DrillBit

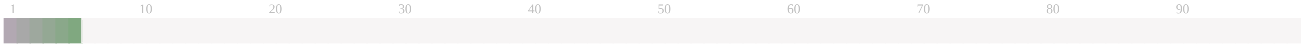
The Report is Generated by DrillBit Plagiarism Detection Software

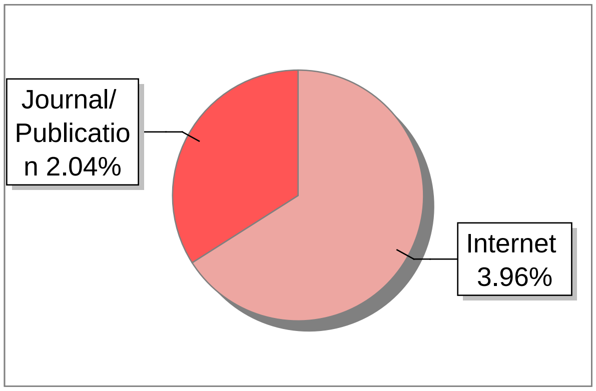
Submission Information

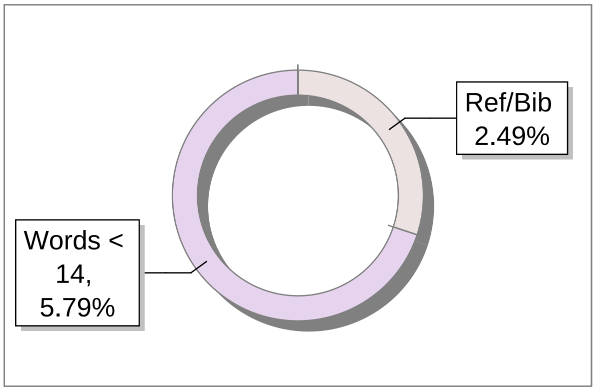
Author Name	GAGAN VADLAMUDI
Title	Finance and Investment with LLM’s and GenAI
Paper/Submission ID	1559817
Submitted by	krc@iiitdwd.ac.in
Submission Date	2024-03-22 10:42:06
Total Pages	34
Document type	Research Paper

Result Information

Similarity **6 %**







Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Sources: Less than 14 Words %	Excluded
Excluded Source	**0 %**
Excluded Phrases	Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

6	**10**	**A**	A-Satisfactory (0-10%) B-Upgrade (11-40%) C-Poor (41-60%) D-Unacceptable (61-100%)
SIMILARITY %	MATCHED SOURCES	GRADE	

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	lablab.ai	1	Internet Data
2	www.ncbi.nlm.nih.gov	1	Internet Data
3	blog.quantinsti.com	1	Internet Data
4	csjmu.ac.in	1	Publication
5	isdsi2023.iimranchi.ac.in	1	Publication
6	www.frontiersin.org	<1	Internet Data
7	Thesis Submitted to Shodhganga Repository	<1	Publication
8	dspace.bracu.ac.bd	<1	Publication
9	Towards a systems-level understanding of development in the marine annelid Platy by Williams-2016	<1	Publication
10	www.linkedin.com	<1	Internet Data

EXCLUDED PHRASES

1 **indian institute of informtion technology dharwad**

Major Project Report

on

Finance and Investment with LLM's and GenAI

Submitted by

Aravind Gangavarapu 20bds010

Balusu Bhanu Prakash 20bds012

Gagan Vadlamudi 20bds019

Peddisetty Venkata Sai Pranay 20bds038

Under the guidance of

Dr. Chinmayananda A

Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

21/03/2024

Certificate

This is to certify that the project, entitled **Finance and Investment with LLM's and GenAI**, is a bonafide record of the Major Project coursework presented by the students whose names are given below during 2023-24 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Data Science and Artifical Intelligence.

Roll No	Names of Students
20bds010	Aravind Gangavarapu
20bds012	Balusu Bhanu Prakash
20bds019	Gagan Vadlamudi
20bds038	Peddisetty Venkata Sai Pranay

Dr. Chinmayananda A
(Project Supervisor)

Contents

List of Figures

List of Tables

1 Abstract

In the rapidly evolving financial landscape, the integration of Large Language Models (LLMs) presents a transformative opportunity to revolutionize investment analysis and decision-making. This research paper delves into the innovative application of LLMs within the financial sector, focusing on their potential to automate complex tasks, predict market trends, and analyze news sentiment, all while incorporating a financial education chatbot to demystify investment basics. By leveraging the power of Natural Language Processing (NLP), this project aims to simplify stock analysis for users, enabling informed investment decisions through the seamless blend of specialized finance knowledge and extensive data analysis.

The study explores the challenges faced by finance LLMs, such as the need for efficient data storage and retrieval, and proposes solutions like Vector Databases (VectorDB) to address these issues. Furthermore, it highlights the promising Retrieval-Augmented Generation (RAG) method as a key advancement in enhancing LLMs' accuracy and reliability. This research paper provides a comprehensive overview of the project, showcasing the potential of LLMs to not only streamline financial processes but also to democratize access to financial education and investment insights.

2 Introduction

The advent of artificial intelligence (AI) has revolutionized various sectors, including finance, by offering innovative solutions to complex problems. In the realm of financial analysis, AI has the potential to transform the way investors and traders make decisions, providing them with insights that were previously inaccessible or too time-consuming to obtain. This paper aims to explore the application of AI, specifically Large Language Generative AI models, in the financial analysis of the stock market. Our focus is on developing an end-to-end analysis system that encompasses a wide range of functionalities, from fundamental analysis to pattern recognition, thereby offering a comprehensive tool for both investors and traders.

Our project is grounded in the belief that leveraging AI can significantly enhance the decision-making process in the financial markets. We have designed a system that incorporates multiple features, including analysis of financial statements, management and sector analysis, and ratio analysis, which are crucial components of fundamental analysis. To ensure the system's efficiency and accuracy, we have employed the latest Retrieval Augmented Generation (RAG) workflow. This approach allows us to store context in vector databases, facilitating faster and semantically rich retrieval of information. Furthermore, we have developed a RAG-based conversational chatbot that is equipped with the context on the financials of different companies, making it a valuable resource for investors and traders alike.

In addition to fundamental analysis, our system also includes a sentiment analysis agent that retrieves market news from various sources, such as Yahoo Finance[7] and Google News, and classifies the news based on sentiment. This feature is particularly useful for investors and traders looking to gauge the market sentiment towards specific stocks. The sentiment analysis agent not only identifies the sentiment (negative or positive) but also provides a percentage of strength, offering a nuanced understanding of the market's perception.

