© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716388

FINENZO: Financial Sentiment Based Analysis System

YOGITA CHAVAN (PROJECT GUIDE)[1], JEET GOR[2], RAHUL JALORA[3], OMIK VICHARE[4]
[1, 2, 3, 4]*Department of Computer Engineering, University of Mumbai, NHITM, Thane (W)*

Abstract- ***Financial reports such as 10-K reports contain extensive information about a company's financial performance, risks, and operations. However, these documents are often lengthy and complex, making manual analysis difficult and time-consuming for investors and analysts. Traditional methods struggle to efficiently extract meaningful insights from unstructured financial text, leading to challenges in effective decision-making. To address these limitations, Finenzo is proposed as a system that utilizes Natural Language Processing (NLP) techniques to analyze financial statements and extract useful insights. The system retrieves 10-K reports from the SEC EDGAR database, performs text preprocessing, and applies TF-IDF for feature extraction. A financial sentiment dictionary is used to identify sentiment patterns and generate quantitative scores from textual data. These insights are combined with financial indicators to support better evaluation of company performance and assist in data-driven decision-making. By converting unstructured financial data into structured information, the system improves analysis efficiency and provides a scalable approach for financial analysis.***

Keywords: Natural Language Processing, Financial Text Analysis, 10-K reports Sentiment Analysis, TF-IDF, Data-Driven Portfolio Management.

I. INTRODUCTION

In the modern financial landscape, the rapid growth of digital data and the increasing complexity of financial information have made efficient data analysis essential for investors and analysts. Among the most valuable sources of financial information are 10-K reports, which provide comprehensive insights into a company's financial performance, risks, and operational activities. However, these reports are often lengthy, unstructured, and difficult to interpret manually, leading to challenges in extracting meaningful insights within a limited time.

As the volume of financial data continues to increase, there is a growing need for automated systems that can process and analyze large amounts of textual information effectively. Traditional methods of financial analysis rely heavily on manual interpretation or basic quantitative metrics, which may fail to capture hidden patterns and sentiment present in financial documents. This limitation often results in incomplete analysis and reduced efficiency in decision-making.

The proposed Finenzo system addresses these challenges by integrating Natural Language Processing (NLP) techniques with financial data analysis. The system processes 10-K reports obtained from the SEC EDGAR database and applies text preprocessing methods such as tokenization and stop-word removal. Feature extraction techniques like TF-IDF are used to identify significant textual patterns, while a domain-specific financial sentiment dictionary is utilized to extract sentiment-based insights from the reports.

By transforming unstructured financial text into structured information and quantitative scores, the system enables more efficient evaluation of company performance. Furthermore, the integration of financial indicators enhances the overall reliability of the analysis. The proposed approach supports data-driven decision-making and provides a scalable solution for automated financial analysis.

1.1.1 Limitation / Existing System / Research Gap

Traditional financial analysis methods primarily rely on numerical data and manual interpretation of financial reports, which can be time-consuming and prone to human error. Investors and analysts often need to review extensive financial documents, making it difficult to identify key insights efficiently. Additionally, existing systems that perform financial analysis may not effectively utilize textual information from financial reports, leading to incomplete understanding of company performance.

Although some approaches use basic sentiment analysis or machine learning techniques, they often

lack domain-specific accuracy or fail to integrate textual insights with financial data in a meaningful way. Many systems do not provide a unified framework that combines financial text analysis, sentiment extraction, and data-driven evaluation for portfolio-related decision-making. This creates a research gap in developing a comprehensive system that can effectively bridge the gap between unstructured financial text and actionable insights.

The Finenzo system overcomes these limitations by combining NLP-based text analysis with financial data processing in a unified framework. By extracting sentiment patterns from 10-K reports and converting them into structured financial scores, the system enhances analysis accuracy and efficiency. This integrated approach improves decision-making and provides a more reliable and scalable solution for financial analysis.

1.1.2 Objectives

The main goal of this project is to develop a data-driven system that utilizes Natural Language Processing (NLP) techniques to analyze financial reports and support effective decision-making.

