Financial Advanced Sentiment Tool (FAST) for Oil and Gas

Ilyas Kussanov, Ben Weideman, Danny Boone, and John Lattal
Rice University - COMP 549 (Summer 2024)
ik25@rice.edu, bw58@rice.edu, db101@rice.edu, and jl351@rice.edu

June 11, 2024

1 Introduction

Understanding the dynamics of stock prices is a significant challenge in financial market research (Shen & Shafiq, 2020). Traditionally, researchers have focused on predicting future stock movements based on historical price data (Fama, 1970; Gordon, 1959). However, the vast amount of data generated daily by the stock market makes it increasingly difficult to consider all current and historical information effectively (Li et al., 2017; Hariri et al., 2019).

Stock market analysis primarily utilizes two methodologies: fundamental and technical analysis. Fundamental analysis evaluates a company's intrinsic value by examining financial statements, economic indicators, and qualitative factors like management quality and competitive positioning. Tools for this analysis include financial reports, earnings statements, balance sheets, and macroeconomic data. This approach often incorporates news articles and competitor performance assessments to gauge a company's growth potential. In contrast, technical analysis focuses on statistical trends and market activity patterns, employing tools such as price charts, moving averages, and volume indicators to predict future price movements (Murphy, 1999).

Different analytical approaches suit varying time horizons. Some methods are effective for identifying long-term trends, while others excel in short-term predictions. The oil industry, characterized by highly dynamic stock prices influenced by numerous factors, exemplifies this complexity (Park & Ratti, 2008).

Both technical and fundamental analyses are extensively used across various industries, including the petroleum energy sector. Technical analysis aids in understanding short-term price movements through market data examination, while fundamental analysis offers insights into long-term trends by evaluating companies' financial health, economic conditions, and qualitative aspects.

Competitor performance and investor relations are critical in many industries, including oil and gas. Factors such as company goals, news sentiment, financial reports, and commodity price changes impact stock prices. Understanding how these elements influence stock performance is essential for making informed strategic decisions and maintaining a competitive advantage.

This capstone project aims to select around 20 industry-related competitors (e.g., Chevron, Exxon, Shell) and analyze how stock prices and/or total market capitalization change based on the aforementioned factors. The specific goals include:

1. Developing a News Sentiment Analysis Tool: Analyzing sentiment from news articles, financial reports, and extensive documents.
2. Determining Feature Importance to Predict Stock Price Movement: Identifying key factors that significantly influence stock price movements.
3. Explaining Stock Price Movements Using News Articles: Finding relevant documents and references explaining stock price movements.
4. Delivering a Comprehensive Analysis Report: Producing detailed reports summarizing the findings.

In publicly traded companies, leadership emphasizes competitive performance relative to industry peers. Success metrics often include financial results and stock performance compared to peers, highlighting the importance of understanding and leveraging the factors driving stock market behavior.

2 Related Work

2.1 Fundamental Analysis

The purpose of this section is to review existing literature and theories that have informed the development of tools and methodologies for analyzing stock price movements, particularly through news sentiment analysis

and feature importance determination. The theories of Efficient Market Hypothesis (EMH), Signaling Theory, Behavioral Finance, Market Microstructure Theory, and Fundamental Analysis provide a comprehensive framework for understanding the interplay between news sentiment and stock prices.

2.1.1 Efficient Market Hypothesis (EMH)

The EMH posits that stock prices fully reflect all available information, suggesting that public news, including sentiment from news articles and financial reports, should be quickly and accurately incorporated into stock prices (Fama, 1970). This theory supports the rationale for developing a news sentiment analysis tool, as it tests the market's efficiency in processing new information.

2.1.2 Signaling Theory

According to signaling theory, companies use public disclosures to convey private information about their future prospects to investors (Spence, 1973). Analyzing the sentiment of these disclosures helps to identify the signals that companies send to the market, providing insights into how different types of news influence stock prices.

2.1.3 Behavioral Finance

Behavioral finance incorporates psychological theories into financial decision making, explaining investor reactions to news sentiment through cognitive bias such as overreaction and herd behavior (Kahneman & Tversky, 1979). This perspective is crucial for understanding how news sentiment can lead to deviations from rational market behavior, affecting stock price movements.

2.1.4 Market Microstructure Theory

Market microstructure theory examines the processes of asset exchange under explicit trading rules, focusing on how information is disseminated and incorporated into prices (O'Hara, 1995). This theory aids in understanding the impact of news timing, frequency, and type on trading behaviors and price formation.

2.1.5 Fundamental Analysis

Fundamental analysis evaluates a company's intrinsic value by analyzing financial statements and economic conditions (Graham & Dodd, 1934; Penman, 2013). By incorporating sentiment analysis into fundamental analysis, this project aims to determine which financial indicators and qualitative factors significantly influence the stock prices.

These theoretical frameworks provide foundation for the methodologies and approaches employed in this study, ensuring comprehensive understanding of the relationship between news sentiment and stock price movements.

2.2 Sentiment Analysis in Financial Markets

Wankhade, M. et. al. in their survey on sentiment analysis methods highlights three main branches of sentiment analysis frameworks, which includes Lexicon Based Approach, Machine Learning Approach, and Hybrid Approach.

2.2.1 Lexicon Based Approach

According to Kiritchenko et al 2014, The lexicon-based approach relies on a predefined list of words (lexicons) with assigned sentiment scores. It does not require training data and is considered an unsupervised method. The sentiment score of a text is computed by aggregating the sentiment scores of individual words.

Yan-Yean et al. 2010 and Moreo et al 2012 in their works described advantages and disadvantages of Lexicon Based approach for Sentiment Analysis. The main advantage of this approach is no training data needed and unsupervised nature of the lexicon-based approach. Meanwhile the main consequence of this is words can have different sentiments in different contexts. Therefore this approach fails to capture nuanced meaning words based on their context. This limitation can affect the accuracy of reasoning of stock price movement.

