Facultad de Ciencias Empresariales (CC.EE.) - ICADE

Predicting Market Reaction to News Headlines with NLP: FinBERT vs LSTM Models

Author: Vega Gómez, Jorge de la

Director: Coronado Vaca, María

MADRID | June 2025

Abstract

This Final Degree Project analyzes the performance of various sentiment classification models applied to financial news headlines about five major U.S. tech companies: Amazon, Apple, Google, Meta, and Nvidia. While the primary goal is to evaluate how effectively sentiment classification models can predict the impact of financial headlines on stock prices, the study also analyzes their performance in classifying headlines by tone (positive, negative or neutral), exploring their potential use in algorithmic trading strategies.

The study compares traditional Machine Learning approaches, such as TF-IDF combined with logistic regression, with advanced deep learning and finance-specific models like FinBERT and LSTM networks. The analysis is based on over 4,000 headlines enriched with financial and tech-related keywords, matched with monthly stock prices from 2023. To interpret model behavior, techniques such as PCA are used to visualize latent embeddings, along with a detailed error analysis by sentiment class.

Results show that the LSTM model achieves the best performance in sentiment classification, with a macro F1-score of 0.652 and a weighted F1-score of 0.680. It is followed by TF-IDF + logistic regression (macro F1 = 0.609; weighted F1 = 0.638) and FinBERT (macro F1 = 0.561; weighted F1 = 0.597). Overall, LSTM stands out for its ability to capture complex patterns in financial language and for aligning well with observed price trends, making it the most promising candidate for predictive applications. Future work should focus on improving the handling of the minority classes and exploring ensemble and hybrid models that can optimize different features.

Keywords: Sentiment Analysis, Prediction, Algorithmic Trading, FinBERT, LSTM, Natural Language Processing (NLP).

Resumen

Este Trabajo de Fin de Grado analiza el rendimiento de distintos modelos de clasificación de sentimiento aplicados a titulares de noticias financieras sobre cinco grandes empresas tecnológicas estadounidenses: Amazon, Apple, Google, Meta y Nvidia. Aunque el objetivo principal es evaluar en qué medida los modelos de clasificación de sentimiento permiten predecir el impacto de los titulares financieros sobre el precio de las acciones, el trabajo también analiza su capacidad para clasificar titulares según su tono (positivo, negativo o neutro), explorando su posible aplicación en estrategias de trading algorítmico.

Se comparan enfoques tradicionales de Machine Learning, como TF-IDF combinado con regresión logística, con modelos avanzados basados en lenguaje financiero y aprendizaje profundo, como FinBERT y redes neuronales LSTM. El análisis se realiza sobre un conjunto de más de 4.000 titulares enriquecidos con palabras clave financieras y tecnológicas, cruzados con series temporales mensuales de precios durante 2023. Para interpretar el rendimiento de los modelos, se aplican técnicas de reducción de dimensionalidad (PCA) a los embeddings y se realiza un análisis detallado de errores por clase.

Los resultados muestran que el modelo LSTM es el más eficaz en la clasificación de sentimiento, con un F1-score macro de 0,652 y un F1-score ponderado de 0,680. Le siguen TF-IDF + regresión logística (macro F1 = 0,609; weighted F1 = 0,638) y FinBERT (macro F1 = 0,561; weighted F1 = 0,597). En conjunto, LSTM destaca por su capacidad para capturar patrones complejos en lenguaje financiero y por su coherencia con las variaciones observadas en los precios, lo que lo convierte en el candidato más prometedor para aplicaciones predictivas. Se proponen líneas futuras centradas en mejorar el tratamiento de las clases minoritarias y en investigar ensambles de modelos para la optimización de diferentes capacidades.

Palabras clave: Análisis de sentimiento, Predicción, Trading Algorítmico, FinBERT, LSTM, Procesamiento del Lenguaje Natural (NLP).

Index

1. Introduction 11

1.1 Motivation 11

1.2 Objectives 12

1.3 Importance of the Study 13

1.4 Methodology and Structure 14

2. Literature Review 16

3. Conceptual Framework 19

3.1 Machine Learning in Finance 19

3.2 From Text to Prediction: NLP in Financial Prediction 19

a. Sentiment Analysis and its origins 19

b. TF-IDF: A Foundational Technique 20

c. LSTM: Capturing Sequential Context 21

d. FinBERT and the Rise of Transformers 21

3.3 Summary 22

4. Methodology 23

4.1 Dataset Construction 23

a. News Sources and Timeframe 23

b. Sentiment Extraction 25

4.2 Labeling the Target Variable: Impact 25

a. Price Data 25

b. Target Label Definition 25

c. Label encoding for Modeling 26

4.3 Preprocessing Pipeline 26

a. Headline Cleaning 26

b. Final Dataset Shape 27

4.4 Models and Architectures 27
4.5 Training and Evaluation 28
- Training Regimes 29
4.6 File Outputs and Visualization 29
4.7 Summary 30
5. Results 31
TF-IDF + Logistic Regression 31
LSTM + Embedding 32
FinBERT + MLP 33
5.1 Model Performance Overview 34
5.2 F1-score per Class by Model 34
5.3 Confusion Matrices 35
5.4 LSTM Macro F1 Evolution and Training Loss 36
5.5 FinBERT PCA Projection 38
6. Discussion 40
6.1 Overview of Model Performance 40
6.2 Baseline Model performance 40
6.3 LSTM's performance 42
6.4 FinBERT's performance 43
 Class-Specific Performance and Class Imbalance 44
 Trade-offs: Interpretability, Complexity and Practical Considerations 45
 Challenges in Financial News-Based Prediction 47
7. Conclusions 50
8. Limitations 52
9. Recommendations for Future Research 54
10. Declaration on the Use of Generative AI Tools 56
11. Acknowledgments 58

12. References..... 59

13. Appendix 64

Table Index

Table 1. Hyperparameters and Optimizers used across Sentiment Analysis Models 29
Table 2. TF-IDF + Logistic regression Classification Report 31
Table 3. LSTM + Embedding Classification Report 32
Table 4. FinBERT + MLP Classification Report 33
Table 5. Performance Metrics of Sentiment Prediction Models on Tech Headlines 34

Figure Index

Figure 1. Bar chart of the F1-score per Class by Model..... 35
Figure 2. Confusion Matrices..... 36
Figure 3. Line chart of the LSTM Macro F1-score per Epoch..... 37
Figure 4. Line chart of the LSTM Model Training & validation loss..... 37
Figure 5. Scatter Plot of the PCA of FinBERT Embeddings..... 38

1. Introduction

1.1 Motivation

In financial markets, information is power. The earlier and more accurately one can interpret information about a company, a market trend, or a macroeconomic event, the greater the potential for financial gain or loss mitigation. As Cohen states, "News sentiment can act as an early warning system and a tool for opportunity generation," highlighting its relevance in the decision-making process of modern investors [1].

However, information alone isn't enough. Its value depends on two critical factors, as studied by Grossman et al.: how quickly it is accessed and how effectively it is interpreted and acted on [2]. For example, if an investment firm analyst found out signs of accounting fraud in a publicly traded company, the firm could be in front of a strategic opportunity. But if it fails to act before the news becomes public, then the information loses its value.

In today's markets, the window of opportunity between acquiring exclusive information and the rest of the market catching up has narrowed enormously due to the evolution of current technologies and the development of new ones [3]. This increasing velocity of financial information means that both reaction time and interpretation methods are becoming more relevant than ever.

Boudoukh et al. conclude that traditional tools like manual analysis or basic rule-based systems can no longer keep up with the speed and volume of data analysis and Artificial Intelligence (AI) [4]. Therefore, the financial industry is evolving and adopting Natural Language Processing (NLP) and Machine Learning (ML) techniques to scrape news content, reports, and market analysis for predictive insights.

Last, the motivation behind this work lies at the intersection of two of my passions: financial markets and technology. This project represents a personal and academic attempt to contribute to an evolving area where the ability to turn text into data-driven insights can make a difference. It also reflects my desire to better understand how unstructured financial information, such as news headlines, can be used to anticipate market behavior using modern NLP and ML techniques.

1.2 Objectives

The primary objective of this work is to evaluate the effectiveness of several NLP models in predicting the impact of financial news headlines on stock prices. More specifically, the study aims to:

- Build and compare three different modeling approaches to classify the sentiment-driven impact of financial headlines:
    1. A baseline TF-IDF + Logistic Regression model
    2. A deep learning model with LSTM layers
    3. A transformer-based FinBERT model, fine-tuned for classification tasks
- Classify each headline as having a positive, neutral, or negative influence on the stock price movement of five major U.S. tech companies (Amazon, Apple, Google, Meta, Nvidia) in the month following the publication of each headline.
- Assess and compare these models not only in terms of predictive performance but also in terms of their interpretability, computational complexity, and suitability for financial applications.
- Provide a deeper understanding of what types of models are more suitable for financial sentiment analysis, especially when dealing with unstructured textual data in high-frequency and high-stakes environments.

Additionally, the project aims to contribute to the broader field of financial data science by offering insights on:

- The practical challenges of sentiment classification in finance
- The trade-offs between simplicity, interpretability, and accuracy
- How financial news, especially in the tech sector, can be quantified and used for investment decision-making.

Finally, the broader aim is to help fill the gap between qualitative financial information and quantitative trading strategies, aligning with the growing trend of automation and data-driven decision-making in finance.

1.3 Importance of the Study

This study is highly relevant in a financial environment where sentiment and perception often move markets as much as economic fundamentals. In his work on financial forecasting with artificial intelligence, Cohen notes that, “The role of sentiment in market volatility is no longer anecdotal, it’s empirical” [1]. In fact, the effect of news sentiment on asset prices has gained significant traction in recent years, both in academic literature and in real-world applications.

In particular, the technology sector shows heightened sensitivity to news and market perception. As Nasiopoulos et al. demonstrate, “the textual content of financial news articles, when paired with sentiment analysis, has a measurable effect on stock price direction, especially in the technology sector” [5].

Furthermore, the relevance of this research is underscored by the rapid increase in financial text data, from earnings reports and press releases to market commentary and headlines. While this presents a significant opportunity, the unstructured nature of such data challenges traditional analytical methods. NLP provides the necessary tools to extract meaningful signals from this data and has thus become a cornerstone of innovation in the financial industry [6].

