School of Social Sciences

2024

Quantitative Text Analysis with Global News Dataset

Using News Data to Predict Stock Market Movements

Student ID: 11320335 | Supervisor: Dr Yan Wang

An extended research project report submitted to the University of Manchester for the degree of Master of Science in Data Science (Social Analytics) in the Faculty of Humanities

List of Contents

List of Tables

List of Figures

Abstract

This study explores the use of Quantitative Text Analysis (QTA) on the global news dataset to forecast stock price movements based on political and economic news. In the modern society, practically every corporate event, procedure, and decision is recorded in legal documents, business and financial news, or internal written communications. Data is frequently called the oil of the twenty-first century (The Economist, 2017). Investors, financial institutions, and other important stakeholders must create sophisticated automated methods that can extract valuable insights from massive data sets because humans find it difficult to recognise pertinent, complicated patterns hidden therein. Unlike the average person who can process 5-10 articles daily, an AI model can analyse thousands to millions, limited only by hardware. The focus is to research whether QTA can help investors make informed stock trading decisions by establishing links between economic news and stock market behaviour.

Declaration

I declare that this dissertation is entirely my own original work, except where explicit references are provided. No part of the work presented in this dissertation has been submitted in support of an application for any other degree or qualification at this or any other university or institution.

Copyright

I. The author of this dissertation (including any appendices and /or schedules to this dissertation) owns certain copyright or related rights in it (the 'Copyright') and she has given the University of Manchester certain rights to use such Copyright including for administrative purposes.

II. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of works in the dissertation, for example graphs and tables ("Reproductions") which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permissions of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see, https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Dissertation restriction declarations deposited in the University Library, and The University Library's regulations (see, https://www.library.manchester.ac.uk/about/regulations/_files/Library-regulations.pdf).

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Yan Wang, for her continuous support and insightful feedback throughout this dissertation. Her expertise and encouragement have been instrumental in shaping the direction of my research. I am grateful for the opportunity to learn under her supervision.

I also extend my sincere thanks to Prof. Mark Elliot, my programme director, for his invaluable guidance throughout this course. His commitment to the program and his dedication to fostering an environment of academic growth have been greatly appreciated.

I am deeply grateful to all my colleagues and friends who have supported me along this journey. Special thanks to those who went beyond academia and assisted me personally while I've been an international student.

Lastly, I want to express my heartfelt appreciation to my family for their unwavering encouragement and belief in me during this rigorous, yet stimulating, period, which contributed to my strength and motivation.

1 Introduction

Various forms of news being consumed everyday greatly influence how we perceive the world and how people make certain decisions. Particularly, in the world of economics and finance, the vast amount of such data being produced regularly creates opportunities for key stakeholders, such as economists and investors, to understand the political communication and economic narratives to make sound decisions. QTA can help the financial trading sector to analyse financial and political news to identify key ongoing themes and topics that provide a macro perspective, unpacking linguistic features of the news and inputting the information into their automated trading systems for the purpose of making better investment decisions. QTA provides the foundation to digest news and textual data at large scale to set grounds for performing forecasting analysis in a way that truly enhances the predictive power of investors. This approach involves filtering relevant news based on their connection to political and economic events and employing Natural Language Processing (NLP) techniques to extract indicators from the news articles that can be specified to assess their potential impact on the stock market. For instance, an announcement by the finance minister regarding an increase in oil prices might escalate supply chain expenses for companies like Walmart, leading to higher product prices. This could negatively affect stock prices due to anticipated inflation, dwindling investor confidence, and reduced consumer spending. Key focus of this research is to understand how certain topics and themes presented in the news promotes market to move a certain way, which can be represented as a proxy of sentiments, and we use topic modelling, to create topical indicators based on dictionaries and a ratio of news volume for stocks, i.e., CBOE Volatility Index (VIX). The project aims to correlate news with historical stock market trends and seek to predict the future outcomes.

2 Background and previous theories

Extensive research in the field of behavioural finance have demonstrated that investors do respond to news. However, they typically have a stronger inclination to make an investment decision based on negative news as opposed to positive news, due to the very nature of human psychology that influences trading attitudes of many investors (Baumeister et al., 2001).

2.1 Evolution of Entity Recognition

The initial idea of this spans back to 90s, when Rau (1991) suggested an algorithm to automatically extract company names from financial news to build a database for querying. With MUC-6, a shared goal to identify not only types, like person, place, and organization, but also numerical mentions, such as time, money, and percentages, the work gained more attention. The purpose of computational linguistics is to use statistical rule-based methods to parse and describe natural language. To achieve this, the unstructured text must be accurately tokenised, part-of-speech tags (POS) must be applied, and a parse tree detailing the relationships and general structure of the sentence must be constructed. This identifies the boundaries of strings in a large text and classifies it into categories such as Organization, Person, or a Location Linguists created rules that characterise common patterns for named things using this knowledge. Machine learning techniques that make use of the previously described tags and so-called surface features quickly supplanted handcrafted rules. Syntactic traits like character count, capitalisation, and other deduced information are described by these surface aspects too. Hidden Markov models and conditional random fields are the most often used supervised learning techniques for such applications, because they can extract probabilistic rules from sequence data (Bikel et al., 1997). Although it works precisely, yet there's a key challenge of gathering and feeding a lot of annotated training data into the supervised learning. Text data can be automatically labelled by bootstrapping techniques by utilising a seed set of entity names. These semi-supervised techniques accomplish this by automatically annotating extra data by utilising contextual information to identify instances of these seed entities in the text. To advance further, deep learning techniques have also been popular recently, being favourable for not needing complex pre-processing, feature engineering, or dependency parsing to extract meaningful patterns from unstructured text.

The work of Yadav and Bethard (2018) is notable as they argue that recurrent neural networks, in particular, have improved the performance of Named Entity Recognition (NER) compared to traditional mechanisms, as RNNs process whole token or letter sequences instead of isolated tokens. Their study revealed that even without using any outside resources or feature engineering, NN models on news corpora outperformed the prior state-of-the-art by 1.59% in Spanish, 2.34% in German, 0.36% in English, and 0.14% in Dutch language.

2.2 Classic technique of Sentiment Analysis

The work of Li (2006) investigated the contemporary correlations, between the qualitative data taken from publicly accessible document texts and future stock returns, from different angles. Based on his research, poor yearly earnings and stock returns are predicted by the terms such as "risk" and "uncertain" in company annual reports. Such information set the tone of qualitative data being presented in the news and is interpreted as a reaction to "risk sentiment" by key stakeholders. Nevertheless, these news articles and other textual sources are based on unstructured formats. Hence, modern Natural Language Processing (NLP) techniques are required to extract business insights in a structured way. Among such techniques is Named Entity linking (NEL), that extends NER and assigns correspondence among the entities to build a knowledge graph that represents relationships among entities and how they change over time (Shen et al., 2015). Traditionally, financial solutions, that can back decision-making for investors, companies, and financial institutions, have been developed based on sentimental analysis by capturing sentiments presented in news data or other textual sources quantifying the economic and political dynamics among financial entities of interest, e.g., price returns, uncertainty, volatility, and alike variable indicators. Arratia et al. (2021) explains the conventional approach to creating such forecasting models using textual data:

1. **Creating and processing textual corpuses**, by gathering textual information (i.e., news articles), preprocessing to remove irrelevant information and applying Named Entity Recognition (NER) methods to classify terms in documents.
2. **Sentiment analysis**, either using unsupervised learning methods such as sentiment lexicon that recognizes grammatical patterns, or by using supervised ML techniques trained on a labelled text.

3. **Aggregating sentiment scores**, to track how people behave over time. The scores are adjusted using weights, to order certain topics by their importance, and filters are applied, i.e., moving averages, to smooth out fluctuations and highlight overall patterns. This can be expressed as:

$$S_t(\lambda, k) = \sum_{n=1}^{N_t} \beta_n S_{n,t}(\lambda, k)$$

Here, β_n are the weights that determine how sentiment scores are aggregated. The resulting sentiment indicators, $\{S_t: t = 1 \ldots T\}$, based on lexicon λ, defines a specific sentiment for target k.

4. **Modelling** the sentimental indicators as exogenous features to finally build forecasting models, that can predict stock price movements and aid investor decision-making.