For traders, our system offers a pattern recognition agent that identifies important patterns from candlestick charts. This feature is trained on a finetuned VGG-16[11] model using transfer learning, showcasing the power of AI in recognizing and interpreting complex financial data. By utilizing multiple Large Language Models (LLMs), embedding models, vector databases, and retrieval strategies, we have successfully linked these components using LangChain[2], creating a cohesive and powerful financial analysis tool.

This paper will delve into the technical aspects of our project, detailing the methodologies used, the challenges encountered, and the solutions implemented. We aim to provide a comprehensive understanding of how AI can be harnessed to enhance financial analysis, offering valuable insights that could shape the future of investment and trading in the financial markets.

3 Related Work

Several studies have been conducted on Large Language Models (LLMs) in the financial sector, highlighting their potential to revolutionize financial processes and decision-making.

3.1 Revolutionizing Finance with LLMs: An Overview of Applications and Insights[13]

The research paper aims to provide a comprehensive overview of the current state of research and applications of Large Language Models (LLMs) in the financial sector. This survey will delve into the key areas of financial engineering, financial forecasting, financial risk management, and financial real-time question answering, highlighting the latest advancements and the challenges faced in each domain.

Financial engineering has been a traditional domain where LLMs have shown promise in automating complex calculations and simulations. The integration of LLMs in this field has been explored through various studies, focusing on the automation of financial modeling and the development of more sophisticated financial instruments. These models have been used to optimize trading strategies, portfolio management, and risk assessment. However, the complexity of financial engineering tasks and the need for high-precision calculations pose significant challenges to the application of LLMs in this domain.

LLMs have been increasingly utilized in financial forecasting, leveraging their ability to process vast amounts of data and generate predictions. Studies have explored the use of LLMs for predicting stock prices, economic indicators, and market trends. The potential of LLMs in this area is vast, with research focusing on improving the accuracy and reliability of forecasts. However, the inherent unpredictability of financial markets and the need for real-time data processing present significant challenges to the application of LLMs in financial forecasting.

The application of LLMs in financial risk management has been a focus of recent research, with a particular emphasis on the development of models that can assess and mitigate financial risks. These models have been used to analyze market volatility, credit risk, and operational risk. The integration of LLMs in risk management has the potential to significantly enhance the accuracy and efficiency of risk assessment processes. However, the complexity of financial risks and the need for regulatory compliance present significant challenges to the application of LLMs in this domain.

The ability of LLMs to process natural language queries and provide instant financial advice and support has been a significant advancement in the financial sector. This capability has been explored through various studies, focusing on the development of models that can understand and respond to complex financial queries. The potential of LLMs in this area is vast, with research focusing on improving the accuracy and reliability of financial advice. However, the need for high-quality training data and the complexity of financial terminology present significant challenges to the application of LLMs in real-time question answering.

Despite the promising advancements in the application of LLMs in the financial sector, several challenges remain. These include the need for high-quality training data, the complexity of financial terminology and regulations, and the requirement for high accuracy and reliability in predictions. Future research in this field will likely focus on addressing these challenges and exploring new applications of LLMs in the financial sector.

In conclusion, the research paper highlights the significant potential of LLMs in revolutionizing the financial sector. While challenges remain, the ongoing research and development in this area are poised to unlock new opportunities and innovations in the financial industry.

3.2 When Large Language Models Meet Vector Databases: A Survey[9]

The research paper focuses on the intersection of Large Language Models (LLMs) and VectorDBs (VecDBs), aiming to address the known shortcomings of LLMs and explore potential solutions through the integration of VecDBs. This survey delves into the challenges faced by LLMs, including hallucination, domain knowledge limitations, bias, high costs, and the oblivion problem, and how VecDBs can serve as a potential solution to these issues.

LLMs, such as GPT, T5, and Llama, have made significant strides in natural language processing, enabling interactive smart assistants that can process, understand, and generate human-like text. However, they face several challenges, including the generation of plausible but factually incorrect or nonsensical information, limitations in domain knowledge, difficulties in real-time knowledge updates, and biases introduced during training. These issues stem from the fact that LLMs are primarily trained on public datasets, which may not cover all domain-specific questions or reflect the dynamic nature of the world. Additionally, the high costs associated with commercial use of LLMs and the oblivion problem, where LLMs tend to forget information from previous inputs, further complicate their application.

VecDBs, or vector databases, offer a promising solution to these challenges by providing a reliable data system for managing and retrieving large amounts of data efficiently. By integrating VecDBs with LLMs, these databases can serve as an external knowledge base, providing domain-specific knowledge for LLMs, acting as a memory to save previous chat contents for each user's dialogue, or as a semantic cache to address the oblivion problem. This integration can enhance the capabilities of LLMs, making them more accurate, reliable, and efficient in various applications.

Despite the potential of VecDBs as a solution to LLMs' shortcomings, there is a lack of research exploring this intersection. This survey aims to fill this gap by providing a comprehensive overview of the current state of research on the integration of VecDBs and LLMs, highlighting the potential benefits and future directions in this area. It also discusses the unique challenges and opportunities that arise from combining these technologies, offering insights into how VecDBs can refine LLMs' known shortcomings and paving the way for future research opportunities in this fertile intersection of technology.

3.3 Retrieval-Augmented Generation for Large Language Models: A Survey[8]

The literature survey for this paper focuses on the evolution and application of RetrievalAugmented Generation (RAG) in the context of Large Language Models (LLMs), such as the GPT series, LLama, and Gemini. RAG represents a significant advancement in the field of natural language processing, addressing the limitations of LLMs in handling domain-specific or highly specialized queries and mitigating the generation of incorrect information, or ”hallucinations.”

RAG, introduced by Lewis et al. in 2020, integrates external data retrieval into the generative process of LLMs, enhancing their ability to provide accurate and relevant responses. This approach involves an initial retrieval step where LLMs query an external data source to obtain relevant information before generating text, ensuring that responses are grounded in retrieved evidence. This process has been rapidly adopted, making LLMs more viable for practical applications and significantly enhancing the accuracy and relevance of their output.

The development of RAG has unfolded through four distinctive phases, starting with the foundational efforts in 2017 to optimize pre-training methodologies. The subsequent introduction of GPT-4 led to a significant transformation in the landscape of RAG technology, evolving into a hybrid approach combining the strengths of RAG and fine-tuning.

Despite the rapid growth of RAG research, there has been a lack of systematic consolidation and abstraction in the field, posing challenges in understanding the comprehensive landscape of RAG advancements. This survey aims to outline the entire RAG process and encompass the current and future directions of RAG research, providing a thorough examination of retrieval augmentation in LLMs.