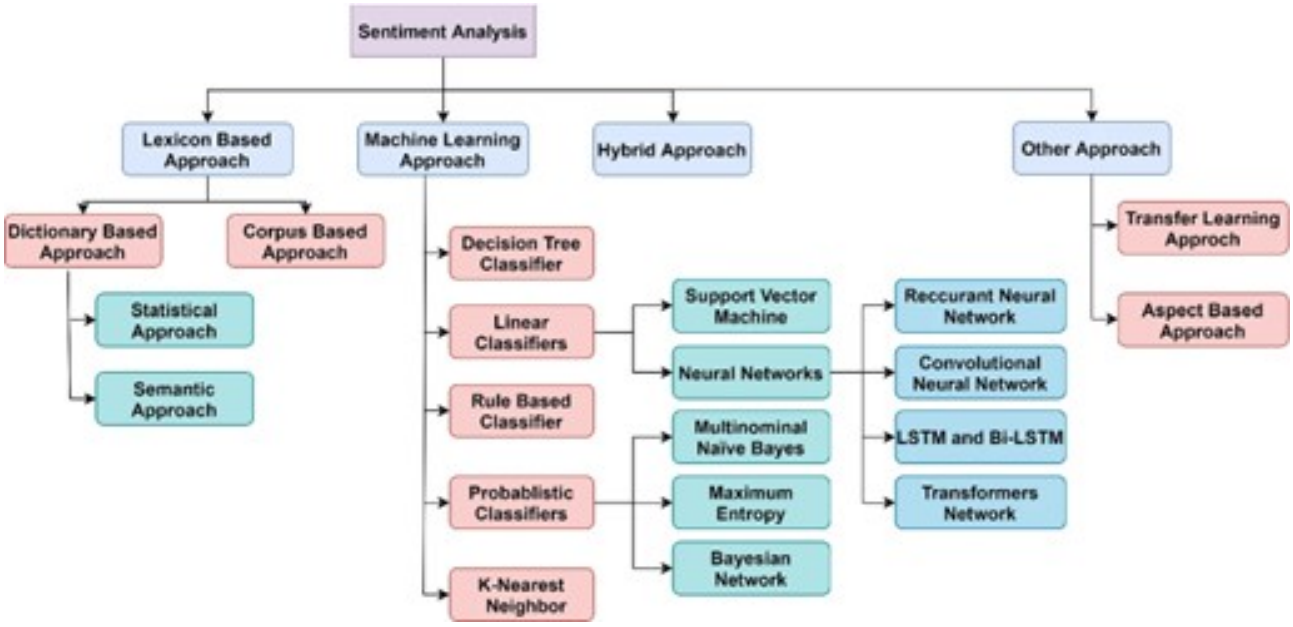

Figure 1: Approach sentiment analysis (Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022))

Table 1: Comparison of Different Classifiers

Method	**Pros**	**Cons**	**Reference**
Decision Tree Classifiers	Simple, interpretable, minimal preprocessing	Prone to overfitting, especially with noisy data	Almunirawi and Maghari (2016)
Linear Classifiers (SVM, Logistic Regression)	Robust, scalable, effective in high-dimensional spaces	Requires feature engineering, may not handle nonlinear relationships as effectively without kernels	Ahmad et al. (2017)
Rule-Based Classifiers	High accuracy in specific tasks, domain knowledge integration	Requires extensive domain expertise, less adaptable to new data	Zhang et al. (2020)
Probabilistic Classifiers (Naive Bayes)	Performs well with large feature spaces, probabilistic outputs	Strong independence assumptions, may not capture complex relationships	Alshamsi et al. (2020)
K-Nearest Neighbors (KNN)	Simple, intuitive	Computationally expensive for large datasets, sensitive to data noise and variance	Bozkurt et al. (2019)

2.2.2 Machine Learning Approach

Machine learning approach is more widely researched and diversified by using different methods. Based on Almunirawi and Maghari, 2016, Ahmad et al., 2017, Alshamsi et al., 2020, Bozkurt et al., 2019 we summarize key machine learning methods applied for news articles.

Overall mentioned traditional ML models often require significant feature engineering, are prone to overfitting, and may struggle with complex, contextual relationships within text. LLMs like BERT and GPT, built in deep learning architectures, address many of these issues by learning from vast amounts of data to understand context, semantics, and intricate patterns without extensive preprocessing or feature engineering.

For financial news articles, LLMs excel in capturing nuanced sentiments and adapting to diverse topics, making them more effective for comprehensive sentiment analysis tasks. LLMs are capable of understanding subtle linguistic nuances, detecting sarcasm, and identifying sentiment shifts within a single document, which are essential for accurately interpreting financial news.

2.3 Industry Specific Studies

The stock market prices of major oil and gas companies are influenced by various factors that are often reflected in financial statements, company disclosures, and news articles. Based on analyzed financial statements

and reports of Chevron, Shell, and Exxon, we selected the following categories found in highly reputable sources such as company release news, news articles in WSJ, NY Times, and government reports.

Financial performance, including revenue and earning reports, profitability metrics, cash flow statements, and dividend announcements, significantly impacts stock prices (Bagirov, M., & Mateus, C. (2019)). Misund, B. and McMillan, D. in their research paper examined how oil and gas companies' reserves growth affects their share price returns. Their study found that positive announcements generally led to an increase in stock prices within the study's timeframe (2007-2009). Announcements of new oil and gas discoveries can lead to stock price movements due to the potential future revenue from these discoveries (Misund, B., & McMillan, D. (2018)). Changes in discovered reserves can affect stock prices as they indicate the company's ability to sustain future production. Degiannakis, S., Filis, G., & Arora, V. highlight in their studies that the effect on companies' stock prices may vary based on the type of reserves. Gas shales reserve revisions generally show a minor impact on stock prices compared to crude oil reserves revision.

Production targets are a general metric for producing companies that are trading in a stock share market. Achieving or setting new production targets can positively affect stock prices.

Recent research has also explored the impact of sentiment analysis on oil companies' stock prices. For instance, a review by Oussalah and Zaidi demonstrated that sentiment analysis of financial news, social media, and company announcements could predict oil price movements with significant accuracy (Energies, 2019). This sentiment analysis includes categorizing news articles and social media posts into positive, negative, neutral, and mixed sentiments, providing insights into how public perception influences market trends (Xiao, Q., & Ihnaini, B. 2023).

Other categories such as New Energy Investments, Environmental Factors, Regulatory / Geopolitical Factors, and Divestments were not individually well studied by researchers. However, we included these categories in our focus, as we see their importance in stock price movements. Sentiment analysis has shown that public mood and news related to these factors can significantly impact stock prices, further emphasizing the need to consider these aspects in market analysis (Wu, B. et. al 2021; Xiao, Q. et. al. 2023, Zhao, L. et. al. 2019).

By integrating traditional financial metrics with modern sentiment analysis, we can achieve a more comprehensive understanding of the factors driving stock price movements in the oil and gas industry.

3 Data Science Pipeline

Our preliminary pipeline design is illustrated in Fig 2. The project is divided into six phases, beginning with data scraping and concluding with the development of an interactive stock performance dashboard. Our paper will delve into each of these aspects in more detail, but a high-level summary is shown below.

