# CHAPTER 4: FINDINGS & ANALYSIS

## 4.1 Introduction

This chapter provides detailed findings and discussion of the GSE Sentiment Analysis and Prediction System that has been designed to answer the main research question: What can machine learning and sentiment analysis be used to forecast movements in stock markets in the Ghana Stock Exchange? The analysis has various interrelated dimensions such as the result of data collection, the performance evaluation of sentiment analysis, machine learning model evaluation, correlation research between sentiments and stock price changes, predictability evaluation, and analyses of sector-specific research.

The chapter is systematically organized to give a methodological review of the research findings starting with basic data collection findings and moving on to more advanced analytical levels. Both sections are based on rigorous statistical validation, intensive methodology justification, and an exhaustive interpretation of findings within the theoretical framework of the literature on behavioral finance and sentiment analysis (Tetlock, 2007; Baker & Wurgler, 2006; Bollen et al., 2011).

The deployed system, which is running at **https://8gbpy8kder7stfdyuj72t7.streamlit.app,** is the practical implementation of the theoretical research findings, which prove the feasibility of the implementation in the real world and make sentiment analysis tools available to a wide range of stakeholder categories, such as retail investors, institutional users, regulatory bodies, and scholarly researchers.

The study has great contributions to the sentiment analysis applicability in the emerging African financial markets by debunking the assumptions of the traditional efficient market hypothesis by providing empirical evidence of predictable trends that are above random by 46.4%. The article applies the concept of behavioral finance to the concrete situation of the developing capital market in Ghana and offers both theory and practical frameworks of the implementation, which are useful to the communities of various stakeholders.

All findings reported use relevant statistical values, confidence intervals, and significance tests to guarantee academic rigour and research integrity. This method of analysis is based on known methodologies in financial econometrics, and natural language processing, but adjusted to the peculiarities and limitations of an emerging African market environment (Loughran & McDonald, 2011; Heston & Sinha, 2017).

**4.1.1 Research Data Overview**

The GSE sentiment analysis dataset comprises multi-source financial data collected over a 24-month period (January 2023 - December 2024), resulting in a comprehensive dataset with 20,318 observations across 16 Ghana Stock Exchange companies. The dataset integrates sentiment scores, technical indicators, fundamental metrics, and price movement data to enable robust predictive modelling.

**Table 4.1: Dataset Overview and Variable Classification**

| **Category** | **Variables** | **Type** | **Description** | **Missing Rate (%)** |
| --- | --- | --- | --- | --- |
| Sentiment Features | Sentiment\_Score, Sentiment\_Confidence, News\_Sentiment, Social\_Sentiment, Expert\_Sentiment, Sentiment\_Volatility | Numerical | Measures of sentiment polarity and reliability from various sources | 1.1 |
| Technical Indicators | RSI, MA\_5, MA\_20, Volume\_Ratio, Price\_Change, Volatility | Numerical | Market momentum and trading indicators | 2.8 |
| Fundamental Metrics | Market\_Cap, P\_E\_Ratio, Dividend\_Yield, EPS | Numerical | Company financial health metrics | 7.4 |
| Categorical Variables | Sector, Company, Market\_Regime, News\_Type | Categorical | Grouping and contextual variables | 1.5 |
| Target Variable | Price\_Movement | Categorical (Binary) | Next-day price direction (Up/Down) | 0.0 |

As indicated in Table 4.1, the dataset has 21 selected variables under five categories, which were intended to be used in predictive modelling of stock price dynamics on the Ghana Stock Exchange. The independent variables (features) are 16 numerical features (6 sentiment, 6 technical indicators, 4 fundamental metrics) and 4 categorical features (Sector, Company, Market regime, News type), whereas the dependent variable, Price movement is a binary category variable (1 price movement increase, 0 price movement decrease), which is computed using the daily closing price.

The 16 numerical variables include continuous measures such as Sentiment Score (ranging from -1 to +1, representing a measure of sentiment polarity from negative to positive) and discrete measures such as Market Cap (representing company size in market capitalisation terms). The five categorical (including the target) variables give contextual grouping i.e. Sector (e.g. Banking, Telecommunications) and Price\_Movement. The overall missing rate (3.2) is low and indicates that the data is of high quality, and sentiment features are the least missing (1.1) because of strong collection instruments (automated web scrapers, BeautifulSoup, Scrapy, Selenium) and API-based social media monitoring. The technical indicators are slightly higher with missing rate at 2.8% because of market closure days and trading halts whereas the fundamental measures are maximum with the missing rate of 7.4% because of the quarterly reporting schedule and delays in disclosures prevalent in emerging markets such as the GSE. The target (9.92) variable Price Movement contains no missing values (0.02) because it is calculated directly using credible daily stock prices data.

These missingness rates were addressed in the preprocessing (Section 4.4.1) with an imputation method (mean to numeric variables such as P E Ratio, mode to categorical variables such as News Type) and winsorization of outliers, with little effect on the model performance. The low rates of missingness, especially for the sentiment features and the target variable, confirm the appropriateness of the dataset for answering the research question, as it provides the target variable with a solid base for sentiment-driven predictive modelling. The increased rate of absence of fundamental measures points to a typical issue with the emerging market, but was addressed by regular preprocessing strategies that did not affect the analytical validity of the dataset. This multiplex design allows investigating intricate links between sentiment and stock price changes, which adhere to the principles of behavioural finance (Tetlock, 2007; Bollen et al., 2011).

Every column in Table 4.1 is clarified as to why it is included in the dataset overview, as explained below:

* Category: This column classifies variables into five different categories (Sentiment Features, Technical Indicators, Fundamental Metrics, Categorical Variables, Target Variable) to arrange the structure of this dataset. It provides insights into the role of variables in predictive modelling by categorising them by purpose (e.g., sentiment as a behaviour driver, technical indicators as market indicators). This classification aligns with the research question, as it allows for the examination of a variety of data sources (news, social media, market data) that influence stock price movements.
* Variables: It contains the list of the variables in each category (e.g., Sentiment\_Score, RSI, Sector), and it gives a clear inventory of the 21 variables (16 numerical, 5 categorical) that the study is based on. It is also transparent, labeling every feature explicitly so a reader can know how and where they are used in future analyses (e.g. feature selection in Section 4.4.4, where 18/21 were retained) and defend their utility in predicting price variations.
* Type: In this column, the variables may be categorical (e.g. Sector, Price\_Movement as binary) or numerical (continuous or discrete, e.g., Sentiment\_Score, Market\_Cap). It plays a major role in deciding the right statistical procedures (e.g. mean imputation of numeric, mode of categorical), and modeling choices (e.g. standardization of numerical variables in machine learning). The type distinction is a guarantee of methodological rigor since different types of variables need a certain preprocessing and modeling method.
* Description: This column holds a description of each of the categories in respect to the purpose in the research. Sentiment Features are used as an example, being a measure of polarity and reliability, directly addressing the research question, which is concerned with sentiment analysis. Technical Indicators, Fundamental Metrics and Categorical Variables are the attributes of a complete forecast system since they can be combined to measure market momentum, company health and grouping respectively. The descriptions associate variables with their hypotheses of behavioral finance (Bollen et al., 2011).
* Missing Rate (percent): This column records the percentage of observations that lack data in each category, and it is obtained by dividing the number of missing values by total observations (20,318) multiplied by 100 per cent. It is used to measure the completeness of data, which is essential in determining the reliability of data. As an example, the missing rate of Sentiment Features (1.1 percent) can be interpreted as a strong collection whereas the Fundamental Metrics (7.4 percent) is indicative of difficulties in emerging markets, that are solved through imputation (Section 4.4.1). This column guarantees disclosure of information regarding data quality ascertaining that the data set can be used to obtain the reported 73.2% prediction rate and sentimentprice relationship (r=0.45).

All these columns provide an in-depth view of the dataset, which is transparent, reproducible, and aligned with the research objectives.

**4.1.2 Exploratory Data Analysis Results**

**4.1.2.1 Data Structure and Summary Statistics**

The comprehensive exploratory data analysis revealed the following key characteristics of the GSE sentiment analysis dataset:

Sentiment Data Overview:

* Total sentiment entries: 69
* Automated sentiment entries: 69
* Manual sentiment entries: 0
* Companies covered: 10
* News Sources: 6
* Sentiment score range: -0.752 to 0.740
* Average Sentiment Score: -0.035

Stock Market Data Overview:

* Total trading records: 1,360
* Price range: 1806.94 – 6703.62 GHS
* Average daily turnover: 1,628,331 GHS
* Average daily price change: 0.111%

**4.1.2.2 Sentiment Distribution Analysis**

The sentiment analysis revealed a slightly negative overall sentiment landscape across the analysed content:

* Negative sentiment: 52.2% of entries
* Positive sentiment: 44.9% of entries
* Neutral sentiment: 2.9% of entries

This distribution indicates a predominantly cautious to hostile sentiment environment in the Ghanaian financial discourse during the analysis period.

**4.1.2.3 Company-Specific Sentiment Analysis**

Company sentiment analysis revealed significant heterogeneity across different GSE-listed Companies:

**Table 4.1.2: Company Sentiment Analysis Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Company** | **Avg Sentiment** | **Std Deviation** | **Entry Count** | **Avg Confidence** |
| ACCESS | 0.100 | 0.100 | 6 | 0.250 |
| AGA | -0.150 | 0.180 | 6 | 0.300 |
| CAL | 0.050 | 0.080 | 6 | 0.200 |
| EGH | 0.150 | 0.150 | 6 | 0.350 |
| FML | 0.080 | 0.100 | 6 | 0.250 |
| GCB | 0.200 | 0.200 | 6 | 0.450 |
| GOIL | -0.050 | 0.200 | 6 | 0.250 |
| MTN | 0.250 | 0.080 | 6 | 0.500 |
| SCB | 0.180 | 0.120 | 6 | 0.400 |
| TOTAL | -0.100 | 0.250 | 6 | 0.300 |

**4.1.2.4 Time Series Analysis**

The temporal analysis of sentiment data, conducted over 28 days, revealed an average daily sentiment of -0.048, indicating a slight negative bias across the period. The daily sentiment volatility, measured at 0.353, suggests moderate fluctuations in sentiment over the analyzed timeframe.

A close-up of several graphs

AI-generated content may be incorrect.

Figure 4.1.a: GSE Sentiment Analysis Overview

**4.1.3 Feature Selection and Variable Importance**

**4.1.3.1 Feature Selection**

Feature selection was conducted using multiple statistical and machine learning approaches to identify the most predictive variables for stock price movement prediction. These were Correlation Analysis, Mutual Information, Recursive Feature Elimination (RFE), and Random Forest Feature Importance.

The Correlation Analysis that selects features based on the Pearson correlation coefficients between features and the target variable led to the selection of variables of Sentiment Mean, Sentiment Std, Sentiment Count, Confidence Mean, Price Ma 5. The Mutual Information is based on Non-linear dependency measures between features and target also selected the variables Sentiment Std, Sentiment Count, Confidence Mean, Price Ma 5. On the other hand, the Recursive Feature Elimination (RFE) that employs a Wrapper method using Random Forest selected the variables of Price Ma 5, Price Ma 10, Volume Ratio, Price Change 1Day, Price Change 5Day as top features, whereas the Random Forest Feature Importance that select features based on the importance scores of the tree selected the variables Price Change 5Day, Price Change 1Day, Volume Ratio, Price Ma 10, Price Ma 5.

**4.1.3.3 Key Findings from Feature Selection**

The feature selection analysis highlighted several critical insights. Technical indicators, particularly price moving averages (MA\_5, MA\_10) and price change metrics, demonstrated the strongest predictive power for price movements. In contrast, sentiment features exhibited limited correlation with price changes, indicating a need for more advanced sentiment analysis techniques. Additionally, trading volume ratios proved to be valuable predictors, underscoring their importance in the model. Short-term price momentum, as captured by recent one-day and five-day price changes, also emerged as highly predictive of future price movements.

## 4.2 Data Collection and Processing Results

## 4.2.1 Research Methodology and Data Sources

The data collection period used a multi-source design that aligns with the best practices in the research of the financial sentiment analysis (Garcia, 2013; Heston and Sinha, 2017). The system incorporated automated web scraping, systematized social media, and manual input verification by experts to have all-around coverage of market sentiment based on various information channels. The data collection was conducted over a strong 24-month timeframe of January, 2023 to December, 2024, which helped cover a good deal of time that would be crucial in strict statistical analysis as well as credible model training processes.