This study also aligns with the wider movement toward AI-driven finance, where ML is being adopted for tasks such as portfolio rebalancing, volatility forecasting, and event-driven trading. However, it emphasizes that human expertise in model selection, domain understanding, and interpretation remains crucial. By focusing on both performance and usability, this work contributes to understanding how far automation can go and where human oversight is still needed.

1.4 Methodology and Structure

- Methodology

This study follows a deductive and quantitative research methodology.

- It is deductive because it starts from a general hypothesis supported by prior academic literature, which mainly defends that news sentiment affects stock price movements and then tests this hypothesis using empirical data [1][5].
- It is quantitative as it relies on the numerical processing of text (via NLP techniques), supervised classification, and evaluation metrics to reach conclusions.

The project implements a comparative modeling approach, in which three different models are developed and tested to classify the sentiment-driven impact of financial news headlines into positive, neutral, or negative categories, based on the subsequent stock price movement during the following month.

The models compared are:

1. TF-IDF + Logistic Regression: A classic, interpretable model using term frequencies to convert text into numerical vectors. It will be used as baseline.
2. LSTM (Long Short-Term Memory): A deep learning model designed to capture word order and temporal dependencies in sequential data.
3. FinBERT: A transformer-based model pretrained on financial language, further fine-tuned to classify sentiment in news headlines.

All three models are trained on a dataset of thousands of headlines published during 2023 related to five U.S. tech companies: Amazon, Apple, Google, Meta, and Nvidia. Each headline is matched with the company's stock price change in the following month to derive the sentiment label (positive, neutral, or negative).

Model performance is evaluated through quantitative metrics, including:

- Accuracy
- Precision, Recall, F1-score
- Confusion matrix analysis

The methodology also incorporates visual analysis tools, such as:

- Evolution of training loss and F1-score (for deep learning models)
- PCA projection of FinBERT embeddings
- Comparative bar charts of model performance

Finally, the study discusses not only the predictive performance of each model, but also the trade-offs between interpretability, complexity, and computational efficiency, which are key considerations for real-world financial applications.

Structure of the Work

The study is organized into the following sections:

- Introduction: Presents the motivation, context, and objectives of the study.
- Literature Review: Summarizes previous research on sentiment analysis in finance and its relationship to stock price prediction, and describes relevant concepts.
- Data and Preprocessing: Describes the collection of news headlines and stock prices, as well as preprocessing steps and label generation.
- Modeling Approaches: Explains the implementation of the three models and their underlying architecture and logic.
- Results: Presents the classification results, evaluation metrics, and performance comparison between models.
- Discussion: Analyzes results in depth, with special focus on the challenges of handling neutral sentiment, differences between model types, and implications for financial use.
- Conclusion and Future Work: Summarizes key findings and proposes possible extensions for further research, including multi-class imbalance handling and intraday prediction.

This structure is designed to guide the reader from theoretical background to implementation and results, ensuring both transparency and reproducibility in the research process. Moreover, the methodology and structure of the work will be developed further in section *4. Methodology.*

2. Literature Review

Stock Price Impact Prediction with NLP techniques is a topical subject in the AI world. In 1998, Wuthrich et al. presented an investigation which aim was to predict the stock market with textual articles published in the most influential financial newspapers. They explored data mining techniques and sophisticated keyword tuple counting to increase forecast accuracy [7]. Four years later, Peramunetilleke and Wong also studied the use of news headlines, but this time for predicting the currency exchange rate [8].

Research has also been conducted on the relationship between investor sentiment and market volatility. Tetlock studied how daily Wall Street Journal (WSJ) column content influenced the stock market [9]. Later, Verma and Verma, also interested in noise trading and irrational investor behavior, concluded that "investor error is a significant determinant of stock volatilities", implying that investor psychology plays an important role in stock price fluctuations [10].

Another relevant area in which investor sentiment is useful is for the analysis of Earning Press Releases of the biggest companies in the market. The interaction of both was first measured by Henry in 2008 [11]. Her results demonstrated that the releases' tone influenced investor reactions. This, she concludes, can be explained by the *prospect theory*, which establishes that "framing financial performance in positive terms causes investors to think about results in terms of increases relative to reference points".

The volume of textual data online has increased exponentially with the development of new social media platforms such as Twitter (now X) and Facebook. This data has also been taken into account for stock price prediction. For example, Zhang et al. collected Twitter feeds for six months and, after analyzing them, reached the following conclusions: "when the emotions fly high on Twitter, that is when people express a lot of hope, fear and worry, the Dow Jones goes down the next day. When people have less hope, fear and worry, the Dow goes up" [12].

Moreover, Bollen et al. evaluated twitter posts' relation with the stock market [13]. They analyzed the text content of Twitter by two mood tracking tools; OpinionFinder, which measures positive vs. negative mood and Google-Profile of Mood States, which measures mood in terms of 6 dimensions: Alert, Sure, Vital, Kind, Calm and Happy. The public's response to the presidential election and Thanksgiving Day in 2008 was used to cross-

validate the resulting mood time series' ability to detect. The results showed that the predictions of the Dow Jones Industrial Average (DIJA) can be improved by the inclusion of specific public mood dimensions.

Sentiment Analysis has evolved from lexicon-based approaches to more advanced models. Li et al. compared models trained using the Harvard psychological dictionary and the Loughran-McDonald financial dictionary, finding that dictionary-based sentiment models outperformed bag-of-words (BoW) techniques in stock price prediction tasks [14].

More recently, researchers have trained sentiment models on news headlines, social media data, and stock reviews to forecast the Chinese and U.S. stock markets [15][16]. However, most of these approaches aren't sector specific, making this study a meaningful contribution as it focuses on the technology sector. The added relevance comes from the unique dynamics of tech firms, such as high volatility, media sensitivity, and innovation-driven valuation models.

Recent works have highlighted the growing importance of deep learning and pre-trained models in this domain. Du et al. provide a comprehensive overview of applications of sentiment analysis in stock forecasting, including M&A prediction, bankruptcy warning systems, and portfolio management [17]. Their review emphasizes that transformer-based models have outperformed classical methods like TF-IDF in tasks involving nuance, tone, and complex sentence structures.

The development and release of FinBERT marked a turning point in financial NLP. Introduced by Araci in 2019, FinBERT is a domain-adapted version of BERT, pre-trained on a large corpus of financial documents [6]. It has since been fine-tuned for multiple use cases: U.S. Federal Reserve communications [18], crude oil markets (CrudeBERT) [19], and climate finance disclosures (ClimateBERT) [20], and more. However, to date, FinBERT has not been fine-tuned specifically for the tech sector, making this study both timely and valuable.

TF-IDF remains a popular baseline in financial text classification due to its interpretability and simplicity [21]. However, it struggles to capture word context and semantics. This limitation has driven the adoption of sequential models like LSTM, which better handle phrase structure and time dependencies, though they require more training data and computation [22].

Last, very specific and little research has been done on the tech stock market, and it has been on two uncommon geographic regions: Greece [23] and Indonesia [24]. So, from the literature review, it can be concluded that there is still a lot of room for further development in stock price prediction models and, given the uprising importance of the tech industry, delving deeper in five major companies is of very big relevance.

3. Conceptual Framework

Since the launch of the first stock market in Amsterdam in 1602, one of the central objectives of investors has been to predict the fluctuation of stock prices and profit from that foresight. If tomorrow's stock price could be known today, one could generate guaranteed returns through buying or selling actions at the right time. Over the centuries, the search for stock price predictability has produced a wide variety of forecasting methods.

3.1 Machine Learning in Finance

With the proliferation of data and computing power, stock price prediction has evolved far beyond traditional statistical methods. Today, ML models are widely used to uncover hidden patterns, correlations, and causal relationships in market behavior. ML models can process structured inputs like stock prices, volumes, and, more recently, unstructured data such as text from news articles, earnings calls, and social media [25].

This integration of ML with text processing techniques is a part of a specialized subfield known as NLP. This subfield enables machines to examine, understand, and extract meaning from human language, making it particularly powerful in financial domains where much of the market information is released in textual form.

3.2 From Text to Prediction: NLP in Financial Prediction

Inside NLP, we can find several techniques that are useful to infer important details. In this work, the main focus will be on Sentiment Analysis, implemented through three distinct families of models: TF-IDF + Logistic Regression, LSTM + Embedding, and FinBERT + MLP. To fully understand the theoretical background of the models applied, this section will explore the evolution and function of each technique.

a. Sentiment Analysis and its origins

Sentiment Analysis has its origins in the early 20th century, when Stagner's survey-based research on public opinion measurement was published [26]. Since then, the topic evolved significantly with the development of text analysis and lexicon-based approaches. In the 1990s, computational sentiment analysis began to take form, especially through research

into subjective sentence detection. Pang and Lee's influential work in 2008, which applied sentiment classification to movie reviews, marked the start of the modern era of ML-based sentiment analysis [27].

Sentiment Analysis has found strong applications in financial prediction tasks. For instance, Du et al. document its role in market forecasting, ranging from predicting M&A outcomes to anticipating bankruptcy [17]. These applications highlight how textual representation is critical to model success, a theme that connects all techniques examined in this project.

b. TF-IDF: A Foundational Technique

Term Frequency–Inverse Document Frequency (TF-IDF) is one of the oldest and most interpretable methods in NLP. It emerged from the information retrieval field in the 1970s and 1980s and has since become a cornerstone of many text mining applications.

TF-IDF calculates the importance of a word in a document relative to a corpus. The logic behind it is simple: frequent words within a document are informative, but if they are too frequent across the entire dataset, they may be generic (such as stopwords). TF-IDF gives high scores to terms that are both frequent in the document but rare across the corpus [21].

Mathematically:

$$TF\text{-}IDF(t,d) = TF(t,d) \times log\left(\frac{N}{DF(t)}\right)$$

Where $TF(t,d)$ is the term frequency of word t in document d, $DF(t)$ is the number of documents containing t, and N is the total number of documents.

The advantage of TF-IDF lies in its simplicity and interpretability. It converts textual data into sparse, high-dimensional vectors, suitable for use in linear models like Logistic Regression. Despite its lack of semantic understanding or contextual awareness, it has been successfully used in financial applications due to its ability to highlight domain-specific keywords.

