



Revista de Estudiantes de Economía / Número 8 / Enero-diciembre 2024

INTERCAMBIO

Stock prediction using topic modeling and sentiment analysis techniques: A machine learning case study on Ecopetrol

***Predicción de acciones mediante
técnicas de modelado de temas
y análisis de sentimiento: un
estudio de caso de aprendizaje
automático sobre Ecopetrol***

.....
**Edwar David Valenzuela Cortés
Santiago Puentes Núñez**

E-ISSN 2619-6131



Stock prediction using topic modeling and sentiment analysis techniques: A machine learning case study on Ecopetrol*

Predicción de acciones mediante técnicas de modelado de temas y análisis de sentimiento: un estudio de caso de aprendizaje automático sobre Ecopetrol

Edwar David Valenzuela Cortés**

Santiago Puentes Núñez***

Abstract

This study introduces a novel technique for predicting market movements using topic and sentiment analysis of financial news about Ecopetrol. News headlines from Hydrocarbons and La República (July 2012 to December 2023) were analyzed using BER Topic, FinBERT, and Vader. The findings show that predictive models based on news headlines are more effective over 3 and 4-week periods compared to shorter periods. The Gradient Boosting model for week 3 achieved a profitability of 49.4% and accuracy of 57%, while a Random Forest model for week 4 yielded a profitability of 33.11% with a 9.71% error, outperforming the buy and hold strategy. These results highlight the advantage of short-term trend predictions in financial decision-making.



Intercamb. Rev. Estud.
Economía. N°. 8
Enero-diciembre 2024
127 pp.
E-ISSN 2619-6131
pp. 89-124

* **Artículo recibido:** 2 de julio de 2024 | **aceptado:** 10 de diciembre de 2024 | **modificado:** 27 de marzo de 2025.

** Estudiante de Administración de Empresas de la Universidad Nacional de Colombia Sede Bogotá. Correo electrónico: evalenzuelac@unal.edu.co

*** Estudiante de Economía de la Universidad Nacional de Colombia Sede Bogotá. Correo electrónico: spuentesn@unal.edu.co

Keywords: stock market prediction, topic modeling, sentiment analysis, Machine Learning, investment strategies.

JEL classification: G17, G14, C55.

1. Introduction

In the current context, stock price prediction has become an area of great interest for both researchers and investors. The ability to foresee market movements can provide a significant advantage in financial decision-making. However, the sheer volume of data generated daily, primarily qualitative in nature such as economic news, can be overwhelming and result in a complex task for humans to analyze to make correct decisions within financial markets.

Various techniques for processing this data to make decisions have been investigated in recent years, such as the use of sentiment analysis to extract the necessary features from qualitative data, which are then fed into machine learning or deep learning models. This generates systems capable of processing these data, allowing investors to make more optimal and rational decisions.

In Colombia, research on financial techniques and the use of data within the local financial market is still in its early stages. Consequently, this article focuses on predicting the stock price of Ecopetrol, the leading oil company in Colombia. Advanced topic modeling and sentiment analysis techniques integrated with machine learning models to predict Ecopetrol's stock price were used.

The study presents a comprehensive approach that combines the extraction and processing of financial news data with the collection of historical stock price data. News headlines were collected from two main sources, Hydrocarbons and La República, using web scraping techniques, covering the period from July 2012 to December 2023. These headlines were preprocessed and analyzed to extract relevant features through topic modeling and sentiment analysis, using tools such as BERTopic, FinBERT, and Vader.

Historical stock price data for Ecopetrol was collected from Yahoo Finance, which included opening and closing prices. This data enabled the calculation of dependent

variables such as trends and returns over various time intervals. This information was then integrated into several machine learning algorithms, including Gradient Boosting, Random Forest, Extreme Gradient Boosting, and K-Nearest Neighbors.

To run these models with the collected data, Scikit-learn was utilized, a Python library that offers machine learning tools. Additionally, Python was used as the programming language for this analysis.

The objective of this work is to provide an innovative technique through topic analysis to identify the most relevant themes related to the Ecopetrol asset, and in combination with the sentiment of the same news, generate robust and profitable models. Thus, the effectiveness of these techniques and the value of qualitative data are evaluated by analyzing their impact through the models on various objective variables related to changes in trend and return 1, 2, 3, and 4 weeks after the news appears. This is achieved by developing and identifying investment strategies based on the predictions generated by the models, contrasting with the buy-and-hold strategy to identify the system's ability to generate higher returns compared to the natural performance of the asset.

This research aims to contribute to the existing literature in the field of stock price prediction, demonstrating how the integration of natural language processing techniques and machine learning can improve the accuracy and utility of financial predictions. The results and discussions presented in the following sections will provide a detailed analysis of the performance of the models and the potential profitability of the proposed investment strategies.

2. Related work

2.1. Sentiment analysis, Machine Learning and Deep Learning in Stock Market

The integration of sentiment analysis with machine learning and deep learning techniques has significantly advanced stock market prediction methodologies. Researchers have demonstrated that incorporating sentiment analysis from financial news into machine learning models can enhance the accuracy of stock price predictions. Their

study in the field underlines the importance of combining historical price data with qualitative data from news and social media to improve predictive performance. This section explores various approaches and models that leverage sentiment analysis and machine learning to address the complexities of stock market prediction.

The research conducted by Maqbool et al. (2023) addresses the complexity of predicting stock prices due to high market volatility influenced by various external and internal factors. They propose a machine learning model that integrates historical price data with sentiment analysis of financial news using algorithms like VADER, TextBlob, and Flair.

The multilayer perceptron (MLP) regression model implemented using Scikit-learn demonstrated high accuracy in predicting trends and future prices, achieving 90% accuracy in 10-day predictions (Maqbool et al. 2023). The research highlights the importance of incorporating sentiment analysis to improve prediction accuracy and suggests the use of additional models, such as LSTM, along with the integration of social media and geopolitical news data for future enhancements.

A similar approach can be seen by Shah, Isah, and Zulkernine (2019). They present a comprehensive review of stock market prediction techniques, emphasizing market volatility due to external and internal factors. This research highlights the importance of integrating historical data with information from microblogs and financial news to enhance predictions. Although similar to the work of Maqbool et al. (Maqbool et al. 2023), it focuses more on the classification and taxonomy of various employed techniques, providing a solid foundation for future research in the field of stock market prediction.

Correia, Madureira, and Bernardino (2022) investigate the application of deep neural networks for sentiment analysis in the stock market, using social media data as a case study. They proposed a Deep Learning-based classification framework, evaluating several models (CNN, LSTM, GRU, and combinations of these with market indicators) and their performance in sentiment analysis.

The methodology included collecting data from various social sources, storing it in DFS, and testing trained models in trading simulations. Results showed that the CNN-LSTM model with market indicators achieved the best accuracy, reaching 73% in training and 69% in

testing, with a return on investment of 4.4% in simulations. The authors identified overfitting issues and suggested improvement strategies such as data batching (Correia et al. 2022).

A hybrid model is developed by Jing, Wu, and Wang (2021) that combines deep learning with sentiment analysis to predict stock prices. They use a Convolutional Neural Network (CNN) to classify hidden investor sentiments extracted from stock forums and apply a Long Short-Term Memory (LSTM) network to analyze technical indicators and sentiment analysis results. Experiments conducted on the Shanghai Stock Exchange demonstrate that this hybrid approach outperforms individual models and those without sentiment analysis, proving the effectiveness of combining advanced machine learning techniques with sentiment analysis in stock prediction, which results highly convenient for the approach proposed in this paper.

Khan et al. (2022) investigate the influence of social media and financial news on stock market prediction. They employ various machine learning algorithms and perform feature selection and spam reduction to improve prediction quality. The results show that social media and financial news significantly impact prediction accuracy, with a maximum accuracy of 83.22% achieved using a random forest ensemble of classifiers. The study also identifies those certain markets, like New York and Red Hat, are more challenging to predict, while others, such as IBM and Microsoft, are more influenced by social media and financial news, respectively.