The survey will present a thorough and systematic review of the state-of-the-art RAG, delineating its evolution through paradigms including naive RAG, advanced RAG, and modular RAG. It will identify and discuss the central technologies integral to the RAG process, focusing on the aspects of ”Retrieval,” ”Generator,” and ”Augmentation,” and delve into their synergies. A comprehensive evaluation framework for RAG will be constructed, outlining the evaluation objectives and metrics, and a comparative analysis will clarify the strengths and weaknesses of RAG compared to fine-tuning from various perspectives. The survey will also anticipate future directions for RAG, emphasizing potential enhancements to tackle current challenges, expansions into multi-modal settings, and the development of its ecosystem.

4 Datasets

The financial data analysis process, inherently complex and multifaceted, necessitates the integration of data from a variety of reliable and trustworthy sources across the internet. This approach allows for the exploration of multiple use cases, each supported by a unique set of inputs derived from distinct sources. A significant portion of the data, predominantly unstructured, encompasses images, text files, and news articles. A central component of our dataset collection is the Vector database, meticulously constructed from the integrated annual reports of various companies. This database is comprehensive, encompassing financial statements, management analyses, corporate information, and the company's vision. The compilation of these reports provides a rich, structured dataset that is fundamental to our financial education chatbot project.

For the development of our financial education chatbot, we have harnessed the power of OpenAI's Whisper model to transcribe interview videos of Warren Buffet available on YouTube. This transcript data serves as a valuable resource for the chatbot, offering insights into the financial strategies and perspectives of one of the most successful investors in history. In addition to the structured data from the Vector database, we have also incorporated market news articles from Yahoo Finance News and Google News into our dataset. We have gathered and utilized multiple other data sources for the purpose of sector analysis, management analysis, and ratio analysis. These additional datasets, while not explicitly detailed here, are integral to our comprehensive financial data analysis framework, providing a broad spectrum of data points to support our research objectives.

For sector analysis, our data sources include the Indian Brand Equity Foundation (IBEF)[12]. The Government of India publishes yearly reports for every industry/sector in the country, which are made public by IBEF. This source is credible and trusted, providing a comprehensive view of the financial health and performance of various sectors within the Indian economy. For pattern recognition, we utilized real-time data from the yfinance library[5]. This dataset contains prices (open, low, close, high), volume, and the number of trades as columns. Out of these columns, we have considered the closing price of the stock as the targeted feature. The data spans a 14-year period (2007-2021), with each price representing the price of one day. This time frame and the specific focus on closing prices are crucial for our pattern recognition analysis, offering a detailed view of market trends over time.

5 Methodology

In this section, we outline the methodology employed to analyze the various functionalities within our application. Our approach involves a detailed examination of each functionality to gain insights into their operations, performance, and user experience.

5.1 Financial Chatbot

In this paper we present the development of a specialized financial chatbot tailored to the Indian Stock Market aiming to streamline processes and provide valuable assistance to users. However, the financial sector poses unique challenges due to the dynamic nature of markets and the need for accurate and timely information. Given the limitations of base Language Models (LMs) in providing comprehensive information regarding financial stocks in the Indian Stock Market, we employed a Retrieval Augmented Generation (RAG) workflow to enhance the capabilities of our chatbot. To enrich the chatbot's knowledge base, we integrated a Vector Database, Qdrant, serving as an external repository of trustworthy and reliable financial information. This external knowledge base supplements the inherent knowledge of the LLMs, enabling them to provide more informed responses to user queries.

Through the incorporation of the Vector Database, the financial chatbot can now access a vast array of relevant data points, including stock prices, market trends, and company performance metrics. Furthermore, the utilization of enhanced retrieval algorithms enables the vector database to offer additional context to the base LLM for text generation. This synergy between retrieval and generation mechanisms ensures that the chatbot delivers relevant and coherent responses to user inquiries, enhancing the overall conversational experience.

5.1.1 Retrieving and Loading Annual Reports

- In the RAG workflow implemented for our financial chatbot, a crucial step involves retrieving and loading annual reports into the vector database. To achieve this, we utilized Google Search as a means of sourcing annual reports, specifying the file type as PDF to target the desired documents. Upon obtaining search results, we extracted the topmost result, which typically corresponds to the most recent annual report available.

- To extract text from the retrieved PDF documents, we employed LangChain's document loader, specifically leveraging the PyPDFLoader module[3]. PyPDFLoader facilitates the extraction of text from PDF files, enabling seamless integration into our workflow for further processing and storage within the vector database.
- This approach not only automates the task of sourcing and retrieving annual reports but also ensures that the extracted text is readily accessible for subsequent analysis and utilization within the chatbot's conversational framework. By employing PyPDFLoader as part of our RAG workflow, we streamline the process of populating the vector database with valuable financial data, thereby enhancing the chatbot's ability to provide accurate and comprehensive responses to user inquiries.

5.1.2 Dividing the data into Chunks

The process of dividing retrieved text into manageable chunks for storage in a vector database is a critical step in the development of our financial chatbot. To facilitate this, we employed LangChain's RecursiveCharacterTextSplitter, a sophisticated tool designed to efficiently partition text into smaller segments. This tool operates with a specific chunk size of 512 tokens, ensuring that each chunk is compact enough for efficient processing while still retaining the contextual integrity of the original document. Additionally, RecursiveCharacterTextSplitter incorporates an overlap of 30 tokens between adjacent chunks, which aids in maintaining continuity and coherence across the text. This overlap mechanism is particularly beneficial in scenarios where the context of a statement spans across two chunks, ensuring that the chatbot can accurately interpret and respond to queries that span multiple chunks. The chunks are then meticulously cleaned using multiple regular expressions as the data extracted from the document loader contains noise.

Furthermore, to enhance the searchability and retrievability of these chunks, we have incorporated metadata for each chunk as payload. This includes the name of the company, the type of document (e.g., annual report), the chunk data itself, and a unique chunk ID (e.g., RELIANCE-AR-2023-2). These identifiers serve as filters during the search process, allowing the chatbot to quickly locate and retrieve the most relevant information based on the user's query. This approach not only optimizes the storage and retrieval of text data but also significantly improves the chatbot's ability to provide accurate and contextually relevant responses.

5.1.3 Extracting Dense and Sparse Embeddings

In our financial chatbot, we employ a dual representation strategy for storing chunks of text in the Qdrant vector database[4], leveraging both dense and sparse representations to optimize data retrieval and processing. Dense embeddings are generated using the ”embed-english-v3.0” model, which produces high-dimensional vectors (1024-dimensional arrays) that capture the semantic meaning of the text. These embeddings facilitate the identification of semantically similar chunks, enhancing the chatbot's ability to provide relevant responses.

Concurrently, we employ sparse representations for the same chunks. This involves creating a vocabulary file in JSON format, which serves as a comprehensive dictionary of all unique words encountered in the chunks. Each chunk is then represented as a sparse vector, consisting of indices and values that correspond to the positions of words in the vocabulary. This sparse representation is particularly efficient for storage and retrieval, as it significantly reduces the memory footprint compared to dense embeddings.