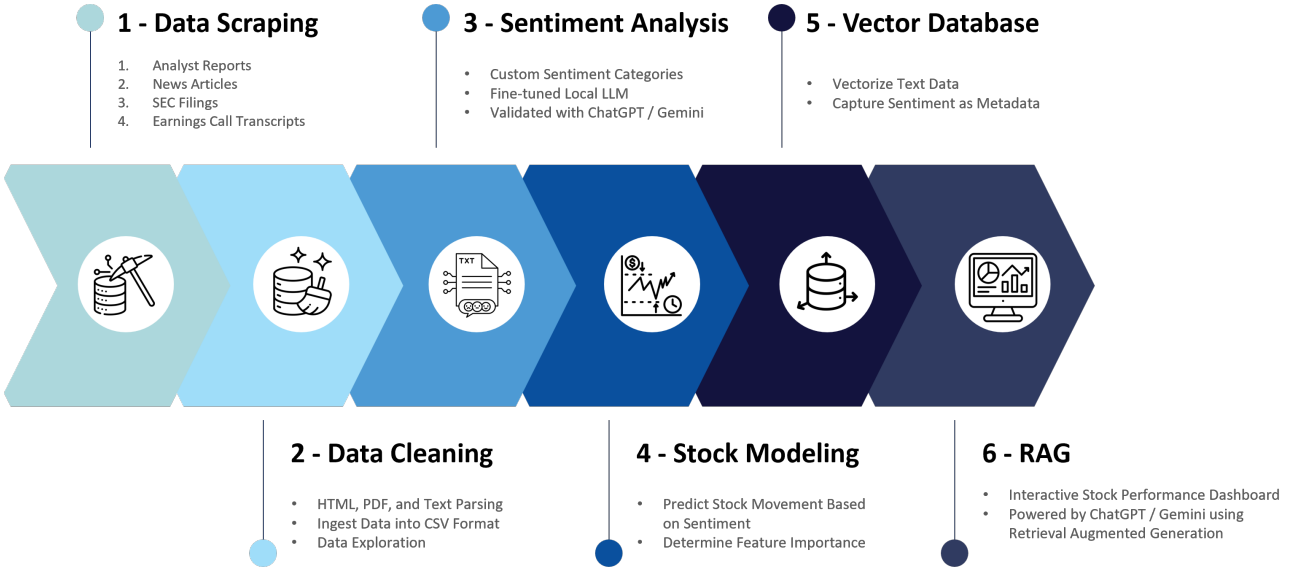

Figure 2: Data Science Pipeline

3.1 Data Scraping

Data collection is a critical aspect of our project. We will collect four categories of data from three different sources, and each category must be scraped, parsed, and cleaned individually due to their varied formats. The categories and source are listed below.

- News Articles from ProQuest
- SEC Filings from the SEC EDGAR database
- Investment Research Analyst reports from S&P NetAdvantage
- Earnings Call Transcripts from S&P NetAdvantage

3.2 Data Cleaning

After scraping, we will clean the data to ensure quality and consistency. This involves parsing text from HTML and PDF formats, and ingesting the data into a structured CSV table. We will also conduct exploratory data analysis to understand the characteristics of each dataset.

3.3 Sentiment Analysis

Following data cleaning, we will conduct sentiment analysis on each document using custom-defined categories. For example, these categories may include sentiments related to financials, production, new energies, acquisitions, regulation, and geopolitical factors. This step also requires conversion of the sentiments into tabular format. This stage is particularly challenging and likely to evolve. We will discuss the challenges in more detail later in this paper.

3.4 Stock Modeling

Using the processed sentiment data, we will model stock movements to identify feature importance. The ultimate goal of this phase is to determine which sentiment categories have the highest impact on stock price, providing valuable insights into market behavior.

3.5 Vector Database

To facilitate Retrieval-Augmented Generation (RAG) in our final dashboard, all documents will be vectorized and stored in a vector database. The text data will be converted into vector representations, and the sentiment analysis results will be captured as metadata. This setup will enhance the performance of our queries and enable more precise and insightful analysis.

3.6 Retrieval-Augmented Generation (RAG)

The final stage involves developing an interactive stock performance dashboard. This dashboard will be powered using retrieval augmented generation with ChatGPT or Google Gemini. It will provide a user-friendly interface for visualizing stock performance and summarize factors contributing to stock movement based on sentiment analysis.

4 Datasets

4.1 Dataset Access

Access to large amounts of pre-aggregated historical news articles and/or research/analyst reports are generally not accessible without a paid subscription. Even with this constraint, we were fortunate enough to have access to utilize the available resources through Rice University's Business Information Center, which offers access to various subject/industry specific databases to Rice students. Business Information Center

Within this group of databases, we identified 2 data sources which met the criteria to compile 3 types of data, with our 4th being directly from the SEC EDGAR database. The news articles were compiled from Proquest One Business, both the Analyst Reports and Earnings Call transcripts were compiled from Standard & Poor's Global NetAdvantage, and lastly the 10-K, 10-Q, 8-K filings were obtained from the SEC EDGAR database. Detailed explanations can be found in the following sections.

4.2 News Articles - ProQuest One Business

In an attempt to categorize news articles by stock ticker or company, the method in which we compiled and organized the news articles from the ProQuest database was done individually per company and narrowed down to the following list in Table 4.2:

Company	Ticker
Chevron	CVX
Exxon Mobil Corp	XOM
Shell PLC	SHEL
ConocoPhillips Co	COP
Phillips 66 Co Inc	PSX
Total SA	TTE
BP PLC	BP
Devon Energy Corp	DVN
EOG Resources Inc	EOG
Hess Corp	HES
Occidental Petroleum Corp	OXY
Marathon Petroleum Corp	MPC
Marathon Oil Corp	MRO
Pioneer Natural Resources Co	PXD
Concho Resources Inc	CXO
Valero Energy Corp	VLO
Equinor ASA	EQNR
PDC Energy	PDCE

This group of companies/tickers was utilized to compile and organize the additional datasets below. News articles were chosen via the respective company's profile within ProQuest One Business, then a selection of "Recent News From Major Publications" with a filter from 2019 to present. Publications include but are not limited to The Wall Street Journal, National Post, RTTNews, Edmonton Journal, Investors Business Daily, The Financial Express, Barron's, Business Insider, Financial Times, Oil & Gas News, and various other highly accredited publications.

4.3 SEC Filings - SEC EDGAR database

4.3.1 10-K Report

An SEC 10-K report is an annual filing that publicly traded companies in the United States are required to submit to the Securities and Exchange Commission (SEC). It provides a comprehensive overview of the company's financial performance, including audited financial statements, management's discussion and analysis (MD&A), market risks, and information about the company's operations, subsidiaries, and executive compensation. The 10-K is a critical document for investors, analysts, and regulators, offering detailed insights into the company's financial health and business activities over the past fiscal year.

4.3.2 10-Q Report

An SEC 10-Q report is a quarterly filing that publicly traded companies in the United States are required to submit to the Securities and Exchange Commission (SEC). It provides an update on the company's financial performance and includes unaudited financial statements, management's discussion and analysis (MD&A), and information about the company's operations for the quarter. Unlike the annual 10-K report, the 10-Q focuses on the most recent quarter and gives investors and analysts timely insights into the company's ongoing financial condition and operational results.

4.3.3 8-K Report

An SEC 8-K report is a filing that publicly traded companies in the United States must submit to the Securities and Exchange Commission (SEC) to announce major events that shareholders should know about. These events can include significant financial developments, changes in management or control, mergers and acquisitions, bankruptcy, or other important occurrences. The 8-K report provides timely disclosure of these events to ensure that all investors have access to important information that may affect the company's stock price or financial status.

4.3.4 11-K Report

An SEC 11-K report is a filing that provides annual information about employee stock purchase, savings, and similar plans. It is submitted by publicly traded companies in the United States to the Securities and Exchange

Commission (SEC). The report includes financial statements and other details about the plan's operations and financial condition, offering transparency to participants and regulators. The 11-K is specifically designed to ensure that these employee benefit plans are managed and reported in accordance with regulatory standards.