2.3 Ascendance of Topic Modelling

However, Pang and Lee (2008) have highlighted the issue of conflicting sentiments in a news article, making it difficult for the model to evaluate between sentiment polarity. Sentiment is often expressed more subtly, and often in a subjective way, making it challenging to identify based solely on individual terms within a sentence or document. While sentiment analysis works well for individual sentences or brief textual segments, they can be problematic for analysing the entire news article if the content covers multiple perspectives surrounding a topic. Since the accuracy of sentiment classification can be influenced by the domain of the content it is applied to, topic modelling offers a more stable approach to holistically view news content. By focusing on the thematic content rather than sentiment alone, topic modelling can better handle domain-specific variations and complexities.

Emphasizing on the paradigm of topic modelling, Dierckx et al. (2021), demonstrates potential for this task at capturing the holistic tone of news as a proxy to sentiments. Topic modelling algorithms can automatically uncover hidden themes and topics within the data, by using statistical features of language to classify words that are related in order to find a set of subjects within a collection of documents. They then evaluate the range of subjects included in a document to provide a description of it. In other words, they reveal how much of each

subject is covered in a particular news article and primarily convey a broader narrative. Linking this back to Li (2006), particular keywords topics that are associated with risk or uncertainty causes the stock market to behave in a specified way because investors respond to the underlying themes and topics within these articles, which shape their sentiment and, ultimately, market behaviour.

2.4 Aims of this research

The presence of certain topics is indicative of the market movements, which sets the foundation for this research, i.e., using global news data to perform topic modelling as part of the quantitative text analysis, alongside historical VIX stock data to explore relationship between news and the movements of financial markets. Here's what the research aims to answer:

- How can advanced topic modelling techniques, such as BERT and LDA, enhance the categorization of economic and political news articles for predictive analysis, and how do they compare to each other?
- How do different text representation methods, including TF-IDF, Bag of Words, Word2Vec, and CountVectorizer impact the performance of machine learning models in NLP tasks for document classification?
- How does the quality and diversity of news data sources impact the accuracy and reliability of predictive models for stock market movements, and is a tailored approach necessary for different types of news articles to achieve a scalable model?
- To what extent does data derived from news sway market sentiment more significantly than external factors, such as environmental disasters or public health crises, and how does this influence the accuracy of predictive models for stock market movements?

3 Methodology & Data

To check whether sentiments drive an impact on the stock market, topics revolving around such news needs to be determined to observe how does stock market perform in response to it, as certain topics, for example, related to violence or bad news may lead to negative sentiments. To achieve this, topic modelling techniques would be utilized to define the main themes of each news article, grouping them by dates and then mapping those themes to the CBOE Volatility Index (VIX) historical data for the date range that news articles exist within. VIX stock is being used for this analysis as its widely acknowledged for measuring market risk and investor sentiments. Hence, it's ideal to be used as a proxy for determining whether certain themes or topics within the news drive investor confidence and sentiments in the stock market.

Python has become increasingly popular for data science and machine learning, and the recent upsurge in the number of libraries available for NLP, topic modelling, and time-series analysis makes it ideal for research analysis on QTA. **Appendix A** lists all the packages used in this research and Figure 3.1 presents the complete workflow of methodology.

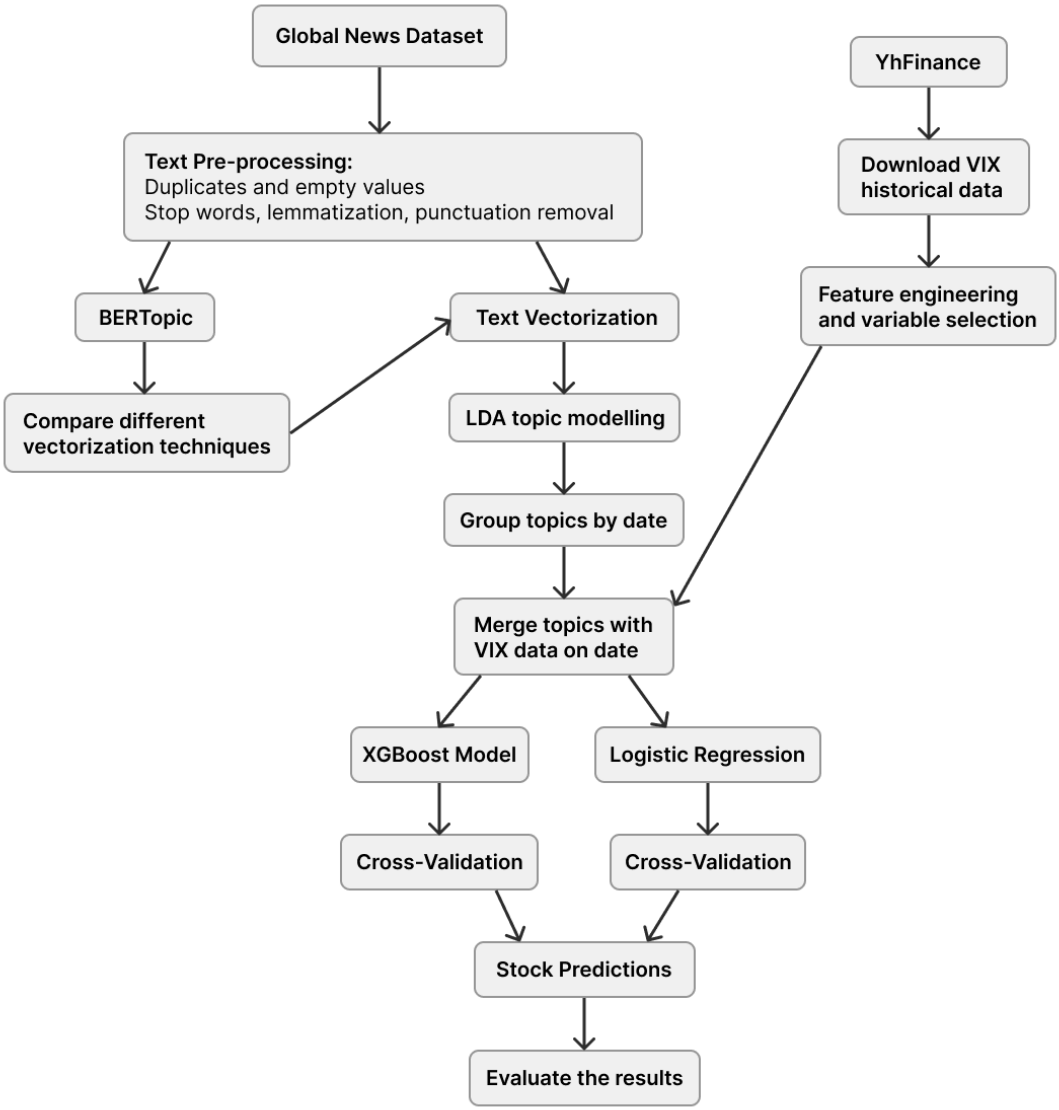

Figure 3.1: - Process flowchart summarizing the methodology

3.1 Data collection and preprocessing

The Global News Dataset is used to conduct this research. This dataset comprises news articles collected over the period of roughly two months, from November 1st 2023 to October 29th 2023, using the NewsAPI by Kumar Saksham (2023). The data is open source and aims to support QTA, sentimental analysis, and other NLP applications (Kumar Saksham, 2023).

The dataset contains 105,375 rows (news articles) and 12 columns (attributes), detailed in **Appendix B**. The relevant columns in question are 'title', 'full_content', 'published_at', and 'description' representing the title, actual content of the news article, datetime it was published, and a short description, respectively. However, the data presented few key issues that required preprocessing before carrying out analysis. Firstly, rows containing any duplicates or missing values were dropped since unlike numerical figures, it's not possible to impute textual data. This left the dataset with 98,488 unique news articles. Secondly, in order for the computer algorithms to understand and interpret, a crucial step was to transform the textual representation of qualitative data into quantitative numerical representation. This starts with identifying relevant news and creating usable information, such as how many times an entity is mentioned in the news or a few keywords that appear frequently, as well as the context in which they are communicated. This is performed by first filtering to remove extraneous words or symbols and organising the article into one or more time series and vectors of numeric integers (Lucas et al., 2015). This process is known as word embeddings and can be achieved in a couple of ways. It also aids addressing one of the research questions for this analysis – "How do different text representation methods, including TF-IDF, Word2Vec, and CountVectorizer, impact the performance of machine learning models in NLP tasks for document classification?"