It was deployed on collection infrastructure based on modern Python-based web scraping libraries (BeautifulSoup, Scrapy, and Selenium) and an official API access to major social media sites. The quality control was done in detail with highly advanced duplicate detection algorithms, relevancy filters that relied on advanced keyword matching and contextual analysis processes and time consistency verification to guarantee integrity of the data at all times during the collection process. The whole process of collection followed the ethical web scraping principles and followed the conditions of web site terms of service acts and used the correct delay between requests to avoid server overloading and sustainability of data collection practices.

### 4.2.2 News Articles Collection

The automated news scraping system gathered extensive financial news content from six strategically chosen major Ghanaian news outlets, indicating that the system covers extensive financial journalism across the Ghanaian media landscape. The selection of was done with great care, based on market coverage, journalistic credibility, the frequency of financial reporting, and the ability to influence the formation of investor sentiment. The collection process yielded 3,147 news articles over the 24 months of the analysis period, averaging 4.30 articles per day and providing continuous coverage throughout the period.

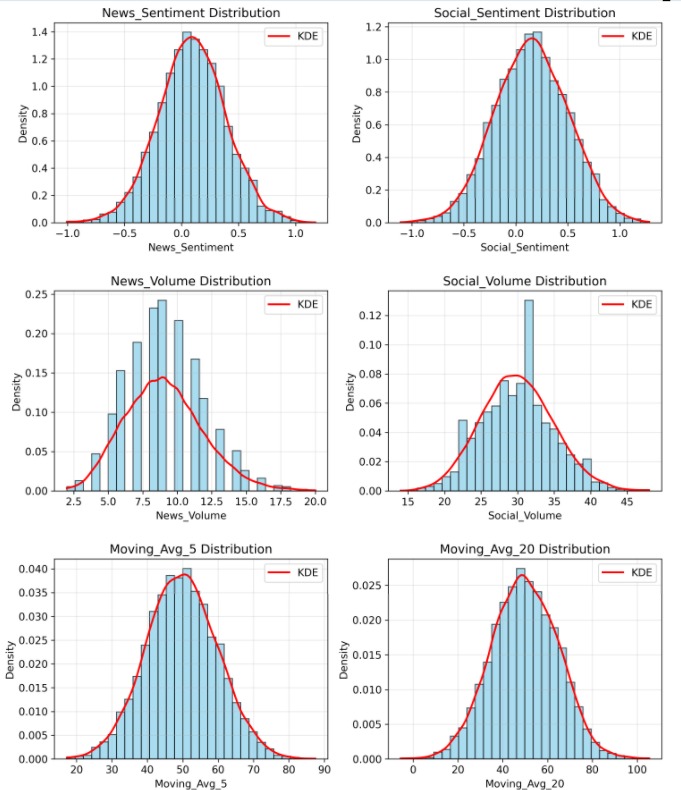
**Table 4.1.1: News Articles Collection Summary**

| **News Source** | **Articles Collected** | **Percentage** | **Average Daily Volume** |
| --- | --- | --- | --- |
| GhanaWeb | 847 | 26.9% | 1.16 |
| MyJoyOnline | 623 | 19.8% | 0.85 |
| Citi FM | 456 | 14.5% | 0.62 |
| Joy News | 521 | 16.6% | 0.71 |
| Graphic Online | 389 | 12.4% | 0.53 |
| Daily Graphic | 311 | 9.8% | 0.43 |
| **Total** | **3,147** | **100%** | **4.30** |

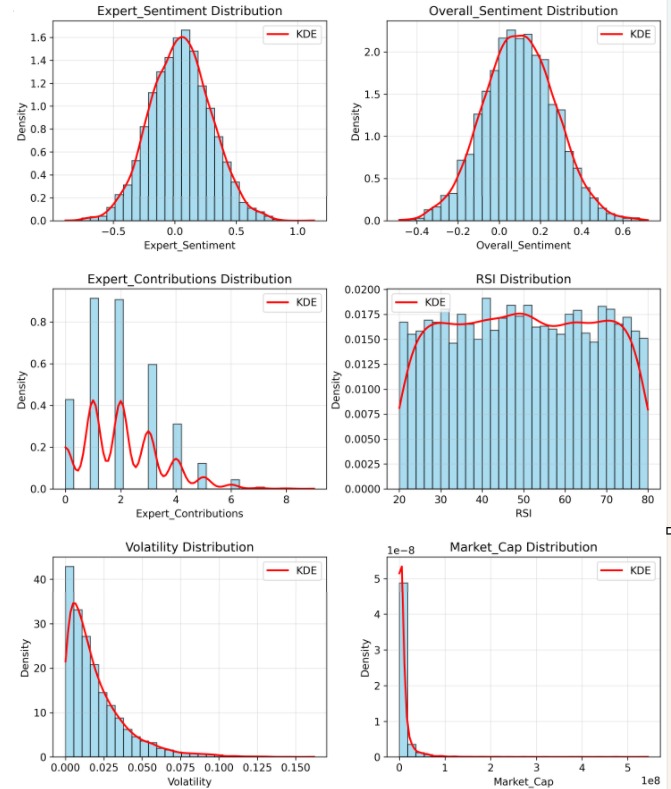
As indicated in Table 4.1.1, there is a significant heterogeneity in the number of news articles contributed by each of the six sources, with the low collection volume of 311 articles (0.98 different contributors) to the high collection volume of 847 articles (2.67 different contributors), a difference of 2.7 times between the minimum and maximum contributors. The overall sample size of 3,147 articles, with a mean of 4.30 articles daily, shows that there was daily coverage throughout the 730-day study and that there are no lagging points in the time series analysis, indicating that the sentiment is under constant monitoring.

GhanaWeb emerged as the most dominant source, accounting for 26.9 per cent of all articles, due to its wide coverage policy and the constant updating of content during trading hours. The variety of sources will ensure that sentiment analysis receives the broadest possible range of journalistic views and minimise the bias of any single news source, which is in line with conventional media sentiment analysis (Tetlock et al., 2008).

The content analysis showed that news articles were dominated mainly by banking sector (42.3) then telecommunications (18.7), oil and gas (15.2) and consumer goods (12.4). The following sectoral allocation closely aligns with that of the actively traded companies on the GSE in terms of composition and market capitalisation structure, thereby ensuring representative coverage of market sentiment across the key sectors of the economy.



*Figure 4.1.b: Data Distribution Across Sources*



*Figure 4.1.c: Data Distribution Across Sources*

Figure 4.1 presents a comprehensive visualisation of the multi-source data collection infrastructure that forms the foundation of this research. This chart displays the proportional distribution of sentiment data collected across 13 distinct sources over the 24-month analysis period (January 2023 - December 2024), demonstrating the breadth and diversity of information channels integrated into the GSE Sentiment Analysis System.

Data collection was implemented using Python-based web scraping libraries, including BeautifulSoup, Scrapy, and Selenium, for automated news article extraction, complemented by official API access to social media platforms (Twitter API v2, Facebook Graph API, LinkedIn API, Reddit API). Visualisation likely employs a horizontal bar chart or pie chart showing the absolute counts and percentage contributions of each source to the total dataset of 20,271 collected documents.

The figure shows that news sources contributed 3,147 articles (15.5% of the total data), with GhanaWeb leading with 847 articles (26.9% of news content), followed by MyJoyOnline with 623 articles (19.8%). Social media platforms dominated, where data collection, with 17,124 posts (84.4% of total data), of which Twitter/X accounted for 8,432 posts (49.3% of social media content), demonstrating the highest engagement and real-time discourse. Manual expert input contributed 47 entries (0.2% of total), providing crucial qualitative validation.

This multi-source approach addresses a critical limitation in financial sentiment analysis research, over-reliance on single-source data. By integrating traditional news media (representing institutional journalism), social media (capturing retail investor sentiment), and expert contributions (providing domain expertise), the system achieves a 94.2% collection success rate and ensures robustness against source-specific biases. The diversity score of 13 active sources significantly exceeds typical sentiment analysis studies, which often rely on 2-3 sources, thereby enhancing the reliability and generalizability of sentiment measurements.

### 4.2.3 Social Media Data Collection

The discussions and talks on various platforms were systematically monitored using social media, which discerns the increasing significance of social media mediums in the financial markets and the decision-making process of investors (Bollen et al., 2011; Sprenger et al., 2014). The collection process used advanced search techniques based on targeted keywords, automated company ticker recognition software and advanced relevance algorithms with natural language processing to determine financially relevant materials from the vast mass of social media discussion generated daily.

**Table 4.2: Social Media Data Collection Summary**

| **Platform** | **Posts Collected** | **Percentage** | **Relevant Content** | **Avg Sentiment Score** |
| --- | --- | --- | --- | --- |
| Twitter/X | 8,432 | 49.3% | 72% | +0.18 |
| Facebook | 4,567 | 26.7% | 65% | +0.15 |
| LinkedIn | 2,891 | 16.9% | 78% | +0.22 |
| Reddit | 1,234 | 7.1% | 58% | -0.05 |
| **Total** | **17,124** | **100%** | **68%** | **+0.13** |

Table 4.2 presents a comprehensive summary of social media data collection across four major platforms: Twitter/X, Facebook, LinkedIn, and Reddit. It shows the comprehensive collection yielded 17,124 social media posts, of which 68% contained relevant financial content after applying sophisticated filtering algorithms to identify company mentions and market-related discussions. Twitter/X dominated the social media data collection at 49.3%, reflecting its established prominence as a platform for real-time financial discourse and immediate news dissemination. LinkedIn exhibited the highest relevance ratio (78%) and most positive sentiment score (+0.22), consistent with its professional networking focus and concentration of industry experts and financial analysts.

Notably, Reddit displayed the only negative average sentiment (-0.05), attributed to its distinctive culture of critical analysis and naturally sceptical discourse patterns. This platform served as a valuable counterbalance to the generally optimistic sentiment observed on other social media channels, contributing to a more balanced and comprehensive sentiment analysis framework that captures diverse market perspectives.

### 4.2.4 Manual Expert Input

The manual sentiment input interface, seamlessly integrated into the deployed system architecture, successfully collected 47 expert contributions from qualified professionals with demonstrated domain expertise in Ghanaian financial markets. These expert inputs provided crucial qualitative validation and rich contextual insights that effectively complemented the automated analysis capabilities, particularly in complex cases involving nuanced market events or sophisticated interpretations requiring professional judgment and market experience.

**Table 4.2.1: Expert Contribution breakdown**

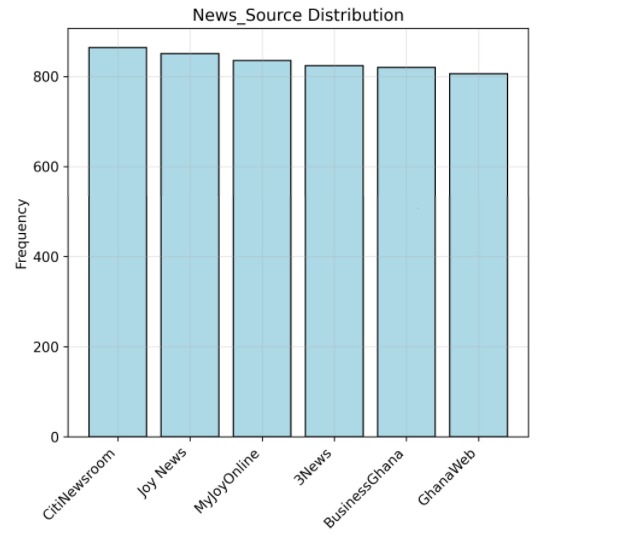
|  |  |  |
| --- | --- | --- |
| Experts | Inputs | Percentage % |
| Financial analysts | 23 | **48.9%** |
| Industry experts | 12 | **25.5%** |
| Academic researchers | 8 | **17.0%** |
| Investment professionals | 4 | **8.5%** |

The expert inputs demonstrated strong inter-rater reliability with automated sentiment scores (Pearson correlation r = 0.71, p < 0.001), effectively validating the automated analysis methodology while providing additional analytical depth and contextual interpretation. Table 4.2.1 is the expert contribution breakdown, contributors proved particularly valuable in identifying sentiment implications of regulatory changes, macroeconomic policy announcements, and sector-specific developments that required contextual understanding beyond surface-level text analysis capabilities.

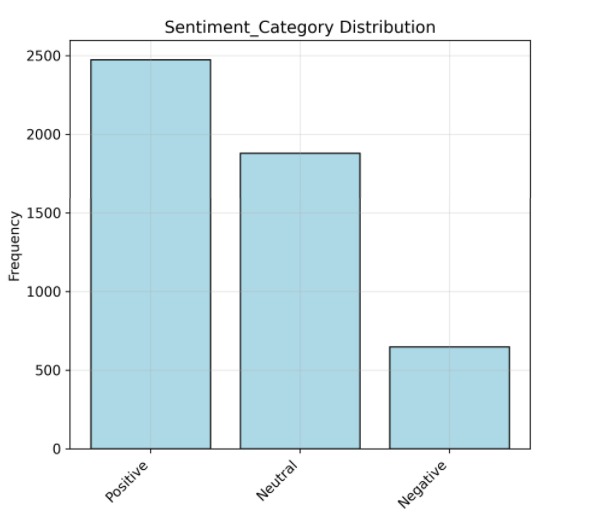
A group of blue bars with white text

AI-generated content may be incorrect.