Specific objectives include:

1. To develop an automated system that leverages Natural Language Processing (NLP) techniques to analyze complex financial reports and extract meaningful insights from unstructured textual data.
2. To design an efficient data acquisition mechanism for collecting 10-K financial reports from the SEC EDGAR database and preparing them for further analysis.
3. To implement robust text preprocessing methods, including tokenization, stop-word removal, and data cleaning, to ensure accurate and reliable processing of financial text.
4. To apply feature extraction techniques such as TF-IDF to convert textual financial data into numerical representations suitable for analysis and interpretation.
5. To utilize a domain-specific financial sentiment dictionary to identify and classify sentiment patterns within financial documents and generate sentiment-based insights.
6. To transform unstructured financial information into structured financial scores that can be used to evaluate company performance effectively.
7. To integrate textual insights with relevant financial indicators in order to enhance the accuracy, depth, and reliability of the overall analysis.
8. To support data-driven portfolio analysis and decision-making by providing interpretable outputs and meaningful financial insights derived from the processed data.

1.2 Proposed System

1.2.1 Analysis/ Framework/ Algorithm

1. The system begins by collecting 10-K financial reports from the SEC EDGAR database for selected companies.
2. The retrieved reports are processed through a text preprocessing pipeline, which includes removal of unwanted characters, tokenization, and elimination of stop-words to clean the textual data.
3. The cleaned text is then transformed into numerical representations using TF-IDF, allowing the system to identify important terms and patterns within the financial documents.
4. A financial sentiment dictionary is applied to classify and extract sentiment-related information from the processed text, enabling the system to capture positive, negative, and domain-specific financial tones.
5. Based on the extracted sentiment features, the system generates quantitative financial scores that represent the overall sentiment and textual characteristics of the reports.
6. These scores are further combined with relevant financial indicators to enhance the reliability and depth of analysis.
7. The processed data is then used to generate insights that assist in evaluating company performance and supporting portfolio-related decision-making.
8. Finally, the system presents the results in a structured format, enabling users to interpret the analysis efficiently and make informed decisions.

1.2.2 Tools and Technologies

Layer	Technology

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716388

Data Acquisition	SEC EDGAR Database, Financial APIs
Processing	Python, NLP Libraries
Frontend	Web-based Interface
Backend	Python-based Processing System
Visualization	Graphical Outputs