To evaluate different methods, the initial step was to preprocess the text before applying any vectorization technique and using data for topic modelling to observe how well each method has performed. To begin with the preprocessing phase, it's important to consider the right attribute among 'title', 'description', and 'full_content', for assessing news article to aid stock market decisions. The average number of words each column contains, 'title', 'description', is 11, 31 and 511, respectively. Considering that 'description' has an ideal text length, i.e., efficient for both time and complexity of the algorithm (Devi et al., 2011), it provides a great starting point to aid the analysis. Here are the preprocessing steps that took place:

1. First step was **tokenization**, breaking text into smaller units called tokens, i.e., words, phrases, or symbols. For example, the sentence “Oil prices expected to decrease”, would break into: “Oil”, “prices”, “expected”, “to”, “decrease”. This is required to convert each word into a numeric representation later using vectorization methods.
2. Next, to reduce each of these words to their root form, **lemmatization** considers the context and meaning of words and converts each word into base form. For example, the words “expecting”, “expected”, “expectation” represent the same meaning and in context of news, having multiple occurrences of same thing drive models focus on it even if it’s not relevant, combining such terms would be helpful at reducing the input’s dimensions (Lucas et al., 2015). So, to convert it to base form, it would become “expect”.
3. To overcome the issue of **stop words**, e.g., "the," "is," "at", and punctuation, that often occur frequently but do not carry significant meaning for analysis, such words are eliminated.

Since the textual data is now cleaned, text can be converted into numerical representations using various vectorization methods and using topic modelling alongside to evaluate the performances of each technique.

3.2 Text vectorization

In our case, Bidirectional Encoder Representations from Transformers (BERT) has been applied using the python library ‘BERTopic’, to find the ideal text vectorization method because it is designed to capture context in the text from both directions (left and right) around a word, hence being ideal for understanding the meaning of words in context. BERT uses self-supervised learning to find and extract topics, hidden themes, from big, unstructured document collections. These algorithms classify materials into discrete subjects by using statistical correlations between words inside the content. The generated topic models can then be applied to automatically classify or summarise texts at a level of detail that is not achievable through manual classification or summarisation (Dierckx et al., 2021).

Three BERT models were trained using different vectorizers, and the results are depicted in the table 3.1 below:

Table 3.1: - Comparison of different vectorizers, performed with BERT

	Topics identified	**Topic diversity**	**Outliers** (non-identified articles)
TF IDF	633	0.550	20718
Word2Vec	491	0.512	21011
CountVectorizer	644	0.460	20743

Based on the facts, CountVectorizer has performed well at identifying a high number of topics, but it has lower topic diversity, suggesting that the topics might be more repetitive or less distinct from one another. TF-IDF offers a balance between the number of topics and topic diversity, achieving balance at identifying a wide range of distinct topics with fewer outliers. Word2Vec identifies fewer topics but maintains a reasonable level of topic diversity. However, it results in a higher number of outliers, indicating that it may struggle more with certain articles, which can be largely attributed to its neural networks architecture; it requires larger datasets to harness higher accuracy and classification output. However, for our case, since a higher topic diversity can add complexity to the models interpretability of sentiment being conveyed in the news, it's ideal to select CountVectorizer for carrying further topic modelling, as it's able to identify the most number of topics while demonstrating less diversity. In other words, providing a clearer macro view rather than getting encumbered in highly complex micro themes of the news, which makes it easy to understand the overall trend or sentiment.

3.3 LDA for topic modelling

Topic modelling helps key stakeholders, i.e., investors to identify market trends by grouping related news articles into topics and aids the monitoring of relevant topics over time to detect shifts in market trends. While BERT topic modelling provides a solid ground for fundamental market analysis, allowing to capture semantic meaning of words in context and visualize the summary of insights in form of topics, yet investors might need to look beyond fundamental analysis and delve into technical analysis for a more nuanced decision-making when considering buying or selling a stock. This brings to the use case of performing time-series analysis to capture historical trends and using Latent Dirichlet Allocation (LDA) topic modelling to make better predictions. The question in place here is whether semantic themes in the news can help with more accurate stock price predictions?

LDA uses the generative probability models to define topics in the news articles by traversing through the distribution of words in each text. In order to assist in classifying words into themes based on co-occurrence patterns, it assigns a probability distribution over topics for each document and over words for each subject. This can be mathematically expressed as:

$$P(\beta 1{:}K, \theta 1{:}D, z1{:}D \mid w1{:}D) = \frac{P(\beta 1{:}K, \theta 1{:}D, z1{:}D, w1{:}D)}{p(w1{:}D)}$$

This approach is based on the idea that in the corpus D, each document d displays a random distribution θ_d over K topics, and each entry $\theta_{d,k}$ expresses the proportion of topic k in document d. For each word w in document d, a topic z is derived from θ_d along with a term that is sampled from its distribution over a fixed vocabulary given by β_z (Dierckx et al., 2021).

This research is carried out using LDA model from the 'SkLearn' library in Python. The LDA model is implemented to extract underlying topics from a corpus of cleaned (pre-processed) text descriptions. First, the text data was vectorized using CountVectorizer. Hyperparameters were set to omit words appearing in more than 90% of the documents or fewer than two times, to eliminate common stop words and rare terms that do not contribute meaningful information to the model. Further, various LDA models were evaluated across different configurations of the number of topics (n) by analysing two critical metrics: perplexity and coherence scores. To identify the most ideal value of n, it was systematically varied the ranging from 1 to 250 in increments of 10. Model's perplexity, that measures how well the model predicts a sample, is an indicator of the model's fit; A lower perplexity suggests better generalization to unseen data (Gurdiel et al., 2021). Conversely, coherence score, computed using the 'gensim' library, assesses the interpretability and semantic coherence of the topics; Higher coherence scores generally indicate more meaningful topics. Figure 3.2&3.3 presents the perplexity and coherence scores for each topic configuration.