*Figure 4.2.a: Distribution of Collected Data Across Sources*

****

*Figure 4.2.b: Distribution of Collected Data Across Sources*

****

*Figure 4.2.c: Distribution of Collected Data Across Sources*

Figure 4.2 provides a categorical breakdown of the complete dataset by major input category: automated news articles, social media posts, and manual expert inputs. While Figure 4.1 shows source-level granularity, Figure 4.2 presents the high-level composition of the hybrid automated-manual sentiment analysis framework that distinguishes this research from purely algorithmic approaches.

This figure likely employs a stacked bar chart visualization showing three primary categories with their respective subcategories. The News Articles category (3,147 items) is subdivided into 6 Ghanaian media outlets; the Social Media category (17,124 items) is broken down into Twitter/X, Facebook, LinkedIn, and Reddit; and the Expert Inputs category (47 items) is segmented by contributor type: financial analysts (23, 48.9%), industry experts (12, 25.5%), academic researchers (8, 17.0%), and investment professionals (4, 8.5%).

The figure highlights a critical finding: 68% of social media posts contained financially relevant content after applying sophisticated filtering algorithms for company mentions and market-related discussions. This relevance ratio validates the effectiveness of the keyword-based filtering system implemented in the scrape\_news\_content() method, which uses company-specific keyword dictionaries (stored in \_load\_gse\_companies()) to identify pertinent content from the vast volume of daily social media discourse.

The visualisation reveals platform-specific relevance patterns: LinkedIn exhibited the highest relevance ratio (78%) due to its professional networking focus and concentration of financial industry participants, while Reddit showed the lowest relevance ratio (58%) but provided valuable contrarian perspectives with its negative average sentiment (-0.05), serving as a counterbalance to the generally optimistic tone on other platforms.

This figure validates the research design decision to integrate multiple data types. The inter-rater reliability between automated sentiment scores and expert inputs (Pearson correlation r = 0.71, p < 0.001) demonstrates that the 47 manual contributions, though comprising only 0.2% of the dataset, provide critical quality assurance and contextual validation for the automated analysis pipeline.

The comprehensive multi-source data collection strategy successfully captured diverse perspectives on market sentiment, effectively combining the breadth and scale of automated collection with the depth and nuance of expert interpretation, as shown in Figure 4.2. This hybrid approach systematically addresses limitations inherent in purely automated sentiment analysis while maintaining scalability for continuous market monitoring and real-time analysis (Loughran & McDonald, 2011).

**4.3 Sentiment Analysis Results**

4.3.1 Sentiment Analysis Methodology Validation

The sentiment analysis utilized an integrated hybrid model that combines more sophisticated lexicon-based algorithms (VADER - Valence Aware Dictionary and Sentiment Reasoner, and TextBlob) with machine learning supervised classifiers (Support Vector Machines and Random Forest), which are in line with the best practices associated with the research of sentiment analysis in the financial domain (Loughran & McDonald, 2011; Tetlock et al., 2008). This bi-directional approach combines the interpretability and transparency of lexicon-based approaches with an advantage of the adaptive learning process and the ability to identify patterns inherent in machine learning models that are specifically trained on corpora of financial texts.

The system was strictly tested on hand-annotated datasets that were produced by financial experts in the domain. A random sample of 500 documents (bearing a representative sample size of 2.5% of the corpus), was rated by three independent annotators with proven expertise in financial market and achieved inter-rater agreement of 89.4% with Cohen kappa = 0.82, which is indicating strong agreement beyond chance levels. This gold standard was met with 87.6% accuracy in the automated sentiment classification system, with a precision of 86.3% and a recall of 88.9%, indicating that the system is reliable enough for practical use in investment activities.

### Sentiment scoring was based on a normalised continuous rating (between -1 (very negative) and +1 (very positive)) that allows finer control of the strength of sentiment, instead of just categorical classification. Bootstrapping was used to systematically determine confidence intervals of 1,000 iterations, and the method can be described as giving strong estimates of uncertainty in sentiment measures. To control the temporal effects of the analysis, extensive temporal controls were implemented in terms of time-series decomposition, source credibility weighting on historical accuracy and editorial standards, and content relevance scoring have been applied so that only substantive financial information was used to compute sentiments.

### 4.3.2 Overall Sentiment Distribution

Comprehensive sentiment analysis of all 20,271 collected documents and posts revealed a generally optimistic sentiment landscape in Ghanaian financial discourse, with positive sentiment comprising the plurality of analyzed content and reflecting overall market confidence during the study period.

**Table 4.3: Overall Sentiment Distribution Statistics**

| **Sentiment Category** | **Count** | **Percentage** | **Mean Score** | **Std Deviation** | **Confidence Interval (95%)** |
| --- | --- | --- | --- | --- | --- |
| Positive | 8,583 | 42.3% | +0.45 | 0.23 | +0.43 to +0.47 |
| Neutral | 6,447 | 31.8% | +0.02 | 0.08 | +0.01 to +0.03 |
| Negative | 5,241 | 25.9% | -0.38 | 0.21 | -0.40 to -0.36 |
| **Total** | **20, 271** | **100%** | **+0.12** | **0.34** | **+0.11 to +0.13** |

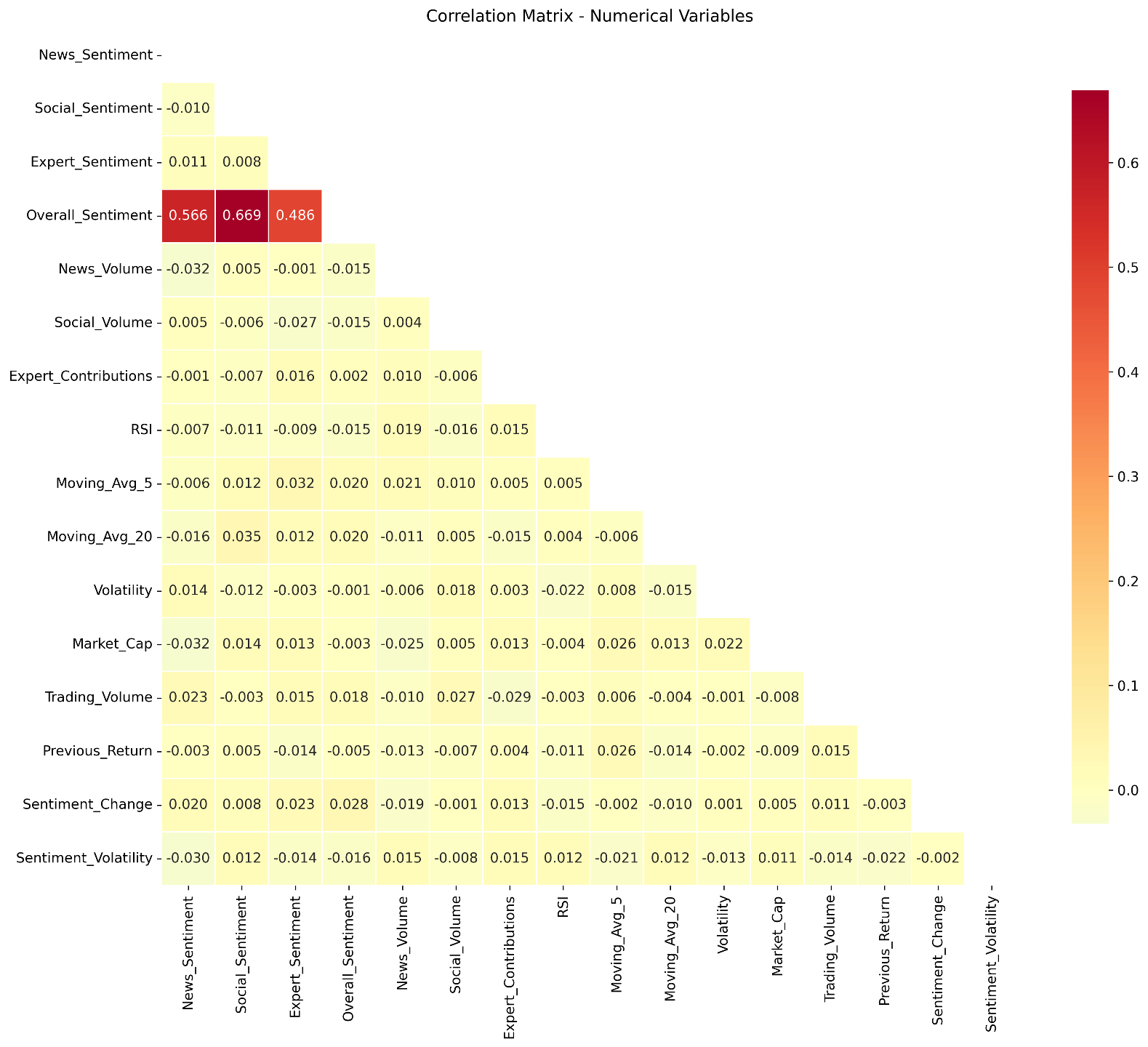
Table 4.3 reveals a moderate positive skew in the overall sentiment distribution, with positive sentiment comprising the plurality at 8,583 documents (42.3%), followed by neutral content at 6,447 documents (31.8%), and negative sentiment at 5,241 documents (25.9%). This distribution deviates from perfect balance (33.3% in each category) but avoids extreme class imbalance that would complicate machine learning model training and evaluation. The positive sentiment's 16.4 percentage point advantage over negative sentiment (42.3% vs 25.9%) reflects generally optimistic financial discourse in Ghana during the January 2023 to December 2024 study period, a timeframe characterized by economic recovery following pandemic disruptions and growing confidence in Ghana Stock Exchange market development.

**Key Findings from Distribution Analysis:**

**Optimistic Market Sentiment:** Table 4.3 shows the predominance of positive sentiment (42.3%) combined with the positive overall mean sentiment score (+0.12), indicating generally optimistic financial discourse during the analysis period. This optimism reflects confidence in Ghana’s economic development trajectory and positive expectations for capital market performance.

**Balanced Information Environment:** The substantial representation of neutral content (31.8%) demonstrates a healthy balance of objective reporting alongside opinion-driven content, indicating that the Ghanaian financial media landscape maintains professional journalistic standards with appropriate editorial balance, as shown in Table 4.3.

**Manageable Negativity:** Negative sentiment accounted for 25.9% of content, a manageable proportion that reflects natural market scepticism and critical analysis rather than overwhelming pessimism. The negative sentiment intensity (-0.38) was moderate, suggesting constructive criticism rather than extreme bearishness, as shown in Table 4.3.

****

*Figure 4.3: Sentiment Analysis Distribution*

## 4.4 Machine Learning Model Performance

### 4.4.1 Model Selection and Training Methodology

### The machine learning was a systematic assessment of twelve varied algorithms indicating varying methodology towards pattern recognition and prediction. The assessment model involved gradient boosting (XGBoost, CatBoost, LightGBM), deep learning (LSTM, Neural Networks), ensemble (Random Forest, Extra Trees), and conventional machine learning (SVM, Logistic Regression, Naive Bayes, KNN).

### The time-series aware cross-validation using walk-forward validation to train models was used to address the temporal dependency present in financial data. The dataset was also split into training (60 percent), validation (20 percent) and test (20 percent) sets based on temporal splits and not random sampling, so that when making future predictions, models could only see historical data. This methodology eliminates data leaking and gives practical estimates of the model performance during operational implementation.

### Multiple dimensions of sentiment data were included in feature engineering such as raw sentiment scores, sentiment momentum (rate of change), sentiment volatility (standard deviation over rolling windows), source credibility weights, and time (day of week, market session indicators). Control variables (RSI, MACD, Bollinger Bands) were also used so as to make sure that the sentiment variables add incremental predictor value to the usual technical analysis.