The vocabulary file is dynamically updated to include new words as they are encountered, ensuring that the sparse representations remain accurate and relevant. This dynamic updating process allows the chatbot to adapt to new information and improve its performance over time. By utilizing both dense and sparse representations, we achieve a balance between computational efficiency and information retrieval accuracy. This dual representation strategy not only enhances the chatbot's ability to process and understand financial data but also ensures that it can provide accurate and contextually relevant responses to user inquiries.

5.1.4 Storing the vectors in Vector Database

In our financial chatbot, each text chunk is ingested into the Qdrant vector database as a distinct point, encapsulating both sparse and dense vector representations alongside comprehensive metadata. This metadata includes detailed information such as the company name, document type, chunk data, and a unique chunk ID, which are crucial for filtering and retrieving relevant information based on user queries. The inclusion of text data within the payload of each point facilitates direct access to the original text, ensuring that the chatbot can provide immediate and accurate responses to user inquiries.

5.1.5 Hybrid Search

In the financial sector, the development of a hybrid search functionality, known as MergedRetriever, has been instrumental in enhancing information retrieval capabilities. This system, powered by LangChain, integrates both Dense and Sparse Retrievers to leverage the strengths of both retrieval methods. The Dense Retriever component employs Hierarchical Navigable Small World (HNSW)[10] technology to efficiently search through a vector database by transforming query text into coherent embeddings. HNSW, known for its ability to approximate nearest dense vectors, significantly reduces search time complexity, making it particularly effective in high-dimensional spaces.

Conversely, the Sparse Retriever component utilizes Qdrant's SparseVectorRepresentation to transform query vectors into sparse vectors, reducing data dimensionality. This process is followed by the retrieval of the nearest sparse vectors using the BM25 algorithm, which prioritizes the most relevant documents based on term frequency and inverse document frequency. BM25's effectiveness in handling textual data makes it an ideal choice for keyword searches, ensuring that the most pertinent information is prioritized.

Following the retrieval process, the documents obtained from both the dense and sparse retrieval components are merged and passed through Cohere's rerank model[1]. This model refines the initial search results, reordering them based on predefined criteria to ensure the final results are both comprehensive and highly relevant. The text from the top-most results, as determined by the Cohere rerank model, is then retrieved and presented to the user. This system not only ensures the accessibility of comprehensive financial information but also supports informed decision-making by providing accurate and relevant results.

5.1.6 Text Generation

The 'gpt-3.5-turbo' model from OpenAI, utilized for text generation, is characterized by its deterministic responses, achieved by setting the temperature parameter to 0, which controls the randomness of the output. This model is augmented with a conversationalRetrievalChain, integrating the Merged Retriever and memory functionalities. The Merged Retriever, comprising both Dense and Sparse Retrievers, efficiently accesses and retrieves relevant information from a vector database, while the ConversationBufferWindow Memory maintains a sliding window of recent interactions, enhancing the chatbot's conversational capabilities. This setup ensures that the chatbot can generate contextually relevant responses by referencing previous interactions.

The entire process, from retrieval to text generation, is automated using LangChain, which orchestrates the integration of various components, including the retrieval chain and memory functionalities. This automation not only streamlines the process but also ensures consistency and reliability in the chatbot's responses. Furthermore, the integration of Qdrant Vector Database as the repository for the vector database significantly enhances the chatbot's ability to offer enhanced insights and support informed decision-making processes. This combination of technologies represents a significant advancement in conversational AI, showcasing the potential to revolutionize the access and utilization of financial information.

5.2 Automation of Sector Analysis

The methodology employed in this study, incorporating the analysis of the IBEF sector reports provided by the government of India. These reports, which are updated annually, offer a comprehensive overview of the Indian economy and its various sectors. To facilitate efficient data retrieval and analysis, these reports are divided into chunks, and coherence embeddings are extracted for each chunk. These embeddings are then stored in a vector database named FAISS (Facebook AI Similarity Search)[6]. FAISS is a library developed by Facebook Research that provides efficient similarity search and clustering of dense vectors. It is designed to handle large-scale datasets and offers significant performance improvements over traditional search methods, making it an ideal choice for storing and retrieving the embeddings derived from the IBEF sector reports.

Two primary functionalities are provided within this framework: a Sector Analysis Chatbot and a brief overview of Sector analysis for a company. The Sector Analysis Chatbot is designed to retrieve relevant information from the FAISS database based on user queries. This functionality leverages the coherence embeddings stored in FAISS to identify and provide the most pertinent information related to the user's query. The chatbot's design allows for dynamic and context-aware responses, ensuring that users receive the most relevant and up-to-date information.

In addition to the chatbot, a brief overview of Sector analysis for a company is provided. This overview considers a range of factors to offer a comprehensive analysis of the sector. These factors include Industry Lifecycle Analysis, which assesses the stage of the industry's development (Pioneering, Growth, Maturity/Saturation, Declining); Michael Porter's five forces analysis, which evaluates the competitive forces within the industry (Level of competition, Threat of new entrants, Threat of substitutes, Bargaining power of suppliers, Bargaining power of buyers); Government Protection and initiatives, which examines the support and policies provided by the government; Additional considerations, which may include environmental, social, and governance (ESG) factors; and a list of the Top 10 companies by market share.

This methodology integrates the analysis of government reports with advanced machine learning techniques to provide users with a detailed and insightful overview of specific sectors. By leveraging the FAISS database for efficient data retrieval and incorporating a range of analytical frameworks, this approach offers a robust and comprehensive tool for understanding the dynamics of the Indian economy and its various sectors.

5.3 Financial Education from Warren Buffet

Our application features a financial education chatbot designed after Mr. Warren Buffet to cater to beginners in investment, aiming to simplify complex financial concepts in a manner that resonates with the industry's giants. To achieve this, we've integrated the Retrieval Augmented Generation (RAG) workflow, where questions and answers are meticulously stored in a CSV file format. This method ensures that the chatbot can provide detailed explanations and insights into financial concepts, making them accessible and understandable to beginners.

The process begins with the utilization of Youtube DL to download audio files of Warren Buffet's interview videos available on YouTube. These audio files are then processed using OpenAI's Whisper model, which extracts relevant textual data from the mp3 files. This extracted textual data is then transformed into a structured format of questions and answers, which are subsequently stored in a CSV file. This step is crucial as it converts the raw audio data into a format that can be easily understood and utilized by the chatbot. Following the transformation of the textual data into questions and answers, textual embeddings are generated using OpenAI's ADA-002 model. These embeddings serve as a numerical representation of the text, enabling the chatbot to understand and process the information more effectively. Once the embeddings are generated, the chatbot is equipped to respond to user queries. When a user poses a question, the relevant information is fed into OpenAI's DaVinci text generation model. This model then generates a response based on the input, providing the user with a detailed explanation or insight into the financial concept they've inquired about.