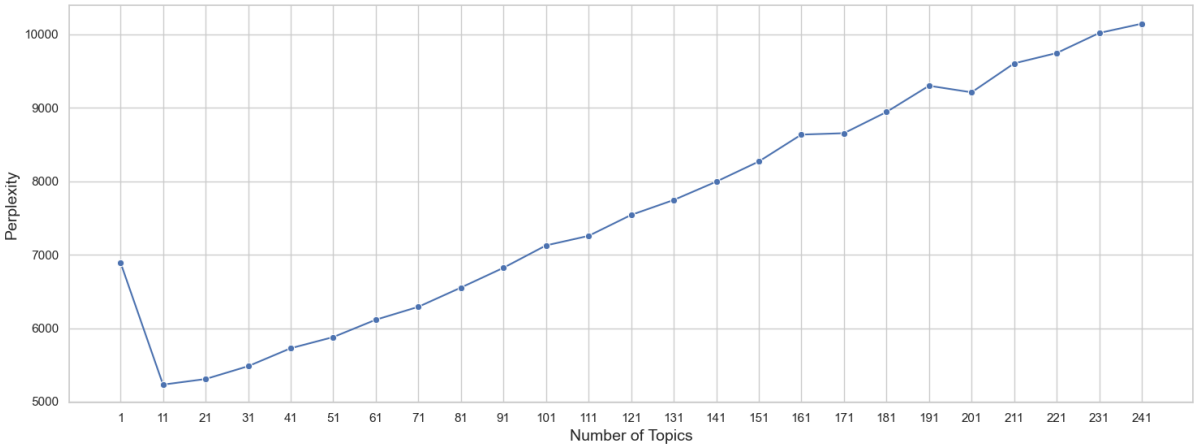

Figure 3.2: - Perplexity computed for n 1 to 250

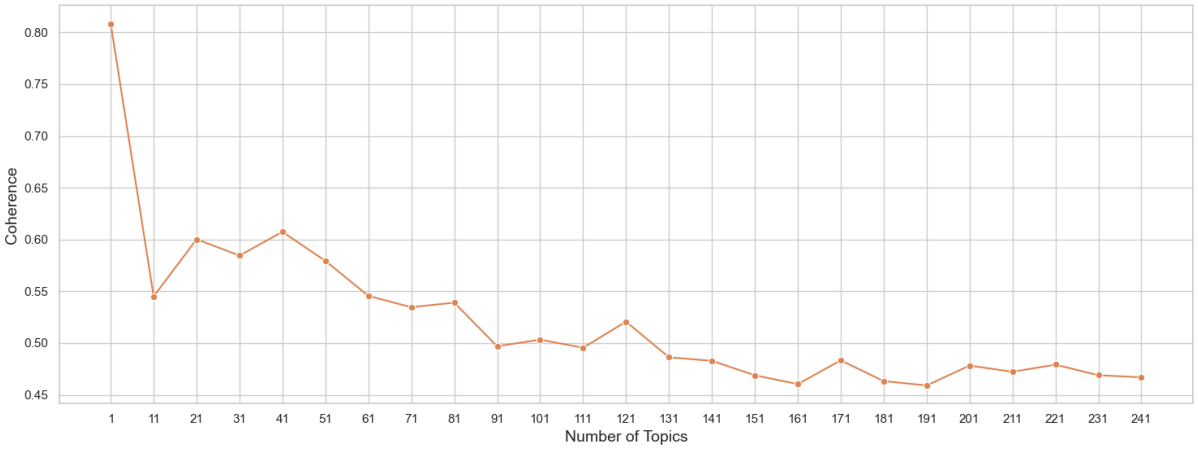

Figure 3.3: - Coherence scores computed for n 1 to 250

As shown in Figure 3.2, perplexity initially decreases, indicating that the model better fits the data with the increasing number of topics. However, after reaching 11 topics, the perplexity begins to increase, signalling a potential to overfitting as the model complexity grows. This suggests that beyond 11 topics, the model starts to capture noise rather than meaningful patterns. The coherence scores, depicted in Figure 3.3 chart, displays a constant decline after 41 topics, with some variability between 10 and 50 topics. Beyond this range, coherence continues to gradually decrease, which could indicate that the topics are becoming less semantically meaningful as more topics are introduced. Given the findings of the perplexity and coherence assessments, a model with 20 topics provides a reasonable balance between interpretability and fit. This arrangement reduces perplexity while preserving a fair amount of coherence, providing a robust interpretable topic model for the given dataset. Though 600

themes could be identified by BERT, this level of granularity is not necessarily useful. Since sentiments typically work at a broad level, it is more useful to gauge the number of topics that are realistically being expressed through the news articles; Micro themes can be aggregated into larger topics. Thus, a lower n value, 20, helps to balance the time and space complexity of the algorithm by combining hundreds of smaller themes into more manageable macro topics.

Figure 3.4 depicts the results visualizing spatial relationships between identified topics by the LDA model on an intertopic distance map (on left) and provides an example of topic 4 to display top 30 most relevant words associated with that topic (on right). For additional research purposes, **appendix C** presents another LDA model for n=200.

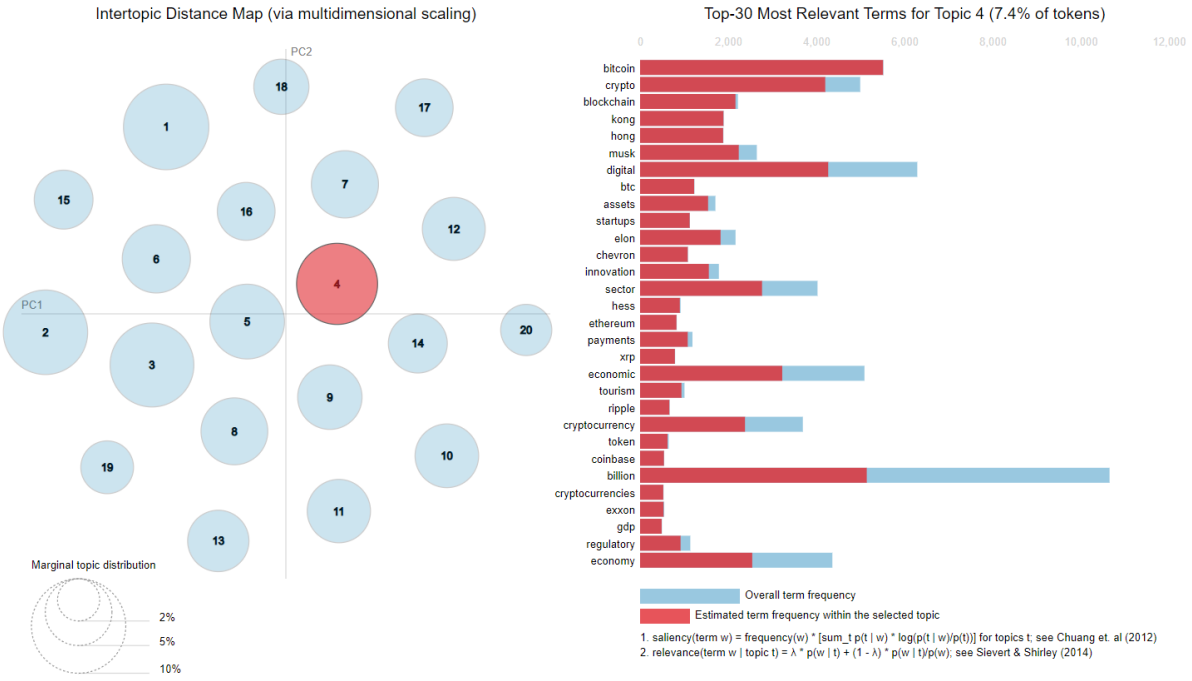

Figure 3.4: - LDA model output at n=20, visualized using 'pyLDAvis' library

Figure 3.4 indicates that, for instance, Topic 4 includes terms related to digital innovation and specifically cryptocurrencies like Bitcoin, suggesting that an investor might encounter mixed sentiments and could potentially forecast future movements in cryptocurrency prices. However, given the difficulty in predicting how these themes will influence stock prices and the challenge of tracking numerous topics in news articles over extended periods, the focus is on exploring the correspondence between news data and stock price data, i.e., analysing

whether identified topics and their associated sentiments affect stock prices over time, based on historical trends.

3.4 Predicting stock price movement

CBOE Volatility Index (VIX) is employed to study market risk. Even though the work of Dierckx et al (2021) takes into account a larger dataset of news, being collected over a longer period of time, this study is limited to new dataset spanning only two months. Therefore, it can only focus on the short-term projections, rather than long-term. The way topic modelling would be used alongside time-series analysis of VIX is by combining textual data with historical stock prices data; The extracted topics from the news articles, using LDA, are grouped based on date and time and merged with the VIX historical data, that is downloaded using 'YhFinance' library in Python, using a left join on 'date' column. Hence, for certain topics surfacing in the news, the associated feature value in the historical market data is mapped to the same time period. The news topics provides additional context and forward-looking information that complements the historical data. Table 3.2 depicts the variables set in place.

Table 3.2: - Variables the data frame being used for predicting stock price movements

Column Name	Description
Open	Price of the stock at the beginning of the trading day
High	highest price the stock reached during the trading day
Low	lowest price the stock reached during the trading day
Close	price of the stock at the end of the trading day
Volume	The number of shares traded during the trading day
Target1…10	The target variable representing the stock's price movement or return for the next 1 to 10 days
Lag1…5	The previous 1 to 5 days' values of a close price used as predictors
Rolling_mean_3/7	The moving average of the stock's price over the past 3 or 7 days, indicating short-term trends
Rolling_std_3/7	The moving standard deviation of the stock's price over the past 3 or 7 days, measuring price volatility.
Topics1…n	Latent topics derived from textual data related to the stock, used as features for prediction

To predict the target variable, two distinct techniques were employed for this study: Classification using Logistic Regression and Gradient Boosting with XGBoost. These methods differ in their approaches to using machine learning for forecasting VIX stock prices based on topic modelling. Logistic Regression is well-suited for binary classification tasks, such as predicting whether the stock price will rise or fall. In contrast, XGBoost is used for predicting the actual stock price by leveraging gradient boosting techniques. The following methodology will compare these approaches to evaluate their effectiveness in forecasting stock price movements.

3.4.1 XGBoost Model

The XGBoost model was initialized with the objective function 'reg:squarederror', being suitable for regression tasks. Different gradient boosting models were trained using a temporally ordered dataset, ensuring that data leakage was avoided by strictly maintaining the chronological order of data. To identify the best performing XGBoost model, a systematic hyperparameter optimization approach was applied for predicting the forward return for each day (Target1 to 10). Model was optimized further using the GridSearchCV in the SkLearn library, that provides a comprehensive search strategy to exhaustively explore a predefined hyperparameter grid. GridSearchCV was configured with 3-fold cross-validation, using the negative mean squared error 'neg_mean_squared_error' as the scoring metric to evaluate model performance across the parameter combinations. The included hyperparameters in the grid and their respective ranges are detailed in the Table 3.3. **Appendix D** corresponds to the functioning of each hyperparameter.