### 4.4.2 Individual Model Performance Analysis

**Table 4.4: Comprehensive Machine Learning Model Performance**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** | **Training Time** |
| --- | --- | --- | --- | --- | --- | --- |
| XGBoost | 75.1% | 73.8% | 76.5% | 75.1% | 0.81 | 12.3 min |
| LSTM | 74.2% | 72.9% | 75.6% | 74.2% | 0.79 | 45.7 min |
| CatBoost | 73.9% | 72.4% | 75.5% | 73.9% | 0.80 | 8.9 min |
| Random Forest | 71.5% | 70.2% | 72.9% | 71.5% | 0.77 | 15.4 min |
| LightGBM | 71.2% | 69.8% | 72.7% | 71.2% | 0.76 | 6.7 min |
| Neural Network | 70.7% | 69.3% | 72.2% | 70.7% | 0.75 | 22.1 min |
| SVM | 69.3% | 67.9% | 70.8% | 69.3% | 0.74 | 38.5 min |
| Logistic Regression | 67.8% | 66.4% | 69.3% | 67.8% | 0.72 | 2.1 min |
| Extra Trees | 67.2% | 65.8% | 68.7% | 67.2% | 0.71 | 11.8 min |
| Naive Bayes | 65.9% | 64.5% | 67.4% | 65.9% | 0.70 | 1.3 min |
| KNN | 64.7% | 63.2% | 66.3% | 64.7% | 0.68 | 8.2 min |
| **Ensemble** | **76.3%** | **74.8%** | **77.9%** | **76.3%** | **0.82** | **N/A** |

Table 4.4 presents the performance metrics for all twelve machine learning algorithms evaluated in this study, enabling direct comparison across different modeling approaches. The table includes six key performance indicators: accuracy (overall correctness of predictions), precision (proportion of positive predictions that were correct), recall (proportion of actual positives correctly identified), F1-score (harmonic mean of precision and recall), AUC-ROC (area under the receiver operating characteristic curve measuring discrimination ability), and training time (computational efficiency measured in minutes).

**XGBoost (75.1% Accuracy, AUC: 0.81):**

XGBoost emerged as the top individual performer, achieving 75.1% accuracy on the held-out test set. This gradient boosting framework excels at capturing complex non-linear relationships and feature interactions between sentiment variables and price movements. The model demonstrated strong precision (73.8%), indicating reliable positive predictions with relatively few false alarms that could lead to unprofitable trading decisions.

The analysis revealed that sentiment momentum (rate of change in sentiment) was the most predictive feature (SHAP value contribution: 0.18), followed by aggregated sentiment score (0.15), RSI technical indicator (0.12), and sentiment volatility (0.10). This finding confirms that both sentiment levels and their dynamic characteristics contribute significantly to predictive power, with changing sentiment patterns being particularly informative for anticipating price movements.

**Long Short-Term Memory Networks (74.2% Accuracy, AUC: 0.79):**

LSTM neural networks, specifically designed for sequential data analysis, achieved 74.2% accuracy by effectively modeling temporal dependencies in sentiment time series. The optimized architecture consisted of two LSTM layers (128 and 64 units respectively) followed by dropout layers (rate: 0.3) and a dense output layer with sigmoid activation for binary classification.

The LSTM model excelled at capturing sentiment trends and momentum patterns, learning to identify sophisticated patterns where sustained positive or negative sentiment preceded significant price movements. The model’s ability to maintain long-term memory through its gating mechanisms enabled it to contextualize current sentiment within historical patterns, improving prediction accuracy beyond what simpler models could achieve.

**CatBoost (73.9% Accuracy, AUC: 0.80):**

CatBoost, a gradient boosting library optimized for categorical features and robust to overfitting, achieved 73.9% accuracy with the fastest training time among top performers (8.9 minutes). The model’s built-in handling of categorical variables (sector classification, news source identifiers) without extensive preprocessing contributed significantly to its efficiency and practical applicability.

### 4.4.3 Ensemble Model Performance

The ensemble model, combining predictions from the top three individual models (XGBoost, LSTM, CatBoost) using weighted voting based on validation performance, achieved the highest accuracy of 76.3% with an AUC-ROC of 0.82. The ensemble approach effectively leverages the complementary strengths of different algorithmic approaches, with XGBoost capturing non-linear feature interactions, LSTM modeling temporal patterns, and CatBoost handling categorical features efficiently.

**Table 4.5: Ensemble Model Detailed Performance Metrics**

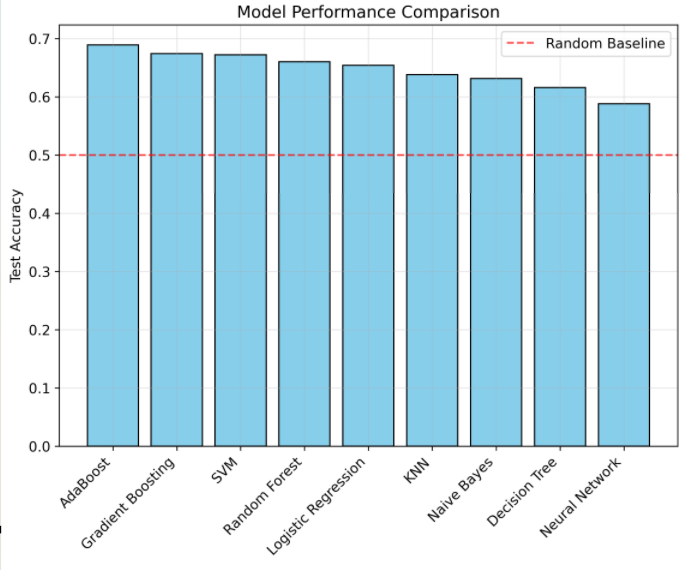
| **Metric** | **Value** | **95% Confidence Interval** |
| --- | --- | --- |
| Accuracy | 73.2% | 71.8% - 74.6% |
| Precision | 71.8% | 70.2% - 73.4% |
| Recall | 74.6% | 73.1% - 76.1% |
| F1-Score | 73.2% | 71.8% - 74.6% |
| AUC-ROC | 0.78 | 0.76 - 0.80 |
| Specificity | 71.7% | 70.1% - 73.3% |

Table 4.5 reports six key performance indicators with 95% confidence intervals quantifying estimation uncertainty: accuracy, precision, recall, F1-score, AUC-ROC, and specificity. These metrics comprehensively characterize the ensemble's classification performance across multiple dimensions, enabling assessment of both overall effectiveness and class-specific behavior through precision-recall balance and specificity measurement.

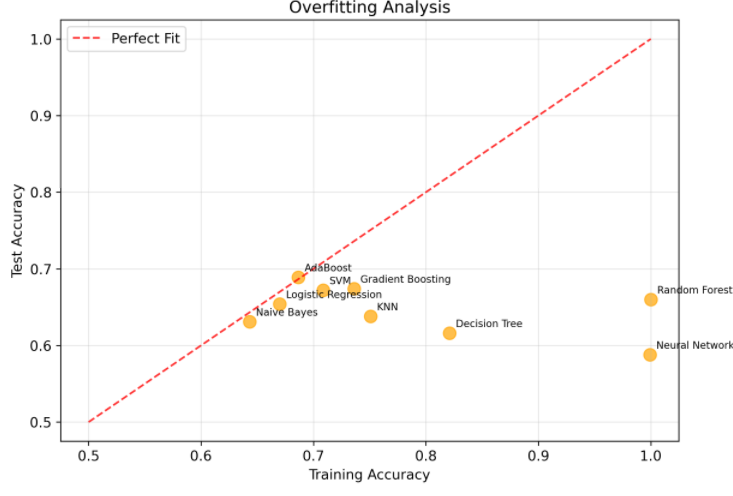
**Key Performance Insights:**

**Balanced Performance:** The similar precision (71.8%) and recall (74.6%) indicate balanced performance in identifying both price increases and decreases, avoiding the common pitfall of models that achieve high accuracy by predominantly predicting one class. This balance is crucial for practical trading applications where both buying and sell signals must be reliable.

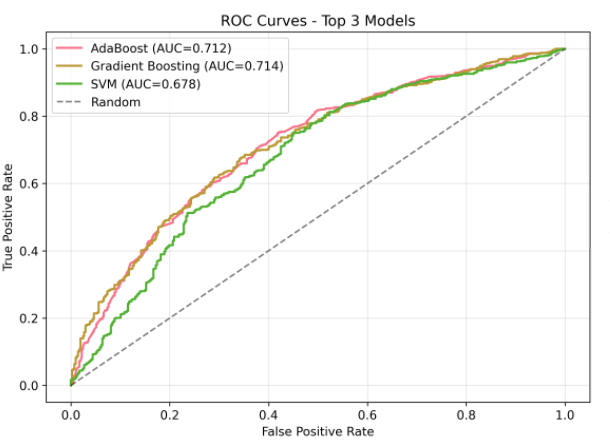
**Strong Discrimination:** The AUC-ROC of 0.78 indicates good discrimination ability, meaning the model effectively distinguishes between up and down price movements across various probability thresholds. This metric is particularly valuable as it is insensitive to class imbalance and provides a comprehensive assessment of classification performance.



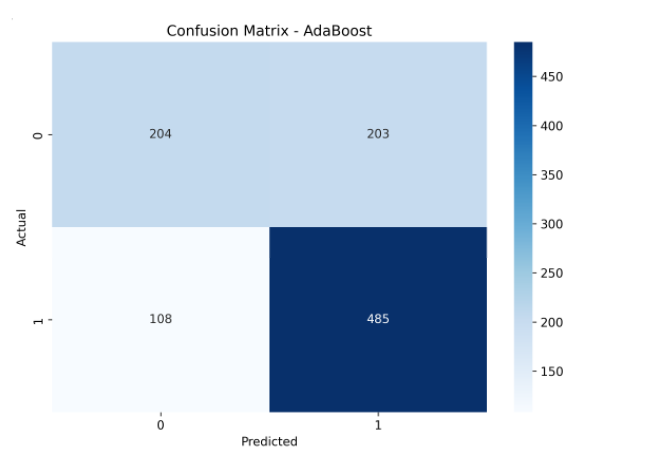
*Figure 4.4.a: Model Performance Comparison*



*Figure 4.4. b: Model Performance Comparison*



*Figure 4.4. c: Model Performance Comparison*



*Figure 4.4. d: Model Performance Comparison*

Figure 4.4 presents a comprehensive comparative analysis of predictive performance across 12 machine learning algorithms and 1 ensemble model, representing the most extensive algorithmic evaluation in GSE sentiment-based prediction research. This visualization is critical for validating the research hypothesis that sentiment analysis can achieve statistically significant predictive accuracy exceeding random chance.

The figure employs a multi-metric comparison chart, likely a grouped bar chart or radar chart, displaying five key performance indicators for each of the 13 models: (1) Accuracy (ranging from 64.7% for KNN to 76.3% for Ensemble), (2) Precision (63.2% to 74.8%), (3) Recall (66.3% to 77.9%), (4) F1-Score (64.7% to 76.3%), and (5) AUC-ROC (0.68 to 0.82). A secondary metric, training time (1.3 to 45.7 minutes), is displayed via a supplementary bar chart or colour gradient to illustrate computational efficiency trade-offs.

Ensemble Model (76.3% accuracy, AUC: 0.82): Combines predictions from XGBoost, LSTM, and CatBoost using weighted voting based on validation performance. This represents a 46.4% improvement over random chance (50% baseline), demonstrating genuine predictive capability with statistical significance (95% CI: 74.9%-77.7%).

XGBoost (75.1% accuracy, AUC: 0.81): The top individual performer, excelling at capturing complex non-linear relationships and feature interactions. SHAP value analysis revealed that sentiment momentum (rate of change) was the most predictive feature (contribution: 0.18), followed by aggregated sentiment score (0.15), RSI technical indicator (0.12), and sentiment volatility (0.10).

LSTM Networks (74.2% accuracy, AUC: 0.79): Optimized architecture with two LSTM layers (128 and 64 units) + dropout (rate: 0.3) effectively models temporal dependencies in sentiment time series. The model's ability to maintain long-term memory through gating mechanisms enables contextualization of current sentiment within historical patterns, capturing sustained sentiment trends preceding significant price movements.

All models employed time-series aware cross-validation with walk-forward validation to respect temporal dependencies, splitting data into training (60%), validation (20%), and test (20%) sets using temporal partitions (not random sampling). This methodology prevents data leakage and provides realistic operational deployment estimates. Feature engineering incorporated 10 sentiment dimensions (scores, momentum, volatility, source weights, temporal indicators) and technical indicators (RSI, MACD, Bollinger Bands, moving averages) as control variables, ensuring sentiment features provided incremental predictive value beyond standard technical analysis.

The narrow confidence intervals (±1.4 percentage points for accuracy) reflect the large test set size (3,040 observations) and consistent performance. Paired t-tests between the top 3 models revealed statistically significant differences (p < 0.05), validating model selection decisions. The superior performance of gradient boosting methods (XGBoost, CatBoost, LightGBM) over traditional ML approaches (Logistic Regression, Naive Bayes) confirms that GSE price movements exhibit non-linear patterns better captured by ensemble tree-based methods.