The introduction of FinBERT, a BERT-based model pre-trained specifically for the financial domain, demonstrates significant improvements in financial sentiment analysis. FinBERT, adapted to a financial corpus, outperforms previous models by 15% in accuracy, showing its effectiveness even with small datasets as they might appreciate in the Araci (2019) approach. The study concludes that pre-trained language models are suitable for sentiment analysis tasks in finance and proposes future applications of FinBERT in stock market data and other natural language processing tasks in the financial area.

Collectively, these studies emphasize the critical role of sentiment analysis in improving stock market predictions. While some research focuses on integrating sentiment data with traditional machine learning models, others explore deep learning approaches

or hybrid methodologies combining multiple techniques. The effectiveness of sentiment analysis varies based on data sources, model architectures, and market conditions. The findings highlight the potential of advanced NLP models like FinBERT and suggest that incorporating diverse data sources, such as social media, financial reports, and geopolitical events, can further enhance prediction accuracy. This review provides a foundation for the approach adopted in this research, which seeks to leverage sentiment analysis to improve stock price forecasting in the context of Ecopetrol.

2.2. Topic Modeling

Topic modeling has emerged as a powerful tool in financial analytics, enabling the extraction of relevant themes and topics from large datasets of financial news and reports. By identifying the underlying topics within textual data, researchers can gain insights into market sentiments and trends that are not immediately apparent from numerical data alone. This section reviews recent advancements in topic modeling techniques, such as BERTopic and Latent Dirichlet Allocation (LDA), and their application in improving the accuracy of topics predictions and classifications. Additionally, this provides the necessary context to introduce the primary technique for processing topic data presented in this article, which has not been explored previously.

In the study by Balaneji and Maringer (2022), the combination of sentiment analysis and topic modeling is explored to improve the prediction of changes in implied volatility (iv30call) in the options market. Using financial news for six Dow Jones index stocks, the authors built text processing and topic modeling pipelines. Results showed that incorporating topic models improves iv30call prediction accuracy for five of the six companies analyzed. The study suggests that integrating topic models and sentiment analysis can enhance financial classifications' accuracy, proposing areas for future research.

García-Méndez et al. (2023) address the challenge of extracting relevant information, including forecasts and predictions, from online financial news. These sources often contain expert opinions on market events in various contexts. The proposed system uses natural language processing (NLP) techniques to segment text, filter less relevant phrases

through Latent Dirichlet Allocation (LDA) analysis and identify predictions through a discursive temporal analysis. This approach results in a concise summary of pertinent information, using techniques like paragraph segmentation, coreference resolution, and discursive temporality analysis. The solution outperformed a rule-based system and was comparable to a supervised system, without requiring manual annotations.

Wang et al. (2023) present an innovative framework for financial market analysis that combines latent topic discovery with investor expectation modeling. This approach is the first to jointly model investor expectations and automatically extract latent relationships between stocks. Experiments conducted on China's CSI 300 market demonstrated that the model consistently achieved annual performance above 10%, outperforming current standards in stock return predictions and trading simulations over several years (back testing).

Grootendorst (2022) introduces BERTopic, a topic model that extends topic modeling as a clustering task. BERTopic generates coherent topic representations using pre-trained transformer-based language models, clustering these representations, and generating topic representations through a class-based TF-IDF procedure. This model is competitive across various benchmarks, demonstrating its ability to generate coherent topics and its utility in various topic modeling applications, highlighting its innovative approach and performance compared to classic models and recent clustering-based approaches.

Chen et al. (2023) compare three state-of-the-art topic models (LDA, Top2Vec, and BERTopic) in the context of analyzing the impact of news on financial markets. Using a framework called "News Impact Analysis" (NIA), the authors analyzed 38,240 news articles to measure their impact on stock prices. Experimental results showed that BERTopic outperformed other models in terms of coherence, interpretability, and computation time, validating the NIA framework's feasibility and usability for financial researchers. This study provides a valuable comparison of topic models applied to financial news, highlighting BERTopic's effectiveness in this scenario.

The reviewed studies highlight the increasing relevance of topic modeling in financial analytics, particularly in stock market prediction. While LDA remains a widely used technique, more recent approaches, such as BERTopic, offer improvements in

coherence, interpretability, and computational performance. Research demonstrates that integrating topic modeling with sentiment analysis can enhance predictive accuracy, especially in market forecasting and volatility estimation. Additionally, novel frameworks incorporating investor expectations and latent topic discovery further extend the applications of topic modeling.

2.3. Colombian Stock Market

The Colombian stock market presents unique challenges and opportunities for predictive modeling, influenced by local economic conditions and market dynamics. Despite the growing interest in applying advanced analytical techniques in this market, research remains relatively nascent. This section examines previous studies and methodologies focused on the Colombian stock market, highlighting the use of machine learning and sentiment analysis to predict stock prices and trends. By reviewing these works, this study aims to provide a contextual background and identify gaps that need to be addressed in the context of Ecopetrol's stock price prediction.

In order to consider the Colombian Stock Market, it seems plausible to analyze some previous works. Monroy-Perdomo et al. (2022) develops an innovative methodology to predict stock trends in the Colombian market, using the sectoral Tobin's Q ratio as a trend index and stock price variation. The methodology is based on a quasi-experimental quantitative analysis of stocks traded up to December 30, 2019, covering at least 90% of trading time over the past five years. The average Q value of the relevant economic sector is adjusted to each stock's value to calculate its estimated price. Disparities between the estimated and actual value at time allow predicting the stock's price transition. Results show a significant influence of sector results on Tobin's Q performance at the corporate level, with significant levels above 50% in all cases and profitability not dropping below 30% in any sector, reaching up to 100%.

López-Gaviria (2019) analyzes the predictability of historical returns of the Colombian stock market in medium and long-term horizons, evaluating whether the risk premium is constant or variable over time and its relationship with other economic

variables. A price, returns, and dividends index is constructed for the period 1995-2017, based on the universe of issuers in the Colombian equity market. It is concluded that fluctuations in the dividend-price ratio are mainly explained by variations in future returns, indicating that the market is subject to cycles and that the risk premium varies over time. Additionally, information on mortgage credits, the real exchange rate, and S&P 500 index returns improves predictive capacity, suggesting that understanding the risk premium in Colombia benefits from considering credit markets and the context of an open economy.

The work of Palacio Roldán (2022) develops a stock recommendation model for the COLCAP index, based on technical analysis and sentiment analysis of the local market, in response to the Second Capital Market Mission in Colombia, which identified a notable decline in stock market participation. The model uses two recurrent neural networks to predict stock prices and classify them by profitability. Data obtained through web scraping and sentiment analysis for news are used, and a final dataset including historical prices and financial statements is built. The models are trained and validated using cross-validation, concluding that the gated recurrent unit (GRU) model is more accurate for long-term forecasts.

Iguarán Cotes (2019) applies feedforward neural networks to predict Ecopetrol's stock price in the Colombian market, considering three prediction time horizons: short, medium, and long term. Results show that it is possible to adequately predict prices one day and one week in advance, although for longer horizons, the results are not applicable. The models are mainly based on Boeing's price, revealing deep interconnections in financial markets that are not directly intuitive. Financial indicators are useful for the models, validating their relevance in professional trading. This study suggests that, with appropriate time horizons, neural networks can be an effective tool for trading in the real market.

The reviewed studies on the Colombian stock market demonstrate the increasing application of advanced analytical techniques, such as machine learning and sentiment analysis, to predict stock prices and trends. Methodologies like Tobin's Q ratio for

trend analysis and the use of recurrent neural networks for stock recommendations show promising results in improving prediction accuracy. Research also highlights the importance of incorporating external economic variables, such as exchange rates and S&P 500 index returns, to enhance predictive capacity. Despite the advances, challenges remain, particularly regarding the predictability of stock prices over longer time horizons. The studies emphasize the need for further exploration into refining predictive models and adjusting time horizons for better forecasting accuracy, providing a foundation for advancing Ecopetrol's stock price prediction in this context.

3. Methodology

This section describes the steps necessary to implement the prediction system for Ecopetrol stock price.