In summary, our financial education chatbot leverages a combination of audio processing, textual data extraction, and advanced text generation models to provide beginners in investment with clear and concise explanations of financial concepts. By incorporating the RAG workflow and utilizing models like Whisper, ADA-002, and DaVinci, we've created a tool that not only simplifies complex financial information but also makes it accessible to those just starting their journey in investment.

5.4 Market News Sentiment Analysis

In the domain of financial market analysis, sentiment analysis plays a crucial role in understanding the prevailing market sentiment based on news and information. However, traditional Large Language Models (LLMs) are often trained on extensive historical data and lack the capability to access real-time market news for sentiment analysis. To bridge this gap, innovative solutions like LangChain's Agents and Tools are employed. These tools are designed to search the internet and gather the most relevant market news for a specific stock, providing the LLM with the latest and most current information.

LangChain's Agents and Tools, such as those integrated with Yahoo Finance and Google News, are instrumental in retrieving market news. These tools are specifically designed to through retrieve vast amounts of data, identifying and extracting news articles that are most relevant to the stock in question. By leveraging the power of these tools, the system can gather a wide range of news sources and articles, ensuring a comprehensive view of the market sentiment.

Once the relevant market news has been retrieved, it is then passed to the GeminiChat model for classification. The GeminiChat model is capable of analyzing the sentiment of the market news, categorizing it as positive, negative, or neutral. Additionally, it provides a percentage indication of the sentiment's strength, offering a nuanced understanding of the market's current mood. This classification process is applied to every piece of news retrieved by the tools, ensuring that the output is not only sentimentally classified but also quantitatively analyzed. This comprehensive approach allows for a more informed and accurate assessment of market sentiment, supporting more informed decision-making in financial markets.

5.5 Pattern Recognition

In the field of financial market analysis, candlestick patterns play a pivotal role in predicting market movements. These patterns, which are visual representations of price activity, can be categorized into two main types: single candlestick patterns and multiple candlestick patterns. Each type requires a distinct approach for recognition and analysis, leveraging different models and techniques to interpret the market's behavior.

For single candlestick patterns, such as Hammer, Pinbar, and Gravestone Doji, a fine-tuned VGG-16 model is employed. The VGG-16 model, originally designed for image classification tasks, is adapted for the specific task of identifying these patterns. This involves converting the candlestick images into numerical arrays through a process that includes resizing, converting to grayscale, and normalizing pixel values. Once processed, these arrays are input into the VGG-16 model, which has been fine-tuned to recognize the unique characteristics of each single candlestick pattern. This approach leverages transfer learning, utilizing the pre-trained weights of the VGG-16 model to significantly reduce training time and computational resources, while ensuring the model is equipped with a solid foundation of knowledge for pattern recognition.

On the other hand, multiple candlestick patterns, including Head and Shoulders, W pattern, and M pattern, are recognized using a pre-trained YOLO v8 model. This model, trained on a wide range of stock market data patterns, is specifically designed to identify and classify these complex patterns. The YOLO v8 model, sourced from Hugging Face, is capable of analyzing the intricate details of multiple candlestick patterns, including their shape, size, and position, to accurately determine their significance. This process involves training the model on a dataset that includes various multiple candlestick patterns, allowing it to learn the key features and relationships that define these patterns. Once trained, the model can be applied to new data to quickly identify and classify these patterns, providing valuable insights into market trends and potential future movements.

Both methods, while distinct in their approaches, serve a common goal: to enhance the ability of financial analysts and traders to interpret market signals and make informed decisions. By leveraging the strengths of the VGG-16 model for single candlestick patterns and the YOLO v8 model for multiple candlestick patterns, these techniques provide a comprehensive framework for analyzing market trends and predicting future price movements.

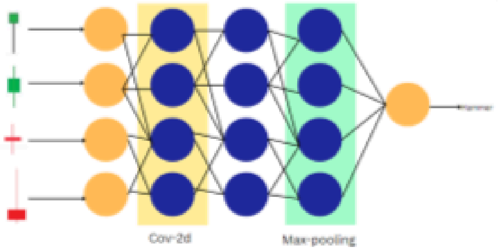


Figure 1. VGG-16 Methodology

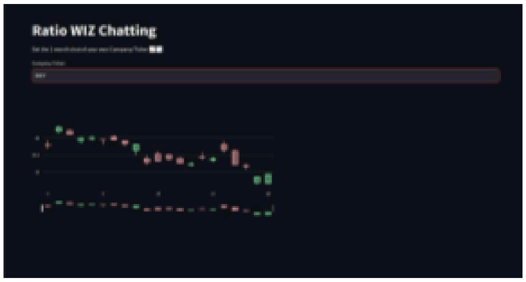

Ratio WIZ Chatting

Figure 2. Candlestick Patterns

5.6 Stock Recommendation

The methodology employed in this study is designed to provide a comprehensive approach to stock recommendation, leveraging a combination of financial data and market sentiment analysis. The process begins with the collection and preprocessing of two distinct datasets: recommendations provided by different private firms from Yahoo Finance and stock market news from Kaggle. These datasets are merged into a single dataset by the date of the market news, ensuring that the analysis is synchronized across both financial recommendations and market sentiment.

Following the merging of datasets, the textual data in the Market News dataset is converted into both TF-IDF vectors and Bag of Words vectors using the Natural Language Toolkit (NLTK) library. This feature extraction process is crucial as it transforms the textual data into numerical representations that can be understood by machine learning models. The TF-IDF method assigns weights to words based on their frequency in the document and their rarity across the entire dataset, highlighting the importance of words in the context of the entire corpus. The Bag of Words approach represents text as a collection of words, disregarding grammar and word order but preserving multiplicity, offering a simpler method that can be useful for certain types of analysis.