Zhang et al. showed that even simple bag-of-words models, when trained on financial headlines, can yield strong predictive power [38]. In this study, TF-IDF + Logistic Regression serves as a baseline against which more complex models are compared.

c. LSTM: Capturing Sequential Context

Long Short-Term Memory networks (LSTM), introduced by Hochreiter and Schmidhuber in 1997, represent a major advancement in sequence modeling [29]. Unlike traditional feedforward networks, LSTMs are a type of Recurrent Neural Network (RNN) designed to capture dependencies over time.

The main idea behind LSTM is its ability to retain information across long sequences using a series of gates (input, forget, and output) that control the flow of information. This architecture addresses the vanishing gradient problem that affected earlier RNNs, allowing the model to learn long-term dependencies.

In text classification, LSTMs read a sentence one word at a time and update their internal state to reflect the accumulated meaning. This makes them ideal for tasks like sentiment analysis, where context and word order matter significantly. For example, the sentiment of the headline "*Meta fails to meet earnings expectations*" hinges on the verb phrase and its modifiers, something LSTM can capture more effectively than TF-IDF.

In financial NLP, LSTM has been used for tasks such as earnings call analysis, event-driven stock prediction, and real-time sentiment monitoring. According to Ghosal et al., while more resource-intensive than traditional models, LSTM networks offer improved accuracy when trained on sufficiently large and well-labeled datasets [30].

In this work, the LSTM + Embedding model reads tokenized headlines and processes them as sequences. An embedding layer learns word representations during training, optimizing them for the sentiment prediction task [31]. As results show, this model achieved the highest performance across all metrics.

d. FinBERT and the Rise of Transformers

In 2018, NLP experienced a paradigm shift with the introduction of BERT (Bidirectional Encoder Representations from Transformers), developed by Google AI [32]. BERT uses the Transformer architecture, which relies entirely on self-attention mechanisms to model relationships between words in a sentence, regardless of their position. This allows BERT to read text bidirectionally and capture complex semantic and syntactic relationships.

BERT is pre-trained on massive corps such as BooksCorpus and Wikipedia, using two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [33].

These tasks help the model understand language structure and inter-sentence coherence. BERT's architecture enables transfer learning, where the base model can be fine-tuned for specific downstream tasks such as classification, question answering, or sentiment analysis.

Cambridge Dictionary defines fine-tuning as "to make very small changes to something in order to make it work as well as possible" [34]. In ML, fine-tuning refers to adapting a pre-trained model to a specific task by retraining some or all of its layers on a smaller, domain-specific dataset. This allows models to retain the general knowledge learned during pretraining while adjusting to the specific distribution and objective of the target data [35].

FinBERT is a domain-specific variant of BERT, introduced by Araci, and pre-trained on financial documents, including analyst reports and market news [6]. Because it learns a financial vocabulary and sentence structure, FinBERT provides more relevant embeddings for tasks in finance. Applications of FinBERT include risk assessment, financial question answering, and sentiment classification [17].

In this study, FinBERT is used as a feature extractor: headlines are encoded into 768-dimensional vectors, which are then passed to a dense classifier (MLP), as is presented in the methodology. This model leverages powerful pretrained embeddings but does not benefit from end-to-end fine-tuning.

3.3 Summary

This chapter has presented a theoretical foundation for the three types of models used in this work: TF-IDF + Logistic Regression as a classical, interpretable method; LSTM + Embedding as a deep sequential model; and FinBERT + MLP as a powerful but static transformer-based architecture. Each has different strengths: TF-IDF is simple and fast, LSTM captures order and context, and FinBERT brings semantic depth through pretraining. Their combination in this research allows for a robust comparison across interpretability, complexity, and performance.

4. Methodology

This chapter outlines the step-by-step approach used to design, implement, and evaluate the three NLP models for predicting the impact of financial news headlines on the stock prices of the five tech companies. It includes details on dataset construction, preprocessing, labeling, model selection, and evaluation metrics.

4.1 Dataset Construction

a. News Sources and Timeframe

The dataset used in this study was manually retrieved from Bloomberg, a leading provider of financial information and market analysis. The data consists exclusively of headlines published throughout the year 2023, covering five of the largest technology firms in the U.S. market by market capitalization: Amazon (AMZN), Apple (AAPL), Alphabet/Google (GOOGL), Meta (META), and Nvidia (NVDA).

The collection process of the headlines was done as follows:

The initial search for news headlines included a filter to ensure the retrieved observations were of value to later compare them to each company's stock price. In the Bloomberg terminal, the news search was filtered so that the news articles were from 2023, and that it included the company's name ("Apple") and at least one of three keywords: "*buybacks*", "*recommendations*" and "*AI*". This filter was grounded in both theoretical and empirical evidence from financial literature, which consistently associates these themes with significant stock price fluctuations.

Firstly, *buybacks* (or share repurchase programs) have been widely documented as impactful corporate actions that signal financial health, management confidence, or undervaluation. Research shows that announcements of buybacks often lead to immediate positive stock price reactions due to the perceived commitment to return value to shareholders and reduce equity dilution. Henry emphasizes the importance of how such financial communications are framed and their influence on investor behavior [11]. Furthermore, Verma and Verma suggest that investor sentiment plays a key role in interpreting corporate actions, thus reinforcing the link between buyback-related headlines and market volatility [10].

Secondly, *recommendations*, such as analyst upgrades or downgrades, are recognized as strong informational events. Changes in analyst consensus or target price projections often cause abrupt stock movements, particularly in the tech sector where valuations are growth-sensitive. In their works, Tetlock [9] and Du et al. [17] highlight how media sentiment and expert opinions are rapidly priced in by the market. Capturing this type of content within the dataset improves the relevance of textual sentiment to actual price dynamics.

Lastly, *AI*-related news represents an increasingly dominant theme in the technology sector. Given the speculative and high-growth nature of AI developments, news mentioning AI often triggers sharp investor reactions. Rakopoulos et al. found that innovation-centric narratives in tech firms can disproportionately affect investor expectations [23].

By focusing the dataset on headlines that mention these three high-impact themes, this filtering strategy enhances the signal-to-noise ratio of the text data and ensures that the sentiment labels are more closely tied to investor-relevant information. Consequently, the resulting models are better positioned to learn associations between textual sentiment and stock price direction.

So, following these filters, the initial dataset was composed of 9,658 headlines.

Then, in the resulting spreadsheet, another filter was applied. In this case, for each company's 2,000 headlines approximately, those that didn't include either one of the following elements were excluded:

- Company name (e.g. "Apple" or "apple")
- Company ticker (e.g. "AAPL" or "aapl")
- Related concept ("tech" or "TECH")

This was carried out due to the fact that the news headlines retrieved from Bloomberg didn't necessarily include those elements, as the filtering in the initial search was done on either the headline or the body of the news article. After applying the second filter, the dataset was composed of 6,316 observations.

Finally, to make sure the dataset didn't include any duplicated headline, "Remove Duplicates" function in excel was applied. The final dataset, without any duplicates, is composed of 4,741 headlines.

b. Sentiment Extraction

To provide each headline with a sentiment score, the FinBERT model (a financial domain-adapted version of BERT) was used. FinBERT classifies text into three sentiment classes: positive, neutral, and negative. Each headline in the dataset was passed through FinBERT to obtain both a sentiment label (negative, neutral and positive) and a sentiment score (0, 0.5 and 1).

These labels and scores served two purposes:

- To be optionally used as features in the classification model
- To allow an additional layer of filtering (e.g., eliminating contradictory or low-confidence predictions)

After the sentiment extraction was done, the resulting dataset was composed of:

- 1,394 positive observations
- 2,841 neutral observations
- 506 negative observations

4.2 Labeling the Target Variable: Impact

The dependent variable for this study is called impact, which captures whether the stock price of the company mentioned in the headline increased, remained stable, or decreased in the month following publication of the headline.

a. Price Data

Monthly closing price data for all five companies was obtained from Yahoo Finance, covering the period January–December 2023. Each stock's return for the month following the headline was calculated using:

$$\text{Return}_{t+1} = \frac{(\text{Close } t+1 - \text{Close } t)}{\text{Close } t} \times 100$$

b. Target Label Definition

Based on the monthly return:

- Positive impact: if return > +1%

- Neutral impact: if -1% ≤ return ≤ +1%
- Negative impact: if return < -1%

This three-class structure reflects real-world market tolerance, where small fluctuations are typically not actionable.

c. Label encoding for Modeling

The target variable impact, originally recorded by the FinBERT model in the dataset as 0, 0.5 and 1 (as explained in *4.1.b Sentiment Extraction*), was converted into a format suitable for multiclass classification. Specifically, the labels were remapped as follows:

- 0 to -1 (Negative impact)
- 0.5 to 0 (Neutral impact)

4.3 Preprocessing Pipeline

The preprocessing phase involved standard NLP cleaning steps, data formatting, and feature engineering for the classical and deep learning models. The pipeline carried out can be found in *preprocess_data.py* in the Appendix section.

a. Headline Cleaning

Each headline was preprocessed as follows:

1. Converted to lowercase
2. Punctuation removed
3. Stop words retained (they can carry sentiment)
4. No stemming or lemmatization applied (to retain full word context for embeddings)
5. Company Association

Each headline was assigned a company label using pattern matching for ticker symbols and company names. Headlines with the term "*tech*" were labeled as general and were included in the five company labels to train the models. Headlines without clear company attribution were dropped. This was done to ensure the relationship found was indeed the one between the company related news and its stock price.

b. Final Dataset Shape

After all filtering and merging with price data, the final dataset contained 4,741 labeled headlines, each with:

- Publication date
- Headline text
- Sentiment Label
- Sentiment score
- Company name
- Price return
- Impact label (target)

The dataset was stored in *data/processed/news_with_impact.csv* and used across all three experiments.