4.4 Earnings Call Transcripts - S&P NetAdvantage

Earnings calls are teleconferences or webcasts conducted by publicly traded companies to discuss their financial performance during a specific reporting period, typically a quarter or fiscal year. These calls are integral to the transparency and communication practices between the company and its stakeholders, including investors, analysts, and the media. These calls often are composed of a presentation of financial results, any new insights or business strategies, future outlooks, and often followed by a Question and Answer session. These transcripts are usually publicly available through the respective company website, however we are fortunate enough to be able to utilize the S&P NetAdvantage subscription through rice to be able to access all of these transcripts from various companies within one source.

These transcripts were compiled for the same list of publicly traded companies listed prior for the same time period (2019 - Current). These transcripts were compiled in PDF format from which text extraction was performed. The transcripts were saved in .csv format after being split between the presentation portion and the Question and Answer Portion. Additional steps moving forward may include the improvement of cleaning/processing data which was presented in tables, charts, or graphs, if needed.

4.5 Analyst Reports - S&P NetAdvantage (Investment Research)

Analyst reports are detailed documents created by financial analysts that provide insights and recommendations about publicly traded companies, industries, and financial markets. These reports are vital tools for investors, offering expert analysis and informed perspectives to aid in making investment decisions. Most importantly for this project use case, it provides a detailed sentiment analysis from a reputable source for a singular company. These reports usually consist of an Investment Thesis or summary giving a recommendation (buy, sell, hold), detailed company analysis including business overview and financial performance, industry/market analysis, and any strategic insights or risks.

These analyst reports were compiled in the same manner as the earnings calls, using the same list of publicly traded companies, and the same time period (2019 - Current). These reports were also compiled in PDF format from which text extraction was performed. The reports were then saved in .csv format for any additional cleaning or pre-processing. As with the Earnings Transcripts, additional steps moving forward may include the improvement of cleaning/processing data which was presented in tables, charts, or graphs, if needed.

4.6 Total Dataset Summary

The current data set consists of 10,274 text extracts from 4 source types and 18 different publicly traded companies over an approximately 5 year period. We believe that the sources chosen are reputable, reliable, and more importantly for the progress of this project, accessible. We also carefully selected companies which were deemed to be comparable from a competitive performance standpoint vs. Chevron. Following further analysis and implementation of our planned pipeline and final model output, it might be necessary to add additional data (new sources, more companies, etc.), this is not expected, however we are prepared if needed.

Below you can see the summary of our current dataset and the distribution of sources and companies.

Table 2: Data Sources and Formats

Source	**Initial Format**	**Count**
News Articles - ProQuest One Business	URL	3708
SEC 10K - SEC EDGAR	HTML	66
SEC 10Q - SEC EDGAR	HTML	176
SEC 8K - SEC EDGAR	HTML	734
SEC 11K - SEC EDGAR	HTML	60
Earnings Report Transcripts - S&P NetAdvantage	PDF	634
Analyst Reports - S&P NetAdvantage	PDF	4896

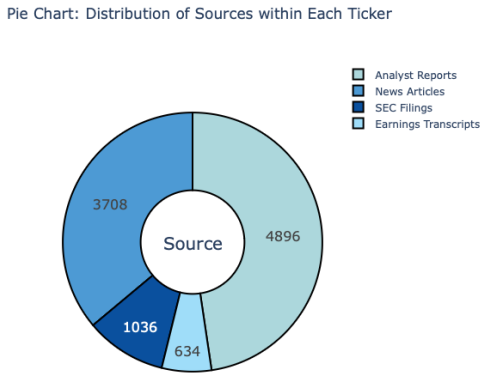

Figure 3: Data Source Counts

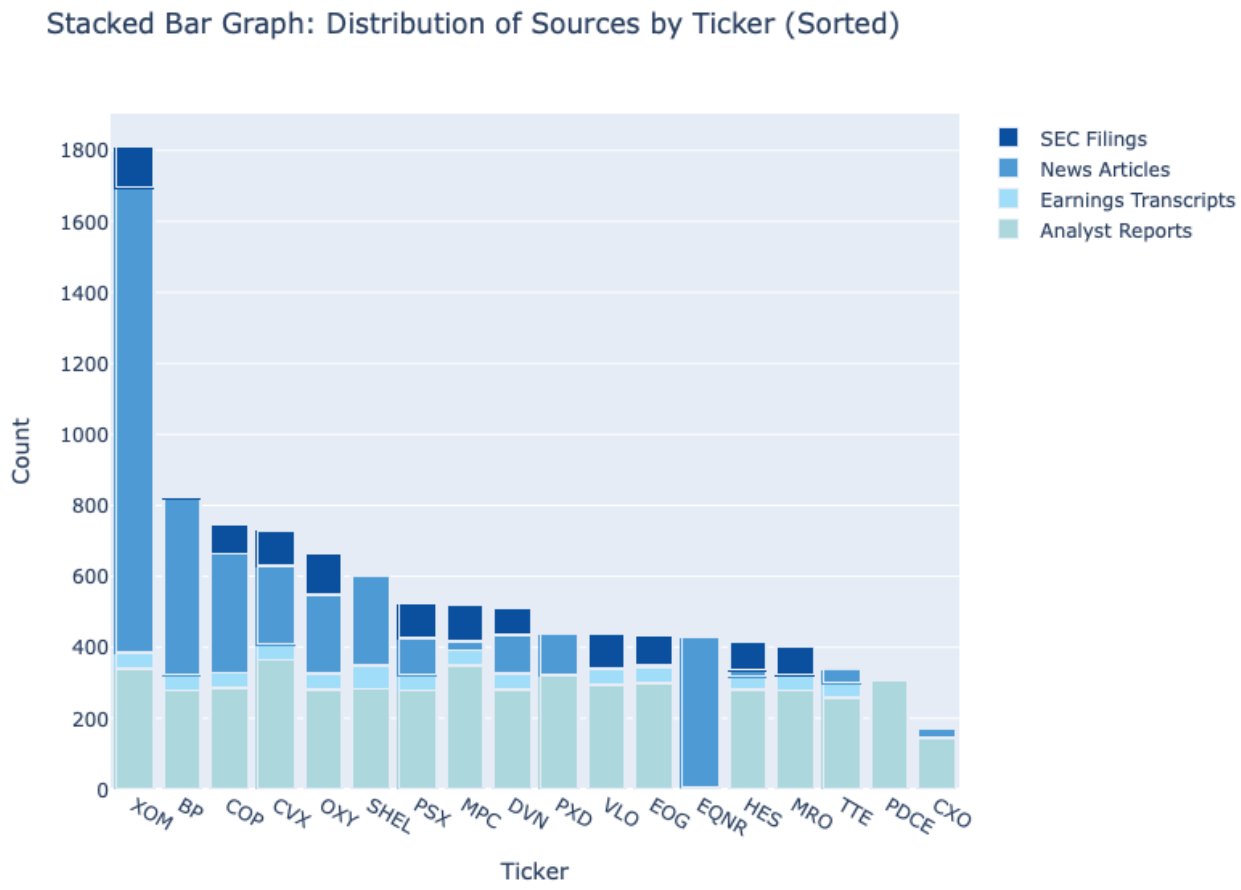

Figure 4: Data Source Company Distribution

5 Sentiment Analysis using LLMs

Sentiment analysis, also known as opinion mining, involves evaluating the sentiments, opinions, attitudes, and emotions expressed in text. It aims to determine whether the expressed sentiment is positive, negative, neutral, or more nuanced. This process is crucial for understanding the author's opinion in various contexts, such as articles, company press releases, and market research. The following section provides a detailed framework for using LLMs such as GPT, BERT, and LLAMA for sentiment analysis, with a specific focus on the oil & gas finance domain.