Table 1.1 - Tools and Technologies

1.2.3 Methodology

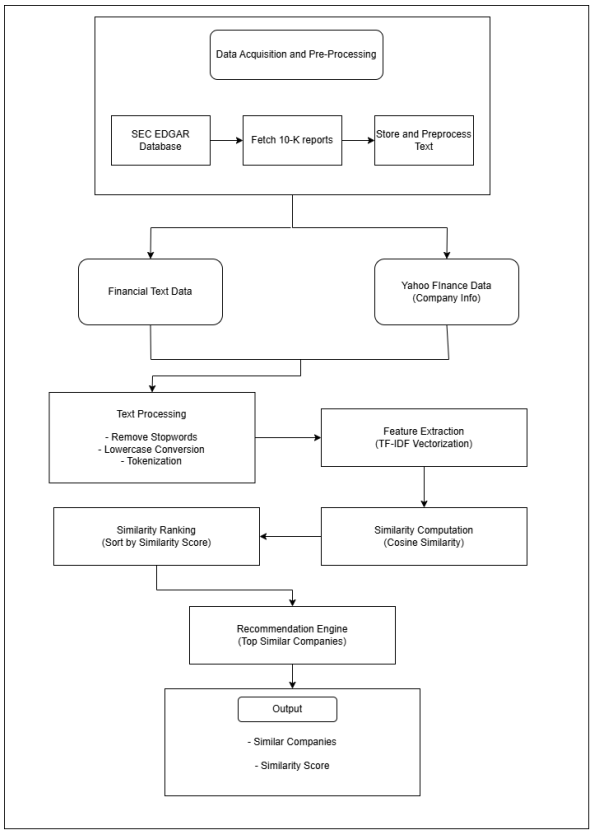

Figure 1.1 – Methodology/ Project Lifecycle

- Data Acquisition: The system begins by collecting financial reports in the form of 10-K reports from the SEC EDGAR database. These reports contain detailed textual information about company performance. Additionally, relevant financial data such as company information and historical metrics are obtained from financial sources like Yahoo Finance to support analysis.
- Data Preprocessing: The acquired financial text is processed to improve quality and consistency. This includes converting text to lowercase, removing stop-words, and performing tokenization to break the text into meaningful units. This step ensures that the data is clean and suitable for further analysis.
- Feature Extraction: After preprocessing, the cleaned textual data is transformed into numerical representations using TF-IDF vectorization. This technique helps in identifying the importance of terms within financial documents and enables comparison between different reports.
- Similarity Computation: The system calculates cosine similarity between processed financial documents to measure the degree of similarity between companies based on their textual disclosures. This step captures relationships and patterns within financial reports.
- Ranking and Analysis: Based on the computed similarity scores, companies are ranked accordingly. Higher similarity indicates closer alignment in financial and textual characteristics, which is useful for comparative analysis.
- Recommendation Generation: A recommendation engine is developed to identify and suggest companies with similar financial behavior or reporting patterns. This assists users in understanding comparable entities and making informed decisions.
- Output Generation: The final output includes a ranked list of similar companies along with their corresponding similarity scores. These results are presented in a structured and interpretable format for easy analysis.
- System Evaluation: The system is evaluated based on the accuracy and relevance of similarity results, as well as overall performance and usability. Improvements are made based on observations to enhance system effectiveness.

1.2.4 Design Details

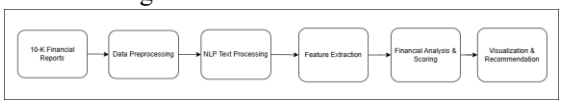

Figure 1.2 – System Design Workflow

1. Data Acquisition

➢ Retrieve 10-K financial reports from the SEC EDGAR database.

➢ Collect relevant company data and financial

information from external sources such as Yahoo Finance.

2. Text Processing
➢ Perform preprocessing operations such as lowercase conversion, tokenization, and stop-word removal.
➢ Clean and normalize financial text to prepare it for analysis.

3. Feature Extraction
➢ Apply TF-IDF vectorization to convert textual data into numerical form.
➢ Identify significant terms and patterns within financial documents.

4. Similarity Computation
➢ Calculate cosine similarity between processed financial reports.
➢ Measure the degree of similarity between companies based on textual disclosures.

5. Ranking & Recommendation
➢ Rank companies based on similarity scores obtained from analysis.
➢ Generate recommendations of similar companies for comparative evaluation.

6. Output & Visualization
➢ Display the list of similar companies along with their similarity scores.
➢ Present results in a clear and interpretable format for user understanding.

1.2.5 Hardware and Software Setup

Software Requirements:
• Operating System: Windows 10/11, macOS, or Linux.
• Programming Languages: Python 3.x.
• Libraries and Frameworks: NLP libraries (NLTK / Scikit-learn), Pandas, NumPy.
• Data Processing: TF-IDF Vectorizer, Cosine Similarity modules.
• Data Sources: SEC EDGAR Database, Yahoo Finance APIs.
• Development Tools: Google Colab (Cloud-based Python Environment).
• Browser: Google Chrome or any modern web browser.

Hardware Requirements:
For Development:
• Minimum: Intel Core i3 processor, 8 GB RAM, 128 GB SSD, stable internet connection.
• Recommended: Intel Core i5 or higher, 16 GB RAM or above, 256 GB SSD, high-speed internet connection.
For End Users:
• Desktop or laptop with internet access.
• Device capable of running modern web browsers.
• Stable internet connection (minimum 5 Mbps).

II. RESULT AND DISCUSSION

2.1.1 Implementation Plan

Phase 1: System Design and Data Processing Setup

• *Requirement Analysis:*

– Identified the need for automated analysis of complex financial reports using Natural Language Processing (NLP) techniques.
– Defined the complete workflow including data acquisition from SEC EDGAR, text preprocessing, feature extraction using TF-IDF, sentiment analysis, and result generation.
– Determined system requirements for handling unstructured financial data and generating meaningful analytical outputs.