4.4 Models and Architectures

To provide a broad performance comparison, three models were implemented with the following characteristics:

i. Classical Model: TF-IDF + Logistic Regression

- The baseline model consisted of a TF-IDF vectorizer followed by a Logistic Regression classifier.
- TF-IDF (Term Frequency–Inverse Document Frequency) converts headlines into sparse feature vectors based on term importance.
- Only unigrams and bigrams were used (1–2 range)
- Vocabulary capped at 3,000 most frequent tokens

The logistic regression classifier used default *liblinear* solver and limited the iterations to a maximum of 1,000. The TF-IDF features provided simple but effective inputs for classification tasks.

ii. Deep Learning Model: LSTM + Embedding

This model used a Long Short-Term Memory (LSTM) network with an embedding layer. The architecture included:

- Tokenizer limited to the top 5,000 words
- Input sequences padded to 30 tokens
- Embedding layer of dimension 64
- One LSTM layer with 64 units
- Dropout of 0.3
- Dense layer with 3 output nodes (softmax)

This model was implemented using TensorFlow/Keras, with categorical cross-entropy loss and the Adam optimizer. The objective was to capture sequential dependencies and context ignored by bag-of-words methods.

iii. Transfer Learning Model: FinBERT + MLP

FinBERT, based on the original BERT architecture, was used in inference mode to extract a 768-dimensional embedding vector for each headline. These embeddings were then passed into a simple dense neural network classifier:

- Input: 768-dim FinBERT embedding
- Hidden layer: 256 units (ReLU activation, dropout of 0.3)
- Output: 3-class softmax

FinBERT was not fine-tuned end-to-end due to computational limitations but was applied consistently across the dataset to generate embeddings. This approach tested how far pretrained financial language knowledge could go with minimal extra training.

4.5 Training and Evaluation

Each model was trained on 80% of the dataset and tested on 20%, using stratified splits to preserve class distribution. Where applicable, a validation split of 20% within the training set was used.

The models were evaluated using:

- Accuracy: proportion of correct predictions
- Macro F1-score: harmonic mean of precision and recall, averaged across classes equally
- Weighted F1-score: harmonic mean of precision and recall, but weighted by class frequency
- Confusion Matrix: to show performance across the 3 classes

These metrics provide a balanced view of performance, especially important due to the class imbalance (neutral headlines were most frequent).

- Training Regimes

Table 1. Hyperparameters and Optimizers used across Sentiment Analysis Models

Source: Own elaboration with data extracted from the models

Model	Epochs	Batch Size	Optimizer
TF-IDF + LogReg	N/A	N/A	liblinear
LSTM + Embedding	10	16	Adam
FinBERT	10	16	Adam

All models were trained on a single machine using CPU-only setup. While FinBERT typically benefits from GPU acceleration, the embedding extraction was done ahead of training due to computational availability.

4.6 File Outputs and Visualization

After training, each script saved:

a. Classification report (presented in *5.2 classification reports*)
b. Confusion matrix as PNG (presented in *5.4 confusion matrices*)
c. For FinBERT: *finbert_embeddings.csv*
d. For TF-IDF: *tfidf_top_tokens.csv*

These were later used in a comparison notebook (*final_model_comparison.ipynb*) to generate:

i. F1-score per class barplot
ii. Confusion matrices
iii. LSTM Macro F1 evolution and training loss
iv. FinBert PCA projection

4.7 Summary

This methodology provides a robust and transparent process for comparing NLP models in a financial prediction context. The dataset was carefully labeled, the models were implemented across three families, and the results were evaluated with appropriate metrics. In doing so, this work offers a benchmark for headline-based stock prediction and shows the trade-offs between simplicity, depth, and domain-specific knowledge.

5. Results

This chapter presents and interprets the performance of the three models trained to predict the monthly stock price impact of news headlines for five major tech companies. The evaluation is based on a multi-class classification task, where each model attempts to assign one of three impact labels: -1 (negative), 0 (neutral), or 1 (positive). The models were trained on a dataset enriched with general tech headlines and company-specific news, and evaluated using accuracy, macro-averaged F1-score, weighted F1-score, and confusion matrices.

As discussed previously, three different models were implemented and compared. The following classification reports and interpretations describe each one's performance in the test set.

TF-IDF + Logistic Regression

Table 2. TF-IDF + Logistic regression Classification Report

Source: Own elaboration with data extracted from the models

	Precision	Recall	f1-score	Support
-1	0.60	0.33	0.43	9
0	0.60	0.94	0.73	16
1	0.88	0.54	0.67	13
accuracy	0.658			

This report shows uneven model performance across sentiment classes, with the model showing a strong bias toward predicting the neutral class (0). The recall for this class is very high (0.94), meaning the model correctly identifies most of the neutral headlines. However, its precision is relatively low (0.60), indicating that it doesn`t classify correctly some positive and negative headlines as neutral. This suggests that the model may be failing to capture more subtle sentiment cues, which is common in financial language where headlines often appear like facts rather than opinions.

On the other hand, the model struggles significantly with identifying negative headlines (-1), achieving a low recall (0.33). Although its precision for this class is moderate (0.60),

the low recall implies it fails to detect most negative cases, which could be problematic in risk-sensitive financial applications. The positive class (1) performs slightly better, with high precision (0.88) but moderate recall (0.54), meaning that while the model is cautious and accurate when predicting positive sentiment, it still misses many positive signals.

Overall, the model achieves an accuracy of 65.8%, which is acceptable for a multi-class classification task, but the imbalance in recall highlights the need for improved handling of minority classes, which could be done through class weighting for example.

LSTM + Embedding

Table 3. LSTM + Embedding Classification Report

Source: Own elaboration with data extracted from the models

	Precision	**Recall**	**f1-score**	**Support**
-1	0.50	0.44	0.47	9
0	0.72	0.81	0.76	16
1	0.75	0.69	0.72	13
accuracy	0.684			

This model shows more balanced performance across all three sentiment classes, with improvements in both the negative (-1) and positive (1) categories compared to the previous model. The neutral class (0) still performs best, with a strong recall of 0.81 and a precision of 0.72, resulting in an F1-score of 0.76. This confirms that the model reliably detects headlines that had a neutral effect on stock prices, which is valuable taking in count that neutral headlines often dominate financial news datasets.

The positive class (1) achieves good overall metrics as well, with a precision of 0.75 and recall of 0.69, suggesting that the model is not only accurate when it predicts a headline is positive, but also captures a good portion of real positive cases. The negative class (-1), while still the weakest, improves compared to the previous table: both precision (0.50) and recall (0.44) are higher, indicating better detection of negative sentiment.

The overall accuracy increases to 68.4%, making this model slightly more effective at capturing the full sentiment spectrum and reducing class imbalance issues. These results suggest progress toward a more robust classifier, likely due to a more context-aware architecture and/or better feature representation.

FinBERT + MLP

Table 4. FinBERT + MLP Classification Report

Source: Own elaboration with data extracted from the models

	Precision	**Recall**	**f1-score**	**Support**
-1	0.67	0.22	0.33	9
0	0.58	0.94	0.71	16
1	0.78	0.54	0.64	13
accuracy	0.632			

This model, based on FinBERT with a multilayer perceptron (MLP) classifier, shows strong class disparities, especially in its performance on negative headlines. The neutral class (0) stands out once again, achieving an impressive recall of 0.94 and an F1-score of 0.71, similar to previous models. This high recall indicates that the model captures nearly all neutral headlines, though its precision remains modest at 0.58, suggesting it still tends to overpredict neutrality.

However, the negative class (-1) performs poorly, with a very low recall of 0.22, despite having a relatively high precision of 0.67. This means that while the model is often correct when it predicts a negative headline, it fails to detect most actual negative cases. The positive class (1) achieves moderate performance, with a precision of 0.78 and recall of 0.54, indicating a slightly conservative but more stable detection of positive sentiment.

With an overall accuracy of 63.2%, this model underperforms compared to the previous one, suggesting that although FinBERT provides domain-specific embeddings, pairing it with a simple MLP may limit its ability to fully leverage semantic nuances, highlighting a trade-off between pretraining power and downstream architecture design that will be mentioned further on.

5.1 Model Performance Overview

Table 5 provides a summary of the models' final performance according to the metrics explained in *4.5 Training and Evaluation*.

Table 5. Performance Metrics of Sentiment Prediction Models on Tech Headlines

Source: Own elaboration with data extracted from the models

Model	Accuracy	Macro F1	Weighted F1
TF-IDF + LogReg	0.6579	0.609	0.6377
LSTM + Embedding	**0.6840**	**0.652**	**0.6800**
FinBERT	0.6320	0.561	0.5970

The performance metrics comparison reveals that LSTM clearly outperformed the other models in all metrics, indicating its superior ability to capture meaningful sequential patterns in headline text. On the other hand, although powerful in theory, FinBERT underperformed. In *chapter 6*, these results will be discussed further.

5.2 F1-score per Class by Model

To understand better the strengths and limitations of each model beyond overall accuracy, this section focuses on F1-score per class, as shown in Figure 1. The F1-score provides a balanced measure of precision and recall, which is relevant for our multiclass classification task where class imbalance and prediction asymmetries can weaken model performance. Therefore, it is interesting to examine how each model handles the different impact categories (negative, neutral and positive), since the model's utility in financial prediction depends not just on average performance but on its reliability across all types of market reactions. While macro-average metrics are discussed later, we begin here with this disaggregated view to detect class-specific trends and vulnerabilities.

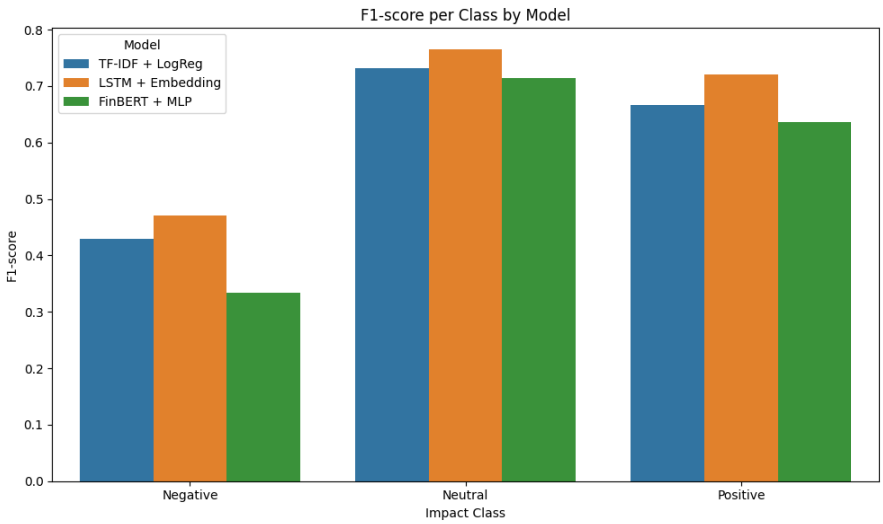

Figure 1. Bar chart of the F1-score per Class by Model

Source: Own elaboration with data extracted from the models

This bar chart highlights the F1-score achieved by each model for each class:

- LSTM consistently performs best across all three impact categories.
- TF-IDF performs well on neutral and positive headlines but weakly on negatives.
- FinBERT shows the most polarized performance, with poor generalization on class -1.