Table 3.3: - Hyperparameters used in GridSearchCV for XGBoost model

Hyperparameter	**Values range**
n_estimators	{50, 100, 200}
learning_rate	{0.01, 0.1, 0.2}
max_depth	{3, 5, 7}
subsample	{0.8, 1.0}
colsample_bytree	{0.8, 1.0}

After fitting the model to 80% of the data, for training, leaving the other 20% for testing, the grid search identified the optimal hyperparameters that minimized the mean squared error on the validation folds. Finally, the best parameters were then used to retrain the XGBoost model, which was subsequently evaluated on the test set. The final model's performance is quantified using the mean squared error (MSE) on test data, providing a ground to evaluate its predictive accuracy.

3.4.2 Logistic Regression

Similar considerations were taken into account for the Logistic Regression model. However, this model was initialized without specifying the solver or regularization, in order for GridSearchCV to determine the best combination through 3-fold cross-validation. Regularization techniques were added to prevent overfitting by adding a penalty term to the loss function, and discouraging the model from fitting too closely on the training data. The parameters in the grid search are outlined in Table 3.4 and further detailed in Appendix D.

Table 3.4: - Hyperparameters used in GridSearchCV for Logistic regression model

Hyperparameter	**Values range**
C	{0.01, 0.1, 1, 10, 100}
solver	{liblinear, saga}
penalty	{L1, L2}
max_iter	{100, 200, 300}

4 Results & Analysis

4.1 Comparing XGBoost results over different prediction horizons

As an overview, the model revealed mixed performances across different prediction horizons. While it struggles with short-term predictions, likely due to the volatility and noise being captured, it performs well in the mid-term range. There is a significant reduction in error at longer horizons, which seems impressive, but indicates overfitting, which warrants further investigation. Figure 4.1&4.2 depicts this on a line graph.

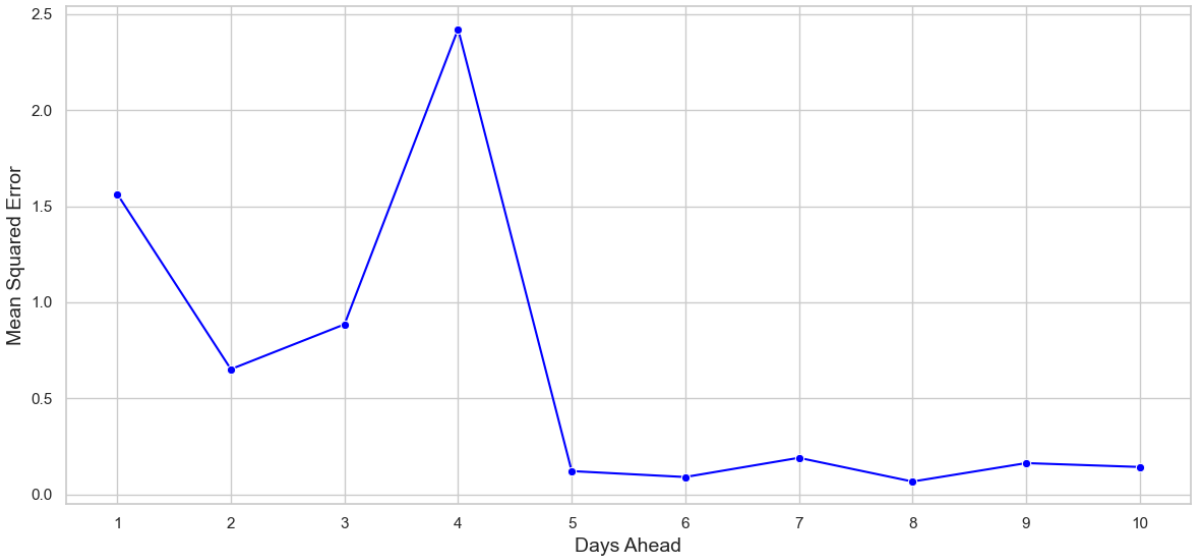

Figure 4.1: - MSE changes over period of predicting n (1-10) days ahead

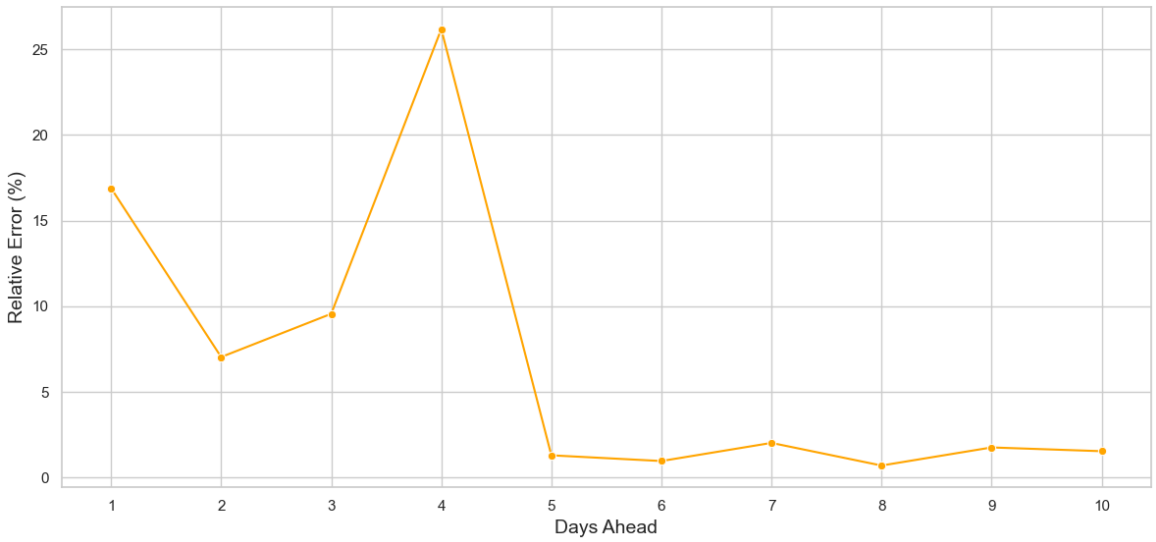

Figure 4.2: - Relative Error changes over period of predicting n (1-10) days ahead

The performances of these models across different time horizons reveal several key insights into their behaviour and accuracy. When predicting 1 day ahead, the model exhibits a relatively high MSE of 1.563 and a corresponding relative error of approximately 16.90%. This higher error rate may indicate that the model struggles to capture the short-term fluctuations inherent in the data. A relatively high error, even though it's predicting only one day ahead, suggests that the model may be overly sensitive to the immediate, possibly erratic, movements in the stock prices. As the prediction horizon extends to 2 and 3 days, the model's performance improves notably. The MSE decreases to 0.652 and 0.885, while the relative error drops to 7.04% and 9.57%, respectively, hinting that the model begins to capture the underlying trend more effectively when it moves beyond the very short-term projection. It is possible that the model smooths out some of the short-term noise, allowing it to generate more accurate predictions over these slightly longer horizons. This trend indicates that the model may be more reliable when forecasting over a few days, where it can balance between capturing the overall direction of the stock prices and avoiding the noise in daily fluctuations.

However, at the 4th day prediction horizon, there is a noticeable spike in the error metrics. The MSE increases sharply to 2.422, and the relative error rises significantly to 26.18%. Interestingly, from the 5th day prediction horizon onwards, the model's MSE drops dramatically, with values as low as 0.067 observed at the 8-day mark and the relative error reaching a low of 0.71%. Although this reduction in error suggests that the model might be highly effective in capturing the overall trend of the stock prices over these mid-to-long-term horizons, but the fact of matter is that such low error rates across these longer prediction intervals hints potential overfitting. Such consistently low error metrics are uncommon in financial forecasting, in cases where data variability is typically higher, thereby risking poor generalization to unseen data.