3.1. Data collection

The data was extracted from two main sources: Hydrocarbons and La República. These databases were selected due to the high volume of news articles published daily and their direct or indirect relevance to Ecopetrol. Web scraping techniques were used to collect the publication dates and headlines of the news articles, which were stored in a data frame. The news data span from July 2012 to December 2023, totaling 6,300 news articles. These sources provide a substantial amount of data to develop the system. Headlines are used because they are simpler and more efficient than using complete bodies of text.

Additionally, the data for Ecopetrol's asset prices were extracted from Yahoo Finance, requiring the use of dates, opening prices, and closing prices to perform the necessary calculations, which will be detailed later.

3.2. Financial news data preprocessing

To properly develop the system, avoid data leakage, potential biases, and increase model efficiency, proper data preprocessing is necessary. The following steps outline the initial treatment of the news articles:

1. *Data Division*: To obtain a suitable dataset for the models, the news data are divided into two sets: a training set, used to develop the topic model, find the most relevant variables, and train the machine learning models. This set will represent 70% of the data, with an initial date of 2012-07-01 and an end date of 2020-07-12. The test set will represent the remaining data, with an initial date of 2020-07-12 and an end date of 2023-12-31, which is used to check the effectiveness of the machine learning models after being trained with the relevant features as well as the performance of the proposed system (which will be detailed later). This step is crucial as it allows for a proper assessment of the system's effectiveness.
2. *Translate*: To enhance the effectiveness of tools for sentiment analysis and topic modeling, it was essential to translate all headlines from Spanish to English, as La República provides data in Spanish. Using a single language simplifies the process. This decision was made to ensure consistent language during the training phase, which facilitates the modeling process.
3. *Data cleaning*: For topic modeling, stopwords, special characters, short words, duplicate headlines, and links are removed. For sentiment analysis, the above is performed but some special characters are retained, as sentiment analysis models require these to determine if a text is positive, neutral, or negative.

3.3. News feature extraction

3.3.1. Topic Modeling

To discover which news has impacted a financial asset, the topic modeling technique can be used. This is a data analysis technique used in natural language processing and text mining to discover and extract latent themes or topics from a corpus of documents. This approach aims to identify underlying patterns in the text that represent coherent themes or areas of interest. By applying topic modeling, it is possible to better understand the structure and content of large document collections, facilitating tasks such as information organization, retrieval of relevant information, and exploration of trends and thematic relationships.

3.3.2. BERTopic

BERTopic is a topic modeling approach that utilizes BERT (Bidirectional Encoder Representations from Transformers), a pre-trained transformer-based language model, to automatically identify and cluster latent topics in a text corpus. Unlike traditional topic modeling methods, BERTopic leverages the power of deep learning to capture complex semantic relationships between words and documents, resulting in a more accurate and contextual representation of topics. This approach allows for efficient and thorough exploration of large volumes of textual data, facilitating the identification and understanding of underlying thematic patterns in various fields of study.

3.3.3. Feature extraction with BERTopic

As mentioned earlier, to use topic modeling in predicting financial assets, it is necessary to train a model with an existing dataset or documents to be able to assign topics to new news articles in the future. To this end, the headlines corresponding to the training set are used to develop the BERTopic model. To simplify the system, a maximum of 38 topics is assigned, based on the use of keywords to find these topics.

After training the BERTopic model, the model is used to make predictions on the headlines corresponding to the test set, assigning a topic to each headline. Since BERTopic uses all headlines without considering that the data are temporal, this topic assignment is applied to both the training and test sets.

3.4. Topics for Machine Learning models data

After assigning a topic to each headline, it is necessary to convert this data into a weekly matrix to later relate these topics to the price data for each week. To do this, each topic is converted into a column with two values:

- If the topic appeared in a given week, a label of “1” is assigned.
- Otherwise, a label of “0” is assigned.

This binary approach was chosen to identify the topics, events, situations, and aspects that may influence investor sentiment toward Ecopetrol stock and how this

is reflected in its stock price. By using a week-by-week measurement, the aim is to capture the presence of specific topics rather than their intensity or frequency. This serves as an initial approach to understanding the impact of news on price movements. Also, this approach allows the models to be trained and tested using the topics extracted from the headlines. Table 1 provides an example of the resulting dataset.

Table 1. A View of how topics data were transformed

Date	Topic1	Topic2	...	Topic _n
1/07/2012				
8/07/2012	1	0	...	0
	0	1	...	1
15/07/2012	1	0	...	0
22/07/2012	1	0	...	1
	0	0	...	0
29/07/2012				

Source: Own Elaboration.

3.4.1. Sentiment analysis

Sentiment analysis is a computational technique that utilizes natural language processing algorithms to determine the emotional attitude associated with a set of text, usually opinions, comments, or online reviews. This approach categorizes text into different classes of sentiments, such as positive, negative, or neutral, enabling researchers to understand public perception about specific topics, products, or services, as well as identify trends and emotional patterns in large textual datasets.

3.4.2. FINBert

To extract sentiment data, FinBERT is employed as a pretrained language model specifically designed for financial and economic text analysis. Based on the transformer architecture, it is one of the leading techniques in natural language processing (NLP). FinBERT has been trained on large-scale financial and economic datasets, allowing it to capture domain-specific language and nuances. Since it is pretrained, it can be fine-tuned for

specialized tasks such as sentiment classification of financial news, market movement prediction, and extracting relevant information from financial reports.

3.4.3. Feature Extraction with FinBERT and Vader

To enhance sentiment analysis, both FinBERT and Vader are utilized in this study. FinBERT processes financial news headlines, leveraging its domain-specific training, while Vader is a lexicon-based model optimized for short texts, making it well-suited for sentiment scoring of news headlines. The combination of these models allows for a more robust sentiment analysis by capturing both the contextual depth of transformer-based models and the rule-based interpretability of lexicon-based approaches.

FinBERT and Vader are applied independently to the entire news dataset, rather than separately to the training and testing sets, as is done with BERTopic. After running these models, sentiment labels are assigned:

- FinBERT: News headlines are categorized as negative (-1), neutral (0), or positive (1).
- Vader: Instead of discrete labels, a continuous sentiment score is generated based on the headline's polarity.

To integrate both models into the dataset, various sentiment-based features are extracted:

- News Sentiment Distribution (FinBERT & Vader): The number of news articles classified as negative, neutral, or positive for the past week.
- Headline Sentiment Score (FinBERT & Vader): The average sentiment score of news headlines for the past week.
- Cumulative Sentiment Trend (FinBERT & Vader): The average sentiment score over the last ten weeks, providing a smoothed representation of sentiment dynamics over time.

3.5. Preprocessing of Price data

This section explains the procedure used to obtain the dependent variables and calculate the performance of the proposed strategy. As mentioned earlier, to properly implement the system and models, the opening and closing prices need to be optimally utilized.

3.5.1. Target variables

Eight variables derived from price are calculated to determine the ability of the data and models to generate predictions and profitability. Four correspond to price changes in terms of trend, and four correspond to returns. All of this is done using the closing prices for each week. Since the system takes data from both prices and news up to the last hour of Sunday, and the model is run before the market opens the following week, calculations are done as follows:

3.5.2. Trend

The trend in n weeks is determined as:

4. Let P_{open} be the opening price of current week.
5. Let P_{close} be the closing price after n weeks.
6. Let $T(n)$ be the trend estimation after n weeks.

Then, the trend estimation function can be defined as

$$T(n) = \{\text{Bullish}, \quad \text{if } P_{close}(n) > P_{open} \quad \text{Bearish}, \quad \text{if } P_{close}(n) \leq P_{open}$$

where n represents the number of weeks ahead for which the trend is being estimated.

Bullish indicates an up-trend and *Bearish* indicates a down-trend.

3.5.3. Returns

Given the parameters exposed in section 3.5.2, the return estimation function can be defined as follows:

$$R(n) = \frac{P_{close}(n) - P_{open}}{P_{open}}$$

where $R(n)$ represents the return or percentage change after n weeks.