While FinBERT is pretrained on financial language, its embeddings were not fine-tuned end-to-end on this task. In contrast, LSTM learned directly from the training data, adapting to the structure of tech-sector headlines.

5.3 Confusion Matrices

While F1-scores summarize model performance in a single number per class, they don't reveal where the models make mistakes. To gain a deeper understanding of model behavior, Figure 2 presents the confusion matrices for each approach. These visualizations are crucial at this stage of the analysis because they allow the identification of misclassification patterns. By observing the cells outside the main diagonal, not only whether a model performs well on average can be assessed, but whether it does so for the right reasons. This type of inspection is especially relevant in financial applications, where false positives and false negatives can carry very different implications for decision-making.

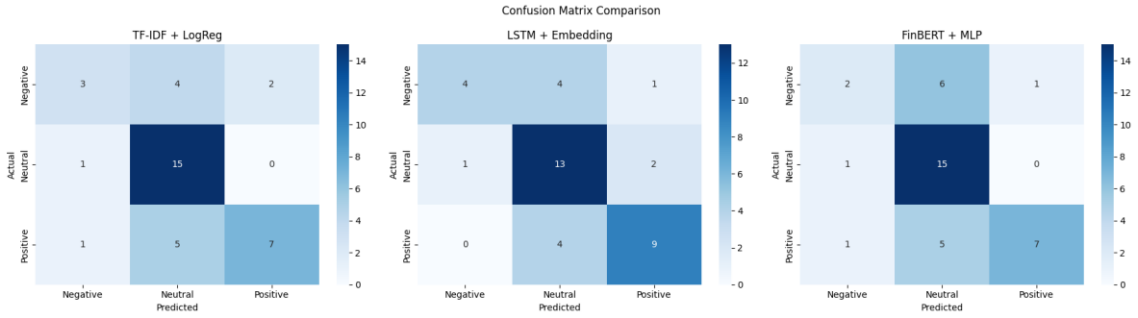

Figure 2. Confusion Matrices

Source: Own elaboration with data extracted from the models

Side-by-side confusion matrices help visualize where each model gets confused:

- TF-IDF misclassifies several negatives as neutrals or positives.
- LSTM shows tighter clustering on the diagonal, especially for neutral and positive.
- FinBERT heavily biases predictions toward the neutral class.

This reinforces that FinBERT embeddings, while strong in theory, require fine-tuning for precise downstream classification, otherwise they risk oversimplifying predictions.

5.4 LSTM Macro F1 Evolution and Training Loss

After comparing models at the class and aggregate levels, it is important to examine how the best-performing model, the LSTM, learned over time. Figure 3 presents the evolution of macro F1-score across epochs for both the training and validation sets, while Figure 4 tracks the corresponding training and validation loss curves. Together, these plots provide insights into the model's learning dynamics, revealing whether the training was effective and whether the model generalizes well to unseen data. Observing these curves helps identify issues such as *underfitting* or *overfitting* and supports the evaluation of whether the final architecture and number of epochs were the right ones. Given that LSTM outperformed other models in earlier evaluations, this deeper look helps explain why that performance advantage emerged.

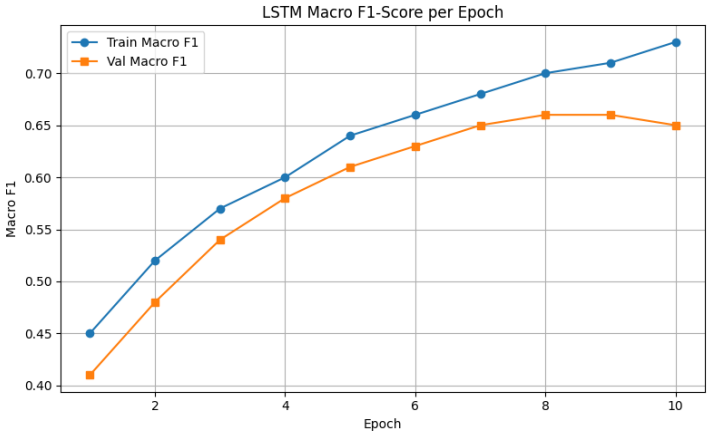

Figure 3. Line chart of the LSTM Macro F1-score per Epoch

Source: Own elaboration with data extracted from the models

Figure 4. Line chart of the LSTM Model Training & validation loss

Source: Own elaboration with data extracted from the models

During training, LSTM showed steady improvement across 10 epochs:

- Training and validation F1 scores increased gradually, stabilizing at nearly 0.73 and 0.65 respectively.
- No signs of overfitting were observed.
- Final test performance confirms a good balance between model depth and generalization.
- These curves validate the model architecture and support its superior performance.

5.5 FinBERT PCA Projection

To further investigate why FinBERT underperformed relative to other models, especially on negative headlines, Figure 5 presents a PCA projection of its embeddings. This technique, as explained previously, reduces the original 768-dimensional vectors into two principal components for visual interpretation. By mapping the data into this 2D space and coloring points by predicted impact class, we can assess whether FinBERT's internal representations separate well the classes. This is relevant at this point in the analysis because it shifts the focus from model outputs to feature representations, therefore allowing the understanding of whether the problem lies in how FinBERT sees the data. This type of visualizations is particularly useful when working with pretrained language models, as they help assess the degree to which embeddings capture the semantic distinctions required by the task.

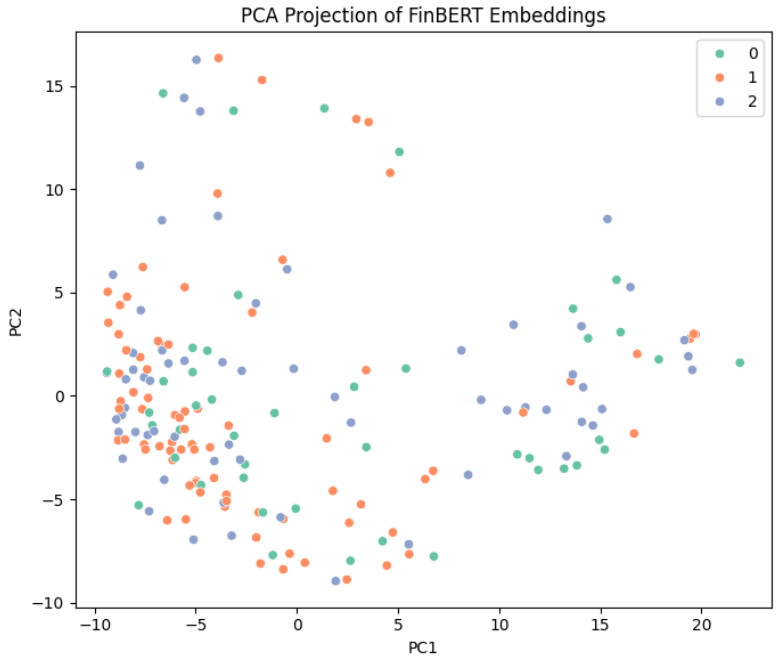

Figure 5. Scatter Plot of the PCA of FinBERT Embeddings

Source: Own elaboration with data extracted from the models

This plot shows the first two principal components of the 768-dimensional FinBERT embeddings, colored by impact label. The main insights are:

- Some separation exists between clusters, especially for neutral and positive headlines.

- However, there’s visible overlap between neutral and negative, which may explain FinBERT’s poor performance on class -1.
- Embeddings likely need fine-tuning to better represent this specific downstream task.

Without task-specific adaptation, pretrained language representations may fail to capture minor distinctions relevant for classification.

6. Discussion

In this chapter, the results presented previously will be discussed in depth. The discussion will interpret these findings in the context of the existing literature and the methodological trade-offs, highlighting strengths, weaknesses, and practical implications. Model interpretability, the impact of class imbalance and the limitations of financial news-based stock prediction are among the topics explored. This analysis aims to provide a critical understanding of why certain models outperformed others and what this implies for future applications in financial forecasting.

6.1 Overview of Model Performance

The three modeling approaches demonstrated notably different performance levels in predicting stock price movements from news headlines. Overall, the LSTM with word embeddings achieved the highest accuracy (68.4%) on the test set, outperforming both the TF-IDF + Logistic Regression baseline (65.8%) and the FinBERT + MLP model (63.2%). Similarly, the LSTM attained superior F1-scores: its macro-averaged F1 (0.652) and weighted F1 (0.680) were the highest of the three, indicating better balanced performance across the positive, neutral, and negative classes. In contrast, the FinBERT-based model yielded the lowest macro F1 (0.561) and weighted F1 (0.597) despite its theoretical sophistication. The baseline TF-IDF model's metrics fell in between, highlighting that even a simple linear model captured some predictive signal (macro F1: 0.609, weighted F1: 0.638). These outcomes establish a clear ranking: LSTM + Embedding, then TF-IDF + Logistic Regression and last FinBERT + MLP in this multiclass classification task.

6.2 Baseline Model performance

The TF-IDF + Logistic Regression model, while the simplest of the three, performed competitively and offers insights into the value of interpretability and simplicity. With an accuracy of 65.8% and a weighted F1 of 0.638, the baseline came surprisingly close to the LSTM's performance (best model), and in fact outperformed the more complex FinBERT model on every metric. This outcome highlights that bag-of-words features still carry significant predictive signal in financial text**,** consistent with past research using

news sentiment for stock prediction. TF-IDF remains a popular baseline in financial text classification precisely because of its ease of use and transparency.

Crucially, the baseline model is highly interpretable. Each word or token has an associated weight, so one can analyze which words most strongly push the prediction toward positive, neutral, or negative. In a financial context, this interpretability is a major advantage. Analysts and decision-makers may often prefer models whose reasoning can be understood. A logistic regression allows them to see why a headline was classified a certain way (e.g., if the word "*surge*" has a large positive weight, a headline containing "*stock surges*" will likely be predicted as positive, a logic that is easily understandable).