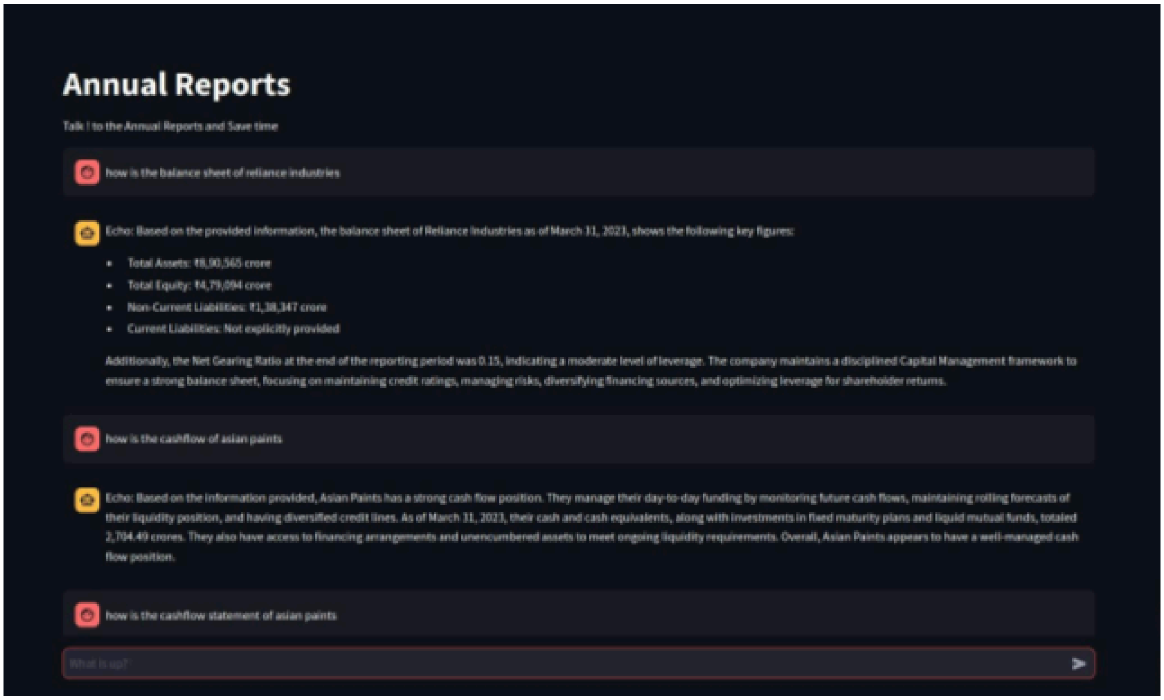
The next stage involves training multiple machine learning models on these vectors. A variety of models are tested, including Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Decision Tree, Support Vector Machine (SVM), and Random Forest. Each model is trained on both the TF-IDF and Bag of Words vectors, allowing for a comprehensive evaluation of their performance. The Random Forest model, in particular, has shown promising results when trained on TF-IDF vectors, achieving an accuracy of 76.31

This methodology provides a robust framework for stock recommendation, demonstrating the potential for accurate and insightful stock recommendations through the use of a variety of machine learning models and feature extraction techniques. The results, particularly the high accuracy achieved by the Random Forest model, underscore the effectiveness of this approach in navigating the complexities of the stock market.

6 Results and Discussions

6.1 Financial Chatbot

Below given is an example of financial chatbot. This includes a search bar where users can type in the name of a company to find its balance sheet or cash flow statement. In the image, we're asking about the annual report of Reliance Industries.



Annual Reports

Talk ! to the Annual Reports and Save time

how is the balance sheet of reliance industries

Echo: Based on the provided information, the balance sheet of Reliance Industries as of March 31, 2023, shows the following key figures:

- Total Assets: ₹8,90,565 crore
- Total Equity: ₹4,79,094 crore
- Non-Current Liabilities: ₹1,38,347 crore
- Current Liabilities: Not explicitly provided

Additionally, the Net Gearing Ratio at the end of the reporting period was 0.15, indicating a moderate level of leverage. The company maintains a disciplined Capital Management framework to ensure a strong balance sheet, focusing on maintaining credit ratings, managing risks, diversifying financing sources, and optimizing leverage for shareholder returns.

how is the cashflow of asian paints

Echo: Based on the information provided, Asian Paints has a strong cash flow position. They manage their day-to-day funding by monitoring future cash flows, maintaining rolling forecasts of their liquidity position, and having diversified credit lines. As of March 31, 2023, their cash and cash equivalents, along with investments in fixed maturity plans and liquid mutual funds, totaled 2,704.49 crores. They also have access to financing arrangements and unencumbered assets to meet ongoing liquidity requirements. Overall, Asian Paints appears to have a well-managed cash flow position.

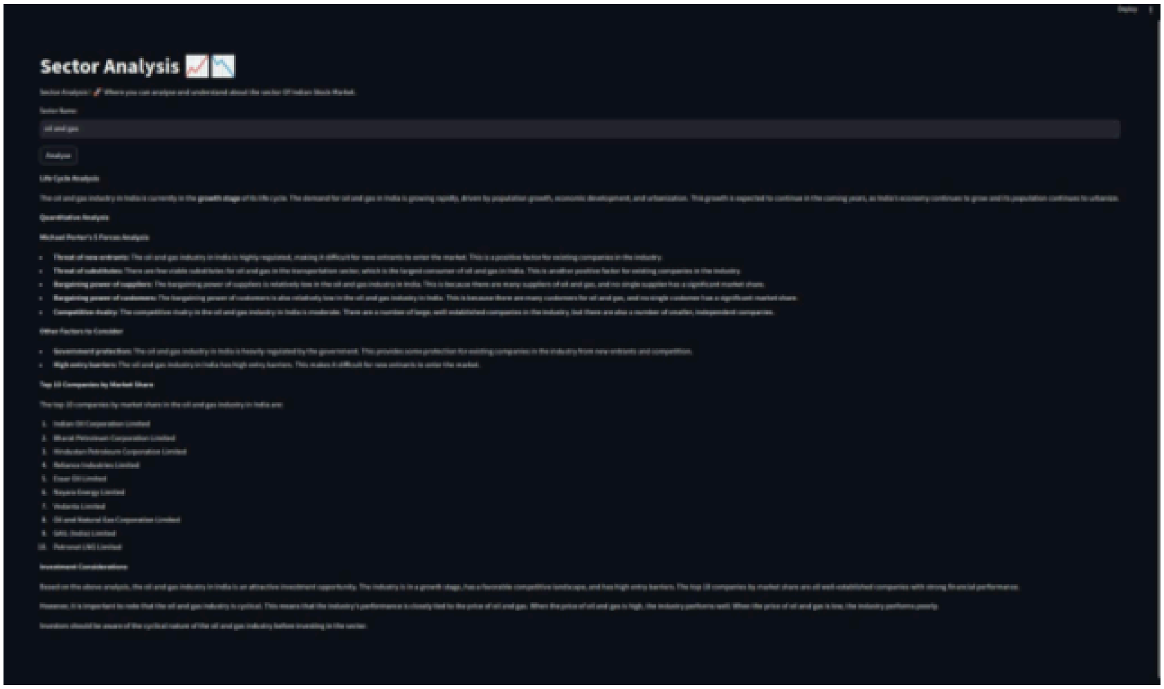
how is the cashflow statement of asian paints

Figure 3. Financial Chatbot

6.2 Sector Analysis

This is the image asking about a sector ”Automobile Industry”. It gives about

- Life Cycle
- Michael Porter’s Five Forces Analysis
- Government Protection
- Additional Points
- Top 10 Companies



Sector Analysis 📈📉

Sector Analysis : 🚀 Where you can analyse and understand about the sector Of Indian Stock Market.