3.6. Implementation of Machine Learning Models

3.6.1. Models

The study employs machine learning algorithms implemented in scikit-learn and XGBoost, including Gradient Boosting, Random Forest, Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). These algorithms are utilized for both classification tasks (trend prediction) and regression tasks (return prediction). The training set is used to fit the models, while the test set evaluates their performance.

3.6.2. Model Evaluation

For regression tasks, mean squared error (MSE) is used to measure return prediction error. For classification tasks, accuracy, precision, and recall are employed to assess trend prediction performance. Precision and recall help determine whether the model is genuinely learning patterns from the data or merely optimizing for the majority class, which could indicate an overfitting issue.

Although cross-validation is a common technique for model evaluation, it is not directly applicable in this study due to the nature of the dataset. A portion of the training set is used to obtain the FinBERT model predictions for new topics in the test set, making traditional time-series cross-validation infeasible.

3.6.3. Hyperparameters Optimization

Hyperparameter optimization is essential for achieving better model performance. In this study, the GridSearchCV method from Scikit-learn was employed on the training set to systematically explore combinations of hyperparameters for each algorithm using tenfold cross-validation. Once the optimal values were identified, the entire training set was used to build the final model, which was then evaluated on the test set. The process was conducted using classification models, and the optimal hyperparameter values obtained were subsequently implemented in the corresponding regression models. Below is a table summarizing the algorithms, the hyperparameters tuned, the range of values considered and the final optimal value:

Table 2. Hyperparameters optimization

Algorithm	Hyperparameter	Range of values	Optimal value
RandomForest	n_estimators	100, 200, 300, 500	300
	max_depth	None, 10, 20, 30	20
	min_samples_split	2, 5, 10	5
	min_samples_leaf	1, 2, 4	2
	max_features	sqrt, log2, None	sqrt
KNeighbors	n_neighbors	3, 5, 7, 9, 11, 15	7
	weights	uniform, distance	uniform
	metric	euclidean, manhattan	euclidean
GradientBoosting	n_estimators	100, 200, 300	200
	learning_rate	0.01, 0.05, 0.1, 0.2	0.1
	max_depth	3, 5, 7	5
	subsample	0.8, 0.9, 1.0	0.9
XGB	n_estimators	100, 200, 300	200
	learning_rate	0.01, 0.05, 0.1, 0.2	0.05
	max_depth	3, 5, 7	7
	subsample	0.8, 0.9, 1.0	1.0
	colsample_bytree	0.8, 0.9, 1.0	0.9

Source: Own Elaboration.

Selection of Relevant Variables

To identify the most relevant independent variables for each dependent variable, the SelectKBest method is applied. This technique evaluates the statistical relationship between each independent variable and the target variable using predefined statistical tests. Specifically:

- For return prediction, an ANOVA F-test is used.
- For trend prediction, a Chi-square test is applied.

The variables are then ranked based on their scores, and the top K features (determined by the user) are selected for each dependent variable. This approach allows for an

independent feature selection process that is not tied to a specific model. Although cross-validation could be used to determine the optimal value of K , in this study, variables with a score greater than 1 are selected and stored for use in the classification and regression models.

3.7. Proposed Strategies

Not only is it necessary to know the accuracy and level of error of the models, but also the ability of the proposed system to make valuable decisions. To this end, strategies were developed for each target.

3.7.1. Buy Decision

If the trend prediction is bullish for any of the following n weeks, the decision to buy is made. Similarly, with return prediction, the decision to buy is made for any of the following weeks if the return is greater than 0. Next function represents this concept.

3.7.2. Profitability Measurement

Moments are selected when trend predictions were bullish or return predictions were greater than 0. The calculated returns mentioned earlier for each weeks ahead relative to the current week are used to measure profitability based on the percentage changes that would have been obtained by buying in that week and selling in any of the following weeks.

As the system is proposed to be run every week, a strategy is evaluated for each week, where a capital is used to invest each week in which the predictions are bullish. For example, if it is invested for the same week (this is represented by trend1 or return1), a capital of 10,000 is used. If it is invested for 2 weeks ahead (trend2 or return2), half of the capital is used, i.e., 5000, with the aim of using the remaining capital the following week to invest again in the case of obtaining a bullish prediction and sell within 2 weeks. Similarly, this is done for 3 weeks ahead where the capital is divided by 3, and for 4 weeks ahead where the capital is divided by 4, this for a purchase every week. Cumulative returns are used to obtain the final return always using simple interest. Next functions represent the calculation for each strategy.

$$\text{Investment per week} = \begin{cases} C & \text{for trend1 o return1} \\ \frac{C}{2} & \text{for trend2 o return2} \\ \frac{C}{3} & \text{for trend3 o return3} \\ \frac{C}{4} & \text{for trend4 o return4} \end{cases}$$

where C is the initial capital (in this case, 10,000).

The cumulative return R is calculated using the simple returns obtained each week:

$$R = \sum_{i=1}^n \left(\frac{\text{Profit}_i}{C_i} \right)$$

where Profit_i is the profit or return given by the asset each week i y C_i is the capital invested i .

3.7.3. Buy and Hold

Finally, the performance that would have represented holding the purchase made at the beginning of the test data set until the end of the test data set is estimated, taking the cumulative return until the last period without executing any buy or sell decisions.

3.8. Dictionary of Final Dataset

Table 3 shows the final variables required to obtain the research results.

Table 3. A View of the dictionary of final dataset

Variable	Meaning
Date	Contains end of weeks dates from 2012-07-01 to 2023-12-31.
open	Open price of current week.
trend1, trend2, trend3 and trend4	Trend in 1, 2, 3, and 4 week(s) with respect to the opening of the current week.
return1, return2, return3 and return4	Return in 1, 2, 3, and 4 week(s) with respect to the opening of the current week
Negative-Neutral-Positive Vader	Quantity of news based on its sentiment with vader of past week (3 variables)

Variable	Meaning
Negative-Neutral-Positive Finbert	Quantity of news based on its sentiment with Finbert of past week (3 variables)
Headline Sentiment Vader	Average news sentiment for the week with vader of past week.
Headline Sentiment Finbert	Average news sentiment for the week with Finbert of the past week.
Average Headline Sentiment Vader	Average news sentiment over the last ten weeks with vader of past week.
Average Headline Sentiment Finbert	Average news sentiment over the last ten weeks with Finbert of past week.
Topic(n)	Topic apparition in past week (38 variables).

Source: Own Elaboration.

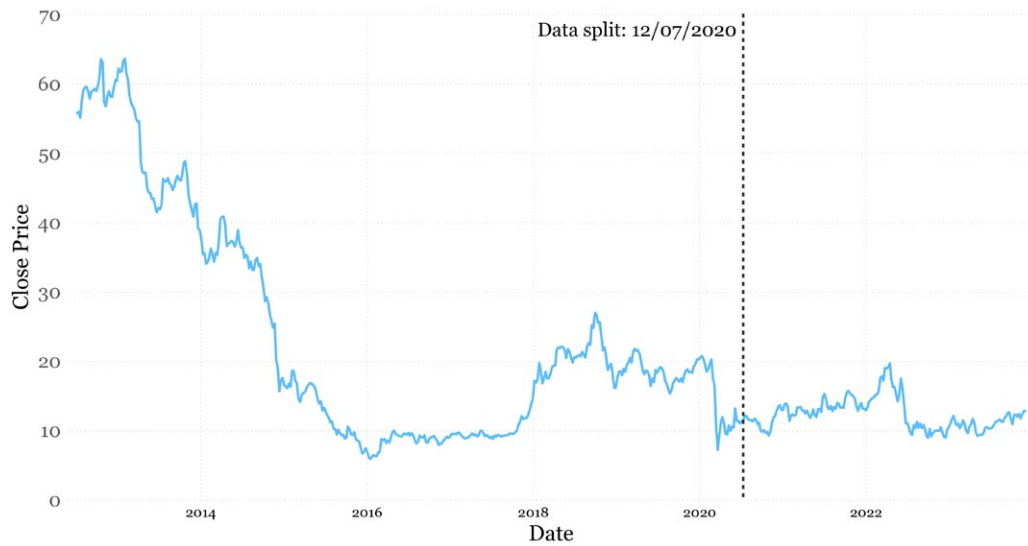
4. Results and Discussion

This section describes the context of the asset price, the topics found, the relevant variables for each target variable, the results obtained from the models and the proposed strategies, both for trend and return variables, and buy and hold strategy.

