model accuracy: ~70%
• Combined model accuracy: ~75%

The combined model outperformed both individual models, confirming the complementary nature of structured and unstructured data.

4. Business Relevance

Regulatory teams or credit analysts can:

- Use sentiment scores as early warnings of potential downgrades.
- Monitor emerging negative topics (Topic 1) across portfolios to take proactive steps.
- Combine structured and unstructured data for automated credit scoring pipelines.

Investors or lenders may:

- Prioritize review of companies with negative sentiment + Topic 1 dominance
- Gain confidence from companies with positive tone and Topic 2 alignment

5.Conclusions

My multimodal analysis reveals that unstructured text in company summaries (like financial news or commentary) contains valuable predictive signals.

Specifically:

- Negative sentiment (e.g., challenges, risk terms) is correlated with lower credit ratings.
- BERT embeddings and Word2Vec features such as bert_24, bert_141, and w2v_22 consistently appeared among top predictors in the text model.
- Topic modeling shows that specific themes (e.g., liquidity concerns, supply chain issues) are more prevalent in lower-rated companies.

Implication:

Regulatory teams, credit risk analysts, or investment firms can use these signals to:

- Automate credit risk alerts
- Pre-empt rating downgrades
- Monitor portfolios based on NLP-driven sentiment shifts