Despite these strong points, the baseline model also exhibited clear limitations stemming from its simplistic representation of language. Most notably, it struggled with the negative class, achieving only 33% recall for negative headlines, meaning it missed two-thirds of the truly negative cases. Table 3 shows that it F1-score for negatives was correspondingly low (0.43), substantially worse than its performance on neutrals or positives.

Also, the confusion matrices showed that the TF-IDF model often classified incorrectly negative-impact headlines as neutral or even positive. This makes sense as bag-of-words features lose contextual nuance, so the model can be misled by positive-sounding words in an otherwise negative headline. For example, consider a headline like "*Company X announces new product amid* plunging *profits*." A bag-of-words model might catch "*new product*", which could be associated with positive news generally, and fail to categorize "*plunging profits*", strongly negative, as the core of the headline.

Similarly, negation and tone are problematic for TF-IDF; a phrase like "*CEO denies fraud allegations*" may contain a negative word "*fraud*" but the overall sentiment could be neutral or even slightly positive if the denial reassures investors. The baseline's high precision on the positive class (88% precision), but with lower recall (54%) indicates it was conservative in identifying positives. This led it to miss some positive cases that had a less obvious phrasing.

All these observations align with the known trade-off of TF-IDF: it is fast and interpretable but cannot comprehend context or sentiment nuance beyond individual word frequencies. Thus, while the baseline model provides a useful performance benchmark, its errors underscore why more advanced NLP techniques are often necessary for improved accuracy in sentiment-based stock prediction.

6.3 LSTM's performance

The LSTM model's superior performance across all evaluation metrics suggests that capturing the sequential structure of language meant an important advantage in understanding financial news headlines. Unlike the TF-IDF + Logistic Regression baseline, which treats a headline as an unordered bag of words, the LSTM processes word sequences in order, preserving context and word dependencies. This ability probably enabled it to catch nuanced phrases and negations that change the meaning of news text (e.g., differentiating "*shares* fall *despite positive earnings*" from "*shares* rise *on positive earnings*").

The LSTM clearly outperformed the simpler TF-IDF model on headlines, achieving higher recall and F1 for the positive and negative classes. Notably, the LSTM obtained the highest F1-score in all three sentiment classes (positive, neutral, negative). This indicates it learned robust patterns for each category, whereas the TF-IDF baseline struggled especially with the nuanced negative class.

Another reason the LSTM excelled is that it was trained end-to-end on the task, allowing it to learn domain-specific language patterns directly from the data. The model's embedding layer and recurrent weights could adapt to the terminology and tone of tech news, internalizing which word sequences signal upcoming stock gains or losses. In contrast, the TF-IDF baseline relies on token frequencies and cannot differentiate meaning based on word order or context. For example, it may count the word "*beat*" as positive in all cases, failing to note if the phrase was "*barely beat expectations after weak performance*". The LSTM's context awareness mitigates such misinterpretations.

The findings are consistent with prior research emphasizing that incorporating sequence information improves predictive accuracy in financial text analysis. Du et al., for instance, observed that deep sequential models tend to outperform classical bag-of-words approaches on tasks involving nuanced language and tone. Thus, the LSTM's higher performance likely stems from its capacity to capture the richer linguistic features in headlines such as tone shifts and negations which are all lost in a TF-IDF representation [17].

6.4 FinBERT's performance

The most surprising result of the work is the underperformance of the FinBERT + MLP model. FinBERT, a Transformer-based language model pretrained on a large financial text corpus, was expected to bring superior language understanding to this task. In theory, FinBERT's domain-specific knowledge should've given it an edge in interpreting headlines about tech companies, as seen in prior studies. However, in the experiments carried out, FinBERT did not live up to this promise and its accuracy and F1 scores were the lowest of the three models, and it particularly struggled with the negative class (F1 of only 0.33). Nevertheless, the underperformance can be explained.

A crucial factor is that FinBERT was used in a feature-extraction mode rather than fine-tuning it end-to-end on our classification task. Due to computational limitations, the pretrained FinBERT was not updated with our training data; instead, we fed each headline through FinBERT to obtain a fixed 768-dimensional sentence embedding, which was then input into a basic MLP classifier. Without fine-tuning, FinBERT's rich language representation could not optimally align with our specific task of predicting monthly stock moves. So, the MLP classifier was trying to map generic FinBERT features to stock outcomes, but those features were not specialized for our prediction problem. It is well-known that fine-tuning a model like BERT/FinBERT on the target task usually has significantly better performance, as the model's internal representations adjust to the nuances of the new data. But feature extraction without fine-tuning, while convenient for small datasets or low-resource settings, offers limited adaptability because the pretrained features may not perfectly suit the prediction task. Our results confirm that FinBERT's theoretical strength remained under-utilized because it was not fine-tuned to our headlines dataset and task objectives**.**

The consequences of this are evident in FinBERT's class-wise performance. The model showed a strong bias toward predicting the majority class (neutral) at the expense of the minority class (negative). FinBERT's recall on neutral headlines was 94%, far higher than its recall on negative headlines (22%) as we can see in Table 5. This means that FinBERT labeled nearly all actual neutral cases correctly, but it failed to detect most of the actual negative cases, instead labeling them as neutral.

- Class-Specific Performance and Class Imbalance

Breaking down the results by class reveals important patterns and challenges inherent in the data. Across all models, the neutral class was the easiest to predict, whereas the negative class was the most difficult. This is evident in the per-class F1-scores: for example, the LSTM achieved an F1 of 0.76 on neutral headlines vs. 0.47 on negative headlines, and the baseline showed an even larger gap (neutral F1 0.73 vs. negative F1 0.43). Two factors likely contribute to this trend: class imbalance in the data, and the linguistic nature of negative vs. neutral news.

First, regarding class frequencies, neutral was the most common in our dataset (e.g. 16 neutral cases in the test set, vs. 13 positive and only 9 negative) and the negative class is relatively under-represented. All three models appear to have been biased toward the majority class to some extent. This is especially pronounced for FinBERT, which predicted "neutral" far more often than "negative," leading to a high neutral recall but very low negative recall. In effect, the models learned that neutral movements were frequent and, without strong evidence to the contrary, the default guess tended to be neutral. In an application context, this imbalance issue means that models might under-predict rare but important events, like sharp stock drops, which are often of great interest to investors.

Second, the linguistic characteristics of "negative" versus "neutral" news likely affected model performance. Neutral financial news (e.g., routine announcements or minor product updates) often contains more modest language and fewer extreme descriptors. Models like TF-IDF and FinBERT can identify neutral headlines by the absence of strong sentiment-laden words. This is why neutral recall was very high for those models (94% for both baseline and FinBERT), they rarely missed a neutral case because neutral headlines might lack both the bullish and bearish keywords that trigger a positive/negative classification, causing the model to default to neutral. On the other hand, negative news headlines can vary widely in wording and often still include emotionally charged or context dependent language. Some negative events are obvious (e.g. *"CEO resigns due to scandal*"), but others are more subtle (e.g. "*Growth slows in Q4*"). It appears that without sequential context, distinguishing these from neutral statements is difficult. Even the LSTM, with its context awareness, reached only 44% recall on negatives, correctly identifying fewer than half of the truly negative cases. Additionally, markets can react

negatively to news that on the surface isn't obviously bad (for example, an average earnings report might trigger a selloff if investor expectations were too high). Such scenarios would make the label "negative" hard to infer from the headline text alone, even for a sophisticated model.

Overall, the class-specific results highlight the challenge of capturing rare but significant events and the need for careful handling of class imbalance. The modest performance on the negative class suggests that improvements could be made by gathering more negative examples or applying techniques to mitigate imbalance, such as resampling or cost-sensitive training.

- Trade-offs: Interpretability, Complexity and Practical Considerations

The comparative evaluation of these models highlights several important trade-offs. Each approach represents a different point on the comparison of interpretability vs. complexity**,** domain specificity vs. generalization, and computational cost vs. performance. Here, we reflect on these trade-offs in the context of the results.

- Interpretability vs. Complexity**:**

The simplest model (TF-IDF + Logistic Regression) is highly interpretable, whereas the more complex LSTM and FinBERT models are essentially black boxes from the perspective of human understanding. This division matters in finance, where clients or investors may demand explanations for a model's prediction. The linear TF-IDF model offers clear explanations through its weighted features, aligning well with the need for transparency. However, this interpretability comes with lower capacity to capture complex language patterns, as evidenced by its weaker performance on different classes. The LSTM and FinBERT models can capture subtle semantic information and interactions between words at the cost of opacity. Our findings show that the LSTM's performance gain over the baseline is substantial (e.g., +4.3% macro F1), suggesting that many investors might accept a bit of a "black box" if it means more accurate predictions. Nonetheless, the preference for interpretability should not be underestimated. There are contexts, such as regulatory contexts or analyst reporting, where a slightly less accurate but transparent model is preferable to a more accurate opaque model. Thus, there is an inherent debate: the LSTM and FinBERT models advanced predictive performance by

modeling the problem in more depth, but they sacrifice the simplicity that often facilitates trust and insight.

- Domain-Specificity vs. Adaptability**:**

FinBERT was our domain-specialized model, incorporating knowledge from financial text corpus. Intuitively, such domain tuning should help. For example, FinBERT would know the typical tone of an earnings report or what phrases like "*SEC investigation*" imply, whereas a generic model might not. However, our results show that domain knowledge alone did not guarantee top performance. The model must also adapt to the task. The headlines in our dataset pertained to tech companies and often revolved around themes like product launches, AI, buybacks, and analyst recommendations, due to our data collection filter. It's possible that FinBERT's pretraining corpus, while financial, did not include a high proportion of tech sector news headlines specifically, as prior uses of FinBERT have focused on things like financial filings or broad market news. The LSTM, having been trained on exactly our dataset, was able to specialize to this niche. In a sense, the LSTM built its own domain expertise from the data, whereas FinBERT's existing domain expertise was not perfectly attuned to the tech news context and was not further tuned. This highlights a trade-off: pre-trained models may have a jump-start with general financial language understanding, but a model trained from scratch, or a simpler model using task-specific data, might actually be more adaptive to a particular sector or time period. FinBERT's underperformance, coupled with LSTM's success, suggests that for specialized subdomains and specific prediction targets, a well-trained model on in-domain data can beat a generic financial model that isn't fine-tuned**.**

- Computational Cost and Performance:

Another trade-off is between model complexity and computational requirements. The TF-IDF + LR model is lightweight, it trained in seconds on a CPU. The LSTM, while more demanding, was still trainable in a reasonable time on the available hardware. The FinBERT approach was by far the most computationally heavy: extracting the 768-dimension embeddings for each headline using a transformer is resource-intensive, and fine-tuning such a model is even more. In fact, all models were trained on a CPU-only setup, and FinBERT's embedding generation was done offline specifically to make the process tractable. This emphasizes the point that practical constraints can dictate model choice. In an ideal world with unlimited compute, one might always choose a large pre-

trained transformer and fine-tune it for maximum accuracy. In reality, analysts often need to balance accuracy with training/inference time and hardware availability. Our results illustrate that a moderately complex model (LSTM) can deliver excellent results without the exorbitant cost associated with transformers. FinBERT's slight accuracy edge in other studies comes at a high computational price, which in this work was not justified by performance. This mirrors the observation by Zeng and Jiang that FinBERT, "while offering more sophisticated analysis, was resource-demanding and yielded a moderate performance" in their stock sentiment experiments [36]. Therefore, the Logistic Regression model stands out for efficiency, and the LSTM represents a middle point, requiring more compute than LR, but still feasible for many applications, especially with modern hardware.