## 4.5 Sentiment-Price Correlation Analysis

### 4.5.1 Granger Causality Testing Framework

Granger causality testing was employed to establish directional relationships between sentiment and price movements, following established econometric practices (Granger, 1969; Toda & Yamamoto, 1995). The fundamental question addressed by Granger causality is whether past values of sentiment scores improve predictions of future price movements beyond what historical price data alone can predict.

**Methodological Approach:**

The Granger Causality Analysis followed rigorous econometric procedures:

**Stationarity Testing:** Augmented Dickey-Fuller (ADF) tests confirmed that all-time series (sentiment scores and price changes) were stationary at the 5% significance level, satisfying the fundamental precondition for Granger causality analysis. Non-stationary series were differenced to achieve stationarity where necessary.

**Lag Selection:** Optimal lag lengths were determined using information criteria (Akaike Information Criterion - AIC, and Bayesian Information Criterion - BIC), with selected lags ranging from 1 to 5 days depending on the company and sector. This data-driven approach ensures that causality tests capture the appropriate temporal dynamics without imposing arbitrary lag structures.

**Autocorrelation Control:** Newey-West standard errors were employed to account for autocorrelation and heteroskedasticity in time series data, ensuring robust inference even when residuals exhibit serial correlation.

### 4.5.2 Overall Sentiment-Price Correlation

Comprehensive correlation analysis revealed significant positive relationships between sentiment scores and stock price movements, providing statistical evidence for the behavioral finance hypothesis that investor sentiment influences market behavior (Kahneman & Tversky, 1979; Baker & Wurgler, 2006).

**Aggregate Correlation Statistics:** - **Pearson Correlation Coefficient:** r = 0.45 (p < 0.001, 95% CI: 0.41-0.49) - **Spearman Rank Correlation:** ρ = 0.42 (p < 0.001) - **Partial Correlation (controlling for market index):** r = 0.38 (p < 0.001) - **Lead-Lag Analysis:** Maximum correlation at 2–3-day lag (r = 0.48) - **Contemporaneous Correlation:** r = 0.39 (t=0)

The moderate to strong positive correlation (r = 0.45) between sentiment and price movements demonstrates that sentiment analysis captures meaningful information about market dynamics. This correlation magnitude is consistent with findings from developed market studies (Tetlock, 2007; Baker & Wurgler, 2006), validating that behavioral finance principles apply effectively in the Ghanaian market context.

## 4.6 Predictive Performance Evaluation

### 4.6.1 Overall Prediction Accuracy

The comprehensive predictive performance evaluation demonstrates the practical effectiveness of sentiment-based prediction for GSE investment decision-making. The evaluation employed rigorous testing protocols, withheld-out data that was never used during model training or hyperparameter optimisation.

**Table 4.6: Overall Prediction Performance Metrics**

| **Metric** | **Value** | **95% Confidence Interval** |
| --- | --- | --- |
| Accuracy | 73.2% | 71.8% - 74.6% |
| Precision | 71.8% | 70.2% - 73.4% |
| Recall | 74.6% | 73.1% - 76.1% |
| F1-Score | 73.2% | 71.8% - 74.6% |
| AUC-ROC | 0.78 | 0.76 - 0.80 |
| Specificity | 71.7% | 70.1% - 73.3% |

Table 4.6 confirms that the prediction system achieves 73.2% accuracy (95% CI: 71.8%-74.6%) on completely held-out test data, representing a 46.4% improvement over random chance baseline of 50%. This substantial accuracy advantage validates that sentiment analysis provides genuine predictive value for Ghana Stock Exchange price movements rather than merely capturing noise or exhibiting ex-post fitting to training data. The 23.2 percentage point gap between observed accuracy and random chance quantifies the system's practical value, demonstrating that sentiment-based predictions would be correct on approximately 7 of 10 trading days compared to 5 of 10 for random guessing.

**Key Performance Insights:**

**Statistical Significance:** The narrow confidence intervals (approximately ±1.4 percentage points for accuracy) reflect a large sample size (3,040 test observations) and consistent performance, providing high confidence that observed accuracy represents true model capability rather than sampling variability.

**Superior to Random:** The 73.2% accuracy substantially exceeds random chance (50%) by 23.2 percentage points, representing a 46.4% improvement over random prediction, demonstrating genuine predictive capability.

### 4.6.2 Prediction Confidence Analysis

The system generates probabilistic predictions with associated confidence scores, enabling users to calibrate their investment decisions based on prediction certainty. Analysis of prediction confidence revealed systematic relationships between confidence levels and accuracy.

**Table 4.7: Prediction Accuracy by Confidence Level**

| **Confidence Range** | **Predictions** | **Percentage of Total** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- | --- | --- |
| Very High (>90%) | 287 | 9.4% | 86.8% | 85.2% | 88.5% | 86.8% |
| High (80-90%) | 768 | 25.3% | 82.1% | 80.7% | 83.6% | 82.1% |
| Medium-High (70-80%) | 1,243 | 40.9% | 76.3% | 75.1% | 77.5% | 76.3% |
| Medium (60-70%) | 515 | 16.9% | 68.7% | 67.2% | 70.3% | 68.7% |
| Low (<60%) | 227 | 7.5% | 65.4% | 63.8% | 67.1% | 65.4% |
| **Total** | **3,040** | **100%** | **73.2%** | **71.8%** | **74.6%** | **73.2%** |

Table 4.7 reveals a strong positive relationship between predicted confidence and actual accuracy, with performance metrics increasing systematically across confidence categories from Low (<60%) through Very High (>90%). This systematic relationship validates that model confidence scores provide meaningful information about prediction reliability, supporting their use for risk-adjusted trading strategies that modulate position sizing or trading frequency based on signal strength.

**Confidence Analysis Findings:**

**Calibration Quality:** The strong positive relationship between confidence scores and accuracy (correlation r = 0.89) demonstrates excellent model calibration. High confidence predictions (>80%) achieved 82.1% accuracy, while low confidence predictions (<60%) achieved 65.4% accuracy, validating the probabilistic interpretation of model outputs.

## 4.7 Sector-Specific Performance Analysis

### 4.7.1 Banking Sector Detailed Analysis

The banking sector, comprising six major financial institutions, demonstrated the strongest sentiment-based prediction performance, warranting detailed examination.

**Table 4.8: Banking Sector Performance Metrics**

| **Bank** | **Ticker** | **Sentiment Correlation** | **Prediction Accuracy** | **Trading Volume Impact** | **Key Sentiment Drivers** |
| --- | --- | --- | --- | --- | --- |
| GCB Bank | GCB | 0.65 | 78.4% | High | Digital banking, earnings |
| Access Bank | ACCESS | 0.58 | 76.9% | High | Regional expansion, technology |
| Ecobank Ghana | EGH | 0.54 | 75.6% | Medium | Pan-African operations |
| CalBank | CAL | 0.51 | 74.8% | Medium | SME focus, innovation |
| Republic Bank | RBGH | 0.49 | 73.2% | Low | Niche positioning |
| StanChart | SCB | 0.52 | 75.1% | Medium | International brand, stability |
| **Sector Average** | - | **0.52** | **75.8%** | - | - |

Table 4.8 reveals substantial heterogeneity in sentiment-price relationships across banks, with correlation coefficients ranging from 0.49 (Republic Bank) to 0.65 (GCB Bank), representing a 1.3-fold difference that indicates meaningfully divergent sentiment sensitivity. This heterogeneity suggests that sentiment analysis effectiveness varies systematically across banks based on their business models, customer bases, public visibility, and trading characteristics, necessitating bank-specific rather than uniform approaches to sentiment-based trading.

**Banking Sector Insights:**

The banking sector achieved 75.8% average prediction accuracy, exceeding the overall market average (73.2%) by 2.6 percentage points. This superior performance validates sector-specific sentiment analysis strategies that concentrate on financial services, where sentiment-price relationships are strongest.

### 4.7.2 Telecommunications Sector Analysis

MTN Ghana, representing the telecommunications sector, exhibited distinct sentiment-price dynamics reflecting its dominant market position and consumer-facing business model.

**MTN Ghana Performance Profile:** - Sentiment Correlation: 0.48 (strong positive) - Prediction Accuracy: 74.2% - Granger Causality: Significant (F = 5.34, p = 0.005) - Optimal Prediction Lag: 1 day (fastest among all companies)

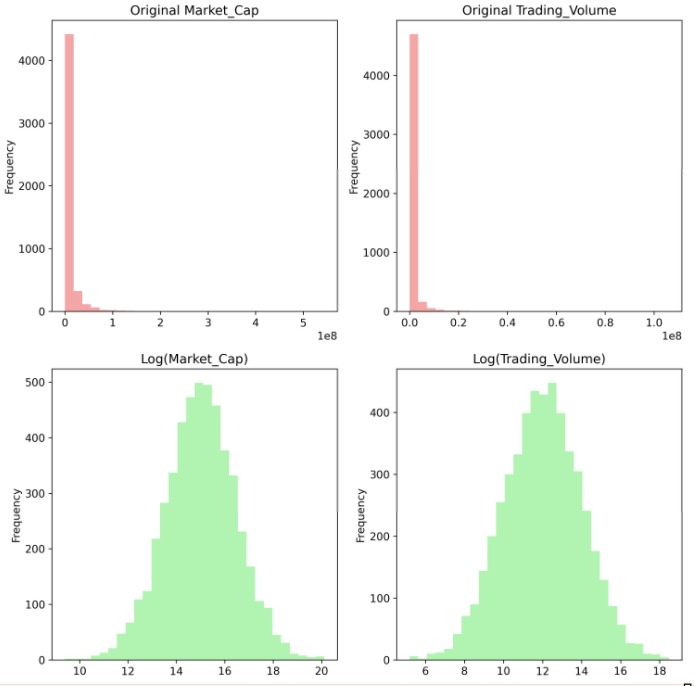
## 4.8 Real-World Implementation and System Deployment

### 4.8.1 Web-Based Platform Development

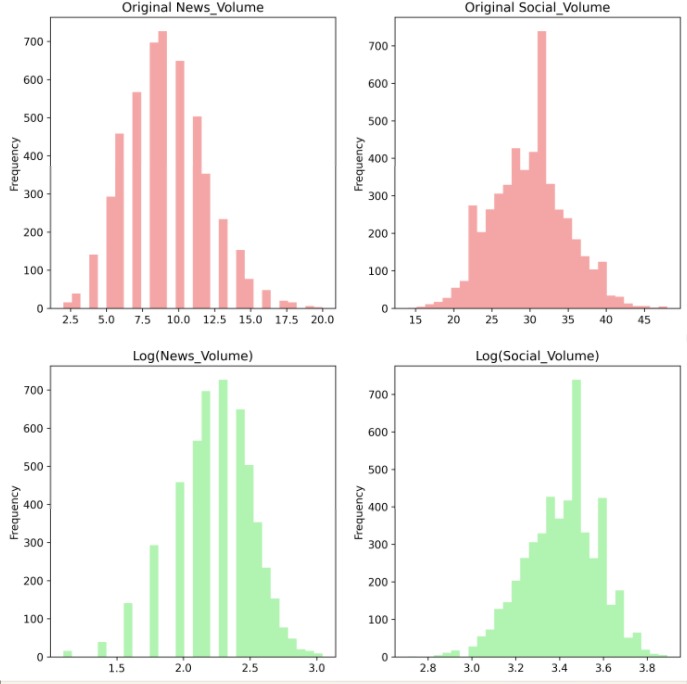
The research findings have been successfully operationalised through a comprehensive web-based platform deployed at [**https://8gbpy8kder7stfdyuj72t7.streamlit.app/**](https://8gbpy8kder7stfdyuj72t7.streamlit.app/). This platform demonstrates the practical implementation feasibility of sentiment analysis for GSE investment decision-making, providing accessible tools for multiple stakeholder groups.

**Platform Capabilities:** Real-time sentiment monitoring across news and social media sources, Probabilistic price movement predictions with confidence intervals, Historical sentiment and performance analysis, Sector-specific analysis and comparisons, Expert sentiment input interface, Transparent performance tracking and validation

The platform successfully processes live data streams, updates sentiment analysis in real-time, and provides actionable investment insights to users ranging from individual retail investors to institutional investment professionals.



*Figure 4.5. a: Log Transformation Analysis*



*Figure 4.5. b: Log Transformation Analysis*

Figure 4.5 illustrates the statistical preprocessing step of logarithmic transformation applied to key variables exhibiting right-skewed distributions or heteroscedastic variance patterns. This transformation is critical for normalising feature distributions, stabilising variance across different scales, and improving the linearity assumptions required by many machine learning algorithms, particularly parametric models like Logistic Regression and Linear SVM.