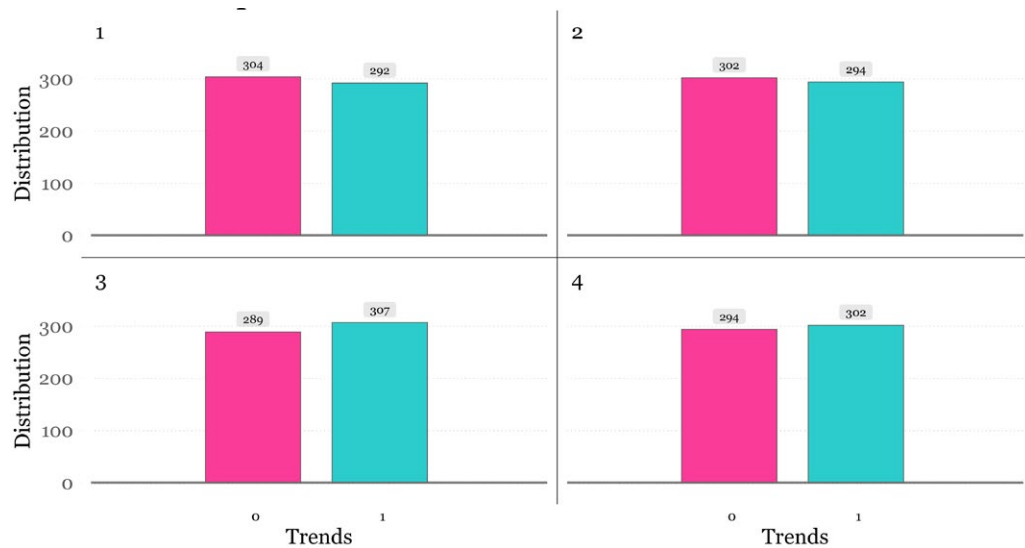
4.1. Ecopetrol Stock price data

The following graphs illustrate the context in which the research was conducted and the characteristics of the dependent variables.

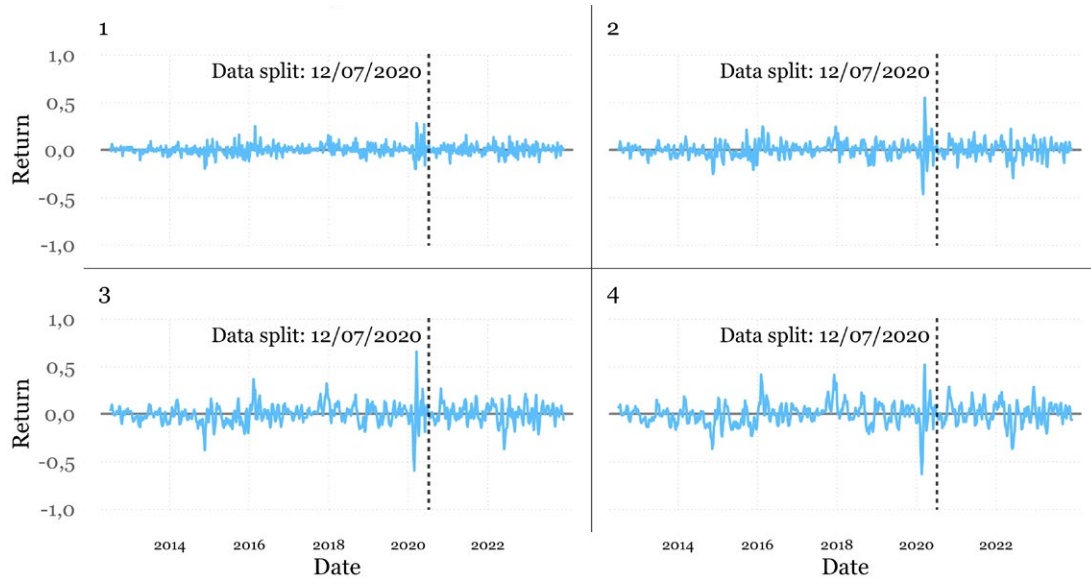
Figure 1 highlights the training and test price data sets, divided by the vertical line at the previously mentioned date. The price data for the training period exhibited a bearish trend until 2016, followed by a phase of consolidation. Similarly, the price data for the testing period also showed a phase of consolidation, providing a well-rounded sample of both bearish and bullish market conditions. As shown in Figure 2, the distribution of trends is balanced, so it was not necessary to use any technique to handle sample imbalance. In Figure 3 returns for each return target is shown and, as could be expected, the higher the time horizon, the price data is more volatile.

Figure 1. Split of train and test of stock price data

Source: Own Elaboration.

Figure 2. Distribution of trend target after weeks 1, 2, 3 and 4

Source: Own Elaboration.

Figure 3. Returns of return target after weeks 1, 2, 3 and 4

Source: Own elaboration with Yahoo Finance data.

4.2 Relevant Features

Table 4 presents the variables selected for each target variable using the SelectKBest method. Since this technique evaluates the statistical relevance of each feature independently of the predictive models, it helps filter out less informative variables before training. The selection of sentiment-related features is particularly crucial, as different sentiment calculation methods may impact prediction outcomes depending on the forecasting horizon or whether the target variable is modeled as return or trend.

In the next section, the impact of topic-related variables is analyzed.

Table 4. Relevant features for each target

Targets	Features
trend1	Average Headline Sentiment Vader, Headline Sentiment Vader, T7, T24, T10, T33, T19, T31, T17
trend2	Average Headline Sentiment Vader, Headline Sentiment Vader, Negative Vader, Neutral Finbert, Neutral Vader, T5, T2, T28, T7, T24, T16, T23, T29, T9, T19, T13, T20, T17
trend3	Average Headline Sentiment Vader, Headline Sentiment Vader, Neutral Finbert, Neutral Vader, T2, T16, T4, T29, T32, T18, T22
trend4	Average Headline Sentiment Finbert, Average Headline Sentiment Vader, Neutral Finbert, Neutral Vader, T14, T24, T16, T11, T23, T26, T29, T19, T37, T18
return1	Headline Sentiment Vader, T5, T7, T21, T14, T24, T12, T4, T23, T29, T33, T19, T20, T30, T32, T17
return2	Headline Sentiment Vader, Negative Finbert, Negative Vader, Neutral Vader, T2, T8, T14, T24, T16, T12, T23, T29, T20, T30, T18, T17
return3	Average Headline Sentiment Vader, Negative Finbert, Negative Vader, Neutral Vader, T2, T7, T14, T24, T16, T12, T11, T23, T29, T38, T33, T19, T20, T17, T22
return4	Average Headline Sentiment Vader, Headline Sentiment Vader, Negative Finbert, Negative Vader, Neutral Vader, T1, T2, T36, T7, T14, T24, T16, T12, T4, T23, T26, T29, T33, T19, T37, T17, T22

Source: Own Elaboration.

4.3 Topics Found

After training the BERTopic model, the topics found are shown in Table 5. The column Topic represents the number of the topic that is related with each T(n) variable shown before. The column Name shows the name of the topic after processing the key words given by BERTopic. Frequency of appearances are related with how many times this topic appears as a relevant feature from SelectKBest method in any of the target variables, which shows those relevant topics for Ecopetrol Stock price in general. Impact in n weeks is related with the time horizon that the topic has had a significant impact in the past.