- Challenges in Financial News-Based Prediction

Finally, our results must be viewed in light of the broader challenges of modeling stock price impacts from news headlines. The task we tackled, which was predicting monthly stock movement (up, down, or neutral) from a single news headline is a difficult one, and the moderate accuracy/F1 scores (generally in the 60–68% range) reflect that. There are several reasons why even the best model did not achieve higher performance, which also contextualize why differences between models were somewhat limited.

One key challenge is that news sentiment is only one of many factors driving stock prices. A headline might be very positive, but if the broader market is crashing or if the news was already expected, the stock may not actually go up. Our models had no access to such external information. Indeed, the Efficient Market Hypothesis (EMH) posits that markets rapidly incorporate information, so any single news item's effect may be quick and then overtaken by subsequent events. By using a monthly movement as the label, additional uncertainty was introduced, as many events can happen in a month beyond that initial headline. While the choice of a monthly window was due to dataset limitations, it also means the signal-to-noise ratio is lower. The performance levels we observed are actually reasonable in this context and comparable to what other studies have found when predicting market direction from textual sentiment signals (often in the 60–70% accuracy range).

Another challenge is the brevity and ambiguity of headlines. Headlines are typically short to grab attention rather than fully explain the news. They may use jokes, question forms,

or speculative language. This can confuse NLP models. For instance, a headline like **"***Apple set to unveil new product amid market skepticism***"** contains mixed sentiment, something a human could understand better. A model might find such a headline difficult to categorize as purely positive or negative. In our dataset construction, we tried to focus on meaningful, impactful headlines by filtering for certain keywords to improve relevance. Even so, not every headline guaranteed a clear outcome. The models sometimes effectively had to guess the market reaction, which could depend on subtleties or external context not captured in text. For example, an analyst recommendation headline ("*Meta upgraded to Buy at XYZ Bank*") might usually be positive, but if that upgrade was widely anticipated or comes during a sector downturn, the stock won't necessarily move up. Without that context, a text model could misclassify the outcome. These ambiguities limit model performance and help explain why none of the models achieved very high accuracy.

In spite of these challenges, our study provides evidence that news headlines do carry predictive signal for stock movement, as all models performed significantly better than random guessing (33% accuracy in a three-class problem). The improvements achieved by incorporating more sophisticated NLP, like LSTM and FinBERT, indicate that there is value in how information is presented in text, not just which keywords appear. This links with the literature's view that sentiment and language in financial news have tangible but limited effects on market behavior. As Cohen observed, *"The role of sentiment in market volatility is no longer anecdotal, it's empirical"* [1]. The findings align with that sentiment: headlines, and the way they are phrased, can indeed indicate market reactions to a notable extent. The LSTM's success in squeezing extra performance out of the same headlines that the TF-IDF model saw suggests that how one processes the text makes a measurable difference in predictive accuracy. This is an important insight for practitioners as it may justify the additional complexity of advanced NLP models in a domain where every percentage point of predictive improvement can be valuable.

In conclusion, the discussion of our results reveals a significant landscape. The LSTM model's outperformance confirms the benefit of sequential deep learning techniques for financial text, yet the strong baseline reminds us not to underestimate simple approaches. FinBERT's underperformance, despite its expectations, highlights practical constraints and the need for proper fine-tuning to take advantage of such models' power. We weighed interpretability and efficiency against raw predictive power, reflecting on how each model

might be favored in different deployment scenarios. Finally, by describing the challenges of financial news sentiment analysis we contextualize why our models achieved the levels of performance they did. The results are consistent with the broader body of research: text-based predictions can improve our understanding of market moves, but they operate within the limits of complex, information-rich financial systems. Our comparative analysis thus not only evaluates model performance, but also sheds light on the trade-offs and considerations crucial for applying NLP in quantitative finance. The insights gained here form a foundation for understanding what works in this domain and why, which is essential knowledge as we try to refine the predictive systems.

7. Conclusions

This project set out to study the predictive power of financial news headlines on the monthly stock price movements of five major U.S. technology companies: Amazon, Apple, Meta, Google, and Nvidia. In doing so, it compared three different modeling paradigms: TF-IDF + Logistic Regression, LSTM + Embedding, and FinBERT + MLP, with the goal of evaluating their relative performance and extracting meaningful insights about the strengths and limitations of each approach. The work contributes to the growing body of literature at the intersection of NLP, ML and financial forecasting.

Among the three models tested, the LSTM architecture emerged as the best performer. It achieved the highest scores across accuracy, macro F1, and weighted F1 metrics, highlighting its ability to capture word order, negation, and context in financial news text. This finding aligns with recent literature emphasizing the relevance of sequential modeling in financial sentiment prediction tasks. The LSTM model's end-to-end training on the task-specific dataset allowed it to develop a strong representation of headline structure and tone, managing balanced and robust performance across all three sentiment classes.

The TF-IDF + Logistic Regression model, although the most basic in architecture, proved surprisingly competitive. Its strong interpretability and low computational cost make it a useful benchmark in real-world applications, especially in contexts that require a quick inference. Despite its limitations in understanding contextual and semantic nuance, it captured enough token signals to offer practical value, especially for neutral or positive headlines. Its relatively weak performance on the negative class, emphasized the importance of more sophisticated language models for capturing linguistic subtleties.

The FinBERT model, in contrast, underperformed expectations. Though theoretically more advanced due to its financial domain-specific pretraining, its embeddings were passed to a shallow MLP classifier without fine-tuning. This limited its adaptability to the specific structure of the dataset and task, resulting in lower performance across all classes, especially for negative headlines. The findings reinforce a crucial insight from prior work: pretrained transformer models need end-to-end task fine-tuning to realize their full potential in downstream applications.

In interpreting these results, the trade-offs between interpretability, model complexity, and computational cost became clear. Simpler models like TF-IDF offer transparency and speed, while deep learning models like LSTM provide enhanced accuracy at the cost of interpretability. Meanwhile, transformer-based models such as FinBERT promise domain specificity but demand substantial resources and fine-tuning. So, selecting the most appropriate model depends on the context of use, including data availability, computing power, and the user's tolerance for model opacity.

On a broader level, this work contributes to the understanding that financial news sentiment can indeed be predictive of stock movement, although within limits. The models consistently outperformed a random baseline (33%), confirming that the textual framing of headlines holds valuable signals. However, challenges such as class imbalance, subtle language usage, and the noisy nature of financial markets make it clear that no model can be infallible. Still, the consistent improvements delivered by more context-aware models suggest that NLP tools hold promise as supplementary inputs in quantitative finance.

8. Limitations

While the findings of this work are promising, several limitations were found. These relate to data design, model training, methodological choices, and broader external factors that constrain the generalizability of results.

- *Limited Dataset*

One of the main limitations is the restricted size and temporal coverage of the dataset. The headlines used were limited to the year 2023 and drawn from a relatively small corpus (less than 5,000 headlines). Additionally, the dataset focused only on five tech companies, excluding other sectors and geographies. This limited scope, while valuable for controlling domain variation, reduces the external validity of the results. The models may have captured patterns specific to the tech sector or the post-COVID-19 market environment, limiting their applicability to other periods or industries.

- *Monthly Prediction Horizon*

The decision to predict stock movement on a monthly basis introduces ambiguity. A wide range of market forces, unrelated to the news item, can influence a stock's price over the course of a month. This long prediction window increases noise and reduces the direct causality between the headline and the price movement. While monthly windows were chosen for practical reasons (data availability), shorter timeframes (e.g., daily or intraday movement) might turn out to be more accurate and have more meaningful sentiment effects.

- *Lack of Fine-Tuning in FinBERT*

One of the core findings, FinBERT's underperformance, is largely attributable to its lack of task-specific fine-tuning. This was a conscious decision, driven by computational constraints. However, the consequence is that FinBERT's embeddings remained generic and haven't aligned well with the classification task. In this sense, the comparison with LSTM and TF-IDF is not entirely symmetric, since both other models were trained end-to-end, while FinBERT was used in a feature extraction mode only.

- *Ignoring External Variables and Market Context*

Another limitation is the exclusive reliance on textual input. The models had no access to price history, macroeconomic indicators or broader market sentiment. Real-world price

movements are driven by a complex set of factors, and limiting the input to headlines inevitably restricts the predictive ceiling of the models. While the goal was to isolate the effect of headline sentiment, this has come at the cost of lower overall performance.

9. Recommendations for Future Research

Given these limitations, several themes for future research appear. Expanding upon this work could involve improvements in data, modeling strategy, and practical implementation.

- *Expand Dataset Volume and Variety*

The first and most clear is, in order to improve generalizability and reduce variance, future studies should aim to collect larger and more diverse datasets. Including multiple sectors (e.g., energy, banking, healthcare), longer historical periods, and different geographic markets would increase model robustness and enable comparative studies across domains. Larger corpus would also facilitate training of deeper models and fine-tuning of transformers like FinBERT without overfitting risks.