Sector Name:

oil and gas

Life Cycle Analysis

The oil and gas industry in India is currently in the **growth stage** of its life cycle. The demand for oil and gas in India is growing rapidly, driven by population growth, economic development, and urbanization. This growth is expected to continue in the coming years, as India's economy continues to grow and its population continues to urbanize.

Quantitative Analysis

Michael Porter's 5 Forces Analysis

- **Threat of new entrants:** The oil and gas industry in India is highly regulated, making it difficult for new entrants to enter the market. This is a positive factor for existing companies in the industry.
- **Threat of substitutes:** There are few viable substitutes for oil and gas in the transportation sector, which is the largest consumer of oil and gas in India. This is another positive factor for existing companies in the industry.
- **Bargaining power of suppliers:** The bargaining power of suppliers is relatively low in the oil and gas industry in India. This is because there are many suppliers of oil and gas, and no single supplier has a significant market share.
- **Bargaining power of customers:** The bargaining power of customers is also relatively low in the oil and gas industry in India. This is because there are many customers for oil and gas, and no single customer has a significant market share.
- **Competitive rivalry:** The competitive rivalry in the oil and gas industry in India is moderate. There are a number of large, well-established companies in the industry, but there are also a number of smaller, independent companies.

Other Factors to Consider

- **Government protection:** The oil and gas industry in India is heavily regulated by the government. This provides some protection for existing companies in the industry from new entrants and competition.
- **High entry barriers:** The oil and gas industry in India has high entry barriers. This makes it difficult for new entrants to enter the market.

Top 10 Companies by Market Share

The top 10 companies by market share in the oil and gas industry in India are:

1. Indian Oil Corporation Limited
2. Bharat Petroleum Corporation Limited
3. Hindustan Petroleum Corporation Limited
4. Reliance Industries Limited
5. Essar Oil Limited
6. Nayara Energy Limited
7. Vedanta Limited
8. Oil and Natural Gas Corporation Limited
9. GAIL (India) Limited
10. Petronet LNG Limited

Investment Considerations

Based on the above analysis, the oil and gas industry in India is an attractive investment opportunity. The industry is in a growth stage, has a favorable competitive landscape, and has high entry barriers. The top 10 companies by market share are all well-established companies with strong financial performance.

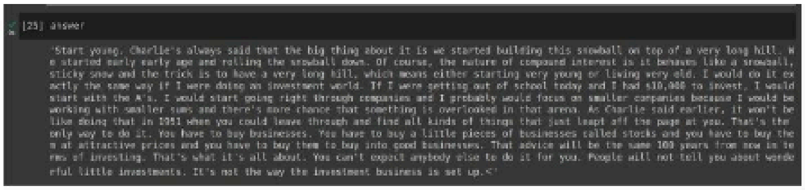
However, it is important to note that the oil and gas industry is cyclical. This means that the industry's performance is closely tied to the price of oil and gas. When the price of oil and gas is high, the industry performs well. When the price of oil and gas is low, the industry performs poorly.

Investors should be aware of the cyclical nature of the oil and gas industry before investing in the sector.

Figure 4. Sector Analysis

6.3 Financial Education from Warren Buffet

This image shows an answer given by Warren Buffet when asked ”Hello Mr. Buffett, what advice would you give to those who are just beginning their investing journey?”



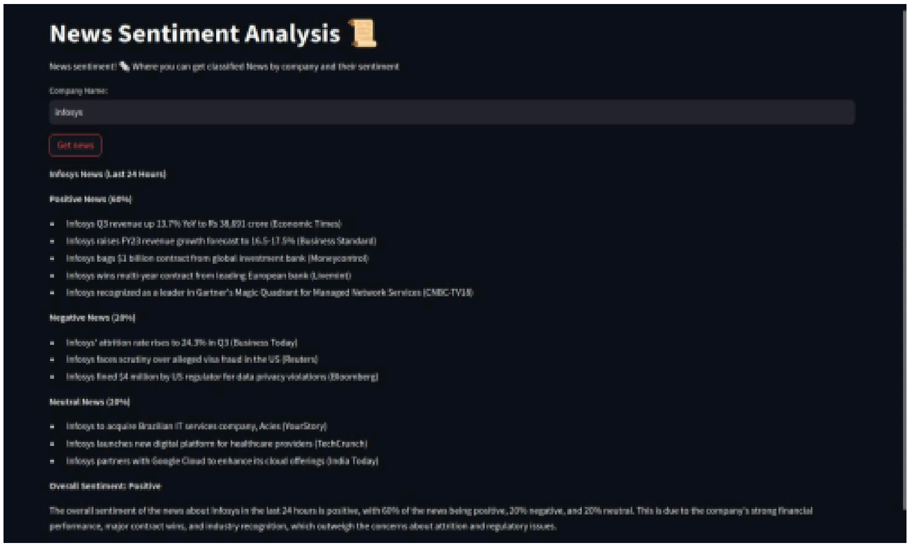
```
[25] answer

'Start young. Charlie's always said that the big thing about it is we started building this snowball on top of a very long hill. W
e started early early age and rolling the snowball down. Of course, the nature of compound interest is it behaves like a snowball,
sticky snow and the trick is to have a very long hill, which means either starting very young or living very old. I would do it ex
actly the same way if I were doing an investment world. If I were getting out of school today and I had $10,000 to invest, I would
start with the A's. I would start going right through companies and I probably would focus on smaller companies because I would be
working with smaller sums and there's more chance that something is overlooked in that arena. As Charlie said earlier, it won't be
like doing that in 1951 when you could leave through and find all kinds of things that just leapt off the page at you. That's the
only way to do it. You have to buy businesses. You have to buy a little pieces of businesses called stocks and you have to buy the
m at attractive prices and you have to buy them to buy into good businesses. That advice will be the same 100 years from now in te
rms of investing. That's what it's all about. You can't expect anybody else to do it for you. People will not tell you about wonde
rful little investments. It's not the way the investment business is set up.<'
```

Figure 5. Financial Advice by Buffet

6.4 Market News Sentiment Analysis

Here we have different news articles from different sources from Company INFOSYS. Our model classifies the news into three categories: Positive, Negative and Neutral and it provides the overall sentiment in the end and the reasons for the prediction.