Table 5. Topics found after building the BERTopic model in the training set

Topic	Name	Frequency of appearances	Impact in <i>n</i> weeks
0	Ecopetrol Financial Performance Analysis	0	-
1	Colombia's Fracking Policy and Stock Market Impact	0	-
2	Oil Refinery Operations and Environmental Concerns	5	2, 3
3	ECP Echeverry Plan for Export Strategy	0	-
4	Fuel Price Fluctuations and Natural Gas Market	0	-
5	USO Strike and Labor Disputes in Putumayo	2	2
6	Corporate Social Responsibility Projects Overview	0	-
7	Rubiales and Pacific Field Reversion Decisions	5	1, 2
8	Pipeline Security and Contingency Planning	1	2
9	Continuing Corruption Scandal at Reficar	0	-
10	Perception and Outlook on Progress	0	-
11	ANLA and ANH License Issues	2	2
12	Geopark's Role in Industry and Community Relations	4	4
13	Security Challenges in Cao, ELN, and Limn Regions	0	-
14	Environmental Concerns along the Magdalena River	5	4
15	Security and Environmental Issues Reporting	0	-
16	Blockades in Barrancabermeja and Mininterior Response	6	2, 3
17	Analysis of Colcap Index and Market Performance	6	3, 4
18	Bayn, ECP, and Bayon Perspective Analysis	3	3, 4
19	Parex Reserves Growth and Agreements	6	4
20	Tax Reform Impact and Fiscal Works	4	4
21	Energy and Mining Sector Transition to Renewables	0	-
22	Lizama Outcrop Scandal and Incident Investigation	3	3
23	Monthly Reports from Inner Circle	6	3, 4
24	Local Hiring Practices and Labor Trends	7	1, 2, 4
25	Referendum in Tauramena and Popular Court Decisions	0	-
26	Reviewing Roads Infrastructure in Casanare	2	2, 4

Topic	Name	Frequency of appearances	Impact in n weeks
27	Royalty System Projects and Departamental Issues	0	-
28	Minminas and EITI Efforts Addressing Challenges	1	2
29	Campetrol's Exploration Forecasts for 2016	7	2, 3, 4
30	ACIPET's Educational Initiatives and Layoff Information	0	-
31	Bioenergy and Ethanol Production Challenges	0	-
32	Hocol's Operations in Macarena, Ballena, and Chuchupa	2	2
33	Smuggling and Fuel Theft Issues	4	1, 2, 4
34	Seismic Exploration and Water Regulation	0	-
35	Canacol's Bet on Shale Exploration in 3Q13	0	-
36	Tierra Gran Llanos Energy Asset Analysis	1	4
37	Field Suspension Decisions and Inactivity	2	2, 4
38	Enit's Achievements and CNE Sale Status	1	4

Source: Own Elaboration.

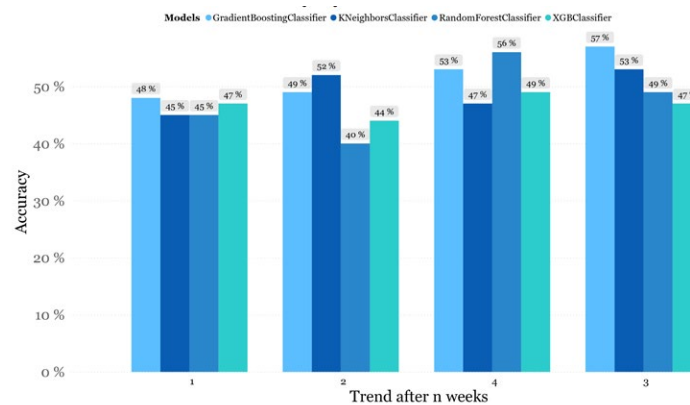
4.4 Trends predictions results

After training and testing the ml classification models, performance results from models were obtained. Figure 4 shows the accuracy of models for each trend target. As can be seen, predictions of trend after 3 and 4 weeks had a better accuracy in comparison with trend after 1 and 2 weeks. Figure 5 shows the profit of the strategy proposed for each target and, similar with accuracy results, investing and selling after 3 and 4 weeks gave a better result that doing it in less time like 1 and 2 weeks. Figure 6 shows the results of the accumulated returns of the proposed strategies using the trends. It can be seen that profits are constant over time and bullish in the case of 3 and 4 weeks ahead, also constant but bearish in the case of 1 and 2 weeks ahead.

Table 6 presents results with more details. The Target column displays the target variables and Model column the model that was trained and tested, the accuracy shows how good the model was. Precision 0 and 1, and also Recall 0 and 1, the results for bullish and bearish trend. Support 0 and 1 show the distribution of trends

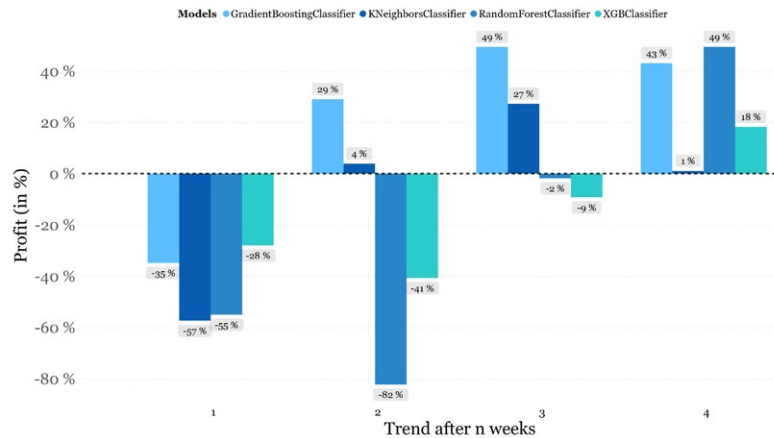
followed by Profit that shows the final return. The best model for predicting trends was the GradientBoostingClassifier in trend3, that was tested with 79 bearish trends and 95 bullish trends, with an accuracy of 0.57, precision for predicting bearish trends of 0.52 and 0.62 for bullish trends. Recall for bearish and bullish trends was 0.59 and 0.55 respectively, which shows that the model had a good performance and was not overfitted, with a final profit of 49.4%. Other models like RandomForestClassifier for trend4 also had a good performance, where its final profit was 49.41%

Figure 4. Accuracy of models for each trend target

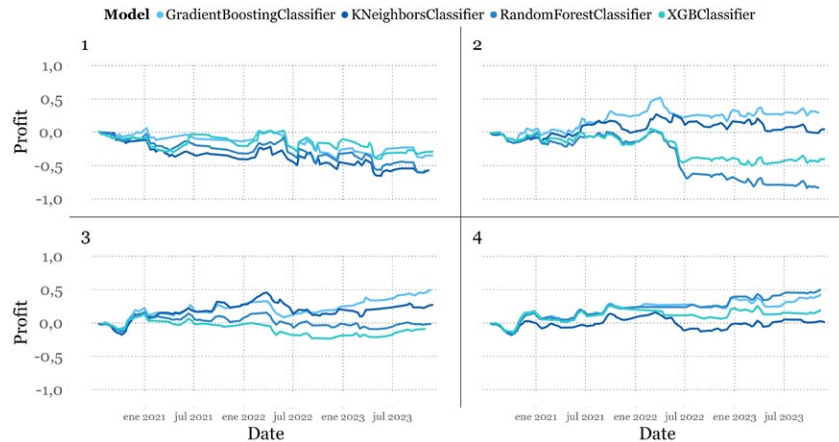


Source: Own Elaboration.

Figure 5. Profit of models for each trend target



Source: Own Elaboration.

Figure 6. Profit over time of models for each trend target

Source: Own Elaboration.