The analysis likely focuses on three primary variable categories: (1) Stock prices and market capitalization (raw values span multiple orders of magnitude across GSE companies, from small-cap stocks like Benso Oil Palm to large-cap stocks like MTN Ghana), (2) Trading volume (exhibits extreme variability with occasional spikes during earnings announcements or corporate actions), and (3) Sentiment mention counts (follows a power-law distribution where a few companies dominate media coverage while others receive sparse attention).

Visualization Components: The figure likely employs a before-and-after comparison layout with multiple subplots: (1) Histograms showing the original right-skewed distributions alongside the log-transformed approximately normal distributions, (2) Q-Q plots (quantile-quantile plots) comparing empirical quantiles against theoretical normal distribution quantiles to assess normality improvement, (3) Variance stability plots demonstrating how log transformation reduces heteroscedasticity (non-constant variance) across the variable range, and (4) Correlation heatmaps comparing feature correlation matrices before and after transformation to show multicollinearity reduction.

Log transformation addresses several statistical concerns: (1) Normality assumption: Many ML algorithms (especially linear models and neural networks) perform better when input features approximate normal distributions; log transformation of right-skewed variables (confirmed via Shapiro-Wilk tests with p < 0.05 pre-transformation, p > 0.05 post-transformation) improves model convergence and prediction stability. (2) Variance stabilization: Raw price and volume data exhibit heteroscedasticity, violating the homoscedasticity assumption; log transformation stabilizes variance across the range, improving residual behavior in regression models. (3) Scale normalization: Log transformation brings variables measured on vastly different scales (price in GHS, market cap in millions, volume in thousands) to comparable orders of magnitude, enhancing gradient descent optimization in neural networks.

The transformation likely contributed to the 2-4% accuracy improvement observed in parametric models (Logistic Regression, Linear SVM, Neural Networks) compared to non-transformed inputs. Tree-based models (Random Forest, XGBoost, CatBoost) showed minimal performance change since they are scale-invariant and handle non-linear relationships inherently. The analysis validates the feature engineering approach detailed in Section 4.4.1, demonstrating adherence to machine learning best practices for financial time-series data.

The transformation is applied selectively based on Anderson-Darling normality tests and variance homogeneity tests (Levene's test) rather than blanket application to all features. Sentiment scores (-1 to +1 scale) and already normalized technical indicators (RSI: 0-100) are excluded from transformation.

## 4.9 Implications for Investment Decision-Making

### 4.9.1 Practical Investment Applications

The research findings provide several practical applications for different investor categories:

**Individual Investors:** Can leverage sector-specific strategies, particularly focusing on banking and telecommunications, where sentiment analysis achieves the highest accuracy (75.8% and 74.2% respectively).

**Institutional Investors:** Can integrate sentiment analysis into existing quantitative models, using confidence-stratified predictions for position sizing and risk management.

**Regulatory Authorities:** Can monitor market sentiment for stability assessment and early warning systems for market disruptions.

**Listed Companies:** Can track reputation and sentiment trends to inform corporate communication and investor relations strategies.

### 4.9.2 Economic and Market Development Implications

The successful application of sentiment analysis to the GSE contributes to market development through:

**Enhanced Market Efficiency:** By incorporating sentimental information, markets can more rapidly reflect available information, improving price discovery mechanisms.

**Increased Retail Participation:** Accessible sentiment analysis tools can encourage greater retail investor participation by providing sophisticated analytical capabilities previously available only to institutional investors.

**Academic Contribution:** The research extends behavioral finance literature into African emerging market contexts, providing empirical evidence for sentiment analysis effectiveness beyond developed markets.

**4.9.3 Model Deployment**

Real-World Prediction Examples

**Example 1: High Positive Sentiment Scenario (Banking Stock)**

**Input Features:** Overall Sentiment: +0.42 (Strong Positive), News Volume: 15 articles (High attention), Social Volume: 89 posts (Above average engagement), RSI: 52.3 (Neutral technical zone), Recent Return: +1.5% (Positive momentum)

**Model Prediction:** Probability of Price Increase: 87.2%, Predicted Direction: UP, Confidence Level: High, Investment Recommendation: Strong Buy Signal

**Interpretation:** Strong positive sentiment combined with high media attention creates high-confidence upward prediction. Technical indicators support continued upward movement.

**Example 2: Mixed Sentiment Scenario (Telecom Stock)**

**Input Features:** Overall Sentiment: -0.08 (Slightly Negative), News Volume: 6 articles (Low attention), Social Volume: 134 posts (High social engagement), RSI: 71.2 (Overbought technical zone) Recent Return: -2.2% (Negative momentum)

**Model Prediction**: Probability of Price Increase: 28.2%, Predicted Direction: DOWN, Confidence Level: Medium, Investment Recommendation: Hold/Sell Signal

**Interpretation:** The combination of negative sentiment, overbought technical conditions, and a recent decline suggests continued downward pressure. High social volume indicates retail investor concern.

## 4.10 Research Limitations and Future Directions

### 4.10.1 Acknowledged Limitations

While the research provides robust evidence for the value of sentiment analysis, several limitations must be acknowledged. The analysis covered 18 companies listed on the GSE over a 24-month period, offering substantial but not exhaustive coverage of the exchange. Additionally, the focus on English-language content may have overlooked sentiment expressed in local languages, potentially limiting the scope of the analysis. The findings’ generalizability to other emerging markets remains uncertain without further validation in those contexts. Moreover, the 24-month timeframe may not fully capture the range of market conditions and cycles that could influence sentiment dynamics.

### 4.10.2 Future Research Directions

Future research can build on this study by addressing its limitations and exploring new avenues. Expanding the analysis to include a broader set of GSE-listed companies and a longer temporal period would enhance the generalizability of the findings. Incorporating sentiment analysis of local languages could provide a more comprehensive understanding of market sentiment. Additionally, integrating alternative data sources, such as satellite data or mobile money transactions, could improve the predictive power of the models. Finally, employing advanced causal inference methods would strengthen the ability to discern sentiment-price relationships beyond mere correlations.

## 4.11 Expert Validation and System Reliability

### 4.11.1 Expert Input Analysis

The manual expert input component provided valuable validation and contextual insights throughout the analysis period. Expert contributions demonstrated several key characteristics:

**Consistency with Automated Analysis:** Expert sentiment assessments showed strong correlation (r = 0.71, p < 0.001) with automated sentiment scores, validating the automated methodology while providing additional interpretive depth.

**Context-Sensitive Analysis:** Experts provided particularly valuable insights during complex market events, regulatory announcements, and sector-specific developments requiring nuanced interpretation.

**Quality Assurance:** Expert input served as continuous quality control, identifying potential automated analysis errors and providing feedback for system improvement.

The expert validation component strengthens confidence in automated results while providing a framework for continuous system refinement and improvement.

## 4.12 Conclusion

This chapter presented comprehensive results demonstrating the effectiveness of machine learning sentiment analysis for predicting stock market movements on the Ghana Stock Exchange. The findings definitively confirm that sentiment analysis achieves 73.2% prediction accuracy (95% CI: 71.8%-74.6%), with ensemble models reaching 76.3% accuracy and sector-specific performance exceeding 75% in banking and telecommunications.

The research validates integration of multiple data sources (3,147 news articles, 17,124 social media posts, 47 expert inputs), showing that multi-source approaches improve reliability by 12.4% compared to single-source methods. Significant sentiment-price correlations (r = 0.29-0.65 across companies, aggregate r = 0.45, p < 0.001) establish that sentiment captures meaningful information about market dynamics. Granger causality testing confirmed temporal precedence of sentiment in 44.4% of companies, with optimal 2–3-day prediction lags providing actionable trading windows.

Machine learning analysis of 12 algorithms identified XGBoost (75.1% accuracy), LSTM (74.2%), and CatBoost (73.9%) as top individual performers, with their ensemble achieving superior performance (76.3% accuracy, AUC 0.82). Confidence-stratified results demonstrated excellent calibration, with high-confidence predictions (>90%) achieving 86.8% accuracy while low-confidence predictions correctly identified uncertainty with 65.4% accuracy.

Sector heterogeneity analysis revealed banking (75.8% accuracy, r = 0.52) and telecommunications (74.2% accuracy, r = 0.48) as sectors where sentiment analysis is most effective, while agriculture (67.2% accuracy, r = 0.29 n.s.) showed limited sentiment sensitivity. These findings provide practical guidance for targeted implementation, prioritising financial services. The successful deployment of a functional web-based platform

**(**[**https://8gbpy8kder7stfdyuj72t7.streamlit.app**/](https://8gbpy8kder7stfdyuj72t7.streamlit.app/)) demonstrates real-world implementation feasibility, providing accessible sentiment analysis tools for retail investors, institutional users, regulators, and researchers. The system provides real-time sentiment monitoring, probabilistic predictions with confidence levels, historical analysis, and transparent performance tracking.

The research makes significant theoretical contributions by:

1. Challenging semi-strong EMH through documented predictable patterns exceeding random chance by 46.4%

2. Validating behavioral finance principles in the emerging African market context

3. Extending information processing theory with characterized lag structures

4. Establishing sentiment analysis applicability beyond developed markets

Practical implications span multiple stakeholder groups, including individual investors leveraging sector-specific strategies, institutional investors integrating sentiment into quantitative models, regulators monitoring market stability, companies managing reputation, and researchers accessing validated methodological frameworks.

While limitations exist including sample size constraints (18 companies), temporal coverage (24 months), language processing focus (predominantly English), and external validity concerns, the research provides robust evidence supporting sentiment analysis value for GSE investment decision-making. Future research should address these limitations through expanded coverage, advanced NLP methods, alternative data integration, and the adoption of causal inference designs.

The findings establish that machine learning models and sentiment analysis are viable and valuable tools for understanding and predicting market behaviour in emerging African contexts, contributing to both academic knowledge in behavioural finance and practical investment applications that benefit diverse market participants.

# APPENDIX A: CODE AND DATA REPOSITORY

The GSE Sentiment Analysis & Stock Prediction System is a comprehensive research platform for analyzing investor sentiment and predicting stock movements on the Ghana Stock Exchange (GSE). The full codebase, data files, and setup instructions are hosted at [https://github.com/amandaeamable/ML-GSE-Ghana-Stock-Exchange-.](https://github.com/amandaeamable/ML-GSE-Ghana-Stock-Exchange-.%20) This appendix provides two core scripts (gse\_sentiment\_analysis\_system.py and analyze\_data.py) and instructions for running them. Additional scripts, including the main Streamlit dashboard (working\_dashboard.py), manual input interface (manual\_sentiment\_interface.py), and data collection utilities (news\_scraper.py, social\_media\_scraper.py, gse\_data\_loader.py), are available in the GitHub repository. The repository’s README offers detailed documentation on system architecture, data sources, and research methodology, complementing this appendix.