News Sentiment Analysis 📜

News sentiment! 🗞 Where you can get classified News by company and their sentiment

Company Name:

infosys

Infosys News (Last 24 Hours)

Positive News (60%)

- Infosys Q3 revenue up 13.7% YoY to Rs 38,821 crore (Economic Times)
- Infosys raises FY23 revenue growth forecast to 16.5-17.5% (Business Standard)
- Infosys bags $1 billion contract from global investment bank (Moneycontrol)
- Infosys wins multi-year contract from leading European bank (Livemint)
- Infosys recognized as a leader in Gartner's Magic Quadrant for Managed Network Services (CNBC-TV18)

Negative News (20%)

- Infosys' attrition rate rises to 24.3% in Q3 (Business Today)
- Infosys faces scrutiny over alleged visa fraud in the US (Reuters)
- Infosys fined $4 million by US regulator for data privacy violations (Bloomberg)

Neutral News (20%)

- Infosys to acquire Brazilian IT services company, Acies (YourStory)
- Infosys launches new digital platform for healthcare providers (TechCrunch)
- Infosys partners with Google Cloud to enhance its cloud offerings (India Today)

Overall Sentiment: Positive

The overall sentiment of the news about Infosys in the last 24 hours is positive, with 60% of the news being positive, 20% negative, and 20% neutral. This is due to the company's strong financial performance, major contract wins, and industry recognition, which outweigh the concerns about attrition and regulatory issues.

Figure 6. News Sentiment Analysis

6.5 Pattern Recognition

We've created an application capable of recognising 6 different candle patterns namely 'Head and shoulders bottom', 'Head and shoulders top', 'M Head', 'StockLine', 'Triangle', 'W Bottom'. The results of this pattern recognition task were highly promising. The model achieved an accuracy of 97.2%, indicating that it was able to correctly identify the candlestick patterns with a high degree of reliability. This high level of accuracy underscores the effectiveness of the modified VGG-16 model in recognizing candlestick patterns, demonstrating its potential as a valuable tool in financial market analysis. The successful application of this model not only validates the approach of leveraging pre-trained models for specific tasks but also highlights the potential of machine learning in enhancing the capabilities of financial analysts and traders.

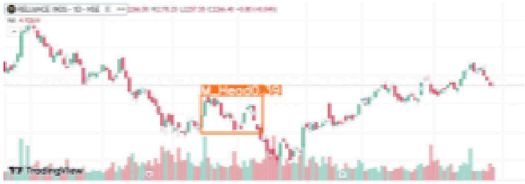


Figure 7. Recognising pattern 'M'

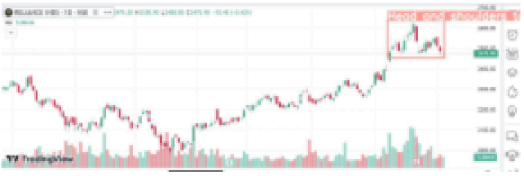


Figure 8. Recognising pattern 'Head and shoulders top'

6.6 Stock Recommendation

From the provided model accuracies for both Bag of Words and TF-IDF vectorization methods, it is evident that the Random Forest model consistently outperforms other models across both vectorization techniques. Notably, the Random Forest model achieves the highest accuracy of 76.3% with TF-IDF vectorization, surpassing the Decision Tree model's accuracy of 69.3% in the same scenario. This performance makes the Random Forest model the most effective choice for Stock Recommendation, particularly when using TF-IDF vectorization.

Model	Bag Of Words	TF-IDF
Gaussian Naive Bayes	49.4%	50.3%
Multinomial Naive Bayes	58.9%	60.1%
Bernoulli Naive Bayes	69.4%	60.1%
Random Forest	62.45%	76.3%
Decision Tree	71%	69.3%

Table 1
Accuracies of Stock Recommendation

7 Conclusion

This research project has explored the transformative potential of Large Language Models (LLMs) and Generative AI (GenAI) within the realm of finance and investment. By leveraging the power of natural language processing and advanced data analysis techniques, we have developed a comprehensive system that offers a range of functionalities designed to empower investors and traders.

Our project demonstrates the ability of LLMs to:

- Democratize financial education: The financial education chatbot, inspired by Warren Buffet's wisdom, simplifies complex financial concepts for beginners, making investment knowledge more accessible.
- Automate complex tasks: The automation of sector analysis and fundamental analysis streamlines the research process, saving time and effort for investors.
- Provide valuable insights: The stock recommendation system, sentiment analysis agent, and pattern recognition tool offer data-driven insights to support informed investment decisions.
- Enhance information retrieval: The integration of vector databases and Retrieval-Augmented Generation (RAG) workflow ensures efficient access to relevant financial information, improving the accuracy and reliability of responses.

While this project showcases the promising capabilities of LLMs and GenAI in finance, it is important to acknowledge the limitations and challenges that remain. These include the need for high-quality training data, the potential for bias, and the evolving regulatory landscape surrounding AI in finance.

Despite these challenges, our research underscores the transformative potential of LLMs and GenAI in the financial sector. As these technologies continue to evolve, we anticipate further advancements in their capabilities, leading to even more sophisticated and insightful financial analysis tools. This project serves as a stepping stone towards a future where AI plays a pivotal role in empowering individuals to make informed and confident investment decisions.

8 Future Work

Advanced Stock Recommendation System: We aim to enhance our stock recommendation system by integrating state-of-the-art natural language processing techniques, specifically Word2Vec, Open AI, and Cohere embeddings. This shift from traditional TF-IDF and Bag of Words models will enable more nuanced and context-aware analysis, significantly improving the accuracy and relevance of our recommendations.

Automated Fundamental Analysis: Building upon our sector analysis, we plan to automate the entire seven-step fundamental analysis process. This includes Management Analysis, P&L Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Ratio and Valuation Analysis. By automating these steps, we aim to provide investors with a comprehensive and up-to-date analysis of a company's financial health, thereby facilitating informed decision-making.

Website Development: To support our research and analysis, we are developing a sophisticated website. This platform will incorporate all the functionalities of our research, including the advanced stock recommendation system and the automated fundamental analysis. The website will be designed to be user-friendly, allowing users to easily navigate through the analysis and make informed investment decisions.

Bibliography

[1] Cohere Documentation. https://docs.cohere.com/, .

[2] Lang Chain Documentation. https://python.langchain.com/docs/get_started/introduction, .

[3] PyPDF Documentation. https://pypi.org/project/PyPDF2/, .

[4] Qdrant Documentation. https://qdrant.tech/documentation/, .

[5] Y Finance Documentation. https://pypi.org/project/yfinance/, .

[6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024.

[7] Yahoo Finance. https://finance.yahoo.com/.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[9] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. When large language models meet vector databases: A survey, 2024.

[10] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2018.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[12] IBEF Website. https://www.ibef.org/.

[13] Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, Ninghao Liu, and Tianming Liu. Revolutionizing finance with llms: An overview of applications and insights, 2024.