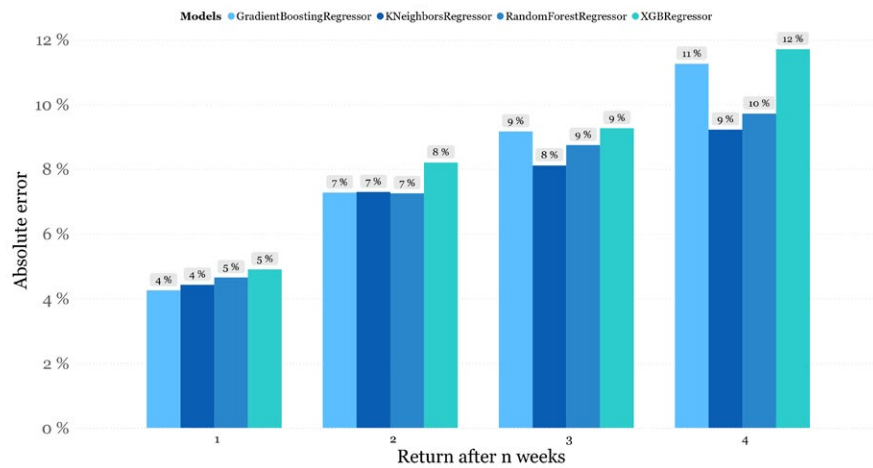
4.5 Returns predictions results

As was mentioned before, it was necessary to see if the data is also useful for predicting returns. The results given by regression models for predicting return targets had a lower performance; however, it is still useful to use this approach for predicting price. Figure 7 shows that error increases if time horizons prediction are longer. In Figure 8 it can be seen that profit is also higher if 3 or 4 weeks' time horizons are used, in comparison with 1 and 2. However, results of profit after 1 and 2 weeks are better in comparison with the trends ones. Figure 9 shows that historical profit from predicting returns after 1 week are slightly more bullish in comparison with after 2 weeks, and 3 and 4 weeks are considerably more bullish.

In Table 7 the results were also presented in more details. Like in table 6, the two first columns represent the models and targets tested. Absolute error represents the average error of predicting results from the models in different time horizons. The profit column also represents the final return of each model tested. In this case, the best model was the RandomForestRegressor for predicting returns after 4 weeks with a profit of 33,11% and an absolute error of 9,33%, followed by the model Gradient-BoostingRegressor with a profit of 22,62% and absolute error of 9,16%. This shows that these last two models can be useful at predicting trends and returns.

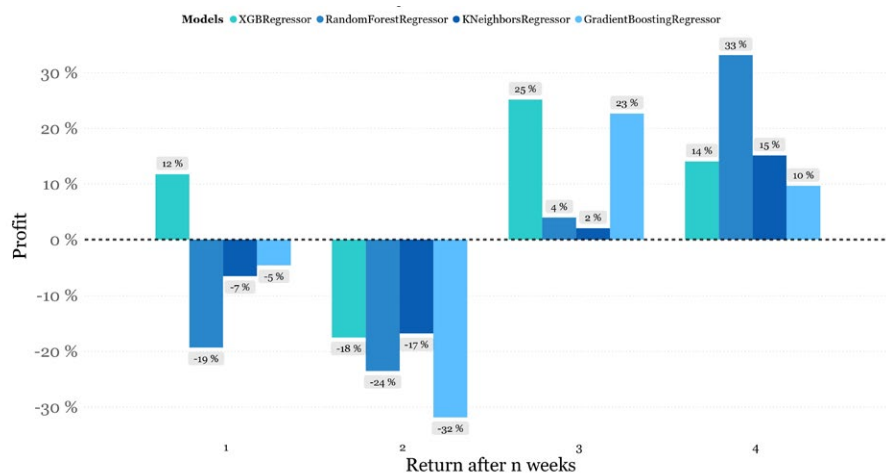
Figure 10 shows historical predicted values and real values over time by each model and target. The number followed by the model's name represents the return after n weeks. As can be seen, the models can capture the returns if time horizons are lower.

Figure 7. Absolute error of models for each return target

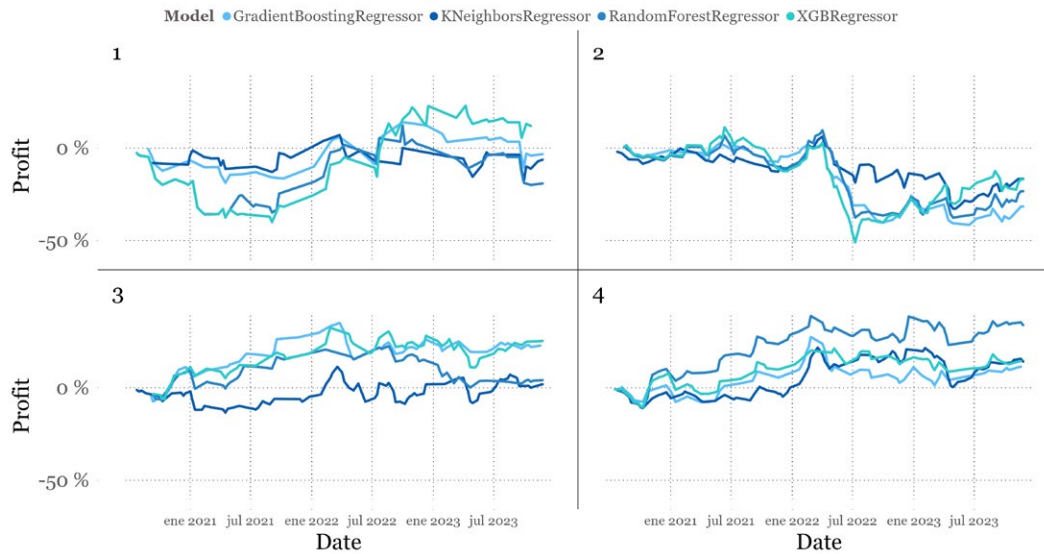


Source: Own elaboration.

Figure 8. Profit of models for each return target



Source: Own elaboration

Figure 9. Profit over time of models for each return target

Source: Own elaboration.

Table 6. Model's Performance Metrics for each trend target variable

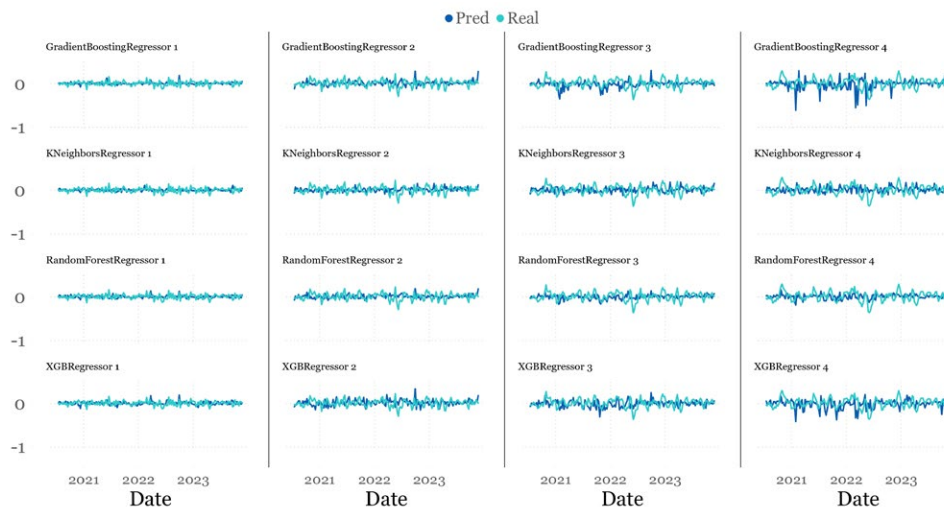
Model	Target	Accuracy	Precision_0	Precision	Recall_0	Recall_1	Support	Support	Profit
RandomForestClassifier	trend1	0.45	0.46	0.43	0.49	0.38	87	87	-55.01
KNeighborsClassifier	trend1	0.45	0.45	0.44	0.49	0.40	87	87	-57.35
GradientBoostingClassifier	trend1	0.48	0.48	0.47	0.64	0.31	87	87	-34.83
XGBClassifier	trend1	0.47	0.46	0.46	0.53	0.39	87	87	-28
RandomForestClassifier	trend2	0.4	0.38	0.39	0.27	0.51	90	84	-82.25
KNeighborsClassifier	trend2	0.52	0.54	0.50	0.46	0.58	90	84	3.83
GradientBoostingClassifier	trend2	0.49	0.52	0.48	0.34	0.65	90	84	28.99
XGBClassifier	trend2	0.44	0.44	0.44	0.30	0.60	90	84	-40.72
RandomForestClassifier	trend3	0.49	0.44	0.53	0.49	0.48	79	95	-1.82
KNeighborsClassifier	trend3	0.53	0.48	0.57	0.49	0.55	79	95	27.21
GradientBoostingClassifier	trend3	0.57	0.52	0.62	0.59	0.55	79	95	49.4
XGBClassifier	trend3	0.47	0.44	0.52	0.59	0.37	79	95	-9.24
RandomForestClassifier	trend4	0.56	0.52	0.60	0.55	0.57	80	94	49.41
KNeighborsClassifier	trend4	0.47	0.42	0.51	0.42	0.50	80	94	1.02
GradientBoostingClassifier	trend4	0.53	0.5	0.57	0.46	0.60	80	94	43.07
XGBClassifier	trend4	0.49	0.45	0.54	0.49	0.51	80	94	18.22