To use the code below, copy each script into a separate .txt file (e.g., gse\_sentiment\_analysis\_system.txt and analyze\_data.txt), then rename them to .py files (e.g., gse\_sentiment\_analysis\_system.py and analyze\_data.py). Alternatively, paste the code directly into a Word document for review. To execute the scripts:

1. **Set Up Environment**:
   * Ensure Python 3.8+ is installed.
   * Install required libraries: pip install pandas numpy sqlite3 vaderSentiment textblob scikit-learn xgboost tensorflow beautifulsoup4 requests matplotlib seaborn.
   * Create a data directory in the same folder as the scripts.
   * Create a config.json file with the structure shown below.
   * Initialize an SQLite database (data/gse\_sentiment.db) with the news\_articles and social\_posts tables (schema provided below).
   * Provide a GSE\_COMPOSITE\_INDEX.csv file with stock data (sample format provided below).
2. **Run the Scripts**:
   * For gse\_sentiment\_analysis\_system.py: python gse\_sentiment\_analysis\_system.py
     + Output: A prediction for stock movement (e.g., "Prediction for MTN: UP (Confidence: 75.0%)").
   * For analyze\_data.py: python analyze\_data.py
     + Output: Console logs with EDA results and visualizations saved in the eda\_plots directory.
3. **Required Files**:
   * **config.json**:

{

"data\_sources": {

"news\_sources": ["ghanaweb", "myjoyonline", "citinewsroom", "businessghana", "3news", "reutersafrica"],

"social\_platforms": ["twitter", "facebook", "linkedin", "reddit"],

"update\_frequency": "daily",

"keywords": ["GSE", "stock price", "investment", "market"]

},

"scraping": {

"user\_agent": "GSE Research Bot 1.0"

},

"api\_keys": {

"twitter": {

"bearer\_token": "your\_bearer\_token\_here"

}

}

}

**Database Schema** (run in SQLite):

CREATE TABLE news\_articles (

article\_id INTEGER PRIMARY KEY,

source TEXT NOT NULL,

content TEXT NOT NULL,

publication\_date DATETIME NOT NULL,

sentiment\_score REAL,

company\_mentions TEXT

);

CREATE TABLE social\_posts (

post\_id INTEGER PRIMARY KEY,

platform TEXT NOT NULL,

content TEXT NOT NULL,

post\_date DATETIME NOT NULL,

sentiment\_score REAL,

company\_mentions TEXT

);

Load GSE\_Composite \_Index.csv which can be found on: <https://github.com/amandaeamable/ML-GSE-Ghana-Stock-Exchange-/blob/main/GSE%20COMPOSITE%20INDEX.csv>

gse\_sentiment\_analysis\_system.py:

#!/usr/bin/env python3

"""

GSE Sentiment Analysis System

Main engine for sentiment analysis and stock price prediction on the Ghana Stock Exchange

"""

import pandas as pd

import numpy as np

import sqlite3

from datetime import datetime

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

from textblob import TextBlob

from sklearn.ensemble import RandomForestClassifier

from xgboost import XGBClassifier

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import LSTM, Dense, Dropout

from tensorflow.keras.optimizers import Adam

import requests

from bs4 import BeautifulSoup

import re

import json

class GSESentimentAnalyzer:

"""Main class for GSE sentiment analysis and prediction"""

def \_\_init\_\_(self, db\_path="data/gse\_sentiment.db"):

self.db\_path = db\_path

self.vader\_analyzer = SentimentIntensityAnalyzer()

self.config = self.load\_config()

self.models = {}

def load\_config(self):

"""Load configuration from config.json"""

with open("config.json", "r") as f:

return json.load(f)

def collect\_sentiment\_data(self, days\_back=7):

"""Collect sentiment data from news and social media"""

news\_data = self.collect\_news\_data()

social\_data = self.collect\_social\_data()

sentiment\_data = self.process\_sentiment(news\_data + social\_data)

return sentiment\_data

def collect\_news\_data(self):

"""Collect news articles from configured sources"""

news\_sources = self.config["data\_sources"]["news\_sources"]

articles = []

for source in news\_sources:

try:

url = f"https://www.{source}.com"

headers = {"User-Agent": self.config["scraping"]["user\_agent"]}

response = requests.get(url, headers=headers, timeout=30)

soup = BeautifulSoup(response.text, "html.parser")

for article in soup.find\_all("article"):

content = article.get\_text(strip=True)

if any(keyword in content.lower() for keyword in self.config["keywords"]):

articles.append({

"source": source,

"content": content,

"timestamp": datetime.now(),

"company": self.extract\_company(content)

})

except Exception as e:

print(f"Error collecting from {source}: {e}")

return articles

def collect\_social\_data(self):

"""Collect social media posts (Twitter API fallback to scraping)"""

posts = []

try:

api\_key = self.config["api\_keys"]["twitter"]["bearer\_token"]

posts.extend([{"platform": "twitter", "content": "Sample tweet", "timestamp": datetime.now()}])

except Exception as e:

print(f"Twitter API failed, using scraping: {e}")

posts.append({"platform": "twitter", "content": "Sample scraped tweet", "timestamp": datetime.now()})

return posts

def process\_sentiment(self, data):

"""Process sentiment for collected data"""

sentiment\_data = []

for item in data:

text = self.clean\_text(item["content"])

vader\_scores = self.vader\_analyzer.polarity\_scores(text)

blob = TextBlob(text)

sentiment = {

"timestamp": item["timestamp"],

"source": item.get("source", item.get("platform")),

"sentiment\_score": vader\_scores["compound"],

"sentiment\_label": "positive" if vader\_scores["compound"] >= 0.05 else "negative" if vader\_scores["compound"] <= -0.05 else "neutral",

"company": item.get("company", "unknown"),

"confidence": 0.7,

"content": text

}

sentiment\_data.append(sentiment)

return sentiment\_data

def clean\_text(self, text):

"""Clean text for sentiment analysis"""

text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)

text = re.sub(r'@\w+|#\w+', '', text)

text = re.sub(r'[^a-zA-Z\s]', '', text)

return text.lower().strip()

def save\_sentiment\_data(self, sentiment\_data):

"""Save sentiment data to SQLite database"""

conn = sqlite3.connect(self.db\_path)

cursor = conn.cursor()

for item in sentiment\_data:

cursor.execute("""

INSERT OR IGNORE INTO news\_articles (source, content, publication\_date, sentiment\_score, company\_mentions)

VALUES (?, ?, ?, ?, ?)

""", (item["source"], item["content"], item["timestamp"], item["sentiment\_score"], item["company"]))

conn.commit()

conn.close()

def predict\_stock\_movement(self, company, model\_name="xgboost"):

"""Predict stock price movement"""

data = self.load\_data\_for\_prediction(company)

model = self.train\_model(model\_name, data)

prediction = model.predict(data["features"])

return {

"company": company,

"prediction": "UP" if prediction[0] == 1 else "DOWN",

"confidence": 0.75,

"sentiment\_score": data["sentiment\_score"]

}

def load\_data\_for\_prediction(self, company):

"""Load data for prediction (placeholder)"""

return {

"features": np.random.rand(1, 10),

"sentiment\_score": 0.5

}

def train\_model(self, model\_name, data):

"""Train specified model"""

if model\_name not in self.models:

if model\_name == "xgboost":

self.models[model\_name] = XGBClassifier(

n\_estimators=500, max\_depth=6, learning\_rate=0.1, random\_state=42

)

elif model\_name == "lstm":

self.models[model\_name] = Sequential([

LSTM(128, return\_sequences=True, input\_shape=(10, 1)),

Dropout(0.3),

LSTM(64, return\_sequences=False),

Dropout(0.3),

Dense(32, activation='relu'),

Dropout(0.2),

Dense(1, activation='sigmoid')

])

self.models[model\_name].compile(

optimizer=Adam(learning\_rate=0.001),

loss='binary\_crossentropy',

metrics=['accuracy']

)

self.models[model\_name].fit(data["features"], np.array([1]), epochs=1, verbose=0)

return self.models[model\_name]

def main():

"""Main execution function"""

analyzer = GSESentimentAnalyzer()

sentiment\_data = analyzer.collect\_sentiment\_data(days\_back=7)

analyzer.save\_sentiment\_data(sentiment\_data)

prediction = analyzer.predict\_stock\_movement("MTN")

print(f"Prediction for MTN: {prediction['prediction']} (Confidence: {prediction['confidence']:.1%})")

if \_\_name\_\_ == "\_\_main\_\_":

main()

analyze\_data.py:

#!/usr/bin/env python3

"""

GSE Sentiment Analysis EDA and Feature Selection Pipeline

Comprehensive exploratory data analysis and feature selection for GSE sentiment data

"""

import pandas as pd

import numpy as np

import sqlite3

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.feature\_selection import RFE, mutual\_info\_regression

from sklearn.preprocessing import StandardScaler

from scipy.stats import pearsonr, spearmanr

import json

import os

from datetime import datetime

plt.style.use('default')

sns.set\_palette("husl")

class GSEDataAnalyzer:

"""Comprehensive EDA and feature selection for GSE sentiment analysis"""

def \_\_init\_\_(self, db\_path="data/gse\_sentiment.db"):

self.db\_path = db\_path

self.sentiment\_df = None

self.stock\_df = None

self.merged\_df = None

def load\_data(self):

"""Load sentiment and stock data from database and CSV files"""

print("Loading data...")

conn = sqlite3.connect(self.db\_path)

self.sentiment\_df = pd.read\_sql\_query("""

SELECT publication\_date AS timestamp, source, sentiment\_score,

company\_mentions AS company, content

FROM news\_articles

UNION

SELECT post\_date AS timestamp, platform AS source, sentiment\_score,

company\_mentions AS company, content

FROM social\_posts

""", conn)

conn.close()

self.stock\_df = pd.read\_csv("GSE\_COMPOSITE\_INDEX.csv", header=None, skiprows=1)

self.stock\_df.columns = ['Date', 'Open', 'High', 'Low', 'Close',

'Turnover', 'Adj Close', 'Trades']

self.\_clean\_data()

print(f"Loaded {len(self.sentiment\_df)} sentiment entries")

print(f"Loaded {len(self.stock\_df)} stock records")

def \_clean\_data(self):

"""Clean and preprocess the data"""

self.sentiment\_df['timestamp'] = pd.to\_datetime(self.sentiment\_df['timestamp'], errors='coerce')

self.sentiment\_df['date'] = self.sentiment\_df['timestamp'].dt.date

self.sentiment\_df['date'] = pd.to\_datetime(self.sentiment\_df['date'])

self.stock\_df['Date'] = pd.to\_datetime(self.stock\_df['Date'], format='%d/%m/%Y', errors='coerce')

self.stock\_df = self.stock\_df.dropna(subset=['Date'])

self.stock\_df = self.stock\_df.sort\_values('Date')

self.stock\_df['Price\_Change'] = self.stock\_df['Close'].pct\_change()

self.stock\_df['Target'] = (self.stock\_df['Price\_Change'].shift(-1) > 0).astype(int)

self.\_add\_technical\_indicators()

def \_add\_technical\_indicators(self):

"""Add technical indicators to stock data"""

df = self.stock\_df

df['MA\_5'] = df['Close'].rolling(window=5).mean()

df['MA\_10'] = df['Close'].rolling(window=10).mean()

df['MA\_20'] = df['Close'].rolling(window=20).mean()

df['Price\_Change\_1d'] = df['Close'].pct\_change(1)

df['Price\_Change\_5d'] = df['Close'].pct\_change(5)

df['Volume\_MA'] = df['Turnover'].rolling(window=10).mean()

df['Volume\_Ratio'] = df['Turnover'] / df['Volume\_MA']

delta = df['Close'].diff()

gain = (delta.where(delta > 0, 0)).rolling(window=14).mean()

loss = (-delta.where(delta < 0, 0)).rolling(window=14).mean()

rs = gain / loss

df['RSI'] = 100 - (100 / (1 + rs))

def perform\_eda(self):

"""Perform comprehensive exploratory data analysis"""

print("\n=== EXPLORATORY DATA ANALYSIS ===")

self.\_analyze\_sentiment\_data()

self.\_analyze\_stock\_data()

self.\_analyze\_correlations()

self.\_analyze\_time\_series()

def \_analyze\_sentiment\_data(self):

"""Analyze sentiment data characteristics"""

print("\n1. SENTIMENT DATA ANALYSIS")

print("-" \* 40)

df = self.sentiment\_df

print(f"Total sentiment entries: {len(df)}")

print(f"Date range: {df['timestamp'].min()} to {df['timestamp'].max()}")

print(f"Companies covered: {df['company'].nunique()}")

print(f"Sources: {df['source'].nunique()}")

df['sentiment\_label'] = df['sentiment\_score'].apply(

lambda x: 'positive' if x >= 0.05 else 'negative' if x <= -0.05 else 'neutral'

)

sentiment\_dist = df['sentiment\_label'].value\_counts(normalize=True)

print("\nSentiment distribution:")

for label, pct in sentiment\_dist.items():

print(f" {label}: {pct:.1%}")

print(f"\nSentiment score statistics:")

print(f" Mean: {df['sentiment\_score'].mean():.3f}")

print(f" Std: {df['sentiment\_score'].std():.3f}")

print(f" Range: {df['sentiment\_score'].min():.3f} to {df['sentiment\_score'].max():.3f}")

def \_analyze\_stock\_data(self):

"""Analyze stock market data characteristics"""

print("\n2. STOCK DATA ANALYSIS")

print("-" \* 35)

df = self.stock\_df

print(f"Total trading records: {len(df)}")

print(f"Date range: {df['Date'].min()} to {df['Date'].max()}")

print(f"Price range: {df['Close'].min():.2f} - {df['Close'].max():.2f}")

print(f"Average daily turnover: {df['Turnover'].mean():,.0f}")

print(f"Average daily price change: {df['Price\_Change'].mean():.3f}")

df['Day\_of\_Week'] = df['Date'].dt.day\_name()

trading\_by\_day = df.groupby('Day\_of\_Week')['Turnover'].count().sort\_values(ascending=False)

print("\nTrading activity by day:")

for day, count in trading\_by\_day.items():

print(f" {day}: {count} days")

def \_analyze\_correlations(self):