4.1.1 Cross-validation results

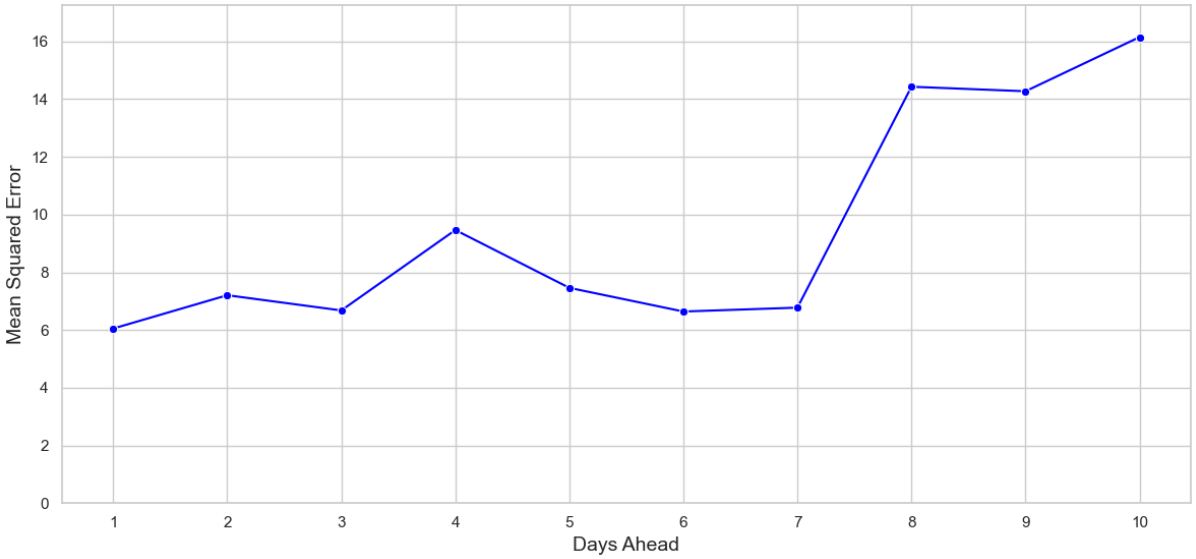

Figure 4.3: - Average MSE for cross-validation folds

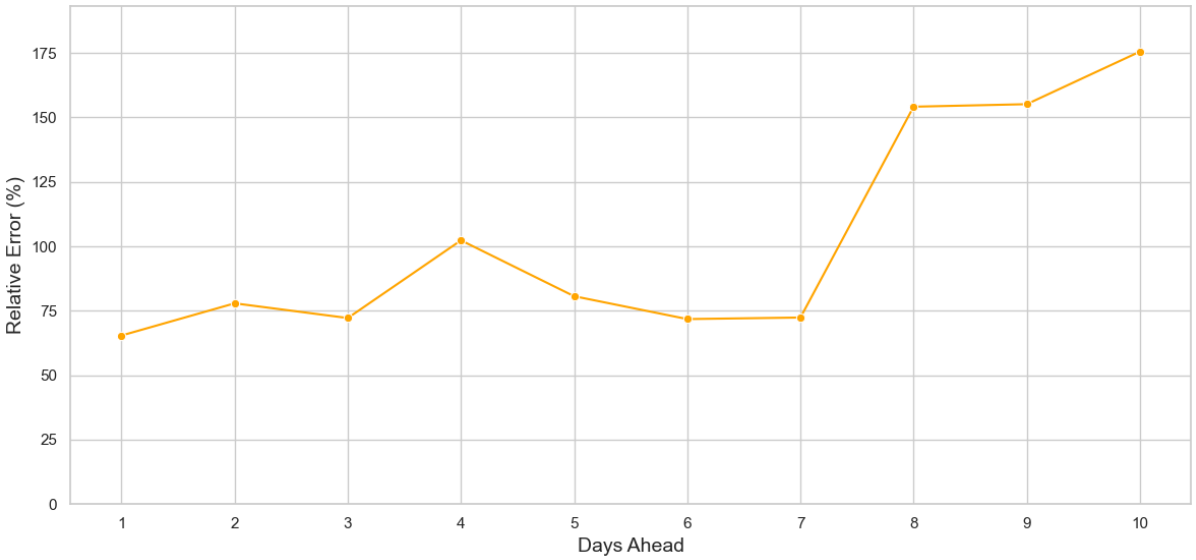

Figure 4.4: - Average Relative Errors for cross-validation folds

The cross-validation results reveal a dramatic shift, as shown in Figure 4.3&4.4, the MSE and relative error values have significantly increased. The discrepancy between initial and cross-validated error rates suggests overfitting and highlights the importance of robust evaluation. Relative error surpassing 100% is indicative of model predicting worse than the actual values by more than the magnitude of the target variable itself. These results suggest that the best

gradient boosting algorithm can get to predicting the stock price is over 1-day horizon, with a relative error of around 65%. Unfortunately, this level of error indicates that the model is not highly accurate, and the market predictions could be significantly off.

4.2 Comparing Logistic regression results over different horizons

Logistic regression is used to predict binary outcomes, i.e., whether the stock price will climb or decline, unlike predicting the stock price itself using XGBoost model. This simplifies the prediction problem for investors by offering a more practical decisions for trading strategies than forecasting exact price levels, which may be less actionable.

The evaluation of the logistic regression model across different time horizons reveals some intriguing patterns in both precision and accuracy. Accuracy is indicative of the overall correctness of the model by calculating the proportion of true positives and true negatives out of all predictions made. Precision, on the other hand, focuses specifically on the relevance of positive predictions by determining the proportion of true positives out of all positive predictions made. It is crucial for understanding how many of the predicted positives are actually correct, which is particularly important when the cost of false positives is high (Gareth et al., 2013). Figure 4.5 presents the models' performance over the horizon of days 1 to 10.

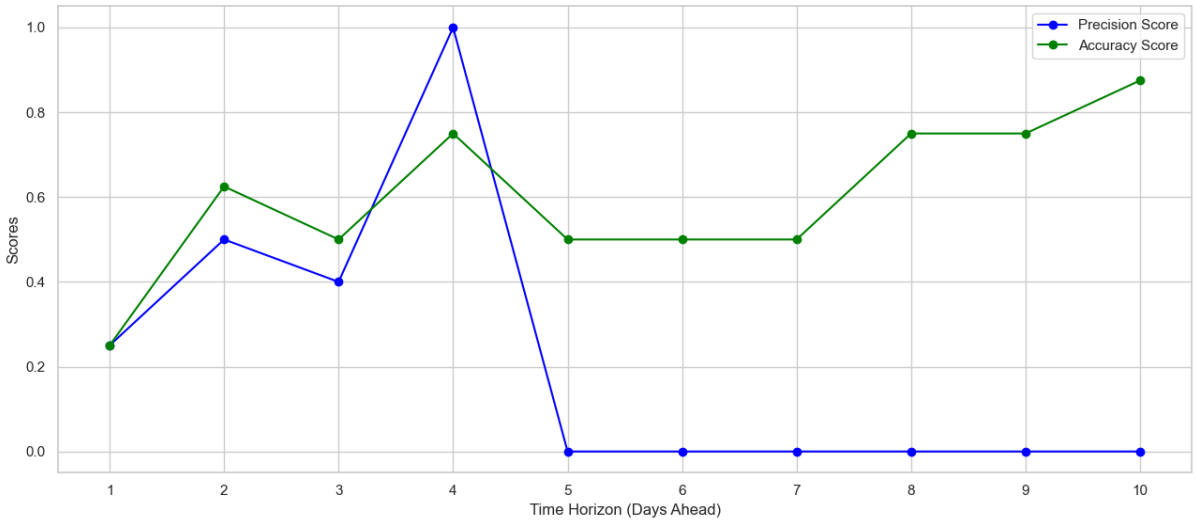

Figure 4.5: - Precision and accuracy scores across different time horizons

The model exhibits moderate predictive power for the first 3 days of horizon. The model's performance is particularly notable on 4-day horizon, with a perfect precision score and an accuracy score of 0.75. This suggests that every positive prediction made by the model on this day was correct, although this may be due to a limited number of positive predictions; The accuracy indicates that the model was relatively reliable for this day. However, from 5-day mark onward, precision drops to 0, and this trend continues infinitely. Interestingly, despite the sharp decline in precision, accuracy scores keep climbing, peaking at 0.88 on day-10, suggesting that the model is increasingly conservative, focusing on correctly predicting the majority class of negatives as the time horizon extends, rather than risking incorrect positive predictions. In other words, there are some correct classifications but fails to identify any true positives.