5.1 Overview of Large Language Models (LLMs)

Large Language Models (LLMs) are built on the transformer architecture (Fig X), which includes an encoder and a decoder. This architecture helps models analyze and understand text by focusing on the most relevant parts

5.1.1 Encoder

The self-attention mechanism (area 1b in fix X) in transformers weighs the importance of different words, capturing relationships and context within the text. For instance, let's consider the sentence: "Rice-Petro demonstrated high performance in a downstream sector despite lower performance in an upstream sector". When processing the word 'despite', the model needs to understand that it indicates a contrast between the performance in 'downstream' and 'upstream'.

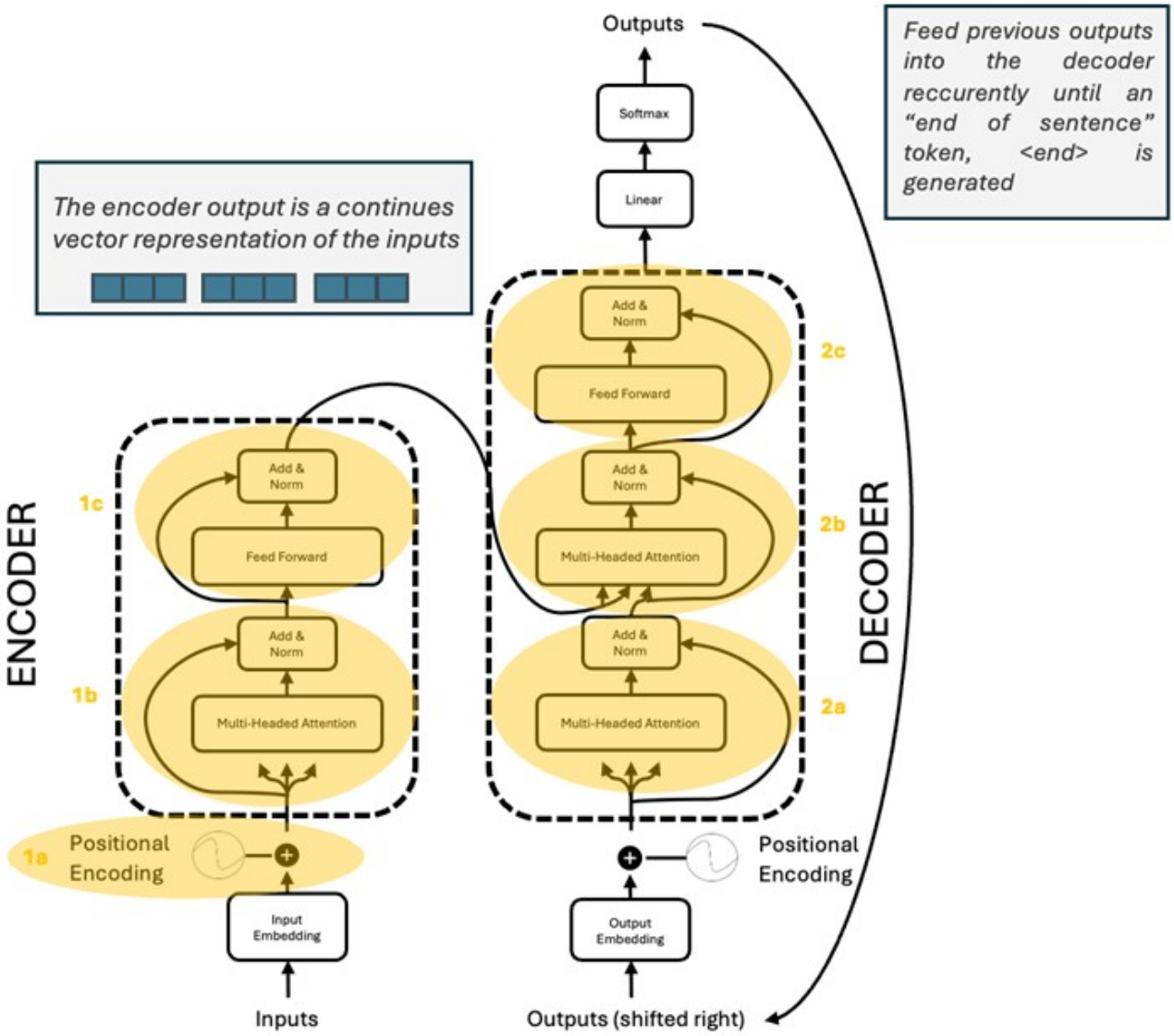

Figure 5: The Transformer Model Architecture (Modified from Vaswani, A et al. 2017)

In addition to self-attention, transformers use positional encoding (1a in Fig x) to retain the order of words in a sequence. Positional encoding injects information about the relative or absolute position of the tokens in the sequence, allowing the model to understand the sequence's structure. This is crucial because, unlike Recurrent Neural Networks, transformers process the entire sequence simultaneously and lack an inherent sense of order.

The self-attention scores help the model focus on relevant words. For 'downstream' and 'upstream', the model uses self-attention to link these terms with their respective performance indicators. Positional encoding ensures that the model recognizes the order of these terms and their relationship to the word 'despite', thus accurately capturing the intended contrast in performance.

The next step in the transformer architecture after self-attention is the Feed-Forward Neural Network (RNN) (area 1c in Fig X). The RNN introduces non-linearity through activation functions like ReLU, enabling the model to capture complex patterns and relationships in the data. It also transforms and refines the features extracted by self-attention, contributing to the model's depth and capacity.

At the end of the encoder, the output is a continuous vector representation of the inputs, which is then passed to the decoder for further processing or directly used for tasks such as classification, translation, or generation.

5.1.2 Decoder

The decoder in the transformer architecture is responsible for generating the output sequence. It works in tandem with the encoder to produce meaningful and coherent text. It starts with a masked self-attention mechanism (area 2a in Fig X). This mechanism is similar to the encoder's self-attention but with one key difference: it prevents attending to future tokens in the sequence. This ensures that predictions for a given position only depend on the known outputs before that position, maintaining the autoregressive property required for sequence generation.