• *Database Design:*

– Designed a structured approach for handling financial reports and processed data without the need for persistent database storage.
– Organized data flow for efficient processing of textual information and integration with financial indicators.
– Ensured smooth handling of intermediate data during preprocessing, feature extraction, and analysis stages.

• *Backend / Processing Development:*

– Implemented the system using Python in Google Colab for efficient data processing and analysis.
– Developed modules for fetching 10-K reports from SEC EDGAR and handling financial data inputs.
– Built preprocessing pipelines for cleaning and preparing textual data.
– Integrated TF-IDF vectorization and cosine similarity techniques for analysis.
– Generated financial scores and structured outputs based on processed data.
– Ensured smooth execution of the complete workflow from data acquisition to result generation.

Phase 2: Data Processing and Analysis Integration

• Data Processing Implementation*:*

– Developed the complete NLP pipeline using Python in Google Colab for processing financial text data.
– Implemented preprocessing techniques including tokenization, stop-word removal, and text normalization.
– Applied TF-IDF vectorization to transform textual data into numerical representations suitable for analysis.

• *Sentiment and Similarity Analysis:*

– Integrated a financial sentiment dictionary to extract sentiment-based features from 10-K reports.
– Implemented cosine similarity to measure relationships between financial documents.
– Generated financial scores based on extracted sentiment and textual patterns.

• *Testing and Validation:*

– Conducted functional testing of each module including data acquisition, preprocessing, feature extraction, and analysis.
– Verified correctness of TF-IDF outputs and similarity calculations.
– Evaluated system performance based on accuracy, consistency, and execution efficiency.

Phase 3: Deployment and Evaluation

• *Execution and Deployment:*

– Executed the system in Google Colab as a cloud-based environment for efficient processing of financial data.
– Ensured smooth end-to-end workflow from data acquisition to output generation.
– Enabled easy access and reproducibility of results through cloud execution.

• *Performance Evaluation:*

– Analyzed system performance based on processing speed, accuracy of results, and scalability.
– Evaluated effectiveness of sentiment-based analysis in extracting meaningful financial insights.
– Assessed the reliability of generated outputs for supporting decision-making.

• *Final Optimization:*

– Optimized preprocessing and feature extraction steps for improved efficiency.
– Reduced execution time and improved handling of large textual datasets.
– Enhanced overall system stability and reliability before final evaluation.

2.2 Result

The developed Finenzo system successfully demonstrates the capability of analyzing complex financial reports using Natural Language Processing (NLP) techniques. The system efficiently processes 10-K reports, extracts sentiment-based insights, and transforms unstructured textual data into structured financial information. The application of TF-IDF and cosine similarity enables effective identification of patterns within financial documents, supporting meaningful analysis.

The integration of textual insights with financial indicators enhances the reliability of the generated results. Performance evaluation shows that the system produces consistent and interpretable outputs, improving the efficiency of financial analysis. The results indicate that the proposed approach can effectively assist in evaluating company performance and support data-driven decision-making. Overall, the system provides a scalable and efficient solution for automated financial analysis.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716388