Source: Own elaboration

Table 7. Results of models for each return target variable

Model	Target	Absolute error	Profit
RandomForestRegressor	return1	4,65	-19,36
KNeighborsRegressor	return1	4,42	-6,55
GradientBoostingRegressor	return1	4,25	-4,62
XGBRegressor	return1	4,9	11,72
RandomForestRegressor	return2	7,25	-23,56
KNeighborsRegressor	return2	7,29	-16,84
GradientBoostingRegressor	return2	7,27	-31,87
XGBRegressor	return2	8,2	-17,56
RandomForestRegressor	return3	8,74	3,97
KNeighborsRegressor	return3	8,11	2,06
GradientBoostingRegressor	return3	9,16	22,62
XGBRegressor	return3	9,26	25,1
RandomForestRegressor	return4	9,71	33,11
KNeighborsRegressor	return4	9,22	15,1
GradientBoostingRegressor	return4	11,25	9,65
XGBRegressor	return4	11,7	14,01

Source: Own elaboration.

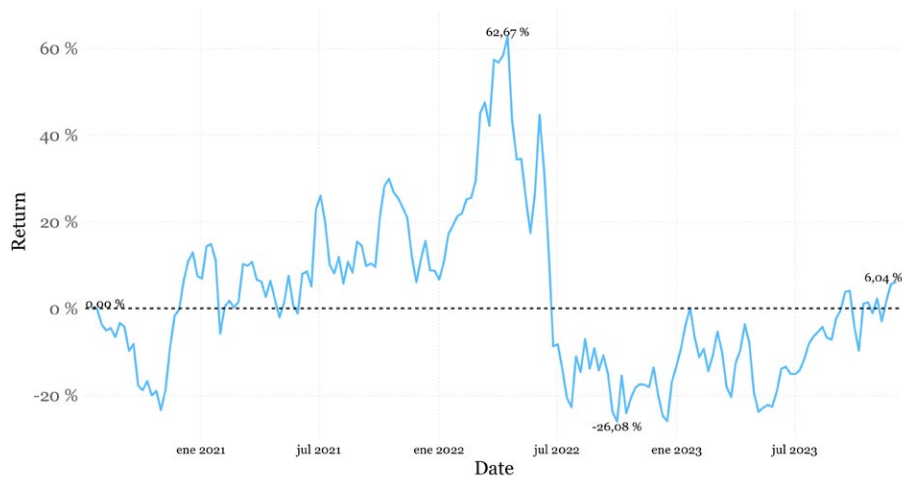
Figure 10. Predictions and real values from results for each return target

Source: Own elaboration.

4.6 Buy and Hold results

In Figure 11, the return without applying any strategy can be observed. A max return of 62.47% was reached and also a minimum of -26.08%. The final return was 6.04%, which is considerably lower than the return provided by the strategies derived from the proposed system.

Figure 11. Profit over time from buy and hold strategy



Source: Own elaboration.

5. Conclusions and future work

This research article presents a novel approach to predicting financial asset movements, specifically focusing on Ecopetrol stock, encompassing both trends and returns over various timeframes. The aim was to identify the potential of data and models, as well as to develop a profitable system within a specified period. By incorporating various sentiment calculations using the Vader and Finbert algorithms, along with the emergence of topics generated by a BERTopic model trained on historical data, relevant variables are identified for each of the target variables.

The metrics gathered from the results of these models and the strategy proposed for each timeframe reveal that utilizing news headlines through the proposed methods demonstrates better performance in 3 and 4-week periods compared to 1 and 2 weeks

(considering the opening price following the appearance of the news in the previous week). The best model found for predicting trends was GradientBoostingClassifier for week 3, with a profitability of 49.4% and an accuracy of 56%. For returns, the model RandomForestRegressor showed a profitability of 33.11% with an error of 9.71%, significantly outperforming the return generated by a buy-and-hold strategy, which was 6.04%. Based on these results, it is concluded that it is more useful, in terms of profitability, to make decisions based on trend predictions rather than returns, and also it is useful to incorporate topics and different calculations of sentiment.

While the overall accuracy of the models is limited—only slightly better than random guessing in a binary classification setting—the main objective of this study is to explore a novel methodology for integrating sentiment analysis into stock trend prediction. The method proposed in this study enhances understanding of topic modeling techniques and sentiment analysis in financial markets. It is recommended for future studies to replicate the methodology on other financial assets and consider not only headlines but also the body of articles. Additionally, conducting a more detailed analysis of the impact of topics on asset prices to identify potential news in the future is suggested.

6. References

1. Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. <https://arxiv.org/abs/1908.10063>
2. Balaneji, F., & Maringer, D. (2022). Applying sentiment analysis, topic modeling, and xgboost to classify implied volatility. *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, 1–8. <https://doi.org/10.1109/CIFEr52523.2022.9776196>
3. Chen, W., Rabhi, F., Liao, W., & Al-Qudah, I. (2023). Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study. *Electronics*, 12(12), 2605. <https://doi.org/10.3390/electronics12122605>
4. Correia, F., Madureira, A., & Bernardino, J. (2022). Deep neural networks applied to stock market sentiment analysis. *Sensors*, 22(12), 4409. <https://doi.org/10.3390/s22124409>

5. García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with latent dirichlet allocation. *Applied Intelligence*, 53(16), 19610–19628. <https://doi.org/10.1007/s10489-023-04452-4>
6. Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. <https://arxiv.org/abs/2203.05794>
7. Iguarán Cotes, J. (2019). Aplicación de redes neuronales para predecir el precio de acciones en la bolsa colombiana. <http://hdl.handle.net/1992/44483>
8. Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178, Article 115019. <https://doi.org/10.1016/j.eswa.2021.115019>
9. Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 13(7), 3433–3456. <https://doi.org/10.1007/s12652-020-01839-w>
10. López-Gaviria, J. I. (2019). Predictibilidad del mercado accionario colombiano. *Lecturas De Economía*, (91), 117–150. <https://doi.org/10.17533/udea.le.n91a04>
11. Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock prediction by integrating sentiment scores of financial news and mlp-regressor: A machine learning approach. *Procedia Computer Science*, 218, 1067–1078. <https://doi.org/10.1016/j.procs.2023.01.086>
12. Monroy-Perdomo, L., Cardozo-Munar, C. E., Torres-Hernández, A. M., Tena-Galeano, J. L., & López-Rodríguez, C. E. (2022). Formalization of a new stock trend prediction methodology based on the sector price book value for the colombian market. *Heliyon*, 8(4), Article e09210. <https://doi.org/10.1016/j.heliyon.2022.e09210>
13. Palacio Roldan, J. (2022). Modelo de recomendación para inversión en acciones colombianas pertenecientes al índice colcap basado en análisis técnico y sentimiento del mercado local.

14. Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, 7(2), 1–22. <https://doi.org/10.3390/ijfs7020026>
15. Wang, L., Huang, C., Gao, C., Ma, W., & Vosoughi, S. (2023). Joint latent topic discovery and expectation modeling for financial markets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13937, 45–57. https://doi.org/10.1007/978-3-031-33380-4_4