The next layer is the encoder-decoder attention (area 2b in Fig X). This layer allows the decoder to focus on relevant parts of the input sequence (encoded by encoder). It helps the decoder align the generated output with the input context, ensuring that the output is coherent and contextually relevant. This is particularly useful in tasks like summarization, translation or crafted prompts, where the generated words must align with the input sentence's meaning and/or provided output template.

Similar to the encoder, the decoder also has a feed-forward neural network (area 2c in Fig X) that processes and refines the attended representations. This layer further enhances the model's ability to capture intricate patterns and relationships.

Finally, the decoder's output is passed through a linear layer followed by a softmax function to generate probabilities for the next word in the sequence. The word with the highest probability is selected as the output, and this process is repeated until the entire sequence is generated.

In summary, the transformer architecture, with its encoder and decoder components, efficiently processes and generates text by focusing on relevant parts of the input sequence and capturing complex patterns and relationships. The self-attention and feed-forward neural network mechanisms in both the encoder and decoder enable LLMs to perform a wide range of natural language processing tasks with high accuracy and coherence.

5.2 Utilizing Pre-trained Models and Prompt Engineering for Sentiment Analysis

To leverage the power of LLMs for sentiment analysis, we utilize pre-trained models like GPT-3, BERT, and LLAMA, which have already been trained on vast amounts of text data. These models can be fine-tuned on specific sentiment analysis tasks, allowing them to classify text into predefined sentiment categories.

5.2.1 Pre-trained Models

We chose to use pre-trained models like FinBERT and LLAMA without fine-tuning due to the benefits of quick deployment, cost efficiency, and leveraging their extensive training. This approach allows us to start our sentiment analysis tasks immediately and avoid the high computational costs associated with fine-tuning. While this means the models might be less tailored to the specific nuances of the oil & gas finance domain and may produce more generalized outputs, we plan to use prompt engineering to provide additional context and improve output consistency.

5.2.2 Prompt Engineering

Prompt engineering involves crafting specific prompts to guide the LLM in generating desired outputs. For sentiment analysis, prompts can be designed to ask the model directly about the sentiment of the text.

Example of the prompt:
"Analyze the following text and categorize the company's performance and related impacts across multiple predefined categories. Each category should be listed with a corresponding sentiment or result derived from the text. If a category is not mentioned or relevant based on the text content, mark it as 'N/A'. Ensure all categories are addressed for a comprehensive summary"

5.3 Objective

The main objective of this phase of our project is to conduct sentiment analysis for custom categories on every document in our dataset. We have established an initial set of categories and sentiment options, as

illustrated in Table 3. These categories have been utilized in our testing so far, and we will continue to refine them as the project advances.

Categories	**Sentiments**				
Financial Targets	Greatly Exceeded Target	Exceeded Target	Achieved Target	Below Target	Significantly below Target
Exploration / Discoveries	Major Discovery	Minor Discovery	Favorable Exploration Results	Unfavorable Exploration Results	Exploration Failure
Reserves	Significant Reserves Add	Minor Reserves Add	Stable Reserve Levels	Small Reserves Loss / Writeoff	Significant Reserves Depletions / Writeoff
Production Targets	Greatly Exceeded Target	Exceeded Target	Achieved Target	Below Target	Significantly below Target
New Energy Investments / Projects	Major Advancements in New Energy Initiatives	Minor Advancements in New Energy Initiatives	Setback in New Energies Project	New Energy Projects Abandoned or Failed	
Acquisitions and Mergers	Major Acquisition or Merger	Minor Acquisition or Merger	Delay in Acquisition or Merger	Cancelled Acquisition or Merger	
Divestments	Major Divestment	Small Divestment			
Public Sentiment	Very Positive	Positive	Neutral	Negative	Very Negative
Regulatory / Geopolitical Factors	Favorable change to Operations	Potential Large Disruption to Operations	Potential Small Disruption to Operations		
Environmental Factors	Very Positive	Positive	Neutral	Negative	Very Negative

Table 3: Categories and Sentiments

We intend to use a pre-trained local LLM to generate sentiment outputs, as there are no existing sentiment tools tailored to our specific needs. We aim to employ prompt engineering and potentially fine-tuning to achieve high-quality, consistent results. The output format is crucial because we will use regular expressions to convert the output data into a tabular format.

5.4 Challenges

So far, our success has been limited. While fine-tuning a model could yield better results, it would require manually curating a dataset. The challenges we face can be divided into two main categories: output consistency and output quality.

5.4.1 Output Consistency

Consistent output is crucial for transforming the data into a tabular format using regular expressions. State-of-the-art LLMs like ChatGPT and Gemini produce outputs that are relatively consistent, but smaller local LLMs are much more inconsistent. The primary driver behind using local LLM's is cost. Putting this much text through ChatGPT or Gemini would likely be cost prohibitive, especially as we will likely change and refine our categories over time.

5.4.2 Output Quality

Output quality is even more important, and we have encountered early challenges with the ProQuest news article dataset. These challenges are primarily due to the non-general nature of these articles. Most of the

articles are about the oil industry in general, or multiple companies. For example, an article about Chevron's recent delay in acquiring Hess due to interference by Exxon references three different companies. Such articles are likely tagged for each company, and the output sentiment should vary for each one. However, even state-of-the-art LLMs struggle to differentiate the outputs for each company. For example, ChatGPT mistakenly stated that Chevron had made a major discovery, when in reality, it was Hess. As we are seeing this behavior with ChatGPT, it will certainly be much harder to get high quality results with a local LLM.

5.5 Path Forward

As we complete cleaning the other datasets, we will begin testing on those as well. News articles focused on a single company do yield better results than the more general articles, so we anticipate higher quality output with the other data sources, as they should all be focused on a single company. However, given our early challenges, we will need to consider alternative approaches if we do not make significant progress in the near future.

6 Stock Predictions

The prediction of stock price movement is key to deriving value and insights from the sentiment analysis. We will take the sentiment analysis and create features to be fed into a stock price movement prediction model. The goal of this model is only to predict up or down movement of the stock price and not the magnitude of the movement. This is because the goal of this analysis is to know based on a certain mix of news sentiment, will that mean good or bad results for the stock price. We believe by having this categorical target it would provide more meaningful results with higher accuracy than trying to do regression analysis to predict that actual price.

Using the sentiment analysis we will create features for the stock prediction. For its analysis we have ignored time dependency or dependency of the price based on previous days. These features will be aggregate to the day of the news or analysis report and the company ticker. These features will include:

- Article Count
- Neutral Sentiment Article Count
- Positive Sentiment Article Count
- Negative Sentiment Article Count
- Avg Neutral Score
- Avg Positive Score
- Avg Negative Score
- Positive Article Ratio (# positive articles/article count)

The target variable will be the change in stock price for the next day from the date of the news or analysis report. This data was accessed using YahooFinance and was merged with the sentiment analysis based on date and company stock ticker. The stock price change was then converted to a binary measure of 1 for positive movement and 0 for negative movement. A threshold value for no movement was tested but was not found to be significant in the prediction.