The observed trends highlight model’s limited effectiveness in predicting positive outcomes, particularly as the forecast horizon extends beyond a few days. The brief improvement in precision and accuracy at day-4 suggests that the model may be effective at capturing some short-term patterns. However, the subsequent drop in precision and the consistent accuracy scores indicate a bias towards predicting negatives, likely as a conservative strategy to avoid false positives.

To determine whether the 4th-day horizon is accurate enough for investors to incorporate the algorithm, the model was retrained using cross-validation to prevent bias and overfitting, ensuring an accurate assessment of the Logistic Regression model's performance.

4.2.1 Cross-validations results

After performing 3-fold cross-validation and plotting the average accuracy and precision scores of each fold on a line graph in Figure 4.6, a dramatic shift in results is observed again.

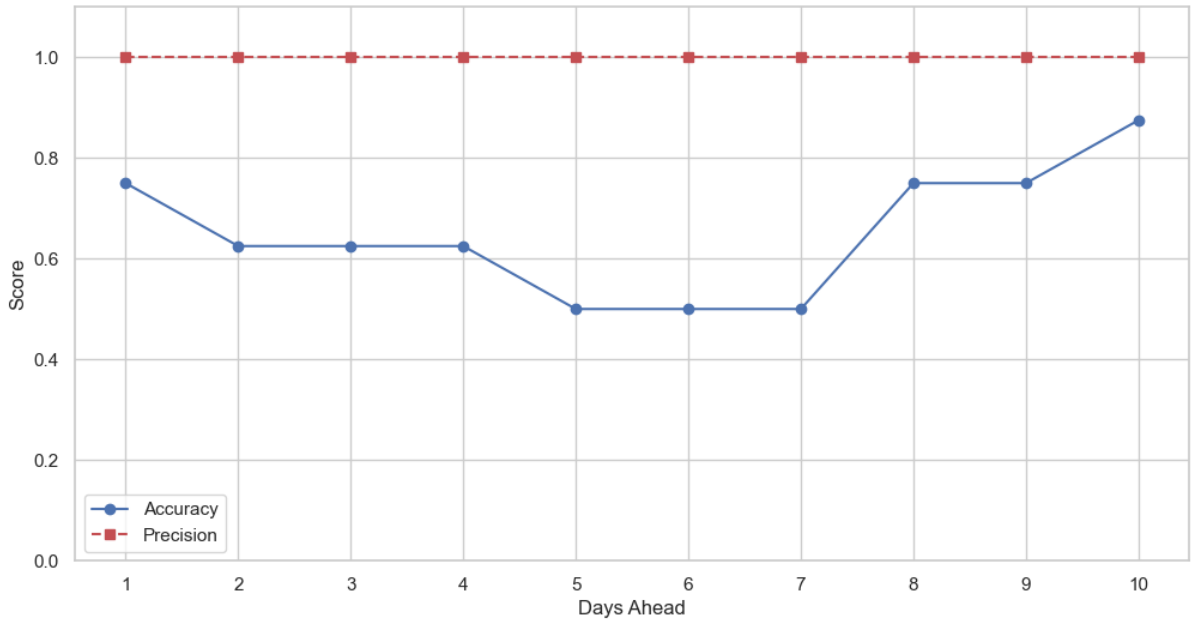

Figure 4.6: - Precision and accuracy scores after cross-validation

For day-1 forecast, the model achieved a high precision of 1.0 and accuracy 0.75, indicating that all positive predictions were accurate and correctly predicted the outcome 75% of the time. Even though, the accuracy tends to decrease until 7th day horizon, the precision remained perfect. This decline in accuracy could be attributed to increased uncertainty and variability in stock movements over a slightly extended horizon. However, the high precision for day-1 suggests that stock movements might be more predictable and less subject to random fluctuations in the very short-term. As the prediction window lengthens, the model's ability to capture short-term patterns may diminish, leading to lower accuracy while still maintaining high precision in positive predictions.

At the 8^{th} day horizon, accuracy improved back to 0.75, similar to the performance of the day-1 forecast. Perhaps, the longer time frame might provide sufficient data for the model to identify meaningful trends, thereby improving accuracy while maintaining high precision. Until 10^{th} day horizon, accuracy keeps on improving and the model achieved the highest accuracy of 0.88 while maintaining perfect precision. The increased accuracy may result from

improved feature interactions and trends being more pronounced and less affected by short-term volatility. Nevertheless, it's important to consider that the model is trained on only two months of news and historical data, and further research might be required to identify any data leakage or overfitting, potentially by using more folds in cross-validation.

5 Discussion

5.1 Reliability of the models

In the context of predicting stock price movements using historical data and features derived from LDA topic modelling and VIX data, both, XGBoost and Logistic Regression, are used with different techniques, across various time horizons and presents contrary results.

XGBoost has shown significant versatility in managing different time horizons, though its performance varied across the days analysed. Predicting stock prices directly appears to be an unreliable method for investor decision-making, largely due to potential overfitting concerns, and thus should be approached with caution. In contrast, binary outcome predictions, such as forecasting whether stock prices will increase or decrease, seem to be more dependable when using a Logistic Regression model. These findings suggest that predicting stock price direction is both feasible and practical. Integrating topic modelling from news datasets with historical stock price analysis can enhance investor decisions, provided that appropriate caution is exercised. However, given that the model was trained on just two months of data, it would be more prudent to extend the training period to capture more intricate patterns that underlie stock market. Additionally, other binary classification techniques, such as Random Forests and Support Vector Machines (SVMs), warrant consideration for future analysis.

The logistic regression model demonstrates varying performance across different time horizons, with precision remaining high throughout and accuracy showing fluctuations. The ten-day prediction horizon stands out as the most effective, reflecting the model's capacity to leverage long-term trends and achieve superior accuracy. This highlights the importance of selecting appropriate forecasting horizons based on model strengths and the nature of the data.

5.2 Causation vs correlation

Unfortunately, considering the given dataset and its research and analysis, it’s still not possible to measure to what extent does data derived from news sway market sentiment more significantly than external factors, but its apparent that the longer a stock or industry remains on the news, it becomes more volatile (Arratia et al., 2021). Since institutional investors rely more on information and employ more automated procedures, they are often thought to be

more logical traders. Unlike individual investors, who are more prone to be swayed by the tone of financial news and to act on it, a divergence in stock prices from their actual value often turns out to be quite apparent. Financial text sentiment research is therefore probably more helpful in developed markets with large numbers of retail investors, such those in the USA and Europe, than it is in emerging markets. It is still possible for institutional investors to profit from fluctuations in stock prices brought on by news reactions from individual investors in these developed markets. However, using time-series analysis alone with topic modelling can be problematic, as it functions like a black box, indicating what might happen but not explaining why it's happening. In the context of finance and news, this lack of transparency can make stakeholders and investors hesitant to base their decisions on such methods. In the world of finance and the stock market, understanding the cause behind any movement is crucial. Without this understanding, predictions may be seen as mere speculation rather than informed measures.

As previously mentioned, predicting binary outcomes, such as whether a stock will rise or fall (as in logistic regression), becomes more actionable if you know whether sentiment and VIX truly drive these movements. However, the models still exhibit noise that makes them prone to overfitting and leading to misleading results. Thus, incorporating models that go beyond understanding the correlation among variables and aid investors with understanding the causal relationships to adjust strategies based on news sentiments and signals can help prevent overcome some of the limitations of these models discussed earlier. Granger (1969) causality present a framework to allow move beyond simple correlations. In context of stock prediction, it can identify whether fluctuations in VIX or sentiment conveyed by news lead market changes rather than just coinciding with them and ensure that the sentiment analysis genuinely contributes to predicting stock movements rather than introducing spurious correlations. This causality adds robustness to the predictive power of existing conventional time-series models. For example, X_t can be considered as a time series derived from the news topic modelling sentiment scores merged with VIX historical data, and Y_t representing the target variable – whether price rises or declines. To perform the granger causality test, two models can be setup for the analysis:

1. A univariate model where Y_t is regressed only on its own lags (what have already been achieved in this research). Expressed as:

$$Y_t = \alpha_0 \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t$$

2. A bivariate model where Y_t regression is not limited to its own lags but also the lags of X_t, i.e., sentiment scores and other independent variables (to extend this research for gaining further insights). Expressed as:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{q} \gamma_j X_{t-j} + n_t$$

If the coefficients γ_j are significant, it reveals what independent variables cause stock returns (dependant variable) to fluctuate, indicating a predictive relationship and validating hypothesis “X does cause Y”. Understanding such relationships among entities and keep track of how they evolve over time can explain the causation, offering a sound decision-making to investors without being ambiguous.