2.2.1 Results/ Outputs (User Interface) :

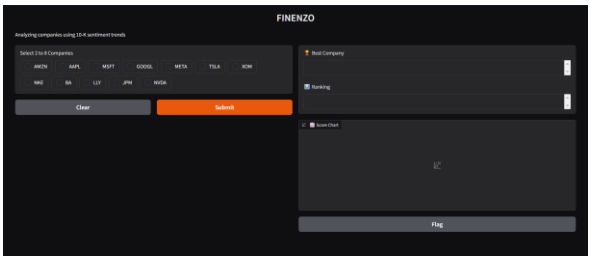

Figure. 2.1 – User Interface

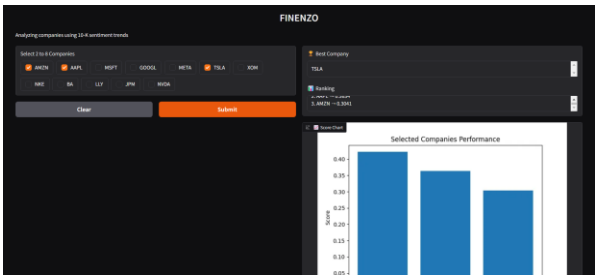

Figure. 2.2 – Company Selection

```
========== FINAL ANALYSIS ==========

📊 Company Ranking:

1. TSLA → Score: 0.4221
2. META → Score: 0.3938
3. JPM → Score: 0.3663
4. AAPL → Score: 0.3634
5. NKE → Score: 0.3518
6. NVDA → Score: 0.3389
7. AMZN → Score: 0.3041
8. XOM → Score: 0.3036
9. BA → Score: 0.2677
10. LLY → Score: 0.1846
11. MSFT → Score: 0.1446

⏳ FINAL RESULT:
Best performing company based on sentiment analysis: TSLA
```

Figure. 2.3 – Company Ranking

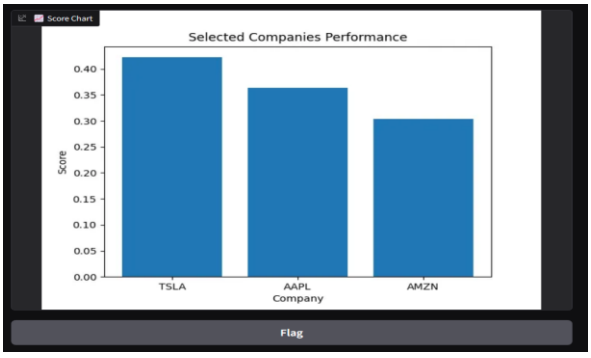

Figure. 2.4 – Company Performance

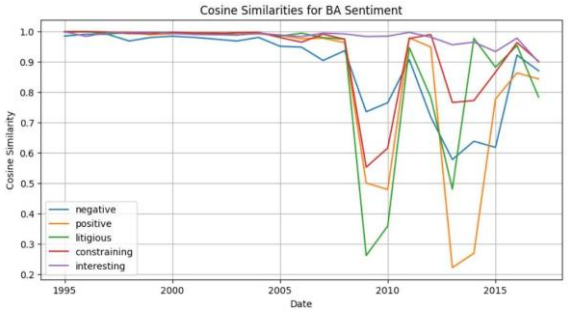

Figure. 2.5 – Cosine Similarities

Figure. 2.6 – Company Ranking

III. CONCLUSION

The Finenzo system enhances financial analysis by integrating Natural Language Processing (NLP) techniques with structured data processing to analyze complex financial reports. It reduces the challenges associated with manual interpretation of lengthy 10-K reports by transforming unstructured textual data into meaningful financial insights. The system streamlines the analysis process by combining data acquisition, text preprocessing, feature extraction, and sentiment-based evaluation within a unified framework.

By utilizing techniques such as TF-IDF and cosine similarity, the system enables efficient identification of patterns and relationships within financial documents. This improves the accuracy and effectiveness of evaluating company performance and supports informed decision-making. The approach demonstrates how NLP and data-driven methodologies can simplify financial analysis and provide scalable solutions for handling large volumes of financial data, ultimately contributing to more reliable and efficient investment analysis.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716388

REFERENCES

[1] Zijie Zhao, “Next-Generation Intelligent Portfolio Management”, MIT Thesis, May (2024).

[2] Prakash K. Aithal, Geetha M., U. Dinesh Acharya, Basri Savitha, and Parthiv Menon, “Real-Time Portfolio Management System Utilizing Machine Learning Techniques”, IEEE Access publication, April (2023).

[3] Dogu Tan Araci, “Financial Sentiment Analysis with Pre-trained Language Models”, Research Publication, (2020).

[4] Purva Singh, “Intelligent Portfolio Management via NLP Analysis of Financial 10-K Statements”, IJAIA publication, Nov (2020).

[5] Tim Loughran and Bill McDonald, “Textual Analysis of Financial Disclosures: A Survey of the Literature”, Journal of Accounting Research publication, (2016).