The data was plotted up to get a sense of the trends. With the count of each article type projected as a bar and positive stock movement as a red dot at 1 and negative stock movement as a red dot at 0.

Several machine learning models were selected for performing the classification prediction of positive or negative stock price movement. The data was split into train and test as an 80/20 split. The models evaluated include the following:

- Random Forest Classifier from sklearn
- XGB Classifier from xgboost
- MLP (multi-layer perceptron) Classifier from sklearn
- Support Vector Machine (SVM) from sklearn

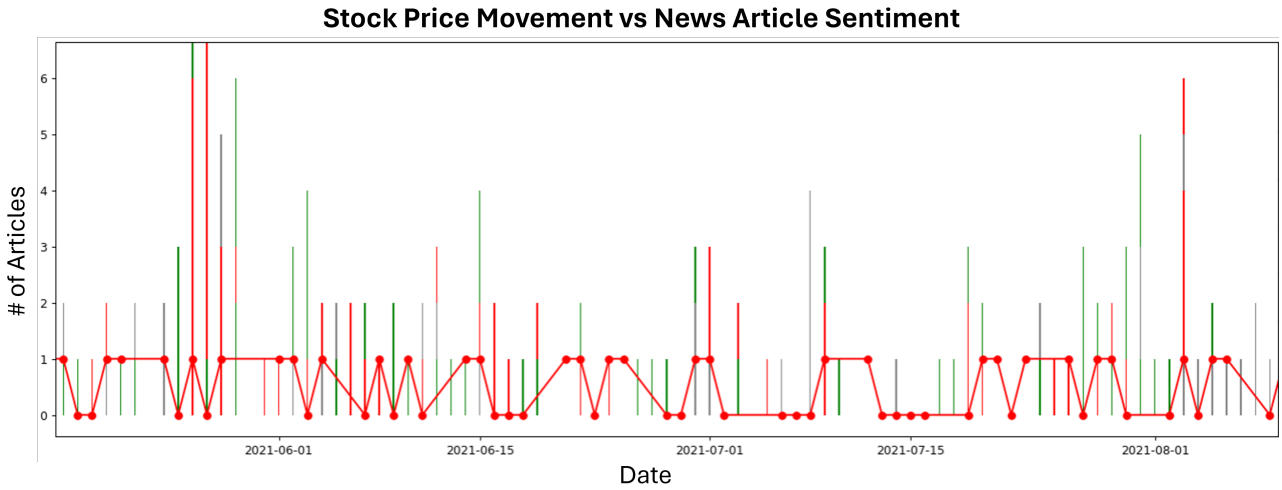

Figure 6: Example Visualization of Sentiment and Stock Movement

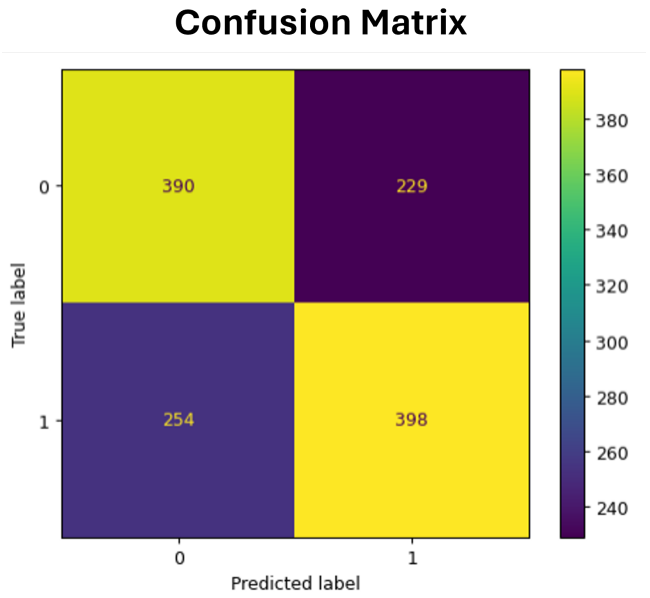

Figure 7: Confusion Matrix for Random Forest

Table 4: Random Forest Feature Importance

Positive Avg Score	Neutral Avg Score	Negative Avg Score	Article Count	Negative Count	Neutral Count	Positive Count	Neg Ratio	Neu Ratio	Pos Ratio
0.3258	0.3112	0.3100	0.0144	0.009153	0.00478	0.00645	0.0069	0.0061	0.0052

As part of the modeling the feature importance was analyzed. The results are presented below which show that the sentiment scoring and ratio of positive documents to total documents were the most significant in the prediction of stock price movement.

Evaluating these 3 most influential features we can filter the data set on a greater than threshold to get a sense for each value if that resulted in more positive or negative movements. It can be observed that as expected as a negative sentiment score increases the likelihood of negative movement increases. But for a positive sentiment score at 0.7 and above seems to have more negative movements. Neutral movements increase almost continuously the likelihood of positive movements as the threshold goes up. Suggesting that neutral may have a more telling trend on positive movements.

These results will be passed into a vector query to help with specifying article segments to be used to respond to the vector queries.

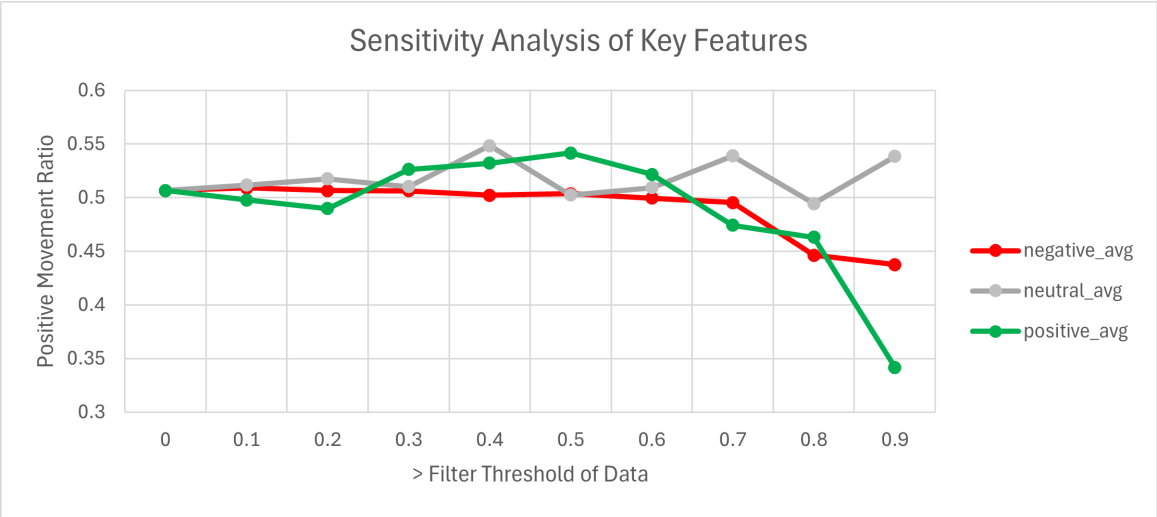

Figure 8: Sentiment Threshold Trends

7 Vector Database

The output from the sentiment analysis plus the article text will be used to support a Retrieval Augmented Generation (RAG). Utilizing a Pinecone vector database we will chunk the article into smaller sections that are more relevant to the provided query.

