Final Report: NLP-Driven Financial Sentiment Analysis and Risk Scoring

NDR Team #4: Brian Adams, Hilung Huang, Luis Tupac
NED Davis Research – Practicum, GT - Spring 2025

March 21, 2025

Abstract

This report presents our project which leverages Natural Language Processing (NLP) to analyze sentiment indicators from financial data, headline news, and SEC filings. The primary goal is to develop a Risk Score for the S&P 500 index that identifies market inflection points. The report covers data collection, exploratory analysis, a literature review, model surveys, preliminary results, and outlines next steps.

1 Introduction

1.1 Project Aim

The project aims to use NLP to analyze sentiment indicators from financial data, headline news, and SEC filings. Ultimately, our objective is to develop a Risk Score for the S&P 500 index to identify market inflection points.

1.2 Background and Motivation

Investment sentiment plays a crucial role in market trends. News articles and SEC filings impact market movements by emphasizing risk-laden keywords such as "bankrupt" and "lawsuit." The goal is to quantify the relationship between sentiment trends and financial health, thereby better predicting market volatility.

1.3 Project Objectives

- **Data Collection:** Collect data from the sponsor.
- **High-Risk Dictionary:** Aggregate high-risk words from news headlines and SEC filings (with emphasis on 10-K reports, particularly Items 1 and 1A).
- **Risk Score Development:** Create a Risk Score to help investors anticipate market drops.

2 Project Summary

2.1 Data Collection and Ingestion

All data are provided by our sponsor. We use DuckDB for seamless Python integration and efficient columnar storage.

2.2 Data Cleaning and Transformation

The data is divided into two schemas:

- **Headline Schema:** Contains 16 tables built from financial news articles.
- **SP500 Schema:** Contains 11 tables built from S&P 500 market data.

Insight: Headline data are daily, while S&P 500 data are quarterly; thus, headlines have greater importance due to higher frequency.

3 Exploratory Data Analysis (EDA)

3.1 Headline Schema

- **Content:** Trading calendar, price and volume data, FinBERT sentiment data, and tokenized/lemmatized text data.
- **Key Transformations:**
  - Separation of sentiment from article titles and descriptions to reduce clickbait effects.
  - Application of a 3-day article window to capture extreme price drops.
  - Mapping articles to adjusted trading dates for improved correlation with market movements.

3.2 SP500 Schema

- **Content:** CIK company information, price and volume data, and FinBERT financial report sentiment data.

- **Key Transformations:**
  - Combination of price and weekly data for all S&P 500 stocks.
  - Merging of financial filings into a single entity for unified sentiment processing.

4 Literature Review

4.1 FinBERT: Financial Sentiment Analysis with BERT

FinBERT is a language model tailored for financial text analysis. It leverages transformers to enable bidirectional processing, although it has some limitations in capturing the specialized language of finance ("trader speak").

4.2 Investor Sentiment and Market Trends

Tetlock's research shows that negative sentiment in news articles (e.g., from the Wall Street Journal) can predict lower future stock returns and heightened market volatility. This insight is reflected in our incorporation of sentiment polarity in the Risk Score computation.

4.3 Event-Based Financial Modeling

Research by Hu and colleagues demonstrates that detecting significant financial events through NLP improves stock return predictions. This approach motivates our keyword-based method for tracking price swings by identifying events such as earnings reports, legal issues, and downgrades.

5 Model Survey and Methodology

5.1 Baseline Models

- **Logistic Regression:** Used for binary prediction of price movement.
- **Support Vector Machines (SVM):** Effective with high-dimensional text features.
- **Random Forest:** Captures nonlinear relationships in the data.

5.2 Sentiment-Enhanced and Article-Level Models

- **FinBERT Integration:** Incorporating finance-specific sentiment to improve upon models like VADER.
- **Article-Level Regression Models:** Treating each article as an observation, using features such as risk terms, sentiment scores, and metadata to predict price changes.

5.3 Hybrid Modeling Approach

Our approach combines daily aggregated signals (e.g., article volume and risk score) with granular article-level features (e.g., minimum sentiment, high-risk flags) to capture both overarching trends and detailed risk indicators.

6 Interesting Observations and Preliminary Results

6.1 Observations

- A surge in articles often precedes market volatility.
- Keywords such as "downgrade," "earnings," and "acquisition" frequently coincide with significant price swings.
- News clustering across multiple sources can amplify market reactions.
- Evening articles sometimes lead to notable price changes the following day.

6.2 Preliminary Modeling Results

- **Initial Modeling:** An XGBoost model using only sentiment scores on selected tech stocks resulted in a negative R-squared, possibly due to the exaggeration in headline news.
- **Improved Modeling:** By incorporating Risk Score, sentiment features (VADER and FinBERT), and article volume, models such as Random Forest and XGBoost performed notably better ($R^2 \sim 0.67$). Linear models (e.g., Ridge) underperformed, indicating the need for capturing complex interactions.

7 Next Steps

7.1 Integration and Enhancement

- **FinBERT Sentiment Integration:** Replace or combine VADER with FinBERT to refine the Adjusted Risk Score, comparing model performance before and after integration.
- **Article-Level Modeling:** Treat each article as an independent observation, extracting features such as high-risk terms, sentiment scores, and earnings mentions.

7.2 Hybrid and Time Series Exploration

- Develop a hybrid model that merges daily aggregation with article-level features (e.g., high-risk article flag, minimum sentiment, count of highly negative articles).
- Incorporate macroeconomic indicators such as the VIX index to assess the impact of negative news on market volatility.
- Extend the analysis to evaluate risk across different market sectors.

8 Project Status and Progress

The project is currently advancing through several key phases:

- **Data Exploration & Pre-processing:** Completed.
- **Identification of High-Risk Words and Risk Score Experimentation:** In progress.
- **Finalization of Risk Score Methodology:** Under development.
- **Modeling and Fine Tuning:** Actively testing various models.
- **Application and Project Wrap-Up:** Upcoming.

Target milestones include data processing completion (3/21/2025) and final model deployment (4/1/2025).

9 Workload Distribution

- **Luis Tupac:** Database creation, data pipelines, and data transformation; FinBERT sentiment classification on financial reports and headline news.
- **Brian Adams:** Exploratory Data Analysis, Risk Score computation, and initial model experimentation.
- **Hilung Huang:** Overall project approach, initial modeling, and coordination of tasks.

10 Conclusion

This report has outlined our approach to using NLP for financial sentiment analysis and risk scoring. With promising preliminary results, the next steps involve refining our models, integrating advanced sentiment analysis, and exploring hybrid approaches to enhance prediction accuracy. We look forward to further validating our methodology and deploying our final risk scoring system.