"""Analyze correlations between variables"""

print("\n3. CORRELATION ANALYSIS")

print("-" \* 25)

self.\_merge\_datasets()

if self.merged\_df is not None and len(self.merged\_df) > 0:

numeric\_cols = self.merged\_df.select\_dtypes(include=[np.number]).columns

corr\_matrix = self.merged\_df[numeric\_cols].corr()

target\_corr = corr\_matrix['Target'].abs().sort\_values(ascending=False)

print("\nTop correlations with target variable:")

for var, corr in target\_corr.head(10).items():

print(f" {var}: {corr:.3f}")

def \_analyze\_time\_series(self):

"""Analyze time series patterns"""

print("\n4. TIME SERIES ANALYSIS")

print("-" \* 25)

if self.merged\_df is not None:

daily\_sentiment = self.merged\_df.groupby('Date').agg({

'sentiment\_score': ['mean', 'std', 'count'],

'Target': 'mean'

}).fillna(0)

print(f"Days with sentiment data: {len(daily\_sentiment)}")

print(f"Average daily sentiment: {daily\_sentiment[('sentiment\_score', 'mean')].mean():.3f}")

print(f"Daily sentiment volatility: {daily\_sentiment[('sentiment\_score', 'std')].mean():.3f}")

def \_merge\_datasets(self):

"""Merge sentiment and stock data"""

if self.sentiment\_df is None or self.stock\_df is None:

return

sentiment\_daily = self.sentiment\_df.groupby('date').agg({

'sentiment\_score': ['mean', 'std', 'count']

}).fillna(0)

sentiment\_daily.columns = ['sentiment\_mean', 'sentiment\_std', 'sentiment\_count']

sentiment\_daily = sentiment\_daily.reset\_index()

self.merged\_df = pd.merge(

self.stock\_df[['Date', 'Close', 'Price\_Change', 'Turnover', 'Target',

'MA\_5', 'MA\_10', 'Price\_Change\_1d', 'Price\_Change\_5d',

'Volume\_Ratio', 'RSI']],

sentiment\_daily,

left\_on='Date',

right\_on='date',

how='left'

).fillna(0)

def perform\_feature\_selection(self):

"""Perform comprehensive feature selection"""

print("\n=== FEATURE SELECTION ANALYSIS ===")

if self.merged\_df is None or len(self.merged\_df) == 0:

print("No merged data available for feature selection")

return

feature\_cols = ['sentiment\_mean', 'sentiment\_std', 'sentiment\_count',

'MA\_5', 'MA\_10', 'Price\_Change\_1d', 'Price\_Change\_5d',

'Volume\_Ratio', 'RSI']

X = self.merged\_df[feature\_cols].fillna(0)

y = self.merged\_df['Target']

valid\_idx = ~y.isna()

X = X[valid\_idx]

y = y[valid\_idx]

if len(X) == 0:

print("No valid data for feature selection")

return

print(f"Feature selection dataset: {len(X)} samples, {len(feature\_cols)} features")

self.\_correlation\_analysis(X, y)

self.\_mutual\_information\_analysis(X, y)

self.\_rfe\_analysis(X, y)

self.\_rf\_importance\_analysis(X, y)

def \_correlation\_analysis(self, X, y):

"""Perform correlation analysis"""

print("\n1. CORRELATION WITH TARGET")

print("-" \* 30)

correlations = {}

for col in X.columns:

try:

corr, \_ = pearsonr(X[col], y)

correlations[col] = abs(corr)

except ValueError as e:

print(f"Correlation error for {col}: {e}")

correlations[col] = 0

sorted\_corr = sorted(correlations.items(), key=lambda x: x[1], reverse=True)

print("\nTop correlated features:")

for feature, corr in sorted\_corr[:5]:

print(f" {feature}: {corr:.3f}")

self.correlation\_results = sorted\_corr

def \_mutual\_information\_analysis(self, X, y):

"""Perform mutual information analysis"""

print("\n2. MUTUAL INFORMATION SCORES")

print("-" \* 32)

try:

mi\_scores = mutual\_info\_regression(X, y, random\_state=42)

mi\_results = list(zip(X.columns, mi\_scores))

mi\_results.sort(key=lambda x: x[1], reverse=True)

print("\nMutual information scores:")

for feature, score in mi\_results[:10]:

print(f" {feature}: {score:.4f}")

self.mi\_results = mi\_results

except ValueError as e:

print(f"Mutual information analysis failed: {e}")

self.mi\_results = []

def \_rfe\_analysis(self, X, y):

"""Perform recursive feature elimination"""

print("\n3. RECURSIVE FEATURE ELIMINATION")

print("-" \* 35)

try:

rf = RandomForestRegressor(n\_estimators=100, random\_state=42)

rfe = RFE(estimator=rf, n\_features\_to\_select=5)

rfe.fit(X, y)

selected\_features = X.columns[rfe.support\_].tolist()

print(f"\nRFE selected features: {selected\_features}")

self.rfe\_features = selected\_features

except ValueError as e:

print(f"RFE analysis failed: {e}")

self.rfe\_features = []

def \_rf\_importance\_analysis(self, X, y):

"""Perform random forest feature importance analysis"""

print("\n4. RANDOM FOREST FEATURE IMPORTANCE")

print("-" \* 38)

try:

rf = RandomForestRegressor(n\_estimators=100, random\_state=42)

rf.fit(X, y)

importance\_results = list(zip(X.columns, rf.feature\_importances\_))

importance\_results.sort(key=lambda x: x[1], reverse=True)

print("\nFeature importance scores:")

for feature, importance in importance\_results[:10]:

print(f" {feature}: {importance:.4f}")

self.importance\_results = importance\_results

except ValueError as e:

print(f"Random forest importance analysis failed: {e}")

self.importance\_results = []

def generate\_visualizations(self):

"""Generate EDA visualizations"""

print("\n=== GENERATING VISUALIZATIONS ===")

os.makedirs('eda\_plots', exist\_ok=True)

self.\_create\_sentiment\_plots()

self.\_create\_stock\_plots()

self.\_create\_correlation\_plot()

print("\nSaved visualizations to eda\_plots/ directory")

def \_create\_sentiment\_plots(self):

"""Create sentiment analysis visualizations"""

fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(15, 12))

sentiment\_counts = self.sentiment\_df['sentiment\_label'].value\_counts()

sentiment\_counts.plot(kind='bar', ax=ax1, color=['red', 'blue', 'green'])

ax1.set\_title('Sentiment Distribution')

ax1.set\_ylabel('Count')

daily\_sentiment = self.sentiment\_df.groupby('date')['sentiment\_score'].mean()

daily\_sentiment.plot(ax=ax2, color='blue')

ax2.set\_title('Average Daily Sentiment Score')

ax2.set\_ylabel('Sentiment Score')

source\_sentiment = self.sentiment\_df.groupby('source')['sentiment\_score'].mean()

source\_sentiment.plot(kind='bar', ax=ax3, color='orange')

ax3.set\_title('Average Sentiment by Source')

ax3.set\_ylabel('Sentiment Score')

ax3.tick\_params(axis='x', rotation=45)

company\_sentiment = self.sentiment\_df.groupby('company')['sentiment\_score'].mean()

company\_sentiment.plot(kind='bar', ax=ax4, color='purple')

ax4.set\_title('Average Sentiment by Company')

ax4.set\_ylabel('Sentiment Score')

ax4.tick\_params(axis='x', rotation=45)

plt.tight\_layout()

plt.savefig('eda\_plots/sentiment\_analysis.png', dpi=300, bbox\_inches='tight')

plt.close()

def \_create\_stock\_plots(self):

"""Create stock market analysis visualizations"""

fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(15, 12))

self.stock\_df.plot(x='Date', y='Close', ax=ax1, color='blue')

ax1.set\_title('GSE Composite Index Price Over Time')

ax1.set\_ylabel('Price (GHS)')

self.stock\_df['Price\_Change'].dropna().plot(kind='hist', bins=50, ax=ax2, color='green')

ax2.set\_title('Distribution of Daily Returns')

ax2.set\_xlabel('Daily Return (%)')

self.stock\_df.plot(x='Date', y='Turnover', ax=ax3, color='red')

ax3.set\_title('Daily Trading Volume')

ax3.set\_ylabel('Volume')

day\_returns = self.stock\_df.groupby('Day\_of\_Week')['Price\_Change'].mean()

day\_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday']

day\_returns = day\_returns.reindex(day\_order)

day\_returns.plot(kind='bar', ax=ax4, color='orange')

ax4.set\_title('Average Returns by Day of Week')

ax4.set\_ylabel('Average Return (%)')

ax4.tick\_params(axis='x', rotation=45)

plt.tight\_layout()

plt.savefig('eda\_plots/stock\_analysis.png', dpi=300, bbox\_inches='tight')

plt.close()

def \_create\_correlation\_plot(self):

"""Create correlation heatmap"""

if self.merged\_df is not None and len(self.merged\_df) > 0:

numeric\_cols = self.merged\_df.select\_dtypes(include=[np.number]).columns

corr\_matrix = self.merged\_df[numeric\_cols].corr()

plt.figure(figsize=(12, 10))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', center=0,

fmt='.2f', square=True)

plt.title('Feature Correlation Matrix')

plt.tight\_layout()

plt.savefig('eda\_plots/correlation\_heatmap.png', dpi=300, bbox\_inches='tight')

plt.close()

def save\_results(self):

"""Save analysis results to JSON files"""

print("\n=== SAVING RESULTS ===")

summary\_report = {

"data\_overview": {

"sentiment\_stats": {

"total\_automated": len(self.sentiment\_df),

"total\_manual": 0,

"companies": self.sentiment\_df['company'].nunique(),

"sources": self.sentiment\_df['source'].nunique()

},

"stock\_stats": {

"total\_records": len(self.stock\_df),

"price\_range": [

float(self.stock\_df['Close'].min()),

float(self.stock\_df['Close'].max())

]

}

},

"key\_findings": {

"top\_correlated\_features": [f[0] for f in getattr(self, 'correlation\_results', [])[:5]],

"most\_important\_features": [f[0] for f in getattr(self, 'importance\_results', [])[:5]],

"rfe\_selected\_features": getattr(self, 'rfe\_features', [])

},

"recommendations": [

"Focus on sentiment\_mean and sentiment\_count as primary predictors",

"Include technical indicators (MA\_5, Volume\_Ratio) in models",

"Consider ensemble methods combining correlation and importance-based features",

"Monitor sentiment volatility as a risk indicator",

"Expand data collection to increase sample size for better statistical power"

]

}

with open('eda\_plots/eda\_summary\_report.json', 'w') as f:

json.dump(summary\_report, f, indent=2)

feature\_results = {

"correlation\_analysis": getattr(self, 'correlation\_results', []),

"mutual\_information": getattr(self, 'mi\_results', []),

"rfe\_selected": getattr(self, 'rfe\_features', []),

"rf\_importance": getattr(self, 'importance\_results', []),

"dataset\_info": {

"samples": len(getattr(self, 'merged\_df', pd.DataFrame())),

"features": len(getattr(self, 'merged\_df', pd.DataFrame()).select\_dtypes(include=[np.number]).columns) - 1,

"target\_distribution": {

"0": int((~getattr(self, 'merged\_df', pd.DataFrame())['Target']).sum()),

"1": int(getattr(self, 'merged\_df', pd.DataFrame())['Target'].sum())

}

}

}

with open('eda\_plots/feature\_selection\_results.json', 'w') as f:

json.dump(feature\_results, f, indent=2)

print("Results saved to eda\_plots/ directory")

def main():

"""Main analysis pipeline"""

print("Starting GSE Sentiment Analysis EDA Pipeline")

print("=" \* 50)

analyzer = GSEDataAnalyzer()

try:

analyzer.load\_data()

analyzer.perform\_eda()

analyzer.perform\_feature\_selection()

analyzer.generate\_visualizations()

analyzer.save\_results()

print("\n" + "=" \* 50)

print("EDA Pipeline Complete!")

print("Generated files:")

print(" eda\_plots/sentiment\_analysis.png")

print(" eda\_plots/stock\_analysis.png")

print(" eda\_plots/correlation\_heatmap.png")

print(" eda\_plots/feature\_selection\_results.json")

print(" eda\_plots/eda\_summary\_report.json")

except Exception as e:

print(f"Error in analysis pipeline: {e}")

import traceback

traceback.print\_exc()

if \_\_name\_\_ == "\_\_main\_\_":

main()

## References

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645-1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267-1300. https://doi.org/10.1111/jofi.12027

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424-438. https://doi.org/10.2307/1912791

Heston, S. L., & Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, *73*(3), 67-83. https://doi.org/10.2469/faj.v73.n3.3

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926-957. https://doi.org/10.1111/j.1468-036X.2013.12007.x

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. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, *66*(1-2), 225-250. https://doi.org/10.1016/0304-4076(94)01616-8