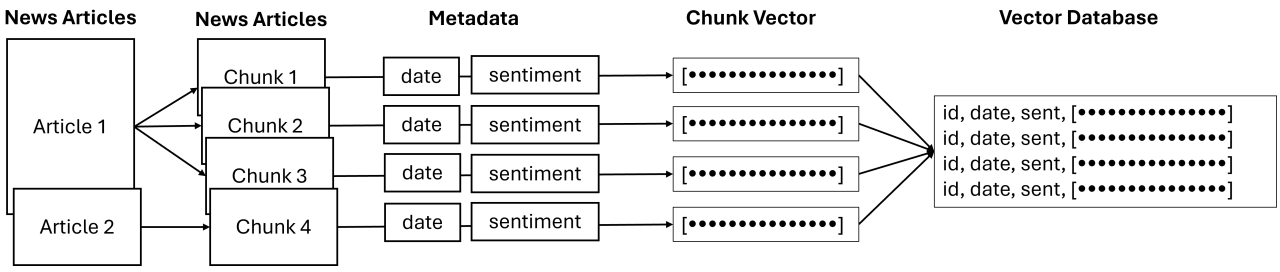

Figure 9: Diagram of Vector Database Set Up

A vector database is set up in the following procedure, pictured in Fig 9:

1. Chunk the article text
   - We used 500 tokens based on the BERT tokenizer
2. Embed the chunk text using an LLM embedding model
   - We used SentenceTransformer “all-mpnet-basev2”
3. Create a Vector database using a commercially available tool
   - We used Pinecone vector database [27]
4. Create an id value and associated meta data
   - Id is the index from the pandas dataframe and meta data is the sentiment analysis features
5. Load records into vector database using API
   - Pinecone has API to insert rows into the vector database

We can then query the vector database by writing a query in plain English and interpreting some of meta data. We will look at a specific date and the stock movement. Using the stock prediction model we know that if the stock moves up or down we have a predetermined sentiment score ranges we are looking for that would most likely explain the movement. We will pass the query embedded using the same LLM model and the associated meta data including date and expected sentiment scores into the vector database. Applying a cosine similarity,

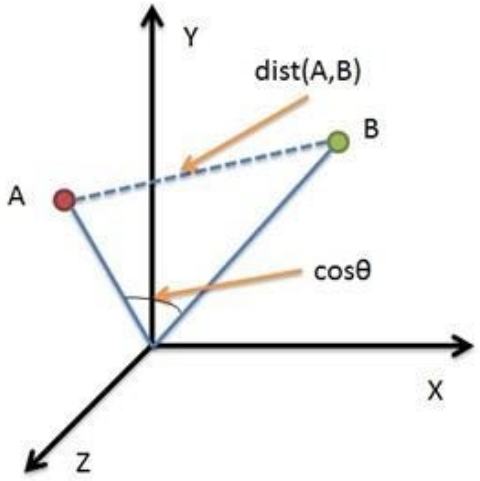

Figure 10: Cosine Similarity vs Euclidean Distance

which looks at the angle between vectors in the database finding the smallest angles between the query and the embedded chunks, returning the most relevant chunks.

These chunks and the plain English query are passed in to an LLM (we used Llama 3) which through a carefully engineered prompt provides the context and query for implementation of a Retrieval Augmented Generation (RAG).

References

[1] Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications.* Penguin.

[2] Park, J., & Ratti, R. A. (2008). Oil Price Shocks and Stock Markets in the U.S. and 13 European Countries. *Energy Economics, 30*(5), 2587-2608.

[3] Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of big Data, 7*, 1-33.

[4] Fama, E. F. (1970). Efficient capital markets. *Journal of finance, 25*(2), 383-417.

[5] Gordon, M. J. (1959). Dividends, earnings, and stock prices. *The review of economics and statistics*, 99-105.

[6] Li, Q., Chen, Y., Wang, J., Chen, Y., & Chen, H. (2017). Web media and stock markets: A survey and future directions from a big data perspective. *IEEE Transactions on Knowledge and Data Engineering, 30*(2), 381-399.

[7] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big data, 6*(1), 1-16.

[8] Spence, M. (1978). Job market signaling. In Uncertainty in economics (pp. 281-306). Academic Press.

[9] Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica, 47*, 278.

[10] O'hara, M. (1998). Market microstructure theory. John Wiley & Sons.

[11] Graham, B., Dodd, D. L. F., Cottle, S., & Tatham, C. (1951). *Security analysis: Principles and technique* (p. 851). New York: McGraw-Hill.

[12] Penman, S. H. (2013). Financial statement analysis and security valuation. McGraw-hill.

[13] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance, 62*(3), 1139-1168.

[14] Bagirov, M., & Mateus, C. (2019). Oil prices, stock markets and firm performance: Evidence from Europe. *International Review of Economics & Finance, 61*, 270-288.

[15] Misund, B., & McMillan, D. (2018). Exploration vs. acquisition of oil and gas reserves: Effect on stock returns. *Cogent Economics & Finance, 6*(1).

[16] Degiannakis, S., Filis, G., & Arora, V. (2017). Oil prices and stock markets. Washington, US: Energy Information Administration.

[17] Wu, B., Wang, L., Wang, S., & Zeng, Y. R. (2021). Forecasting the US oil markets based on social media information during the COVID-19 pandemic. *Energy, 226*, 120403.

[18] Xiao, Q., & Ihnaini, B. (2023). Stock trend prediction using sentiment analysis. *PeerJ Computer Science, 9*, e1293.

[19] Zhao, L. T., Zeng, G. R., Wang, W. J., & Zhang, Z. G. (2019). Forecasting oil price using web-based sentiment analysis. *Energies, 12*(22), 4291.

[20] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731-5780.

[21] Almunirawi, K. M., & Maghari, A. Y. (2016). A comparative study on serial decision tree classification algorithms in text mining. *International Journal of Intelligent Computing Research, 7*(4), 754-760.

[22] Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment analysis of tweets using SVM. *Int. J. Comput. Appl, 177*(5), 25-29.

[23] Alshamsi, A., Bayari, R., & Salloum, S. (2020). Sentiment analysis in English texts. *Advances in Science, Technology and Engineering Systems Journal, 5*(6).

[24] Bozkurt, F., Çoban, Ö., Baturalp Günay, F., & Yücel Altay, Ş. (2019). High performance twitter sentiment analysis using CUDA based distance kernel on GPUs. *Tehnički vjesnik, 26*(5), 1218-1227.

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*

[26] U.S. Securities and Exchange Commission. (n.d.). Retrieved June 3, 2024, from `https://www.investor.gov/introduction-investing/investing-basics/glossary`

[27] H. Yang, J. Guo, J. Qi, J. Xie, S. Zhang, S. Yang, N. Li, and M. Xu. A method for parsing and vectorization of semi-structured data used in retrieval augmented generation, 2024.