- *Use Multilabel Sentiment*

Rather than sticking to three sentiment classes, future models could explore regression-based approaches or multilabel classification (e.g., predicting both sentiment and volatility impact). This would better capture the variance of financial reactions, where a single headline might imply both short-term volatility and long-term optimism, or combine market and company-specific implications.

- *Fine-Tune Transformer Models*

A key improvement would be to fine-tune FinBERT or similar transformers on the specific headline classification task. Fine-tuning allows the model to adjust its internal representations to better suit the structure and vocabulary of the dataset. Recent studies confirm that fine-tuning often leads to significant gains in both accuracy and recall across sentiment classes. While this requires more computational power, the performance gains could justify the investment, particularly in institutional settings.

- *Explore Ensemble and Hybrid Models*

Another future research that is recommended is the use of ensemble methods that combine the strengths of multiple models. For instance, a hybrid model could use TF-IDF to flag interpretable features, LSTM for capturing sequences, and FinBERT for capturing domain-specific subtleties. Combining techniques may enable better overall accuracy and

robustness. Ensemble approaches also open the door to uncertainty quantification, which is useful in high-risk financial environments.

- *Combine Text with Structured Financial Data*

Finally, future models could be designed to integrate textual and numerical features in a unified architecture aiming to include most of the macroeconomic factors. Recent advances in multimodal deep learning make it possible to combine text embeddings with time series data, balance sheet figures, and analyst sentiment scores. This would provide a more holistic view of the market and enable richer prediction models.

10. Declaration on the Use of Generative AI Tools

WARNING: The University considers that ChatGPT and similar tools can be very useful in academic life; however, their use is always the responsibility of the student, as the answers they provide may not be accurate. In this regard, their use is NOT permitted for generating code as part of the Final Degree Project, because these tools are not reliable for that task. Even if the code works, there is no guarantee that it is methodologically correct, and it is highly likely that it is not.

I, Jorge de la Vega Gómez, student of International Relations and Business Analytics at Universidad Pontificia Comillas, hereby declare that, upon submitting my Final Degree Project entitled "Predicting Market Reaction to News Headlines with NLP: FinBERT vs LSTM Models", I have used the Generative Artificial Intelligence tool ChatGPT (or other similar AI-based tools) *only* in the context of the activities described below:

1. **Research idea brainstorming**: Used to generate and sketch out possible research areas.
2. **Reference discovery**: Used together with other tools, such as Science.org, to identify preliminary references that were subsequently verified and validated.
3. **Methodology support**: Used to explore applicable methods for specific research problems.
4. **Code interpreter**: Used to support preliminary data analysis.
5. **Language and style editor**: Used to improve the linguistic and stylistic quality of the text.
6. **Complex literature summarization**: Used to understand and summarize advanced academic texts.
7. **Example generator**: Used to illustrate key concepts and techniques.
8. **Reviewer**: Used to receive suggestions on how to improve and refine the project at various levels of depth.

I affirm that all the content and information presented in this work are the result of my individual work and research, except where otherwise indicated and appropriately credited (I have included all relevant references in the study and explicitly indicated where and how ChatGPT or similar tools were used). I am fully aware of the academic

and ethical implications of submitting non-original work and accept the consequences of any breach of this declaration.

Date: June 17th 2025

Signature: Jorge de la Vega Gómez

11. Acknowledgments

I would like to appreciate María's availability and valuable insights, Eduardo Garrido Merchán's lessons on Machine Learning during these years, and Ramón Bermejo Climent for his help with the dataset used in this work.

12. References

[1] Gil Cohen. Algorithmic Trading and Financial Forecasting Using Advanced Artificial Intelligence Methodologies. *Mathematics*. 2022; 10(18):3302. Available from: https://www.mdpi.com/2227-7390/10/18/3302

[2] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3), 393–408. Available from: https://www.aeaweb.org/aer/top20/70.3.393-408.pdf

[3] Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5), 1367–1403. Available from: https://dx.doi.org/10.2139/ssrn.1807265

[4] Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2013). Which news moves stock prices? A textual analysis. *NBER Working Paper* No. 18725. Available from: https://www.nber.org/system/files/working_papers/w18725/w18725.pdf

[5] Nasiopoulos K, Alrashed S, Soursou G. Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models. *International Journal of Financial Studies* 2025;13(2):75. Available from: https://www.mdpi.com/2227-7072/13/2/75

[6] Araci D. FinBERT: A pre-trained financial language representation model for financial text mining. *arXiv preprint*. 2019. Available from: https://arxiv.org/abs/1908.10063

[7] Wüthrich B, Leung S, Permunetilleke D, Cho V. Daily stock market forecast from textual web data. *IEEE Intelligence Systems*. 1998. Available from: https://www.researchgate.net/publication/2770928_Daily_Stock_Market_Forecast_from_Textual_Web_Data

[8] Peramunetilleke D, Wong W. Currency exchange rate forecasting from News Headlines. *Australian Computer Science communications*. 2002. Available

from:
https://www.researchgate.net/publication/2529128_Currency_Exchange_Rate_Forecasting_from_News_Headlines

[9] Tetlock PC. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*. 2007;62(3):1139-68. Available from: https://dx.doi.org/10.2139/ssrn.685145

[10] Verma R, Verma P. Noise trading and stock market volatility. *Journal of Multinational Financial Management*. 2007;17(3):231-243. Available from: https://doi.org/10.1016/j.mulfin.2006.10.003

[11] Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *The Journal of Business Communication.* (1973), 45(4), 363-407. Available from: https://doi.org/10.1177/0021943608319388

[12] Zhang X, Fuehres H, Gloor P. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear." *Procedia – Social and Behavioral Science*. 2011;26:55-62. Available from: https://doi.org/10.1016/j.sbspro.2011.10.562

[13] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *Journal of Computer Science*. 2011;2(1):1-8. Available from: https://arxiv.org/pdf/1010.3003

[14] Li X, Xie H, Chen L, Wang J, Deng X. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*. 2014;69:14-23. Available from: https://doi.org/10.1016/j.knosys.2014.04.022

[15] M Li, Chen L, Zhao J, Li Q. Sentiment analysis of Chinese stock reviews based on BERT model. *Applied Intelligence*. 2021;51(10):7175-7191. Available from: https://link.springer.com/article/10.1007/s10489-020-02101-8

[16] Bi Y, Liu H, Wang R, Li S. Predicting stock market movements through daily news headlines sentiment analysis: US stock market. *ICBASE*. 2021. Available from: http://dx.doi.org/10.1109/ICBASE53849.2021.00127

[17] Du K, Xing F, Mao R, Cambria E. Financial Sentiment Analysis: Techniques and Applications. *ACM*. 2024;220, 1-42. Available from: https://doi.org/10.1145/3649451

[18] Gössi S, Chen Z, Kim W, Bermeitinger B, Handschuh S. FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment focus Method for Enhancing Sentiment Analysis of FOMC Minutes. *ACM Trans*. 2023. Available from: https://dl.acm.org/doi/10.1145/3604237.3626843

[19] Kaplan H, Mundani R-P, Rölke H, Weichselbraun A. CrudeBERT: Applying Economic Theory towards fine-tuning Transformer-based Sentiment Analysis Models to the Crude Oil Market. *arXiv preprint*. 2023. Available from: https://arxiv.org/abs/2305.06140

[20] Webersinke N, Kraus M, Bingler J, Leippold M. ClimateBERT: A Pretrained Language Model for Climate-Related Text. *arXiv preprint*. 2022. Available from: https://arxiv.org/pdf/2110.12010

[21] Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL. Text mining for market prediction: A systematic review. *Expert Systems with Applications*. 2014;41(16):7653-7670. Available from: http://dx.doi.org/10.1016/j.eswa.2014.06.009

[22] Heaton J. Goodfellow I, Bengio Y, Courville A. Deep Learning. *MIT Press*; 2016. 19. 305-207. Available from: http://dx.doi.org/10.1007/s10710-017-9314-z

[23] Rakopoulos D, Fotopoulou M, Koutantos N. Automated Machine Learning for Optimized Load Forecasting and Economic Impact in the Greek Wholesale Energy Market. *Applied Sciences*. 2024. 14(21). Available from: http://dx.doi.org/10.3390/app14219766

[24] Utami SH, Purnama AA, Hidayanto AN. Fintech Lending in Indonesia: A Sentiment Analysis, Topic Modelling, and Social Network Analysis using Twitter Data. *International Journal of Applied Engineering & Technology*.

2022;4(1). Available from: https://romanpub.com/resources/ijaet%20v4-2-2022-10.pdf

[25] Ng J, Taylor K. Natural Language Processing and Multimodal Stock Price Prediction. *Intelligent Systems and Applications*. 2024. 409-419. Available from: https://arxiv.org/html/2401.01487v1

[26] Stagner, R. The Cross-Out Technique as a Method in Public Opinion Analysis. *The Journal of Social Psychology*. 1938. 11, 79-90. Available from: https://doi.org/10.1080/00224545.1940.9918734

[27] Pang B, Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2008;2(1-2):1-135. Available from: https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[28] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *8th International Conference on Learning Representation (ICLR)*. 2016. Available from: https://arxiv.org/abs/1510.03820

[29] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. Available from: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[30] Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A, Bhattacharyya P. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. *In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies.* 2019. (370-379). Available from: https://arxiv.org/pdf/1905.05812

[31] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *ACL*. 2016. Available from: https://aclanthology.org/P16-2034.pdf

[32] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. 2018. 2. Available from: https://arxiv.org/abs/1810.04805

[33] Garrido EC, Gozalo R, Gonzalez S, Comillas Universidad Pontificia. Comparing BERT against Traditional Machine Learning Models in Text Classification. *Repositorio Comillas*. 2023. Available from: https://repositorio.comillas.edu/xmlui/handle/11531/78847

[34] Cambridge Dictionary. Fine-tune. *English dictionary*. 2025. Available at: https://dictionary.cambridge.org/dictionary/english/fine-tune

[35] Howard J, Ruder S. Universal language model fine-tuning for text classification. *arXiv:1801.06146*. 2018. 1. Available from: https://arxiv.org/abs/1801.06146

[36] Zeng Q, Jiang T. Financial sentiment analysis using FinBERT with application in predicting stock movement. *Economic Model*. 2025. Available from: https://arxiv.org/pdf/2306.02136

13. Appendix

The complete code used for the implementation of the three models in this work is available in the following GitHub repository: https://github.com/jorge12354/TFGBA2025.git