5.3 Reconnecting with QTA and NER in market analysis

While the analysis shows that it’s still difficult to definitively quantify how news-derived data sways market sentiment relative to external factors like environmental or health crises, one pattern is evident: extended media coverage of a stock or industry often leads to increased volatility. However, merely linking topic modelling and time-series analysis may fall short, as these techniques tend to operate as "black boxes"—predicting outcomes without providing the reasoning behind them. In finance, where the cause-and-effect relationship is crucial for decision-making, this lack of transparency can hinder the adoption of such models. This is where integrating QTA and NER allow for structured extraction of key themes and sentiments from news articles, which can be crucial inputs into predictive models.

Central to this goal is refining how text data is annotated and processed, which involves building comprehensive knowledge graphs that map out relationships between entities, sentiments, and their market impacts (Repke and Krestel, 2021). One promising direction lies in enhancing entity recognition and sentiment extraction through more sophisticated models

like BERT, which captures contextual nuances better than traditional methods like LDA. BERT's ability to grasp the deeper semantics of text could lead to more accurate identification of financial sentiment and more precise entity linking, which in turn improves the predictive power of models that rely on these features (Devlin et al., 2019). However, despite BERT's effectiveness, its computational demands often necessitate trade-offs in research environments with limited hardware resources. This is why simpler models like LDA are still widely used; they offer interpretability and ease of implementation, though at the cost of finer semantic understanding (Blei et al., 2003).

The potential for QTA lies not only in detecting sentiment but also in capturing the interplay between multiple themes and their cumulative effects on market behaviour. This involves going beyond simple sentiment analysis to model the interaction between news topics, how they evolve, and how they collectively influence investor behaviour. A simpler example of a company network can be seen in Figure 5.1 and advanced entity semantic relationships captured in Figure 5.2.

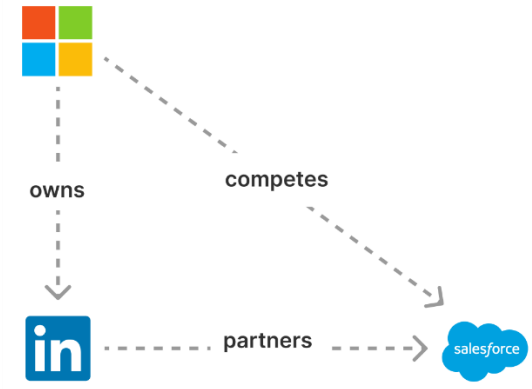

Figure 5.1: - Network graph based on the news: "Microsoft acquired LinkedIn in 2016, while LinkedIn will continue to partner with Salesforce for improving CRM features" (Taylor, 2016).

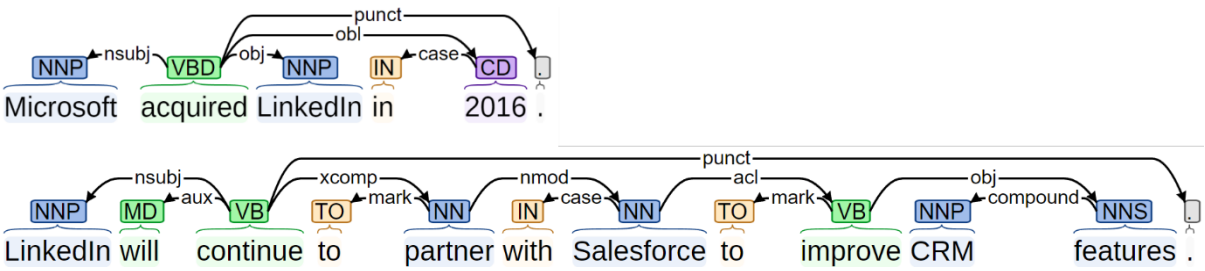

Figure 5.2: - Enhanced dependency parser demonstrating relationships among the entities, visualised using CoreNLP (https://corenlp.run)

Such a multi-faceted approach could bridge the gap between correlation and causation in predictive models, paving the way for more informed investment strategies that account for both immediate and longer-term market reactions. Beyond just model improvements, integrating QTA with time-series analysis in stock prediction requires more robust mechanisms to handle the dynamic and context-dependent nature of financial language. For instance, using domain-specific sentiment lexicons and fine-tuning models for financial contexts could enhance the relevance of extracted features. Moreover, advancements in graph-based QTA approaches could enable the tracking of narrative shifts and evolving market sentiments in near real-time, providing a more nuanced view of how news impacts stock movements across different time horizons.

6 Conclusion

This research provided empirical evidence supporting the notion that news narratives carrying a thematic content does apparently influence stock market behaviour. It demonstrated that quantified narratives extracted from news articles, characterized by sets of topics, can be predictive of future movements in the CBOE Volatility Index (VIX) across different time horizons. The study utilized BERT to identify an appropriate text vectorizer, while Latent Dirichlet Allocation (LDA) combined with CountVectorizer proved effective for topic modelling.

Additionally, the findings reveal that the predictive power of these narrative features varies depending on the prediction horizon. Although gradient boosting mechanisms, in general, predicting the stock prices does not offer reliable decision-making tools for investors, predicting a binary outcome using algorithms like logistic regression with optimized configurations and hyperparameters can outperform the baseline models for day-1, and day-10 ahead predictions. This suggests that narrative-driven features impact market behaviour either immediately or after a delayed period. Specifically, at the 1-day horizon, the immediate sentiment captured by news can affect market dynamics, whereas the mid-term predictions show diminished influence. After this initial emotional response, there appears to be a latent period where the consequences of news narratives gradually manifest, triggering another wave of market movement later. This may indicate that market participants do not always react instantly to financial news narratives, perhaps the full impact of such information takes time to materialize in the market.

Further research with datasets bearing extended timeframes could be necessary to validate this hypothesis. Nevertheless, it’s evident that such predictive models are valuable, but should be viewed as supplementary tools for financial stakeholders, based on the fact they are complementing rather than replacing human analysis. The integration of quantitative models with human intuition and expertise remains essential for more rational and informed decision-making in financial markets.

Bibliography

Arratia, A., Guarniz, G.O., Cabaña, A., Duarte-López, A. and Martí Renedo-Mirambell (2021). Sentiment Analysis of Financial News: Mechanics and Statistics. pp.195–216.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C. and Vohs, K.D. (2001). Bad Is Stronger than Good. Review of General Psychology, [online] 5(4), pp.323–370.

Bikel, D.M., Miller, S., Schwartz, R. and Weischedel, R. (1997). Nymble. *Proceedings of the fifth conference on Applied natural language processing.*

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, pp.993–1022.

Dierckx, T., Davis, J. and Wim S. (2021). Quantifying News Narratives to Predict Movements in Market Risk. Springer eBooks, pp.265–285.

Devi, S.G., Selvam, K. and Rajagopalan, S.P. (2011). An abstract to calculate big O factors of time and space complexity of machine code. *International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011)*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Gareth, J., Daniela W., Trevor H., Robert T. (2013). An Introduction to Statistical Learning : with Applications in Python. New York :Springer.

Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3), pp.424–438.

Gurdiel, L.P., Mediano, J.M., Quintero Cifuentes, J.A. (2021). A comparison study between coherence and perplexity for determining the number of topics in practitioners interviews analysis.

Kumar Saksham (2023). Global News Dataset. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/everydaycodings/global-news-dataset/data [Accessed 7 Jul. 2024].

Li, F. (2006). Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? *SSRN Electronic Journal*.

Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A. and Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. Political Analysis, 23(2), pp.254–277.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, [online] 2(1–2), pp.1–135. doi:https://doi.org/10.1561/1500000011.

Rau, L.F. (1991). *Extracting company names from text*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CAIA.1991.120841.

Repke, T. and Krestel, R. (2021). Extraction and Representation of Financial Entities from Text. *Springer eBooks*, pp.241–263.

Shen, W., Wang, J. and Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, [online] 27(2), pp.443–460.

Taylor, H. (2016). *Why Microsoft beat Salesforce to acquire LinkedIn*. [online] CNBC. Available at: https://www.cnbc.com/2016/11/15/why-microsoft-beat-salesforce-to-acquire-linkedin.html.

The Economist (2017). The world's most valuable resource is no longer oil, but data. [online] The Economist. Available at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the International Conference on Computational Lin- guistics*. pp. 2145–2158.